Priorization of the good character in perceptual matching depends on self-relevance

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@gmail.com

Abstract

Morality is central to social life, moral character is the center of morality. Researchers had long assumed moral character information is prioritized in perceptual process, yet the evidence is scarce. In a series of experiments, we examined the effect of immediately acquired moral character information on perceptual matching. Participants first learned the association between moral characters (labels) and visual cues (shapes), then performed a shape-lable perceptual matching task. The results showed that shapes associated with good moral character were prioritized, as compared to shapes associated with neutral or bad moral characters. This pattern was robust after chaning the words for label or using diagnostic behavioral as an proxy of mroal character. Also, this pattern was robust when changing simultaneous presentation to sequential presentation. We then examined two approximate explanations for this effect: value-based prioritization or social-categorization based prioritization. We manipulated the identity of different moral character explicitly and found that the good character effect was strong when it is self-referential but weak or non-exist when it is other-referential. In further experiments where identity or moral character information were task-irrelevant stimuli, we found that participant care about the character when identity information is salient while the valance is irrelevant, but when valence is task-relevant, people seems only care about the good character's identity but not so much about the neutral or bad character. Together, these results suggested that people are senstive to who is the good character, but less sensitive to who is the neutral or bad character. When the identity information is ambiguous, people may spontaneously referring the good character as self. These results added new evidence for the social vision and suggested the advantage of good character depends on the self-relevance.

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

Word count: X

Priorization of the good character in perceptual matching depends on self-relevance

Alternative title: The good person is me: Spontaneous self-referential may explain the prioritization of good moral character

Introduction

[quotes about moral character]

social vision –> moral vision –> two competing explanations (value-based vs. true-self-based) –> true-self is not perspective free but self-centered.

[morality is central to social life, moral character is the central of morality] People experience a substantial amount of moral events in everyday life (e.g., Hofmann, Wisneski, Brandt, & Skitka, 2014). Whether we are the agent, target, or a third party of a moral event, we judge the behaviors as "right" or "wrong," and by doing so, we judge moral character of people as "good" or "bad" (Uhlmann, Pizarro, & Diermeier, 2015). Moral character is so important in social life that a substantial part of people's conversation are gossiping others' moral character (or, reputation) (e.g., Dunbar, 2004). Also, evidence from studies of person perception and social evaluation revealed that morality is a basic dimension for social evaluation and it is weighed more than traits from other dimensions such as competence and sociability (Abele, Ellemers, Fiske, Koch, & Yzerbyt, 2020; Goodwin, 2015; Goodwin, Piazza, & Rozin, 2014; Willis & Todorov, 2006). The importance of moral character may have been internalized to individuals' self-concept and the positive moral self is the most important aspect of identity (e.g., Strohminger, Knobe, & Newman, 2017), and moral character is a standard we used to evaluate our in-group members and distinguish out-group members (Ellemers, 2018).

[No real perceptual studies on moral character] Given the importance of moral character, it is not surprised that people often assume that moral character related information are prioritized in human information processing system, especially 'bad' agents (e.g., the introduction part of Siegel, Mathys, Rutledge, & Crockett, 2018). A scrutiny of the evidence, however, revealed few direct evidence. For example, while Schupp et al. (2004), Ohman, Lundqvist, and Esteves (2001) are cited as the evidence, they are using facial expressions as stimuli that do not contain any moral meaning; Skowronski and Carlston (1989), Fiske (1980), and Baumeister, Bratslavsky, Finkenauer, and Vohs (2001) are often cited but they were not referring to moral character in specific but using negative social traits, which include competence and sociability as well. For instance, Pratto and John (1991) focused on the desirability of personal traits, which is not specific to moral character either. While Vanneste, Verplaetse, Van Hiel, and Braeckman (2007) tried to study the attentional grabbing effect of the facial expression when the agent decided not to cooperate, the mechanism of the effect, however, can not be attributed to uncooperativeness *per se*, because participants who performed the dot-detection task have no idea about the moral character or behavior of that facial expression and therefore can have different interpretations of that facial expression. In short, though researchers in the field assumed that moral character, especially the bad one, is prioritized in information processing , direct evidence is scarce. This issue is not limited to moral character but common in person perception studies. As Freeman and Ambady (2011) put it, most studies didn't try to explain the perceptual process itself, rather, they are trying to explain the higher-order social cognitive processes that come after. Therefore, it remains unclear (1) whether moral character related information are prioritized in information processing (e.g., perception) and, if yes, (2) what are the underlying mechanisms of the prioritization effect.

[Challenge: operationalization of moral character in laboratory settings] The scarcity of studies on low-level information process of moral character is not without reasons. When trying to study moral character's effect on information process (e.g., perception), one big challenge lies in the difficulty to operationalize the moral character. Morality is defined by context. Whether a behavior should be judged as moral or immoral depends on a number of factors such as intention, consequences (Cushman, Young, & Hauser, 2006; Young, Cushman, Hauser, & Saxe, 2007). Also, whether a behavior is moral relevant depends on cultural and social norm (Haidt, 2007; Rai & Fiske, 2011). These contextual factors, when studied in laboratory settings, have to be carefully controlled and manipulated by providing complex verbal information. However, in most classic cognitive paradigms, the stimuli are presented shortly and participants are required to make quick decisions, therefore long sentences or complicated scenarios are usually prohibited.

To solve this issue, two approaches emerged in the last decade. The first approach used direct associative learning. For example, Shore and Heerey (2013) asked participants first interact with a stranger, who was represented by a face on the screen. By manipulating the behavior associated with the face, participants can form an impression of the person through interaction and judge it as trustworthy or not. After getting such impression, participants then finished a attention blink task where the faces were used as stimuli. Shore and Heerey (2013) found that faces associated with cooperative interaction history are preferentially processed in the pre-attention stage.

Another approach used indirect associative learning, where participants first associate visual stimuli (e.g., faces) with descriptions of a person's moral behaviors, then perform a task that examine the differences between visual stimuli that associated with different moral behaviors. For example, E. Anderson, Siegel, Bliss-Moreau, and Barrett (2011) associated faces with different behaviors (both negative and neutral behaviors from both social and nonsocial domains) and then asked participants to perform a binocular rivalry task, where a face and a building were presented to each eye and participant need to report the content of their vision by pressing buttons. They found that faces associated negative social behaviors were dominant for longer time in the visual awareness than faces associated with other types of behaviors (but see Stein, Grubb, Bertrand, Suh, & Verosky, 2017). Eiserbeck and Abdel Rahman (2020) combined indirect associative learning with attention blink paradigm, where neutral faces were associated with sentences about trust behaviors and asked participants to perform a attention blink task. They also found that neutral faces associated with negative trust behavior were processed preferentially. This indirect associative learning paradigm had been developed primarily for affective meanings, and these studies found that building such association requires minimal behavioral information (Bliss-Moreau, Barrett, & Wright, 2008; Falvello, Vinson, Ferrari, & Todorov, 2015; Todorov & Olson, 2008). A similar approach has been used to explore the prioritization of self-referential effect (Sui, He, & Humphreys, 2012), where even more abstract information (conceptual words, e.g., "self," "friend," "stranger") and even simpler visual cues (geometric shapes, e.g., triangle, circle, or square) were used. This simpler shape-label associative learning task produced robust self-prioritization effect.

Both direct and indirect associative learning paradigms are consistent with the dynamic interactive model of person perception (Freeman & Ambady, 2011). The dynamic interactive model proposed that the perceived personal traits are interactively linked with behavior and sensory stimuli. By activating some sensory stimuli, some person traits can be activated. The associative learning task reverse engineering the process and linking the personal traits (moral character in our case) with new visual stimuli, therefore created a temporary but direct link between personal traits and visual stimuli. After creating such associations between different traits and different visual cues, we can then test the newly established trait-cue associations by different cognitive tasks and examine the instantly learned associations' influence on cognitive processing.

[The current study] The current study was designed to investigate the perceptual process of moral character the shape-label associative learning task. This paradigm has two major advantages over the other associative learning tasks. First, it only used a few number of labels that represent different moral characters, therefore control extra factors and avoid the semantic priming problem (c.f., Firestone & Scholl, 2015). Second, it uses non-social visual stimuli as cues, avoided the idiosyncratic features brought by using faces. Besides, the simplicity of the task allows it to be combined with other cognitive tasks. Using this shape-label associative learning and perpetual matching task, the current study aimed at answering two questions mentioned above: (1) whether moral character related information are prioritized, if so, what is the exact pattern; (2) what is the potential explanation for such pattern.

To investigate the first issue and validate that moral character concepts activated moral character as a social cue, we designed four experiments to explore and validate the paradigm. The first experiment directly adopted associative paradigm from Sui, He, and Humphreys (2012) and changed labels from "self," "friend," and "stranger" to "good-person," "neutral-person," and "bad-person." In the follow-up studies, we tested other character labels that have similar moral meaning ("kind-person," "neutral-person," and "evil-person"). In the third experiments, as in E. Anderson, Siegel, Bliss-Moreau, and Barrett (2011), we asked participants to learn associations between three different diagnostic behaviors and three different names, and then use the names as character labels for the associative learning. Finally, we also tested that simultaneously present shape-word pair and sequentially present word and shape. All of these four experiments showed a consistent pattern of effect, that is, the visual cues that associated with positive moral character were prioritized.

Although the available studies agree that social/moral information can enhance the saliency of the sensory stimuli, yet the reported direction of the effect is not consistent. For instance, there are two studies reported a negativity effect where neutral faces associated with negative social behavioral were processed better than neutral faces that associated with neutral behaviors (E. Anderson, Siegel, Bliss-Moreau, & Barrett, 2011; Eiserbeck & Abdel Rahman, 2020; but see Stein, Grubb, Bertrand, Suh, & Verosky, 2017). The underlying reasons for the prioritization was usually attributed to affective meaning of the negative behaviors, which, in essence, is a threatening effect. In contrast, there was one study reported a positivity effect, where faces associated with positive interaction history were prioritized over faces associated with neutral or negative interaction history. And the positivity effect was attributed to a value-based information process (Shore & Heerey, 2013).

The direction of the effect leads to different underlying explanations. The negativity effect usually explained by the evolutionary adoptive mechanism where the threatening feature were prioritized. However, accumulating evidence supported the view that negativity effect, especially those related to affective stimuli, are prioritized because of the low-level physical features, e.g., low frequency feature in the facial expression (see a recent review, Pool, Brosch, Delplanque, & Sander, 2016). This is reflected in the pattern that threatening stimuli are prioritized in detection task, e.g., dot-probe task. In the current study, because all visual stimuli share similar physical features and we did not using detection task but matching task, therefore, it's not surprising that we didn't found a threatening effect.

The positivity effect, on the other hand, appeared later in the processing stage and were attributed to its rewarding value (we limit the value-based account to rewarding value). The value based account is an appealing explanation, there were strong evidence showing that positive emotional stimuli are prioritized (Pool, Brosch, Delplanque, & Sander, 2016). For example, Brian A. Anderson, Laurent, and Yantis (2011) found that stimuli associated with higher reward could be found more easily in a visual search task. The follow-up studies confirmed that value-based prioritization if a robust effect (Brian A. Anderson, 2019). In our experiments, the positive moral character label "good person" may represent an indirect and positive value. The moral value of others was found to be profitable in previous survey (Abele & Wojciszke, 2007).

When applying to social information such as moral character, both the value-based account and the threatening model ignored the social meaning of the stimuli. Increasing evidence supported the view that social categorization also impact our information processing, including perception (Freeman & Ambady, 2011; Xiao, Coppin, & Bavel, 2016). Social categorization theory stated that we perceiving others based on whether or not belong to "us" (Turner, Hogg, Oakes, Reicher, & Wetherell, 1987). In other words, we may view a person with positive moral character as an in-group member, while a bad person as them (Ellemers, 2018). If moral judgment is a social categorization process (DeScioli, 2016; McHugh, McGann, Igou, & Kinsella, 2019), and if social categorization impact our visual perception (Xiao, Coppin, & Bavel, 2016), then we can infer positive moral character may also be prioritized because we regard good person as an natural extension of the self.

However, the above four experiments could not distinguish between these possibilities, because the concept "good-person" creating an unexpected ambiguity for the identity of the good person. it can both be rewarding and be categorized as in-group member, and previous studies using associative learning paradigm revealed that both rewarding stimuli (e.g., Sui, He, & Humphreys, 2012) and in-group information (Enock, Hewstone, Lockwood, & Sui, 2020) are prioritized.

[Distinguish two explanations by make self salient, exp3a, 3b, 6b] Though both two the value-based attention and moral-based categorization accounts can explain the positivity effect found in first four experiments (i.e., prioritization of "good-person," but not "neutral person" and "bad person"), they have different prediction if the experimental design include both identity and moral valence where the valence (good, bad, and neutral) conditions can describe self or other. In this case the identity become salient and participants are less likely to spontaneously identify a good-other as the extension of self, but the value of good-person still exists. Actually, the rewarding value of good-other might be even stronger than good-self because the former indicate potential cooperation and material rewards, but the latter merely confirmed one's personal belief. This means that the social categorization theory predicts participants prioritize good-self but not good-other, while reward-based attention theory predicts participants are both prioritized, or maybe good-other are even more prioritized. Also, as in Hu, Lan, Macrae, and Sui (2020), people may also only identify with good-self instead of bad self. That is, people will show a unique pattern of self-identification: only good-self is identified as "self" while all the others categories were excluded.

We introduced identity (self vs. other) as an addition independent variable in exp 3a, 3b, and 6b. Now the moral valence is orthogonal to the identity. We found that (1) good-self is always faster than neutral-self and bad-self, but good-other only have weak to null advantage to neutral-other and bad-other. which mean the social categorization is self-centered. (2) good-self's advantage over good other only occur when self- and other- were in the same task. i.e. the relative advantage is competition based instead of absolute. These three experiments suggest that people more like to view the moral character stimuli as person and categorize good-self as an unique category against all others. A three-level Bayesian generalized linear mixed effect model showed that there was no effect of valence when the identity was other. This results showed that value-based attention was not likely the mechanism behind the pattern we observed in first four experiments. However, it is still unclear why good-self was prioritized. Besides the social-categorization explanation, it's also possible that good self is so unique that it is prioritized in all possible situation and therefore is not social categorization *per se*.

[what we care? valence of the self exp4a or identity of the good exp4b?] We went further to disentangle the good-self complex: is it because the special role of good-self or because of social categorization. We designed two complementary experiments. in experiment 4a, participants only learned the association between self and other, the words "good-person," "neutral person," and "bad person" were presented as task-irrelevant stimuli, while in experiment 4b, participants learned the associations between "good-person," "neutral-person," and "bad-person," and the "self" and "other" were presented as task-irrelevant stimuli. These two experiment can be used to distinguish the "good-self" as anchor account and the "good-self-based social categorization" account. If good-self as an anchor is true, then, in both experiment, good-self will show advantage over all other stimuli, and there will be no other effects. More specifically, in experiment 4a, where only the self-relevance is task-relevant, there will be advantage for good as task-irrelevant condition than the other two self conditions, while there is no other effects; in experiment 4b, in the good condition, there will be an advantage for self as task-irrelevant condition over other as task-irrelevant condition, and no other effects. If good-self-based social categorization if true, then, the prioritization effect will depends on whether the stimuli can be categorized as the same group of good-self. More specifically, in experiment 4a, there will be good effect in self conditions, this prediction is the same as the "good-self as anchor" account, but also, it predicts a reverse good effect in other condition because good and other a conflict in terms of social-categorization, this prediction is different from the "good-self" anchor account; however, for experiment 4b, it predicts no identity effect in the good-person condition because both self and other are in the good group.

[Good self in self-reported data] As an exploration, we also collected participants' self-reported psychological distance between self and good-person, bad-person, and neutral-person, moral identity, moral self-image, and self-esteem. All these data are available (see Liu et al., 2020). We explored the correlation between self-reported distance and these questionnaires as well as the questionnaires and behavioral data. However, given that the correlation between self-reported score and behavioral data has low correlation (Dang, King, & Inzlicht, 2020), we didn't expect a high correlation between these self-reported measures and the behavioral data.

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.

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. To have a better estimation of the effect size, we included the data from unreported data in our three-level models (experiment 5, 6a, 6b) (See Table S1 for overview of these experiments).

All participant received informed consent and compensated for their time. These experiments were approved by the ethic board in the Department of Tsinghua University.

General methods

Design and Procedure

This series of experiments studied the perceptual process of moral character, using the social associative learning paradigm (or tagging paradigm)(Sui, He, & Humphreys, 2012), in which participants first learned the associations between geometric shapes and labels of person with different moral character (e.g., in first three studies, the triangle, square, and circle and good person, neutral person, and bad person, respectively). The associations of the shapes and label were counterbalanced across participants. After remembered the associations, participants finished a practice phase to familiar with the task, in which they viewed one of the shapes upon the fixation while one of the labels below the fixation and judged whether the shape and the label matched the association they learned. When participants reached 60% or higher accuracy at the end of the practicing session, they started the experimental task which was the same as in the practice phase.

The experiment 1a, 1b, 1c, 2, and 6a shared a 2 (matching: match vs. nonmatch) by 3 (moral character: good person vs. neutral person vs. bad person) within-subject design.

Experiment 1a was the first one of the whole series studies and found the prioritization of stimuli associated with good-person. To confirm that it is the moral character that caused the effect, we further conducted experiment 1b, 1c, and 2. More specifically, experiment 1b used different Chinese words as label to test whether the effect only occurred with certain familiar words. Experiment 1c manipulated the moral valence indirectly: participants first learned to associate different moral behaviors with different neutral names, after remembered the association, they then performed the perceptual matching task by associating names with different shapes. Experiment 2 further tested whether the way we presented the stimuli influence the effect of valence, by sequentially presenting labels and shapes. Note that part of participants of experiment 2 were from experiment 1a because we originally planned a cross task comparison. Experiment 6a, which shared the same design as experiment 2, was an EEG experiment which aimed at exploring the neural correlates of the effect. But we will focus on the behavioral results of experiment 6a in the current manuscript.

For experiment 3a, 3b, 4a, 4b, 6b, 7a, and 7b, we included self-reference as another within-subject variable in the experimental design. For example, the experiment 3a directly extend the design of experiment 1a into a 2 (matchness: match vs. nonmatch) by 2 (reference: self vs. other) by 3 (moral valence: 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). The experiment 6b was an EEG experiment extended from experiment 3a but presented the label and shape sequentially. Because of the relatively high working memory load (six label-shape pairs), experiment 6b were conducted in two days: the first day participants finished perceptual matching task as a practice, and the second day, they finished the task again while the EEG signals were recorded. Experiment 3b was designed to separate the self-referential trials and other-referential trials. That is, participants finished two different types of block: in the self-referential blocks, they only responded to good-self, neutral-self, and bad-self, with half match trials and half non-match trials; in the other-reference blocks, they only responded to good-other, neutral-other, and bad-other. Experiment 7a and 7b were designed to test the cross task robustness of the effect we observed in the aforementioned experiments (see, Hu, Lan, Macrae, & Sui, 2020). The matching task in these two experiments shared the same design with experiment 3a, but only with two moral character, i.e., good vs. bad. We didn't include the neutral condition in experiment 7a and 7b because we found that the neutral and bad conditions constantly showed non-significant results in experiment 1 ~ 6.

Experiment 4a and 4b were design to explore the mechanism behind the prioritization of good-self. In 4a, we used only two labels (self vs. other) and two shapes (circle, square). To manipulate the moral valence, we added the moral-related words within the shape and instructed participants to ignore the words in the shape during the task. In 4b, we reversed the role of self-reference and valence in the task: participant learnt three labels (good-person, neutral-person, and bad-person) and three shapes (circle, square, and triangle), and the words related to identity, "self" or "other," were presented in the shapes. As in 4a, participants were told to ignore the words inside the shape during the task.

E-prime 2.0 was used for presenting stimuli and collecting behavioral responses, except that experiment 7a and 7b used Matlab Psychtoolbox (Brainard, 1997; Pelli, 1997). For participants recruited in Tsinghua University, they finished the experiment individually in a dim-lighted chamber, stimuli were presented on 22-inch CRT monitors and their head were fixed by a chin-rest brace. The distance between participants' eyes and the screen was about 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 the shape or the word and the fixation cross was 3.5° of visual angle. For participants recruited in Wenzhou University, they finished the experiment in a group consisted of 3 ~ 12 participants in a dim-lighted testing room. Participants were required to finished the whole experiment independently. Also, they were instructed to start the experiment at the same time, so that the distraction between participants were minimized. The stimuli were presented on 19-inch CRT monitor. The visual angles are could not be exactly controlled because participants' chin were not fixed.

In most of these experiments, participant were also asked to fill a battery of questionnaire after they finish the behavioral tasks. All the questionnaire data are open (see, dataset 4 in Liu et al., 2020). See Table S1 for a summary information about all the experiments.

Data analysis

We used the `tidyverse` of r (see script `Load_save_data.r`) to exclude the practicing trials, invalid trials of each participants, and invalid participants, if there were any, in the raw data. Results of each experiment were then analyzed in two Bayesian approaches and reported in supplementary materials.

Bayesian hierarchical model. We first tested the effect of experimental manipulation using Bayesian hierarchical model. More specifically, we used the Bayesian hierarchical model (BHM, or Bayesian generalized linear mixed model, BGLMM) to model the reaction time and accuracy data. We used Bayesian hierarchical model because BHM provided three advantages over the classic NHST approach (repeated measure ANOVA or t-tests): first, BHM estimate the posterior distributions of parameters for statistical inference, therefore provided uncertainty in estimation (Rouder & Lu, 2005). Second, BHM, as generalized linear mixed models, can use distribution that fit the distribution of real data instead of using normal distribution for all data. Using appropriate distributions for the data will avoid misleading results and provide better fitting of the data. For example, Reaction times are not normally distributed but right skewed, and the linear assumption in ANOVAs is not satisfied (Rousselet & Wilcox, 2019). Third, BHM provided an unified framework to analyze data from different levels and different sources, avoid the information loss when we need to combine data from different levels.

We first used the `r` package `BRMs` (Bürkner, 2017), which used Stan (Carpenter et al., 2017) to sample from the posterior, to build the model for RTs and accuracy separately. Using the Bayesian hierarchical model, we can directly estimate the over-all effect across similar experiments with similar experimental design, instead of using a two-step approach where we first estimate parameters, e.g., d' for each participant, and then use a random effect model meta-analysis to synthesize the effect (Goh, Hall, & Rosenthal, 2016). We also we used HDDM to model RTs and accuracy data together using drift diffusion model as the data generative model.

Accuracy. We followed practice of previous studies (Hu, Lan, Macrae, & Sui, 2020; Sui, He, & Humphreys, 2012) and used signal detection theory approach to analyze the accuracy data. More specifically, the match trials are treated as signal and the non-match trials are noise. As we mentioned above, we estimated the sensitivity and criterion of SDT by BHM (Rouder & Lu, 2005). Because the BHM can model different level's data using a single unified model, we used a three-level HBM to model the valence effect, which include five experiments: 1a, 1b, 1c, 2, and 6a. Also, we modelled the experiments with both identity and moral valence with a three-level HBM model, which includes 3a, 3b, and 6b. For experiment 4a and 4b, we used two-level models for each separately. However, we compared the posterior of parameters directly because we have full posterior distribution of the effect and can directly compare the posteriors.

We used the Bernoulli distribution to model the accuracy data. For a single participant, we assume that the accuracy of ith trial is Bernoulli distributed (binomial with 1 trial), with probability p_i that $y_i = 1$.

$$y_i \sim Bernoulli(p_i)$$

and the probability of choosing "match" p_i at the ith trial is a function of the trial type:

$$\Phi(p_i) = \beta_0 + \beta_1 IsMatch_i$$

therefore, the outcomes y_i are 0 if the participant responded "nonmatch" on the ith trial, 1 if they responded "match." We then write the generalized linear model on the probits (z-scores; Φ, "Phi") of ps. Φ is the cumulative normal density function and maps z scores to probabilities. In this way, the intercept of the model (β_0) is the standardized false alarm rate (probability of saying 1 when predictor is 0), which we take as our criterion c. The slope of the model (β_1) is the increased probability of responding "match" when the trial type is "match," in z-scores, which is another expression of d'. Therefore, c = -zHR = $-\beta_0$, and $d' = \beta_1$.

In our experimental design, there are three conditions for both match and non-match trials, we can estimate the d' and c separately for each condition. In this case, the criterion c is modeled as the main effect of valence, and the d' can be modeled as the interaction between valence and match, and we explicitly removed the intercept:

$$\Phi(p_i) = 0 + \beta_0 Valence_i + \beta_1 IsMatch_i * Valence_i$$

In each experiment, we had multiple participants. We can estimate the group-level parameters by extending the above model into a two-leve model, where we can estimate parameters on individual level and the group level parameter simultaneously. The probability that the jth subject responded "match" ($y_{ij} = 1$) at the ith trial p_{ij}. In the same vein, we have

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

The the generalized linear model can be re-written to include two levels:

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

We again can write the generalized linear model on the probits (z-scores; Φ, "Phi") of ps.

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

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

For experiments that had 2 (matching: match vs. non-match) by 3 (moral character: good vs. neutral vs. bad), i.e., experiment 1a, 1b, 1c, 2, 5, and 6a, the formula for accuracy in `BRMs` is as follow:

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

For experiments that had two by two by three design, we used the follow formula for the BGLM:

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

In the same vein, we can estimate the posterior of parameters across different experiments. We can use a nested hierarchical model to model all the experiment with similar design:

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

the generalized linear model is then

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

The outcomes y_{ijk} are 0 if participant j in experiment k responded "mismatch" on trial i, 1 if they responded "match."

$$\begin{bmatrix} \beta_{0jk} \\ \beta_{1jk} \end{bmatrix} \sim N(\begin{bmatrix} \theta_{0k} \\ \theta_{1k} \end{bmatrix}, \sum)$$

and the experiment level parameter mu_{0k} and mu_{1k} is from a higher order distribution:

$$\begin{bmatrix}\theta_{0k} \\ \theta_{1k}\end{bmatrix} \sim N(\begin{bmatrix}\mu_0 \\ \mu_1\end{bmatrix}, \sum)$$

in which mu_0 and mu_1 means the population level parameter.

Reaction times. For the reaction time, we used the log normal distribution (https://lindeloev.github.io/shiny-rt/#34_(shifted)_log-normal) to model the data. This means that we need to estimate the posterior of two parameters: μ, σ. μ is the mean of the `logNormal` distribution, and σ is the disperse of the distribution. Although the log normal distribution can be extended to shifted log normal distribution, with one more parameter: shift, which is the earliest possible response, we found that the additional parameter didnt' improved the model fitting and therefore used the logNormal in our final analysis.

The reaction time of the jth subject on ith trial is a linear function of trial type:

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

while the log of the reaction time 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.

$$\mu_j \sim N(\mu, \sigma)$$

$$\sigma_j \sim Cauchy()$$

Formula used for modeling the data as follow:

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

or

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

we expanded the RT model three-level model in which participants and experiments are two group level variable and participants were nested in the experiments.

$$log(y_{ijk}) \sim N(\mu_{jk}, \sigma_{jk})$$

y_{ijk} is the RT of the ith trial of the jth participants in the kth experiment.

$$\mu_{jk} \sim N(\mu_k, \sigma_k)$$

$$\sigma_{jk} \sim Cauchy()$$

$$\mu_k \sim N(\mu, \sigma)$$

$$\theta_k \sim Cauchy()$$

Effect of moral character. We synthesized effect size of d' and RT from experiment 1a, 1b, 1c, 2, 5, and 6a for the effect of moral character. We reported the synthesized the effect across all experiments that tested the valence effect, using the mini meta-analysis approach (Goh, Hall, & Rosenthal, 2016).

Effect of moral self. We further synthesized the effect of moral self, which included results from experiment 3a, 3b, and 6b. In these experiment, we directly tested two possible explanations: moral self as social categorization process and value-based attention.

Implicit interaction between valence and self-relevance. In the third part, we focused on experiment 4a and 4b, which were designed to examine two more nuanced explanation concerning the good-self. The design of experiment 4a and 4b are complementary. Together, they can test whether participants are more sensitive to the moral character of the Self (4a), or the identity of the morally Good (4b).

Finally, we explored correlation between results from behavioral results and self-reported measures.

For the questionnaire part, we are most interested in the self-rated distance between different person and self-evaluation related questionnaires: self-esteem, moral-self identity, and moral self-image. Other questionnaires (e.g., personality) were not planned to correlated with behavioral data were not included. Note that all questionnaire data were reported in (Liu et al., 2020).

For the behavioral task part, we used three parameters from drift diffusion model: drift rate (v), boundary separation (a), and non decision-making time (t), because these parameters has relative clear psychological meaning. We used the mean of parameter posterior distribution as the estimate of each parameter for each participants in the correlation analysis. We used alpha = 0.05 and used bootstrap by `BootES` package (Kirby & Gerlanc, 2013) to estimate the correlation.

Hierarchical drift diffusion model (HDDM). To further explore the psychological mechanism under perceptual decision-making, we used a generative mode drift diffusion model (DDM) to model our RTs and accuracy data. As the hypothesis testing part, we also used hierarchical Bayesian model to fit the DDM. The package we used was the HDDM (Wiecki, Sofer, & Frank, 2013), a python package for fitting hierarchical DDM. We used the prior implemented in HDDM, that is, weakly informative priors that constrains parameter estimates to be in the range of plausible values based on past literature (Matzke & Wagenmakers, 2009). As reported in Hu, Lan, Macrae, and Sui

(2020), we used the stimulus code approach, match response were coded as 1 and nonmatch responses were coded as 0. To fully explore all parameters, we allow all four parameters of DDM free to vary. We then extracted the estimation of all the four parameters for each participants for the correlation analyses. However, because the starting point is only related to response (match vs. non-match) but not the valence of the stimuli, we didn't included it in correlation analysis.

Part 1: Perceptual processing moral 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 situation. 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 valence. Shapes associated with positive moral valence ("good person," "kind person" or a name associated with morally good behavioral history) has higher sensitivity (mean = , 95% HDI =) than shapes associated with neutral condition (mean = , 95% HDI =), but we did not find differences between shapes associated with negative moral label (mean = , 95% HDI =) and neutral condition.

For the reaction times, we also found robust effect of moral valence. Shapes associated with positive moral valence has faster responses (mean = , 95% HDI =) than shapes associated with neutral condition (mean = , 95% HDI =). We also found that the responses to shapes associated with negative moral valence (mean = , 95% HDI =) were slower as compared to the neutral condition. See Figure 1.

Part 2: interaction between valence and identity

In this part, we report 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 participants.

See Figure 2.

Part 3: Implicit binding between valence and identity

In this part, we reported two studies in which the moral valence or the self-referential processing is not task-relevant. We are interested in testing whether the task-relevance will eliminate the effect observed in previous experiment.

For the task relevant part, we found self-related conditions were performed better than other-related conditions, on both *d* prime and reaction times.

Most importantly, we found evidence, albeit weak, that task-irrelevant moral valence also played an role. The *d* prime is greater when shapes were associated with good self condition than with neutral self (BF = 4.4) or bad self (3.1), but shapes associated with bad self and neutral self didn't show differences. In contrast the *d* prime was smaller when shapes were associated with good other than with neutral other or bad other. See Figure 3.

In this task, we found shapes associated with good person conditions were performed better than other-related conditions, on both *d* prime and reaction times.

Most importantly, we found evidence, that task-irrelevant self-relevance also played an role. For shapes associated with good person, the *d* prime was greater when shapes had an "self" inside as task-irrelevant stimuli than with "other" inside (mean_diff = 0.14, 95% credible intervals [-0.02, 0.31], BF = 12.07, p = 0.92), but this effect did not happen when the target shape where associated with "neutral" (mean_diff = 0.04, 95% CI [-.11, .18]) or "bad" person (mean_diff = -.05, 95% CI[-.18, .09]). The same trend appear for the RT data. For shapes associated with good person, an “self” inside will reduce the RTs as compared with when a “other” inside the shape (mean_diff = -55 ms, 95%CI[-75, -35], $p < 0.0001$), but this effect did not occur when the shapes were associated neutral (mean_dfiff = 10, 95% CI [1, 20]) or bad (mean_diff = 5, 95%CI [-16, 27]) person. See Figure 3.

Self-reported personal distance

See Figure **??**.

Correlation analyses

The reliability of questionnaires can be found in (Liu et al., 2020). We calculated the correlation between the data from behavioral task and the questionnaire data. First, we calculated the score for each scale based on their structure and factor loading, instead of sum score (McNeish & Wolf, 2020). Then, we used SEM to estimate the correlation because it can include measurement model and statistical model in a unified framework.

To make sure that what we found were not false positive, we used two method to ensure the robustness of our analysis. first, we split the data into two half: the data with self and without, then, we used the conditional random forest to find the robust correlation in the exploratory data (with self reference) that can be replicated in the confirmatory data (without the self reference). The robust correlation were then analyzed using SEM

Instead of use the exploratory correlation analysis, we used a more principled way to explore the correlation between parameter of HDDM (v, t, and a) and scale scores and person distance.

We didn’t find the correlation between scale scores and the parameters of HDDM, but found weak correlation between personal distance and the parameter estimated from Good and neutral conditions.

First, boundary separation (a) of moral good condition was correlated with both Self-Bad distance ($r = 0.198$, 95% CI [], $p = 0.0063$) and Neutral-Bad distance ($r = 0.1571$, 95% CI [], $p = 0.031$). At the same time, the non-decision time is negatively correlated with Self-Bad distance ($r = 0.169$, 95% CI [], $p = 0.0197$). See Figure **??**.

Second, we found the boundary separation of neutral condition is positively correlated with the personal distance between self and good distance ($r = 0.189$, 95% CI [], $p = 0.036$), but negatively correlated with self-neutral distance($r = -0.183$, 95% CI [], $p = 0.042$). Also, the drift rate of the neutral condition is positively correlated with the Self-Bad distance ($r = 0.177$, 95% CI [], $p = 0.048$).a. See figure **??**

We also explored the correlation between behavioral data and questionnaire scores separately for experiments with and without self-referential, however, the sample size is very low for some conditions.

Discussion

References

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2020). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review*. https://doi.org/10.1037/rev0000262

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, *93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Anderson, Brian A. (2019). Neurobiology of value-driven attention. *Current Opinion in Psychology*, *29*, 27–33. https://doi.org/10.1016/j.copsyc.2018.11.004

Anderson, Brian 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

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion*, *8*(4), 479–493. https://doi.org/10.1037/1528-3542.8.4.479

Brainard, D. H. (1997). The psychophysics toolbox [Journal Article]. *Spatial Vision*, *10*(4), 433–436.

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%20http://dx.doi.org/10.18637/jss.v080.i01

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

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

Science, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

DeScioli, P. (2016). The side-taking hypothesis for moral judgment. *Current Opinion in Psychology*, *7*, 23–27. https://doi.org/10.1016/j.copsyc.2015.07.002

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.

Enock, F. E., Hewstone, M. R. C., Lockwood, P. L., & Sui, J. (2020). Overlap in processing advantages for minimal ingroups and the self. *Scientific Reports*, *10*(1), 18933. https://doi.org/10.1038/s41598-020-76001-9

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition*, *33*(5), 368–386. https://doi.org/10.1521/soco.2015.33.5.368

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

Fiske, S. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906. Retrieved from insights.ovid.com

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279. https://doi.org/10.1037/a0022327

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

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

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002. https://doi.org/10.1126/science.1137651

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

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

Kirby, K. N., & Gerlanc, D. (2013). BootES: An r package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, *45*(4), 905–927. https://doi.org/10.3758/s13428-013-0330-5

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/

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. (2019). *Moral judgment as categorization (MJAC)*. PsyArXiv. https://doi.org/10.31234/osf.io/72dzp

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01398-0

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

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies [Journal Article]. *Spatial Vision*, *10*(4), 437–442.

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin*, *142*(1), 79–106. https://doi.org/10.1037/bul0000026

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*(3), 380–391. https://doi.org/10.1037//0022-3514.61.3.380

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation:

Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. https://doi.org/10.1037/a0021867

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

Schupp, H. T., Ohman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: An ERP analysis. *Emotion (Washington, D.C.)*, *4*(2), 189–200. https://doi.org/10.1037/1528-3542.4.2.189

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

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

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

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

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

Todorov, A., & Olson, I. R. (2008). Robust learning of affective trait associations with faces when the hippocampus is damaged, but not when the amygdala and temporal pole are damaged. *Social Cognitive and Affective Neuroscience*, *3*(3), 195–203. https://doi.org/10.1093/scan/nsn013

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Cambridge, MA, US: Basil Blackwell.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691614556679

Vanneste, S., Verplaetse, J., Van Hiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior*, *28*(4), 272–276. https://doi.org/10.1016/j.evolhumbehav.2007.02.005

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00014

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

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

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. https://doi.org/10.1073/pnas.0701408104

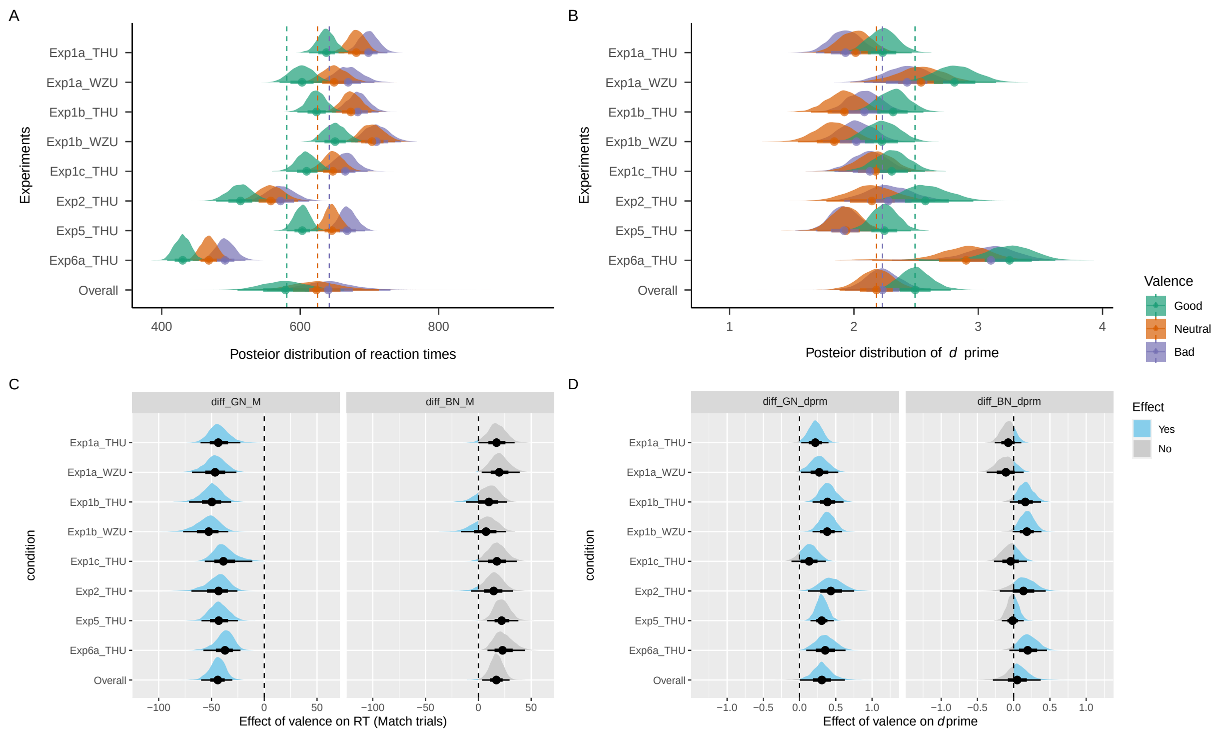

Figure 1. Effect of moral valence on RT and d'

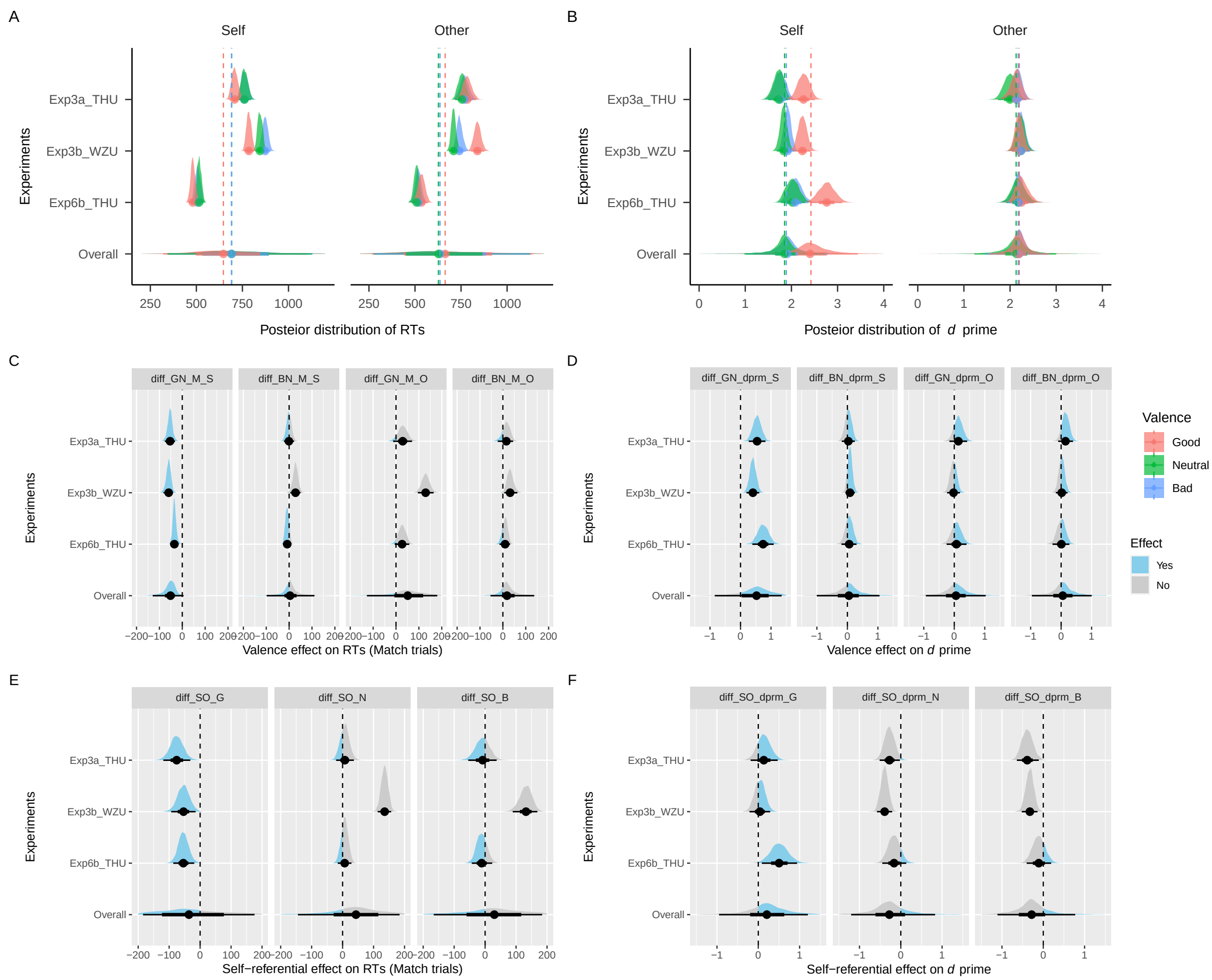

Figure 2. Interaction between moral valence and self-referential

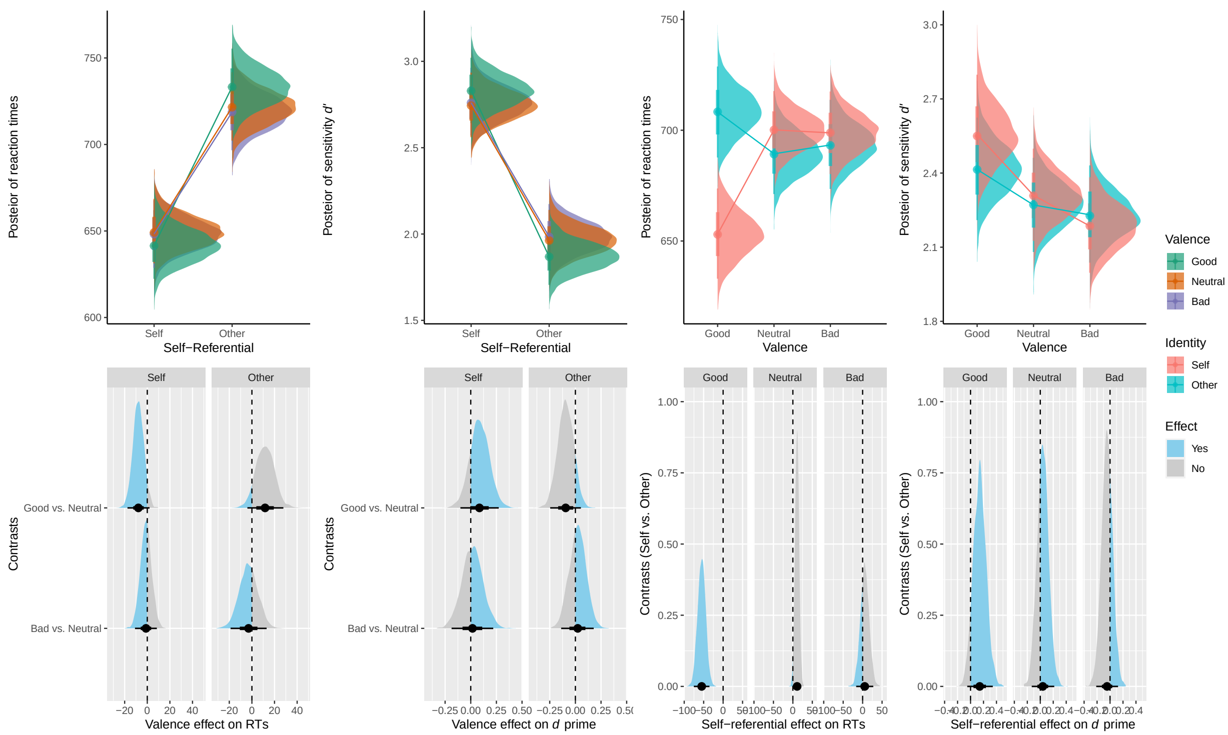

Figure 3. exp4: Results of Bayesian GLM analysis.