Oxytocin modulates social value representations in the amygdala

Yunzhe Liu[1,2,3], Shiyi Li[1,2,3#], Wanjun Lin[1,2,3#], Wenxin Li[4#], Xinyuan Yan[1,2,3], Xuena Wang[4], Xinyue Pan[4], Robb B. Rutledge[5,6], Yina Ma[1,2,3]*

1. State Key Laboratory of Cognitive Neuroscience and Learning Beijing Normal University, Beijing, China
2. IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China
3. Beijing Key Laboratory of Brain Imaging and Connectomics, Beijing Normal University, Beijing, China
4. School of Psychological and Cognitive Sciences, Peking University, Beijing, China
5. Wellcome Centre for Human Neuroimaging, University College London, London, UK
6. Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

[#] These authors contributed equally to this study

* Correspondence should be addressed to Yina Ma, Ph.D.
State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University,
19 Xin Jie Kou Wai Da Jie, Beijing, 100875, China
Phone/Fax: 8610-5880-2846
Email: yma@bnu.edu.cn

ABSTRACT [149 words]

Humans exhibit considerable variation in how they value their own interest relative to the interests of others. Deciphering the neural codes representing potential rewards for self and others is crucial for understanding social decision-making. Here we integrate computational modeling with fMRI to investigate the neural representation of social value and the modulation by oxytocin, a nine-amino acid neuropeptide, in participants evaluating monetary allocations to self and other (self-other allocations). We found that an individual's preferred self-other allocation serves as a reference-point for computing the value of potential self-other allocations. In more-prosocial participants, amygdala activity encoded a social-value-distance signal, i.e. the value dissimilarity between potential and preferred allocations. Intranasal oxytocin administration amplified this amygdala representation and increased prosocial behavior in more-individualist participants but not in more-prosocial ones. Our results reveal a neurocomputational mechanism underlying social-value representations and suggest that oxytocin may promote prosociality by modulating social-value representations in the amygdala.

INTRODUCTION

Humans live in complex social environments and rely heavily on social reciprocity. Many of our important decisions are made in social contexts where the costs and benefits to both ourselves and other people need to be considered[1]. Deciphering the neural codes that represent potential rewards to oneself and others is crucial for understanding social reciprocity and social decisions[2]. Recent studies of social decision-making find that people are rarely purely self-centered or altruistic: they care about both themselves and others' interests, but with considerable individual variation in how they weigh equity of self-other gain[3] and cooperation with others[3,4] during their decision-making. Individuals with prosocial preferences tend to prefer allocations considering the interests of both self and other and often seek to minimize the self-other difference (henceforth prosocials). In contrast, individuals with selfish preferences tend to maximize resources for themselves and generally prefer self-centered allocations (henceforth individualists).

Individual differences in social preference may stem from individual variation in preferred social allocations and differences in neural representations of potential relative to preferred allocations[5-7]. It remains unclear how the difference between potential and preferred self-other allocations is computed and represented in the brain and how these computations and neural representations are related to social decision-making. Here, we propose that the preferred self-other allocation (i.e., what an individual hopes the allocation will be) serves as a social reference-point against which potential allocations are represented and that quantity can guide social value-based decisions (social reference model). The deviation from the preferred allocation generates an "error" signal that could drive adaptive actions to reduce the size of the deviation. In much the same way as reward prediction errors[8] represent differences between expected and actual rewards and provide a basis for value-based decisions[9,10], this social error signal could represent deviation from the preferred allocation and serve as a basis for value-based decisions in the social domain.

The amygdala, with a large number of oxytocin and dopamine receptors[11,12] and strongly implicated in social cognition and social decision-making[3,13], is a prominent candidate to encode deviations from a social reference-point. Recent studies have

shown that amygdala activity tracks the subjective values of rewards and punishments[14] and reflects individual preferences[15]. Notably, the amygdala has been suggested to encode error signals that represent the differences between expectations and outcomes[9], a quantity that is fundamental for value-based decision-making[2]. Amygdala activity has also been shown to encode "aversive" signals to absolute inequality when evaluating reward pairs for self and other[3] and in response to dishonest behavior[16]. One untested possibility is that amygdala activity encodes the deviation of a potential self-other allocation from a reference-point that depends on individual-specific social preferences.

Prosocials and individualists, who should differ in their social reference-point, would be expected to engage different neural substrates for social value representations. For prosocials, the distance from their social reference-point could signal deviation from normative social principles such as inequity aversion, which is associated with the amygdala[3]. On the other hand, individualists employing a self-interest maximizing strategy in their decision-making could represent deviation from this reference-point mainly as conflict with self-interest, engaging lateral prefrontal regions associated with inhibition of self-interest and self-other allocation trade-off[17], such as lateral orbitofrontal cortex (lOFC)[18,19] and dorsolateral prefrontal cortex (dlPFC)[20].

It has also been suggested that individual differences in social behavior result in part from differences in the neuromodulatory regulation of neural circuits[6]. The neuropeptide oxytocin, an evolutionarily conserved hormone, is a potential candidate[21]. Oxytocin has been found to play an important role in social interaction and social decision-making[22], including promoting social motivation[23], increasing trust and cooperation with own-group members[24], and reducing social distance[25]. Whether and how oxytocin modulates the basic computations of social preference and social value representations remains largely unexplored. Individual differences in social preferences in nonhuman primates have been shown to be due in part to oxytocinergic regulation of amygdala-related neural circuits[6,7]. In nonhuman primates, exogenous inhaled oxytocin promotes social donation behavior[26] and focal infusion of oxytocin into the amygdala significantly increases prosocial decisions[7]. In humans, it has been suggested that individual differences in oxytocin effects are adaptive depending on an

individual's social disposition[21] such that intranasal oxytocin produces stronger effects on cooperation in less socially proficient individuals[4]. Oxytocin differentially impacts cooperative and aggressive choices in individuals with different pre-existing beliefs in prosociality[27]. We therefore predicted differential effects of oxytocin on regulating prosocial behavior between prosocials and individualists by selectively increasing prosociality in individualists via amplification of amygdala social value representations.

Here, we set out to test whether intranasal administration of oxytocin differentially modulates the neural representation of social values in prosocials and individualists performing a monetary outcome-pair evaluation task during fMRI scanning in a double-blind placebo-controlled between-subjects design. We first show that our social reference model most parsimoniously explains behavior consistent with social values being encoded as distance to an individual-specific reference-point. While prosocials represent social values relative to a more prosocial reference-point than individualists, oxytocin selectively increased prosociality in individualists and not prosocials in both competitive and non-competitive contexts. Moreover, these findings were replicated in two additional behavioral experiments. Using model-based fMRI analysis, we found that under placebo, amygdala activity in prosocials encodes a social value distance reflecting the degree to which a potential self-other allocation deviates from an individual-specific reference-point. Oxytocin selectively amplified the neural representation of social values in the amygdala in individualists, suggesting a link between oxytocin and prosociality via modulation of social value representations in the amygdala.

RESULTS

Experimental settings

In the fMRI experiment, we first invited participants (n=282) to a behavioral session to identify their social dispositions (i.e., prosocials vs. individualists) using the triple dominance[5] and social value orientation (SVO)[28] decision-making tasks (*Supplementary Fig. 1*). Thereafter, eligible prosocials and individualists (n=127) administered oxytocin or placebo performed a monetary outcome-pair evaluation task

during fMRI scanning (**Fig. 1a**). On each trial, participants were presented with pairs of monetary outcomes for himself and another participant (referred to as the partner) and evaluated his preference on each pair.

All monetary allocations were evenly sampled on the circumference of a circle centered at the origin (0, 0) in the Cartesian coordinate space spanned by social values (with monetary outcomes for self as the x-axis and outcomes for the partner as the y-axis, radius=5, **Fig. 1a**). Monetary outcomes for oneself and the partner define an angle θ, which samples the space from -90° to 180°. The angle between any two potential allocations is both necessary and sufficient to quantify their relationship. Both positive and negative values were included for a comprehensive investigation of social value representations, except for those in the third quadrant due to invariant preference rating shown in an independent sample (Methods).

We also ran two additional behavioral experiments, one experiment with a large sample (behavioral online-replication experiment, n=315) providing a replication for our finding that the social reference model outperforms other models and one experiment providing a replication of the oxytocin effect (oxytocin-replication experiment, n=80 males, within-subjects design, 40 prosocials and 40 individualists). To improve the ability to distinguish between different models, both additional experiments were run on a modified design where monetary pairs were sampled on 3 circles of different circumference (radius=5, 6, 9), with θ ranging from -90° to 180° with different intervals (5°, 17°, 23°). These specifications were identified based on model recovery analysis, which suggested that the combination of these parameters would lead to maximal discriminability between our social reference model and an inequality aversion model. The task design for these additional experiments was otherwise identical to the fMRI experiment.

Representing social values according to an individual-specific reference-point

In the fMRI experiment, we first plotted z-scored preference ratings for each social allocation across participants for visualization purposes (**Fig. 1b**). In general, participants most preferred self-gain/other-gain pairs and least preferred self-loss/other-gain pairs, suggesting that participants considered the interests of both self and partner. Based on preference ratings for all allocations, we computed an

individual-specific *reference-point*, referred to as φ. The principle of φ calculation was consistent with the "mean orientation" measure in a map-like structure[29,30] (**Fig. 1c**). The degree of φ indicated how much a participant preferred the potential outcome for the partner in relation to himself, with larger angles corresponding to stronger preference for allocations that benefit the partner relative to oneself and thus greater prosociality.

We then calculated cosine similarity between each allocation θ and the individual-specific reference-point φ to compute *social value distance*, the dissimilarity distance of that allocation to the participant's preferred allocation (also the deviation from the social reference-point, calculated as 1-cosine(θ-φ)). This measurement allowed us to quantify the difference between the second (self-loss/other-gain pairs) and fourth (self-gain/other-loss pairs) quadrants (**Fig. 1c**), which is not feasible when only including gains or using absolute value differences[3,28]. The *social reference model* was consistently the most parsimonious model across all studies: the fMRI during-scan experiment, the post-scan behavioral experiment, the behavioral online-replication experiment and the oxytocin-replication experiment (supported by model comparisons using variational free energy as the model selection criteria, *Supplementary Fig. 2*).

More prosocial reference-points for social value representations in prosocials

We quantified the difference in the estimated reference-point between prosocials and individualists under placebo. We found significantly higher values of φ in prosocials than individualists ($F(1,59)=33.49$, $p=2.91\times10^{-7}$, $\eta^2=0.36$, **Fig. 2a**), which was replicated in the large sample online experiment ($F(1,313)=92.14$, $p=2.71\times10^{-19}$, $\eta^2=0.23$, **Fig. 2b**). Moreover, in the online-replication experiment, the social reference-point derived from the social reference model was correlated with individuals' SVO scores ($r=0.55$, $p=4.38\times10^{-26}$, 95% CI=[0.47, 0.62], **Fig. 2c**), suggesting a more prosocial reference-point in prosocials both at a group and individual level. The pattern of more prosocial reference-points in prosocials than individualists was similarly observed in the competitive context in the post-scan experiment ($F(1,59)=12.59$, $p=7.69\times10^{-4}$, $\eta^2=0.18$, **Fig. 2d**) and in the oxytocin-replication experiment ($F(1,78)=28.27$, $p=9.78\times10^{-7}$, $\eta^2=0.27$, **Fig. 2e**).

Under placebo, φ was significantly correlated with independent measures of previously established prosocial behavior (Methods), with positive correlations between φ and the amount of contribution in a public goods game (r=0.52, p=2.44×10^{-5}) and in a dictator game (r=0.44, p=0.0006). The degree of φ was also correlated with the degree of absolute inequality aversion (which reflected a general preference for fairness and resistance to inequalities, with higher values indicating higher inequality aversion, r=0.65, p=1.54×10^{-8}), measured in an independent task[28](Methods). Note that a φ of 45° indicates a preference for equal offers. Within a fairly small range of φ corresponding to most of our participants (-30 to 45°), the larger the φ, the more prosocial a participant.

It has been suggested that decision time reflects perceived conflicts between prior expectation and current choice, with faster responses for preferred and less conflicted choices[31,32]. Thus, we would expect faster decision-making when conflict is minimal and the potential allocation is close to the individual-specific reference-point. As the social value distance increased, longer decision times were predicted. We indeed found that decision time for a potential allocation increased as a function of deviation from individual-specific reference-point, and such correlation was stronger in individualists than prosocials under placebo (independent-samples t-test on the Fisher z-scored correlation coefficients, individualists vs. prosocials: 0.18±0.03 vs. 0.09±0.02; t(59)=2.33, p=0.023 in the fMRI experiment, with a similar trend in the oxytocin-replication experiment, paired-samples t-test: t(78)=1.86, p=0.067, *Supplementary Fig. 3*). The greater the dissimilarity between potential and preferred allocations, the longer individualists took to evaluate potential allocations.

Selective oxytocin effects on promoting prosociality in individualists

We evaluated the oxytocin effect on the social value representations. First, we checked the relationship between baseline salivary oxytocin and social value representations across all participants. The individual-specific reference-point φ was independent of baseline salivary oxytocin (r=0.03, p=0.75) (*Supplementary Fig. 4*). We also measured participants' social perceptions of their partner by rating the first impression, likeability

and attractiveness of the partner. There was no significant difference across all groups on any of these measures (*Supplementary Fig. 5*). Therefore, any significant effect of Social Disposition and/or Treatment on the social value representation cannot be attributed to baseline oxytocin or social perception differences.

We conducted ANOVA on φ, with Social Disposition (prosocial vs. individualistic) and Treatment (oxytocin vs. placebo) as between-subjects factors. There was a significant main effect of Social Disposition ($F(1,121)=28.09$, $p=5.29\times10^{-7}$, $\eta^2=0.19$), with prosocials (vs. individualists) using a more prosocial reference-point to evaluate potential allocations. Interestingly, we found a significant Social Disposition x Treatment interaction ($F(1,121)=6.35$, $p=0.013$, $\eta^2=0.05$, **Fig. 2a**), as intranasal oxytocin significantly increased the reference-point φ towards a preference for more prosocial allocations in individualists (independent-samples *t*-test, $t(57)=2.21$, $p=0.031$), but not in prosocials ($t(64)=-1.54$, $p=0.13$, **Fig. 2a**), indicating that oxytocin selectively increased prosociality in individualists. Furthermore, there was no effect of scanning order or partner type on φ (*Supplementary Fig. 6*).

We replicated the selective oxytocin effect on promoting prosociality in individualists in the independent oxytocin behavioral experiment where we employed a within-subjects design and included monetary pairs sampled on 3 circles of different circumferences (Social Disposition: $F(1,78)=19.51$, $p=3.19\times10^{-5}$, $\eta^2=0.20$; Social Disposition x Treatment interaction: $F(1,78)=6.73$, $p=0.011$, $\eta^2=0.079$, **Fig. 2e**). Moreover, the within-subjects design, where each participant was invited to both oxytocin and placebo sessions, allowed us to examine whether the oxytocin effect varied as a function of individual scores in social value orientation. We expected a negative correlation between SVO scores and the oxytocin effect on prosociality, and indeed found a significant negative correlation between SVO scores and the size of oxytocin effect on social reference-point ($r=-0.23$, $p=0.041$, **Fig. 2f**), suggesting that the more selfish the individual, the stronger the effect of oxytocin on promoting a prosocial reference-point.

Finally, to determine whether the lack of oxytocin effect on prosocials was due to a ceiling effect (i.e., prosocials already care about others' outcomes), we introduced a competitive social context in the post-scan behavioral task where self-interest and other-interest were in direct competition. In the *competitive* context, we framed the

payoff in a "winner takes all" manner, so that one would be motivated to make selfish decisions to gain more than the partner. If oxytocin can promote prosociality in prosocials, which is masked by a ceiling effect in the *non-competitive* setting, we would expect oxytocin to affect prosocials in the *competitive* context. We conducted an ANOVA on φ in prosocials, with Treatment as between-subjects factor and Context (competitive vs. non-competitive) as within-subjects factor. There was a significant main effect of Context ($F(1,64)=68.61$, $p=1.02\times10^{-11}$, $\eta^2=0.52$), but no Treatment effect ($F(1,64)=0.856$, $p=0.358$), or interaction with Context ($F(1,64)=1.33$, $p=0.254$), suggesting that the lack of a prosocial effect of oxytocin in prosocials was not due to a ceiling effect given their relative "self-centered" social value preference in the *competitive* compared to *non-competitive* context. A similar ANOVA on individualists showed that oxytocin increased prosociality for individualists across both competitive and non-competitive contexts (Treatment: ($F(1,57)=8.56$, $p=0.005$, $\eta^2=0.131$; Treatment x Context: $F(1,57)=0.018$, $p=0.894$). Furthermore, the ANOVA on φ (**Fig. 2a, d**), with Social Disposition and Treatment as between-subjects factors and Context as a within-subjects factor, revealed the expected significant main effect of Context ($F(1,121)=145.92$, $p=1.59\times10^{-22}$, $\eta^2=0.55$), with decreased prosociality in the competitive context. There was a significant Social Disposition x Treatment interaction ($F(1,121)=5.28$, $p=0.023$, $\eta^2=0.042$). Moreover, this interaction was not affected by Context ($F(1,121)=0.697$, $p=0.406$), suggesting that the oxytocin effect on promoting prosociality was selective to individualists across multiple contexts.

<u>Amygdala in prosocials represents a social value distance signal</u>

Based on the behavioral model, we looked for brain regions that encoded the social value distance between potential and preferred allocations. We created a parametric modulator for the social value distance[29,30]: $1-\cos(\theta(t)-\varphi)$, based on our *social reference model*, where $\theta(t)$ is the angle of a current allocation at trial t and φ is the individual-specific reference-point. This measure reflects the degree to which an allocation deviates from the reference-point, with higher values indicating greater distance (i.e., lower desirability) between potential and preferred allocations. At the second-level analysis, we found that the posterior cingulate cortex (peak coordinate in MNI space: -6/-64/42) and middle frontal gyrus (peak coordinate in MNI space: -34/8/54) encoded social value distance when collapsing over the four groups

(whole-brain significant at a voxel-wise threshold $p<0.001$ and a cluster-wise FWE correction with $p<0.05$).

We then searched for brain regions that encoded the social value distance respectively for individualists and prosocials. We found that, under placebo, amygdala activity encoding social value distance was significantly stronger in prosocials than individualists (voxel-wise threshold $p<0.001$ and a cluster-wise FWE correction with $p<0.05$, peak MNI coordinate: 20/-10/-12, **Fig. 3a**, *Supplementary Table 1*). Previous studies have linked inhibition of self-interest and top-down control of selfish behavior with right lateral prefrontal cortex[17], such as right lOFC[18,19] and right dlPFC[20]. We hypothesized that individualists would represent deviation from preferred allocations mainly as a conflict with self-interest and a social value distance representation might be present in these areas. Comparison of individualists and prosocials under placebo revealed that right lOFC activity encoded social value distance to a greater degree in individualists than prosocials (voxel-wise threshold $p<0.001$, small volume correction $p<0.05$ for an anatomically defined right lOFC mask, using combined connectivity-based parcellations 8-11 covering right lOFC[33], **Fig. 3b**). This relationship was not present for right dlPFC. Further region of interest (ROI) analysis revealed a significant interaction between brain areas (right lOFC vs. amygdala) and Social Disposition (individualist vs. prosocials) under placebo ($F(1, 58)=11.210$, $p=0.0014$, $\eta^2=0.162$). Here, the amygdala and right lOFC ROIs employed anatomically defined masks.

We extracted beta estimates associated with encoding social value distance from an anatomically defined amygdala ROI. We found that the strength of the amygdala social value distance representation was correlated with the degree of inequality aversion ($r=0.296$, $p=0.0012$, *Supplementary Fig. 7*), whereas right lOFC activity bore no relationship to inequality aversion ($r=-0.13$, $p=0.59$). A moderation analysis revealed that this positive correlation was significantly stronger in prosocials than individualists under placebo (R^2 change=0.06, $p=0.003$, *Supplementary Fig. 7*).

Oxytocin modulates social value representations in the amygdala in individualists

We then searched for the main effect of Treatment and the interaction effect of Social Disposition and Treatment in the whole brain. There was no significant main effect of Treatment. The Social Disposition x Treatment interaction *F* contrast revealed a significant cluster in the amygdala (**Fig. 4a**, peak voxels in the right amygdala survived voxel-wise FWE correction: $p<0.05$). Intranasal oxytocin selectively amplified the neural representation of social value distance in the amygdala of individualists, but not prosocials. A similar interaction pattern was found in other brain regions, including the right temporoparietal junction (TPJ) and ventral striatum (*Supplementary Fig. 8*). Moreover, as illustrated in *Supplementary Fig. 9a-d*, amygdala activity increased as a function of deviation from an individual-specific reference-point in prosocials under placebo (slope estimate of the linear fit=0.222, $p=0.001$) and this pattern was not found under oxytocin (slope estimate=0.010, $p=0.88$). In contrast, amygdala activity increased as a function of deviation from an individual-specific reference-point in individualists under oxytocin (slope estimate=0.232, $p=0.003$) and this pattern was not found under placebo (slope estimate=0.042, $p=0.50$). Amygdala responses were not related to absolute value differences or deviations from the allocentric reference (*Supplementary Fig. 9e-l*, all $p>0.5$).

We then examined the relationship between neural responses and evaluations of monetary allocations on a trial-by-trial basis to test whether the amygdala or right lOFC activity explained trial-by-trial variation in subjective preference ratings that was independent of the predicted social value distance. At the first-level (individual-subject-level) analysis, we modelled each trial separately and extracted beta estimates for amygdala and right lOFC for each trial. We then regressed the trial-by-trial amygdala and right lOFC responses (as x in the regression) respectively onto the evaluation made for each monetary allocation on each trial (as y in the regression), while controlling for the deviation of each potential allocation from the individual-specific reference-point. In doing so, we ensure that the beta estimates associated with the trial-by-trial amygdala and right lOFC activity reflect unique variance in predicting preference ratings based on amygdala and right lOFC responses above and beyond variance explained by predicted social value distance. We conducted a Social Disposition-by-Treatment ANOVA on the trial-by-trial correlation coefficient and found a significant interaction between Social Disposition and Treatment ($F(1,115)=6.722$, $p=0.011$, $\eta^2=0.057$, **Fig. 4b**). The amygdala responses in prosocials

362 negatively predicted trial-by-trial preference ratings under placebo, which was reduced
363 by oxytocin. In contrast, amygdala activity negatively predicted trial-by-trial
364 preference ratings in individualists under oxytocin vs. placebo. The negative
365 correlation indicated that the stronger the amygdala activity encoding social value
366 distance, the lower the preference. This pattern of results was consistent with the
367 amygdala providing an input signal for preferences, and fluctuations in the amygdala
368 responses can explain trial-by-trial deviations from average preferences. No such
369 results were found for right lOFC activity.

370 Furthermore, we conducted a general linear model (GLM) with preference rating for
371 each monetary allocation as a parametric modulator to identify any neural activity
372 sensitive to subjective preference ratings. We found activity in the medial prefrontal
373 cortex (mPFC) and lOFC, brain regions typically associated with value-coding[34,35],
374 correlated with subjective preference ratings for monetary allocations collapsing across
375 the four groups. Moreover, there was a significant interaction between Social
376 Disposition and Treatment for the mPFC and lOFC activity that encoded preference
377 ratings (height threshold $p<0.001$, cluster-based FWE correction, $p<0.05$;
378 ***Supplementary Fig. 10***). No significant Social Disposition x Treatment interaction was
379 found in the amygdala encoding the preference rating at the whole-brain or ROI level,
380 suggesting that the amygdala activity encoding social value distance does not simply
381 reflect the reverse of the preference signal.

382 We performed a generalized psycho-physiological interaction (gPPI) analysis with
383 anatomically defined bilateral amygdala as the seed region at the whole-brain level. We
384 found that amygdala activity encoding social value distance was coupled with ventral
385 mPFC activity. Moreover, amygdala-vmPFC coupling was significantly stronger in
386 prosocials than in individualists under placebo (height threshold $p<0.001$, uncorrected,
387 ***Supplementary Fig. 11***). Further, the ROI analysis suggested that oxytocin increased
388 the strength of functional connectivity between amygdala and vmPFC in encoding
389 social value distance in individualists (independent-samples *t*-test, $t(54)=2.69$, $p=0.009$)
390 but not in prosocials ($t(58)=0.067$, $p=0.95$, ***Supplementary Fig. 11***). Given that the
391 vmPFC is typically associated with value computation and value-guided choice[34,35],
392 these results may suggest a potential amygdala pathway linked to social preferences
393 and social value-based decisions, which can be modulated by oxytocin in

394 individualists.

395 **DISCUSSION**

396 Our results suggest that the representation of social values is a relational map that
397 encodes the distance between potential values for oneself and others on the same
398 coordinate system. This representation can guide how we interact with others and how
399 we respond to perceived unfairness. We provide empirical evidence that social value
400 representations are constructed in relation to individual-specific social preferences,
401 with the distance between potential and preferred allocations determining the value of
402 social allocations. Prosocials represent social values relative to a more prosocial
403 reference-point than individualists, even in a competitive social context where self- and
404 other-interest are in direct competition. Moreover, the social reference-point derived
405 from our *social reference model* accounts for individual variation in prosocial
406 behaviors (e.g., cooperation, generosity and inequality aversion) and therefore could
407 serve as a compact description of social decision-making.

408 This dissimilarity distance measure bears some similarity to variables in value-based
409 decision-making frameworks[2]. Social value distance is encoded by the amygdala in
410 prosocials: the more dissimilar a potential allocation to the individual-specific
411 reference-point (i.e., their preferred allocation), the greater the amygdala response. Our
412 results offer a mechanistic account of how social value representations contribute to
413 decision-making in prosocials. Trial-by-trial amygdala activity encoding social value
414 distance reflects how attractive potential social allocations are judged to be by
415 prosocials (i.e., the stronger the amygdala response, the less attractive the allocation).
416 Our control analyses show that amygdala activity is better explained by
417 individual-specific reference-points than by egocentric or allocentric frames of
418 reference (Methods). Thus, amygdala activity might encode the difference between
419 potential and preferred allocations (i.e., a "surprise" signal) much like dopamine firing
420 represents reward prediction errors reflecting the difference between outcomes and
421 expectations[10].

422 We found an amygdala representation of social value distance that reflects a deviation
423 from the most preferred allocation (what an individual hopes the allocation to be). This
424 result is consistent with studies in both nonhuman primates[7] and in human

neuroimaging studies[16] suggesting that the amygdala represents how undesirable an outcome is. Providing further support, we found that trial-by-trial amygdala activity was negatively correlated with the desirability of potential social allocations. Moreover, we found evidence that social value distance is also encoded by neural responses in the ventral striatum and TPJ in prosocials, consistent with previous findings linking prosocial decisions with several hubs in the social brain network[7,36], including the amygdala, ventral striatum and TPJ. For example, TPJ activity encodes the subjective value of altruistic choice and the value of generosity[36-38].

We also found some evidence for a social value distance representation in the right lOFC in individualists relative to prosocials, which may reflect a distinct coding scheme for representing social values, although this activity did not survive whole-brain cluster-level correction and was not predictive of trial-by-trial preference ratings or modulated by oxytocin. However, this pattern is consistent with previous studies related to right lOFC function[39]. Right lOFC activation is associated with inhibition of self-interest, reward-guided decisions and detecting and evaluating threats to self-interest[18,19]. The right dlPFC, another region implicated in social decision-making and top-down control of selfish behavior[20], did not have any significant social value distance representation in either individualists or prosocials.

Taken together, individuals may consider or calculate self-interest, altruistic values and evaluation of threat to self-interest when comparing potential and preferred allocations. These processes differ among individuals with different social orientations. Social value distance signals in prosocials may also relate to mentalizing about the needs of others[36,37], integrating social information into estimates of subjective value[40], and calculating altruistic values[41] in the social brain network. However, it is possible that individualists perceive deviations from their preferred allocation mainly as conflicts with self-interest, consistent with studies related to inhibition of self-interest conflict and evaluating threats to self-interest[18,19].

Oxytocin is believed to facilitate social approach and to increase the salience of social cues in promoting adaptive social behaviors[21]. While individualists focus on self-interest and personal goals when making decisions[3,4], oxytocin may increase prosociality by shifting reference-points to more prosocial allocations and increasing the weights of outcomes for others, possibly through amplifying the amygdala

representation of social value distance. This is consistent with studies showing that the amygdala plays a critical role in allocating attention to other people[7] and in integrating social information[42] and social emotions[13] into decision-making. However, oxytocin fails to show a prosocial effect in prosocials. This does not necessarily mean that oxytocin makes prosocials greedy, as we found that prosocials still have greater prosociality than individualists under oxytocin. This is also not likely to be a ceiling effect, as there is no oxytocin effect on increasing prosociality even when prosocials employ a more "self-centered" reference-point in a competitive context.

We found that oxytocin significantly reduces the strength of amygdala social value distance representations in prosocials and this was associated with a trend towards reduced prosociality. More sensitive changes in neural responses have often been observed in previous studies of prosocial behavior[43]. In the current study, one possible account is that social desirability or social pressure prevents prosocials from engaging in more self-centered performance. Although oxytocin significantly reduces amygdala representations of social values, consideration of both reputation[44] and others' approval[45] may prevent neural effects from translating into explicit changes in behavior. Finally, in the within-subjects oxytocin replication experiment, we showed that the oxytocin effect on shifting the social reference-point towards greater prosociality varied as a function of an individual's disposition of social value orientation, suggesting that the dichotomous comparison in the between-subjects design may prevent identification of an effect that depends on individual disposition.

Oxytocin has been implicated in many social behaviors, from promoting trust, generosity and cooperation[21,22,46,47] to aggravating mistrust and aggressive behavior[48,49]. The variable nature of oxytocin effects on prosociality is increasingly recognized. Seemingly contradictory oxytocin effects may be moderated by poorly understood individual and contextual differences[21,24]. Our finding of distinct oxytocin effects on the social reference-point for prosocials and individualists provides evidence for the underlying computational and neural basis for oxytocin's effect on prosociality and helps reconcile conflicting results in the literature. A concern in previous studies is that post hoc explorations of different modulations by individual differences risk inflating the rate of Type I errors[50]. The current study examined only a single a priori specified modulator as we screened participants for social disposition before the oxytocin

experiment, which allowed us to specifically test for different effects of oxytocin in prosocials and individualists. We consistently found across multiple studies that the selective effect of oxytocin on promoting prosociality in individualists was present in both competitive and non-competitive contexts, in both within- and between-subjects designs and with different experimental task designs.

Taken together, our results reveal a neural mechanism that underlies social value representations, providing new insight into the processes that influence human social decisions. Our results demonstrate that oxytocin adaptively modulates social value representations in the amygdala and imply a fundamental role of oxytocin in social decision-making. These insights and the identification of a selective effect of oxytocin on prosociality in individualists may have implications for treating neuropsychiatric disorders with social deficits, including autism and sociopathy.

METHODS

Methods, including additional references, Nature Research reporting summaries, statements of data availability and any associated accession codes are available in the online version of the paper.

Acknowledgments. We thank H. Zhang, D. Wang, X. Zou and H. Wang for their assistance in data collection. We also thank M. Crockett and three anonymous reviewers for their constructive suggestions on an earlier draft of this manuscript. This work was supported by the National Natural Science Foundation of China (Projects 31722026;91632118; 31771204; 31661143039, to YM), the Fundamental Research Funds for the Central Universities (2017XTCX04; 2018EYT04, to YM), Open Research Fund of the State Key Laboratory of Cognitive Neuroscience, Beijing Normal University (to YM), startup funding from the State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University (to YM), and by the Max Planck Society and a MRC Career Development Award (MR/N02401X/1) to RBR.

Author Contributions. Y.M. conceived and designed the project, designed the fMRI experiment; Y.M. and Y.L. designed the replication experiments; W.Li., X.W., X.P.

and Y.M. performed the fMRI experiment; S.L. and X.Y. performed the replication experiments; Y.L., S.L., W.Lin. and Y.M. analysed the data and interpreted the results of the fMRI and behavioral experiments; and Y.L., W.Lin., R.B.R. and Y.M. wrote the paper.

Competing Interests Statement. The authors declare no competing interests.

References

1. Sanfey, A.G. Social decision-making: insights from game theory and neuroscience. *Science* **318**, 598-602 (2007).
2. Rangel, A., Camerer, C. & Montague, P.R. A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545-556 (2008).
3. Haruno, M. & Frith, C.D. Activity in the amygdala elicited by unfair divisions predicts social value orientation. *Nat. Neurosci.* **13**, 160-161 (2010).
4. Declerck, C.H., Boone, C. & Kiyonari, T. The effect of oxytocin on cooperation in a prisoner's dilemma depends on the social context and a person's social value orientation. *Soc. Cogn. Affect. Neurosci.* **9**, 802-809 (2014).
5. Van Lange, P.A. The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *J. Pers. Soc. Psychol.* **77**, 337 (1999).
6. Chang, S.W., *et al.* Neuroethology of primate social behavior. *Proc. Natl. Acad. Sci. USA* **110**, 10387-10394 (2013).
7. Chang, S.W., *et al.* Neural mechanisms of social decision-making in the primate amygdala. *Proc. Natl. Acad. Sci. USA* **112**, 16012-16017 (2015).
8. Will, G.-J., Rutledge, R.B., Moutoussis, M. & Dolan, R.J. Neural and computational processes underlying dynamic changes in self-esteem. *eLife* **6** (2017).
9. Gottfried, J.A., O'Doherty, J. & Dolan, R.J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104-1107 (2003).
10. Schultz, W., Dayan, P. & Montague, P.R. A neural substrate of prediction and reward. *Science* **275**, 1593-1599 (1997).
11. Boccia, M.L., Petrusz, P., Suzuki, K., Marson, L. & Pedersen, C.A. Immunohistochemical localization of oxytocin receptors in human brain. *Neuroscience* **253**, 155-164 (2013).
12. Haber, S. N., & Knutson, B. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacol*, **35**, 4 (2010).
13. Bickart, K. C., Dickerson, B. C., & Barrett, L. F. The amygdala as a hub in brain networks that support social life. *Neuropsychologia* **63**, 235-248 (2014).
14. Paton, J.J., Belova, M.A., Morrison, S.E. & Salzman, C.D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865-870 (2006).

559 15. De Martino, B., Camerer, C.F. & Adolphs, R. Amygdala damage eliminates
560 monetary loss aversion. *Proc. Natl. Acad. Sci. USA* **107**, 3788-3792 (2010).
561 16. Garrett, N., Lazzaro, S.C., Ariely, D. & Sharot, T. The brain adapts to dishonesty.
562 *Nat. Neurosci.* **19**, 1727-1732 (2016).
563 17. Knoch, D., & Fehr, E. Resisting the power of temptations. *Ann. N. Y. Acad.*
564 *Sci.* **1104**, 123-134 (2007).
565 18. Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. & Fehr, E. The neural
566 signature of social norm compliance. *Neuron* **56**, 185-196 (2007).
567 19. Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., & Murray, E. A.
568 Prefrontal mechanisms of behavioral flexibility, emotion regulation and value
569 updating. *Nat. Neurosci.* **16**, 1140 (2013).
570 20. Haruno, M., Kimura, M., & Frith, C. D. Activity in the nucleus accumbens and
571 amygdala underlies individual differences in prosocial and individualistic economic
572 choices. *J. Cogn. Neurosci.* **26**, 1861-1870 (2014).
573 21. Ma, Y., Shamay-Tsoory, S., Han, S. & Zink, C.F. Oxytocin and social adaptation:
574 Insights from neuroimaging studies of healthy and clinical populations. *Trends. Cogn.*
575 *Sci.* **20**, 133-145 (2016).
576 22. Zink, C.F. & Meyer-Lindenberg, A. Human neuroimaging of oxytocin and
577 vasopressin in social cognition. *Horm. Behav.* **61**, 400-409 (2012).
578 23. Alvares, G.A., Hickie, I.B. & Guastella, A.J. Acute effects of intranasal oxytocin
579 on subjective and behavioral responses to social rejection. *Exp. Clin.*
580 *Psychopharmacol.* **18**, 316 (2010).
581 24. Ma, Y., Liu, Y., Rand, D.G., Heatherton, T.F. & Han, S. Opposing oxytocin
582 effects on intergroup cooperative behavior in intuitive and reflective minds.
583 *Neuropsychopharmacol* **40**, 2379-2387 (2015).
584 25. Scheele, D., *et al.* Oxytocin modulates social distance between males and females.
585 *J. Neurosci.* **32**, 16074-16079 (2012).
586 26. Chang, S.W., Barter, J.W., Ebitz, R.B., Watson, K.K. & Platt, M.L. Inhaled
587 oxytocin amplifies both vicarious reinforcement and self reinforcement in rhesus
588 macaques (Macaca mulatta). *Proc. Natl. Acad. Sci. USA* **109**, 959-964 (2012).
589 27. Lambert, B., Declerck, C.H., Boone, C. & Parizel, P.M. A functional MRI study
590 on how oxytocin affects decision making in social dilemmas: Cooperate as long as it

pays off, aggress only when you think you can win. *Horm. Behav.* **94**, 145-152 (2017).
28. Murphy, R.O., Ackermann, K.A. & Handgraaf, M. Measuring social value orientation. *Judgm. Decis. Mak.* **6**, 771-781 (2011).
29. Constantinescu, A.O., O'Reilly, J.X. & Behrens, T.E. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464-1468 (2016).
30. Doeller, C.F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657-661 (2010).
31. Evans, A.M. & Rand, D.G. Cooperation and decision time. *Curr. Opin. Psychol.* **26**, 67-71 (2018).
32. Krajbich, I., Bartling, B., Hare, T. & Fehr, E. Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nat. Commun.* **6**, 7455 (2015).
33. Neubert, F.X., Mars, R.B., Thomas, A.G., Sallet, J. & Rushworth, M.F. Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron* **81**, 700-713 (2014).
34. Lim, S. L., O'Doherty, J. P., & Rangel, A. The decision value computations in the vmPFC and striatum use a relative value code that is guided by visual attention. *J. Neurosci.* **31**, 13214-13223 (2011).
35. Jocham, G., Hunt, L. T., Near, J., & Behrens, T. E. A mechanism for value-guided choice based on the excitation-inhibition balance in prefrontal cortex. *Nat. Neurosci.* **15**, 960 (2012).
36. Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C. & Fehr, E. Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron* **75**, 73-79 (2012).
37. Hutcherson, C.A., Bushong, B. & Rangel, A. A Neurocomputational Model of Altruistic Choice and Its Implications. *Neuron* **87**, 451-462 (2015).
38. Strombach, T., *et al.* Social discounting involves modulation of neural value signals by temporoparietal junction. *Proc. Natl. Acad. Sci. USA* **112**, 1619-1624 (2015).
39. Kringelbach, M. L. The human orbitofrontal cortex: linking reward to hedonic experience. *Nat. Rev. Neurosci.* **6**, 691 (2005).
40. Ruff, C.C. & Fehr, E. The neurobiology of rewards and values in social decision making. *Nat. Rev. Neurosci.* **15**, 549 (2014).

41. Telzer, E. H., Fuligni, A. J., Lieberman, M. D., & Galvan, A. Neural sensitivity to eudaimonic and hedonic rewards differentially predict adolescent depressive symptoms over time. *Proc. Natl. Acad. Sci. USA* **111**, 6600–6605 (2014).
42. Mosher, C. P., Zimmerman, P. E., & Gothard, K. M. Neurons in the monkey amygdala detect eye contact during naturalistic social interactions. *Curr. Biol.* **24**, 2459-2464 (2014).
43. Hein, G., Morishima, Y., Leiberg, S., Sul, S. & Fehr, E. The brain's functional network architecture reveals human motives. *Science* **351**, 1074-1078 (2016).
44. Ariely, D., Bracha, A., & Meier, S. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *Am. Econ. Rev.* **99**, 544-55 (2009).
45. Bénabou, R., & Tirole, J. Incentives and prosocial behavior. *Am. Econ. Rev.* **96**, 1652-1678 (2006).
46. Aydogan, G.*, et al.* Oxytocin promotes altruistic punishment. *Soc. Cogn. Affect. Neurosci*. **12**, 1740-1747 (2017).
47. Yan, X., Yong, X., Huang, W., & Ma, Y. Placebo treatment facilitates social trust and approach behavior. *Proc. Natl. Acad. Sci. USA* **115**, 5732-5737 (2018).
48. Ne'eman, R., Perach-Barzilay, N., Fischer-Shofty, M., Atias, A. & Shamay-Tsoory, S.G. Intranasal administration of oxytocin increases human aggressive behavior. *Horm. Behav*. **80**, 125-131 (2016).
49 Radke, S. & De Bruijn, E. The other side of the coin: oxytocin decreases the adherence to fairness norms. *Front. Hum. Neurosci*. **6**, 193 (2012).
50. Nave, G., Camerer, C., & McCullough, M. Does oxytocin increase trust in humans? A critical review of research. *Perspect Psychol Sci* **10**, 772-789 (2015).

Figure legends.

Figure 1. Experimental design. (a) Participants were presented with monetary outcome pairs specifying potential amounts of money ("+" indicated gain and "–" indicated loss) received by themselves (labeled as self) and another player (labeled as other). Participants had 3 s to rate their preferences from 1 (least preferable) to 4 (most preferable) for each monetary outcome pair, followed by a 1-5 s jittered inter-trial interval. Monetary outcomes for the self and other define an angle θ, which samples the space from -90° to 180°. (b) Mean standardized preference ratings across all

participants were plotted against monetary allocations for the self and other. (c) Based on the preference ratings, we computed the "*reference-point*" in the social value representation φ, closely related to the preferred allocation that best accounts for the participant's social preferences over all allocations. Higher values of φ correspond to a stronger preference for allocations that benefit the partner relative to oneself, indicating greater prosociality.

Figure 2. Oxytocin boosts prosociality selectively in individualists. More prosocial reference-points for social value representations in prosocials than individualists in the fMRI during-scan experiment (**a**, 31 prosocials and 30 individualists, $p=2.91\times10^{-7}$, under placebo), in the online-replication experiment (**b**, n = 315, 160 prosocials and 155 individualists, $p=2.71\times10^{-19}$), in the fMRI post-scan experiment (**d**, in a competitive context via framing the payoff in a "winner takes all" manner, $p=7.69\times10^{-4}$, under placebo), and in the oxytocin-replication experiment (**e**, 40 prosocials and 40 individualists, $p=9.78\times10^{-7}$, under placebo). In the online-replication experiment, the estimated social reference-point was positively correlated with individual SVO scores: the more prosocial the disposition, the higher the φ (**c**, n = 315, Pearson's r = 0.55, $p = 4.38\times10^{-26}$). Moreover, intranasal oxytocin increased prosociality in individualists but not in prosocials, by moving their *reference-point* φ towards a preference for more prosocial allocations in the fMRI during-scan experiment (**a,** n = 125, individualists under placebo: n = 30 males, under oxytocin: n = 29 males; prosocials under placebo: n = 31 males, under oxytocin: n = 35 males, p = 0.013) and in the oxytocin-replication experiment (**e**, 40 prosocials and 40 individualists, p = 0.011). Moreover, in the oxytocin-replication experiment, SVO scores were negatively correlated with the effect of oxytocin on social reference-point (**f**, n = 80, Pearson's r = -0.23, p = 0.041). The individual specific reference-point $\varphi<0°$ indicates a preference for pairs with self-gain and other-loss; $\varphi=0°$ indicates a preference for self-gain without consideration of other's outcome; $0°<\varphi<90°$ indicates a preference for self-gain/other-gain pairs; $\varphi > 90°$indicates a preference for self-loss/other-gain pairs). Error bars represented standard error of the mean across participants within each group (*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$; n.s, not significant).

Figure 3. Amygdala activity in prosocials (n = 30 males) and right lOFC activity in individualists (n = 30 males) encode social value distance relative to an individual-specific reference-point. (**a**) Coronal view of activations in the amygdala that encode the social value distance, the difference between potential and preferred social allocations, was significantly stronger in prosocials than individualists under placebo ($P < 0.05$, FWE-corrected at the cluster level after voxel-wise thresholding at $P < 0.001$). (**b**) Axial view of right lOFC activity encoding social value distance to a greater degree in individualists than prosocials (voxel-wise threshold $p < 0.001$, small volume correction $p<0.05$ for an anatomically defined right lOFC mask). Independent-samples t test was used in (**a**) and (**b**) on the beta estimates from ROIs of amygdala ($t(58) = 2.75$, $p = 0.008$, 95% CI = [0.05, 0.33]) and right lOFC ($t(58) = -2.46$, $p = 0.017$, 95% CI = [-0.04, -0.004]). The amygdala and right lOFC ROIs employed anatomically defined masks (amygdala based on AAL bilateral anatomical mask, an anatomically defined right lOFC mask, using combined connectivity-based parcellations 8-11 covering right lOFC[33]).

Figure 4. Oxytocin promotes amygdala activity in representing social values in individualists ($n_{placebo} = 30$, $n_{oxytocin} = 26$ in individualists, $n_{placebo} = 30$, $n_{oxytocin} = 30$ in prosocials). (**a**) Coronal view of amygdala showing interaction effect between Treatment and Social Disposition (peak voxels in right amygdala, FWE-corrected p=0.02). Social Disposition-by-Treatment ANOVA showed significant interaction on the beta estimates from anatomically defined amygdala ($F(1, 112) = 12.536$, $p = 5.83\times10^{-4}$, $\eta^2 = 0.057$). (b) Trial-by-trial amygdala responses in predicting preference rating, after controlling for predicted social value distance for each trial. The Social Disposition-by-Treatment ANOVA on the trial-by-trial correlation coefficient showed a significant interaction between Social Disposition and Treatment ($F(1, 115) = 6.722$, $p = 0.011$, $\eta^2 = 0.057$). Error bars represented standard error of the mean across participants within each group, $*p < 0.05$ and $**p < 0.01$.

METHODS

Participants

For the oxytocin-fMRI and behavioral oxytocin-replication experiments, we recruited only male participants to avoid potential confounds of sex differences in oxytocin effects[21,51], consistent with previous studies examining oxytocin effects on social cognition. All participants had normal or corrected-to-normal vision and reported no history of neurological or psychiatric diagnoses, or medication, drug or alcohol abuse. Participants provided informed consent after the experimental procedure had been fully explained and were informed of their right to withdraw at any time during the study. The experimental protocol was in line with the standards of the Declaration of Helsinki and approved by the research ethics committee at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University (Beijing, China).

Oxytocin fMRI experiment. There were 282 male college students (mean age = 22.3 ± 2.12 years) that participated in this study as paid volunteers. Participant's disposition in social value orientation was measured in the behavioral session (149 prosocials and 83 individualists were identified). Among these, 127 participants were qualified and willing to participate in the fMRI experiment (at least 7 days after the behavioral session). Two participants (1.6%) were excluded due to technical issues during scanning, leaving 125 participants in the behavioral analysis (individualists under placebo: $n = 30$ males, mean age 22.2 ± 2.35 years, under oxytocin: $n = 29$ males, mean age 21.7 ± 2.39 years; prosocials under placebo: $n = 31$ males, mean age 22.1 ± 2.70 years, under oxytocin: $n = 35$ males, mean age 22.1 ± 2.70 years). An additional 9 participants (7.2%) were excluded from further fMRI analysis due to excessive head movement during scanning (> 3 mm), leaving 116 participants for fMRI data analysis. In the end, there were 60 prosocials including 30 administered placebo (mean age, 22.7 ± 2.61 years) and 30 administered oxytocin (mean age, 21.8 ± 2.38 years), and 56 individualists including 30 administered placebo (mean age, 23.0 ± 2.29 years) and 26 administered oxytocin (mean age, 22.5 ± 3.03 years) in the formal fMRI data analysis. Prosocials and individualists receiving oxytocin or placebo were matched on state and trait anxiety, depression, subjective well-being and happiness ratings (all $p > 0.05$ on both the main and interaction effects of Treatment and Social Disposition; *Supplementary Table 3*).

752 The sample size of the fMRI study was determined prior to data collection. We
753 conducted sample size estimation using G*Power 3.1[52] to determine the number of
754 participants sufficient to detect a reliable effect. Based on an estimated average
755 small-to-medium effect size of oxytocin effect on social behaviors (Cohen's d = 0.28)
756 [53]. 104 participants were needed to detect a significant effect ($\alpha = 0.05$, $\beta = 0.80$,
757 two-by-two mixed ANOVA interaction effects). We planned to recruit 125 participants
758 (assuming 10-20% participants would be removed from the fMRI data analysis due to
759 excessive head movement). In the end, we recruited 127 participants because the 41th
760 and 42th participants did not complete the experiment due to technical issues during
761 scanning. For comparison, we also considered the 58 oxytocin-fMRI studies published
762 at the time we initiated our experiment in June 2015, of which 23 employed
763 between-subject design recruiting healthy individuals. On average, the sample size was
764 50.89 in total, 25.82 for the placebo group and 25.47 for the oxytocin group. Thus, our
765 planned sample size of 125 participants was a decent sample size compared to the
766 average across oxytocin-fMRI studies. Moreover, the sample size of 116 participants
767 (after removal of subjects due to technical issues and excessive head movement) was
768 adequate to reveal reliable effects, exceeding the 104 participants needed for 80%
769 power.

770 *Oxytocin-replication experiment.* We conducted an additional behavioral experiment
771 for the replication of the oxytocin effect using a double-blind, randomized,
772 placebo-controlled, within-subjects crossover design. The sample size was
773 predetermined based on the effect size (Cohen's d = 0.45) from our original finding in
774 the fMRI study. The G*Power calculation suggested that 40 participants (20 for each
775 group) were required to detect a reliable effect ($\alpha = 0.05$, $\beta = 0.80$ for a within (oxytocin
776 vs. placebo)-between (prosocial vs. individualist) interaction). To obtain a better sense
777 of the robustness of the original findings, we doubled the estimated sample size, aiming
778 to enroll 40 participants per group, with corresponding power equal to 98%. We
779 replicated the selective oxytocin effects on promoting prosociality (i.e., φ) in
780 individualists in the whole sample (40 prosocials and 40 individualists), as well as in
781 the first 20 prosocials and 20 individualists (as estimated by the G*Power analysis).
782 140 males (mean age, 22.33 ± 3.35 years) were invited to a behavioral session to
783 identify their disposition in social value orientation. Among these, 82 participants were

784 qualified and willing to participate in the oxytocin experiment (at least 7 days after the
785 first behavioral session). Two participants did not show up for the second session. Thus,
786 80 participants (40 prosocials, mean age, 22.08 ± 3.47 years; 40 individualists; mean
787 age, 21.54 ± 2.42 years) were included in the final data analysis.

788

789 *Online-replication experiment.* We conducted an online experiment with a large
790 sample (n = 315, 132 males, 160 prosocials, mean age = 22.40 ± 3.27; 155
791 individualists, mean age = 22.48 ± 3.30) to provide a replication for our finding that the
792 social reference model outperforms other models. Prosocials and individualists did not
793 differ in their ratings on the first impression, likeability and attractiveness of the online
794 partner (independent-samples t test, impression: $t_{313} = 1.03$, p = 0.305; likeability: t_{313}
795 = 0.98, p = 0.328; attractiveness: $t_{313} = 0.73$, p = 0.465.

796

797 ***Procedure.***

798 Participants were first invited to the behavioral session to identify their social
799 disposition and be screened for eligibility of the fMRI and oxytocin behavioral
800 experiments. Participants recruited in the fMRI experiment were randomly assigned to
801 the intranasal administration of oxytocin or placebo in a double-blind
802 placebo-controlled between-subjects design. In the oxytocin experiment, participants
803 received either oxytocin or placebo intranasally in two separate sessions, with a
804 5-7-day washout period between two sessions. The order of oxytocin and placebo
805 treatment was counterbalanced across participants. All participants were instructed to
806 abstain from cigarette, alcohol and caffeine during the 24 hours prior to the experiment,
807 and to refrain from eating or drinking anything except water for 2 hours before the
808 experiment. Participants self-administrated oxytocin or placebo 35 min[54] before the
809 main task, i.e., a monetary outcome-pair evaluation task (a revised one with monetary
810 pairs sampled on 3 circles of different circumference was used in the
811 oxytocin-replication experiment).

812 *Social disposition measurements.* In the fMRI and the oxytocin-replication
813 experiments, participants were first invited to a behavioral session to identify their

dispositions in social preference. In the behavioral session, all participants provided demographic information and completed the triple dominance (TD)[5] and social value orientation (SVO)[28] tasks, which were conventional measurements of one's stable disposition in social value orientation. To incentivize authentic responses during social interactions, participants were recruited in groups of 8-10 individuals (all were strangers to each other). For each economic game, participants were paired with a new, mutually anonymous partner.

The TD task is a 9-item measure of one's social disposition by asking participants to choose from 3 types of hypothetical self-other monetary allocation options (e.g., prosocial option: self = 100, other = 100; individualistic option: self = 110, other = 60; and competitive option: self = 100, other = 20). Based on their decisions to the 9 items, participants were classified as prosocial (who chose prosocial options on 6 or more items), individualist (who chose individualistic options on 6 or more items), or competitor (who chose competitive responses on 6 or more items). Participants who failed to choose the same type of options on at least 6 items were referred to as "unidentified". In the current study, we referred to both "individualist" and "competitor" as "individualists" in comparison to "prosocial".

The SVO slider measure included 6 primary items and 9 secondary items. For each item, participants were asked to choose the most preferred one from 9 monetary allocation choices over a well-defined continuum of joint payoffs. Based on the inverse tangent of the ratio between mean allocations for the self and the paired partner, the 6 primary items yielded a measure that categorized participants into: altruist, prosocial, individualist, and competitor[28]. Here, we referred to both "altruist" and "prosocial" as "prosocials" (i.e., SVO° > 22.45°), and both "individualist" and "competitor" as "individualists" (i.e., SVO° < 22.45°). The scores of 9 secondary items of SVO are used to calculate an independent measure of inequality aversion, i.e., the general preference for fairness and resistance to inequalities. To ensure a reliable measure of social disposition, only participants who were consistently classified by the TD and SVO tasks were deemed "qualified" (either "prosocial" or "individualist").

Prosocial behavior measures. Participants were also invited to the public goods game (PGG) and the dictator game (DG). The contribution participants made in these two

games have been separately used as indicators of the levels of cooperation and altruism - two key characteristics of prosocial behaviors[24,55-57].

In the 4-player PGG, participants initially receive 80 experimental monetary units (MU) and decide the amount of MU to contribute to a 4-player common project vs. to keep for themselves. The money contributed to the common project would be doubled and evenly divided among the 4 players. The final payoff is equal to the sum of money they keep for the self and money split from the common project. The amount of money contributed to the common project reflects cooperative behavior.

In the DG, "the dictator" (i.e., the participant), determines how to split 80 MU between himself and another player. The other players, "the recipient", simply receive the remainder of the endowment left by the dictator. The recipient's role is entirely passive and has no input into the outcome of the game. The amount the dictator sent to the recipient indicates his altruistic behavior.

fMRI session. A pair of participants, who were strangers to each other, was invited to the fMRI experiment at the same time. Upon arrival, participants' mood was measured using Positive and Negative Affect Scale (PANAS), which was later measured again after the experiment to quantify potential mood change. There was no significant mood change overall and no significant interaction effect with division of Social Disposition or Treatment (*Supplementary Table 4*). We measured participants' salivary oxytocin baseline levels by collecting their salivary samples before oxytocin or placebo administration (*Supplementary Fig. 4*). There was no significant main effect or interaction effect between Social Disposition and Treatment on the salivary oxytocin level. Each pair of participants was given 5 min to introduce themselves to each other to strengthen the oxytocin effect on social cognition[58]. We ensured that participants introduced their names to each other, which were also presented on the screen for each monetary allocation. Participants in each pair were scanned in sequence and randomly treated with oxytocin or placebo. The procedure of oxytocin or placebo administration was similar to previous research[24]. A single dose of 24 IU oxytocin or placebo (containing the active ingredients except for the neuropeptide) was intranasally self-administered by nasal spray approximately 35 min before the fMRI scanning under an experimenter's supervision. The spray was administered to participants three times

with each administration consisted of one inhalation of 4 IU into each nostril. The choice of 24 IU oxytocin and its effect on brain oxytocin level is explained in *Supplementary Note 1.* After scanning, participants were asked to perform a similar post-scan monetary outcome-pair evaluation task in a competitive context. The duration of the fMRI scanning and the post-scan test were carefully controlled within the time frame of the oxytocin peak response in the brain[54].

Monetary outcome-pair evaluation task during MRI scanning. In the MRI scanner, participants were presented with pairs of monetary outcomes assigned to the self and the paired participant (referred as partner). Participants evaluated their preference of each monetary allocation on a 4-point Likert scale (1=least preferable to 4=most preferable) by a button press. To encourage genuine responses and minimize the influences of social norms or social pressure, the preference ratings were unknown to the other player. Participants were told that their preference rating for each monetary outcome pair would determine the overall gains for self (G_s) and the partner (G_p), i.e.,$G_s = \sum ms_i * p_i$, and $G_p = \sum mp_i * p_i$, where p_i is participants' preference rating for the monetary outcome pair, i; ms_i/mp_i is the monetary amount for self or the partner in monetary pair, i. In each trial, the monetary allocation was presented for 3 s, followed by a jittered time interval, pseudo-randomized from 1s to 5 s (with mean interval of 3 s; **Fig. 1a**). There were two sessions with 90 trials per session, presented in a random order.

To determine appropriate monetary allocations for the fMRI scanning, we first conducted a pilot behavioral experiment on an independent sample (n = 60), where we included the full space of monetary outcome pairs and asked participants to rate their preference for each allocation on a 9-point Likert scale (1 = least preferable; 9 = most preferable). We found that participants reported invariably with the least preferable for pairs in the third quadrants and along the negative X- or Y-axis, where both self and the partner lose money (average preference rating of 1.8 on a 1-9 scale, with no rating scores higher than 3). Therefore, these pairs (i.e., \$Self ≤ 0 and/or \$Other ≤ 0) were not included in the fMRI task. The monetary outcome pairs for self and the partner, as illustrated in **Fig. 1a**, were designed in a way as to form angles that evenly sampled from -90° to 180° with an interval approximately 5°, based on an egocentric reference-point (i.e., 0°, which indicates perfect alignment with the positive X-axis).

We also included pairs where only the self or the partner gained money (evenly sampled along the positive X/Y-axis) while the opponent received zero. These pairs were used to generate functional masks for fMRI ROI analysis and were not included in formal behavioral analyses.

Monetary outcome-pair evaluation task in a competitive context. Participants completed a post-scan behavioral experiment largely the same as that in the fMRI task, but with two key differences. First, participants reported their preference of each monetary outcome pair on 10 instead of 4 levels (0 = least preferable to 9 = most preferable). Second, we induced a self-interest and other-interest conflict situation by framing the payoff in a competitive context, where participants would get a bonus reward if and only if the sum of gains to the themselves was larger than that to the partner. Otherwise, they gained nothing (i.e., "winner takes all"). Therefore, the self and partner's interest were in direct competition in this context. There was one session with 90 trials presented in a random order.

Monetary outcome-pair evaluation task in oxytocin-replication and online replication experiments. The task design for the replication experiments was identical to the fMRI experiment except that the monetary pairs were sampled on 3 circles of different circumference (radius=5, 6, 9), with θ ranging from -90° to 180° with different intervals (5°, 17°, 23°). There was one session with 82 trials presented in a random order.

Behavioral analysis. We constructed 8 behavioral models based on theoretical considerations (*Supplementary Fig. 2*). For the *social reference model* (the winning model, *Supplementary Fig. 2*), we modeled z-scored preference ratings (P) for each participant: $P=\beta1*\cos(\theta) + \beta2*\sin(\theta)$. In this model, β1 and β2 are weights for how much people care about the value of a potential payoff for themselves ($Self) and for the partner ($Other), respectively. The angle θ depends on the difference between those values. We then computed a single individual-specific reference-point φ for each participant based on the ratio of β1 and β2: $\varphi=\text{atan}(\beta2/\beta1)$[29,30] (**Fig. 1c**). The social value distance reflects the difference between a potential self-other allocation θ and a preferred allocation φ that reflects an individual-specific reference-point against which potential allocations are compared.

When all monetary pairs lie on the circumference of a circle, cos(θ) can be seen simply as the amount offered to the self and sin(θ) as the amount offered to the other, divided by the radius of the circle. However, when including monetary pairs from circles with different radii (i.e., the modified design used in additional experiments), cos(θ) and sin(θ) provide a compact index that permits investigation of the relationship between self and other in a value-insensitive way (since $\cos(\theta)^2 + \sin(\theta)^2=1$).

fMRI acquisition and preprocessing.

Imaging acquisition. Whole-brain imaging data was collected on a GE 3-Tesla MR scanner with a standard head coil (HDx, Signa MR 750 System; GE Healthcare, Milwaukee, WI). Functional images were collected using an echo-planar imaging sequence (axial slices, 32; slice thickness, 4 mm; gap, 1 mm; TR, 2000 ms; TE, 30 ms; voxel size, 3.75 × 3.75 × 5 mm; flip angle, 90°; FOV, 240 × 240 mm; and 285 volumes for each session, two sessions in total). Structural images were acquired through 3D sagittal T1-weighted magnetization-prepared rapid gradient echo (180 slices; TR, 8.208 ms; TE, 3.22 ms; slice thickness, 1 mm; voxel size, 0.47 × 0.47 × 1.0 mm^3; flip angle, 12°; inversion time, 450 ms; FOV, 240 × 240 mm).

Imaging preprocessing. Brain imaging data was preprocessed using Statistical Parametric Mapping (SPM12; http://www.fil.ion.ucl.ac.uk/spm). The first 5 functional images from each session were discarded for signal equilibrium and participants' adaptation to scanning noise. Remaining images were corrected for slice acquisition timing and realigned for head motion correction. Subsequently, functional images were coregistered to each participant's grey matter image segmented from corresponding high-resolution T1-weighted image, then spatially normalized into a common stereotactic Montreal Neurological Institute (MNI) space and resampled into 2-mm isotropic voxels. Finally, images were smoothed by an isotropic 3D Gaussian kernel with 8-mm full-width at half-maximum.

GLM analysis. After preprocessing, we estimated parameters of different general linear models (GLMs). All models included regressors for monetary outcome pair presentation separately for trials on and off the X/Y axis, button press, instructions, six nuisance regressors for motion-related artefacts, and various parametric modulations

associated with these regressors (detailed below). Parametric regressors were not orthogonalized in the design matrix, ensuring that parameter estimates were not confounded by spurious correlations due to signals related to other regressors[59]. All regressors (parametrically modulated or not) were convolved with the canonical hemodynamic response function (HRF) in SPM before entering the GLM. Data were high-pass filtered at 1/128 Hz. We controlled for decision times for all fMRI analyses.

We created parametric regressors that were associated with the value distance or value difference between self and other, at the monetary outcome-pair presentation to search for brain regions that encoded the subjective distance in social value representation. In the fMRI analysis, we included 10 GLM models: a social value distance from an individual-specific reference-point (i.e., $1 - \cos(\theta(t) - \varphi)$, $\theta(t)$ is the angle of a potential allocation at trial t and φ is the individual-specific reference-point derived from our social reference model), an egocentric reference (i.e., $\cos(\theta(t))$), an allocentric reference (i.e., $\sin(\theta(t))$), an objective equality reference-point (i.e., $\cos(\theta(t) - 45°)$), monetary outcome for the self (i.e., $Self), monetary outcome for the partner (i.e., $Other), absolute value difference (|$Self - $Other| or advantageous (i.e., max (0, $Self -$Other)) or disadvantageous inequality aversion (i.e., max (0, $Other -$Self)) separately, or with preference rating as parametric modulator in the GLM.

In building different GLMs, we are not arguing that the social reference model is *superior* to other models in all environments nor claiming the dissimilarity measure is the best measure for capturing amygdala responses as this was not our aim[60]. Rather, we aimed to identify brain regions that could specifically represent deviations from the reference-point of social preference.

Coefficients for each regressor were estimated for each participant using maximum likelihood estimates to account for serial correlations in the data. Statistical significance was determined at the group level using a random-effects analysis. Significant clusters from second-level analyses were determined using a height threshold of $P < 0.001$ and an extent threshold of $P < 0.05$ with cluster-based family-wise error (FWE) correction. We also applied voxelwise inference using the FWE-corrected threshold of $P = 0.05$ on the whole-brain analysis, given recent concern over cluster-wise inferences. For the relationship between value distance ($1-\cos(\theta(t) –$

φ) and neural responses during monetary outcome-pair presentation, the peak voxels in right amygdala survived voxelwise FWE correction ($P = 0.02$).

Control analysis of amygdala responses. One alternative hypothesis of the amygdala activity pattern is that it encoded \$Other or \$Disadvantageous Inequality, instead of the dissimilarity distance to the "reference-point" in our winning model, we reran the fMRI analysis from the first level controlling for \$Other and \$Disadvantageous Inequality as regressors in the GLM (without orthogonalization between regressors). We looked for the unique variance that can be explained by the dissimilarity distance to social reference-point over and beyond the variance explained by \$Other or \$Disadvantageous Inequality. The main result of a significant Social Disposition x Treatment interaction on the amygdala activity in coding deviations from the social reference-point was unchanged.

We further tested another possibility of the amygdala response: it encoded \$Other in proportion to its importance to the individual. To test this possibility, we built another GLM model with the parametric regressor: β2*\$Other, where β2 represented the estimated weight of \$Other on social preference, reflecting the individual-specific importance of \$Other in social preference evaluation for each participant. However, the amygdala activity did not simply encode \$Other to the extent that it predicts social preferences, either at the whole brain level (height $p < 0.001$, cluster-wise FWE, $p < 0.05$) or ROI (anatomically-defined) level.

Statistics. The oxytocin-fMRI and oxytocin-replication experiments were double-blind, i.e., both participants and experimenters were blind to experimental conditions (both treatment and social disposition conditions). Data analysis was not performed blind to the conditions of the experiments. We first conducted one-way ANOVA with Social Disposition as a between-subjects factor to compare the social reference-point between individualists and prosocials under placebo. To evaluate the oxytocin effect on the social value representation, we conducted ANOVAs on behavioral and fMRI data, with Social Disposition (prosocial vs. individualistic) and Treatment (oxytocin vs. placebo) as between-subjects factors, followed by planned two-tailed *t* tests to examine oxytocin effect separately in individualists and prosocials

(independent-samples *t* test for fMRI study and paired-samples *t* test for the oxytocin-replication study). Data distribution was assumed to be normal but this was not formally tested. All correlations were performed by Pearson's correlation coefficient analysis.

Data Availability Statement. The data that support the findings of this study and the analysis code are available from the corresponding author upon reasonable request.

Code availability. Analysis code to model the social value representation based on preference rating data is provided in the Supplementary Software.

Life Sciences Reporting Summary. Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Method-only References

51. Gao, S., *et al.* Oxytocin, the peptide that bonds the sexes also divides them. *Proc. Natl. Acad. Sci. USA* **113**, 7650-7654. (2016).

52. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods.* **41**, 1149-1160 (2009).

53. Walum, H., Waldman, I.D. & Young, L.J. Statistical and Methodological Considerations for the Interpretation of Intranasal Oxytocin Studies. *Biol. Psychiatry.* **79**, 251-257 (2016).

54. Paloyelis, Y., *et al.* A Spatiotemporal Profile of In Vivo Cerebral Blood Flow Changes Following Intranasal Oxytocin in Humans. *Biol. Psychiatry.* **79**, 693-705 (2016).

55. Rand, D.G., Greene, J.D. & Nowak, M.A. Spontaneous giving and calculated greed. *Nature* **489**, 427-430 (2012).

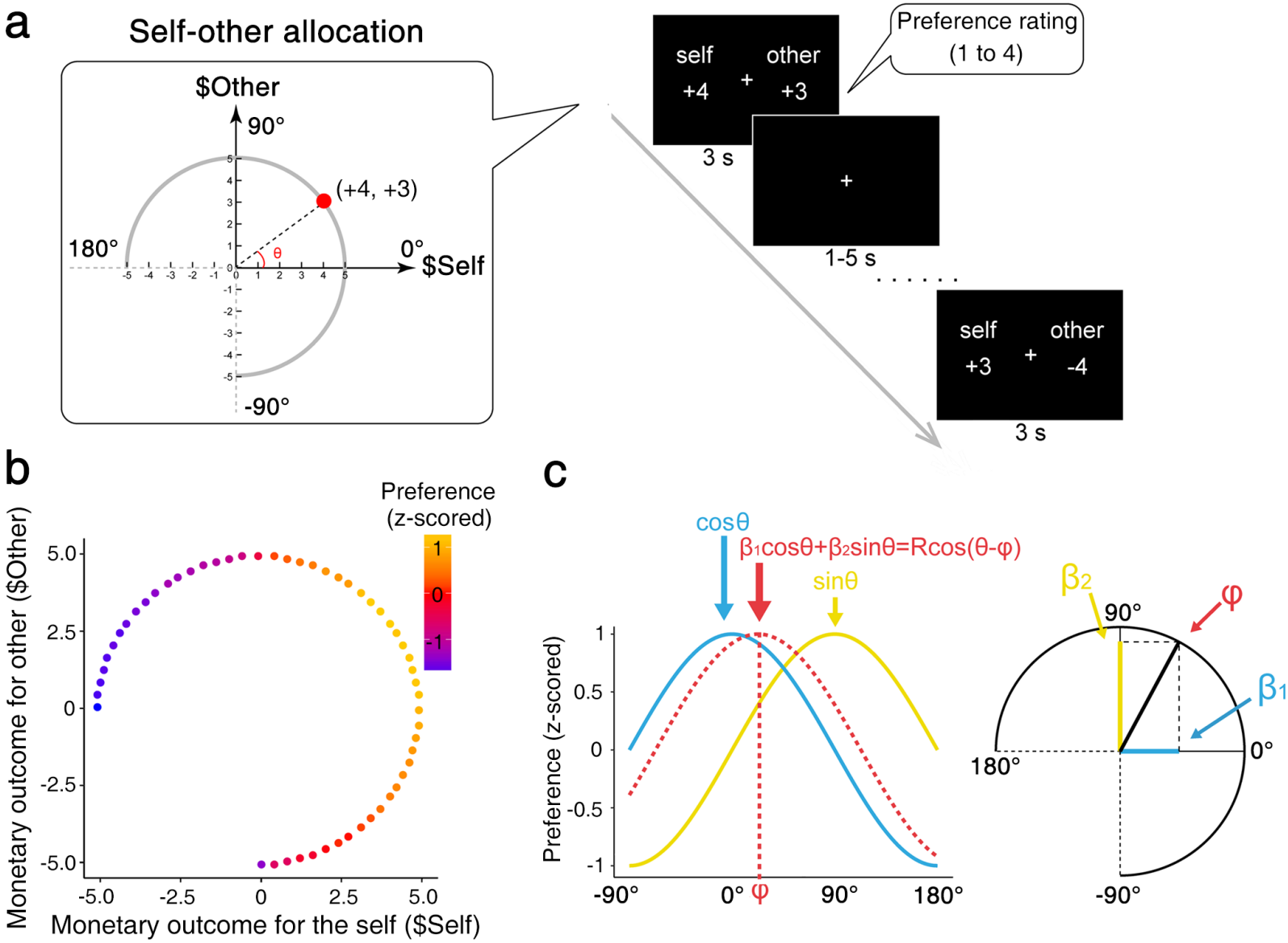
56. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785-791 (2003).

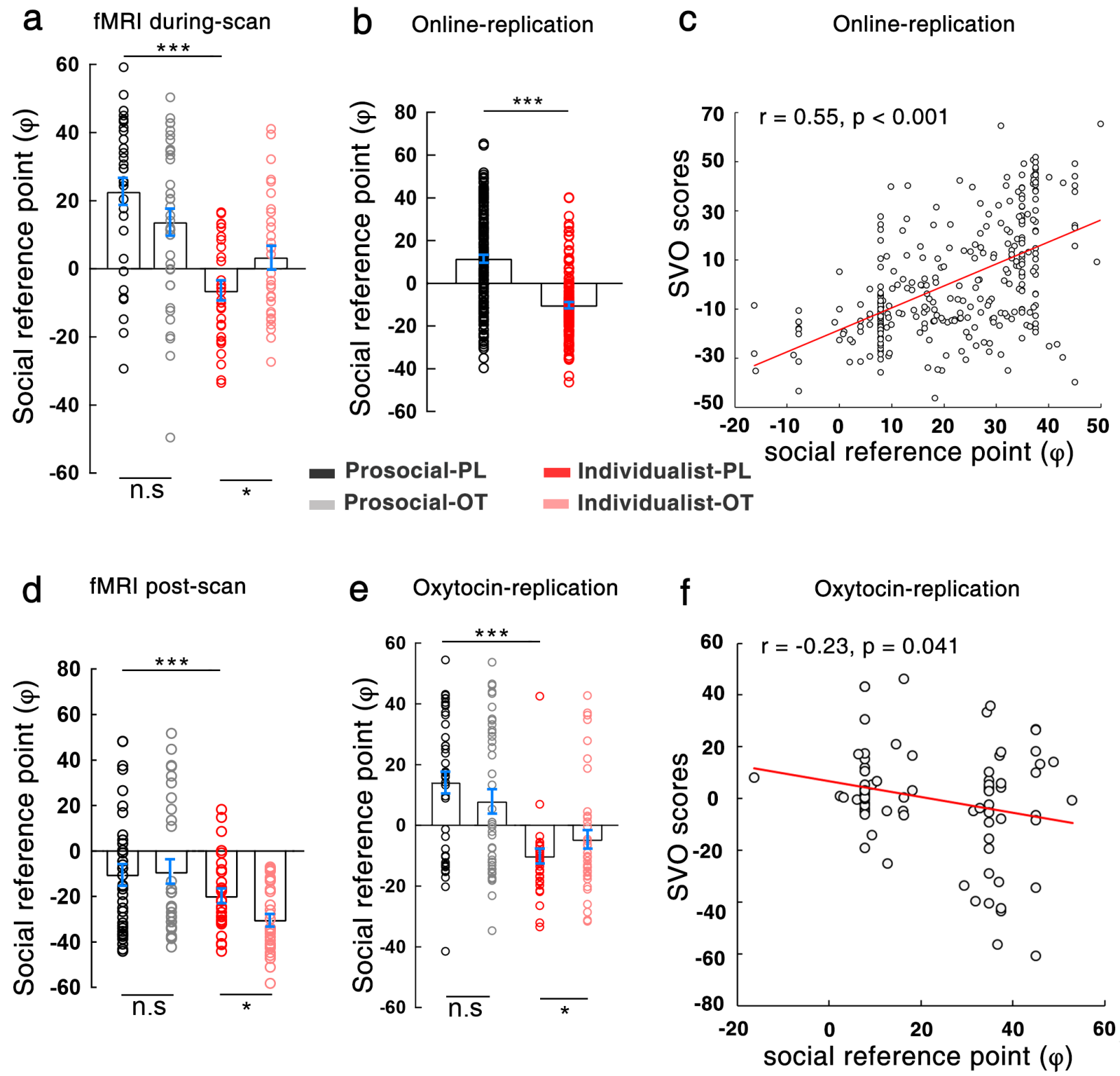
57. Harbaugh, W.T., Mayr, U. & Burghart, D.R. Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* **316**, 1622-1625 (2007).

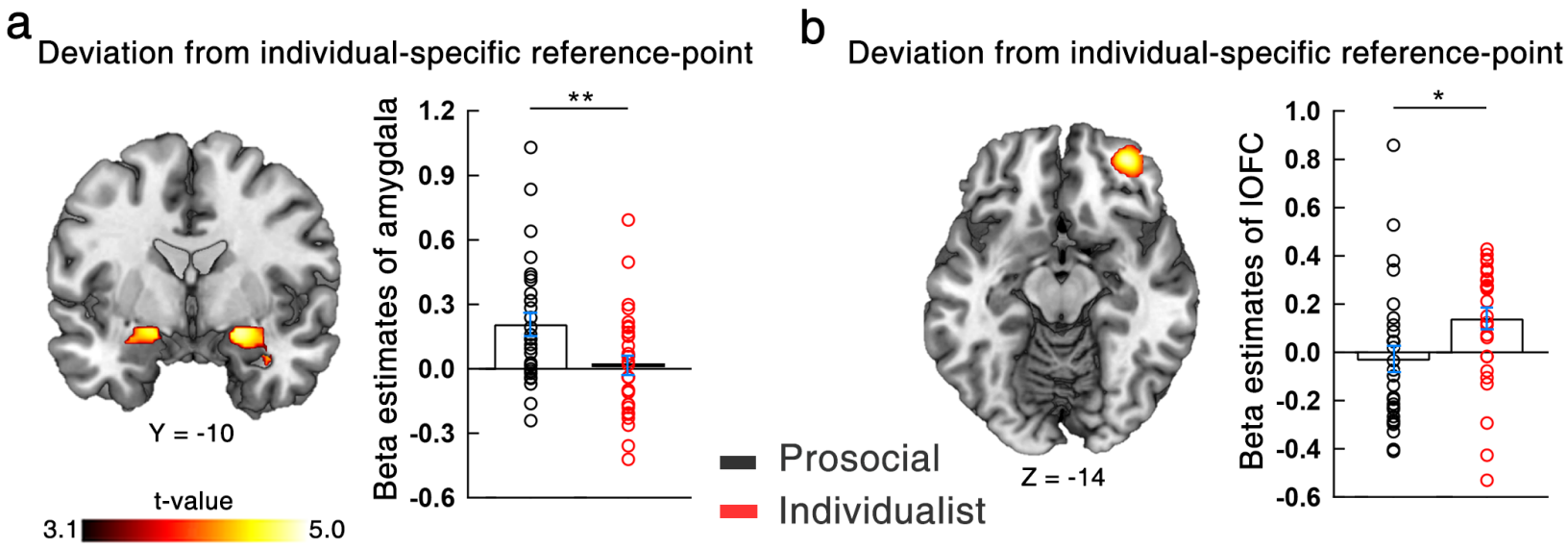
58. Declerck, C.H., Boone, C. & Kiyonari, T. Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. *Horm. Behav.* **57**, 368-374 (2010).

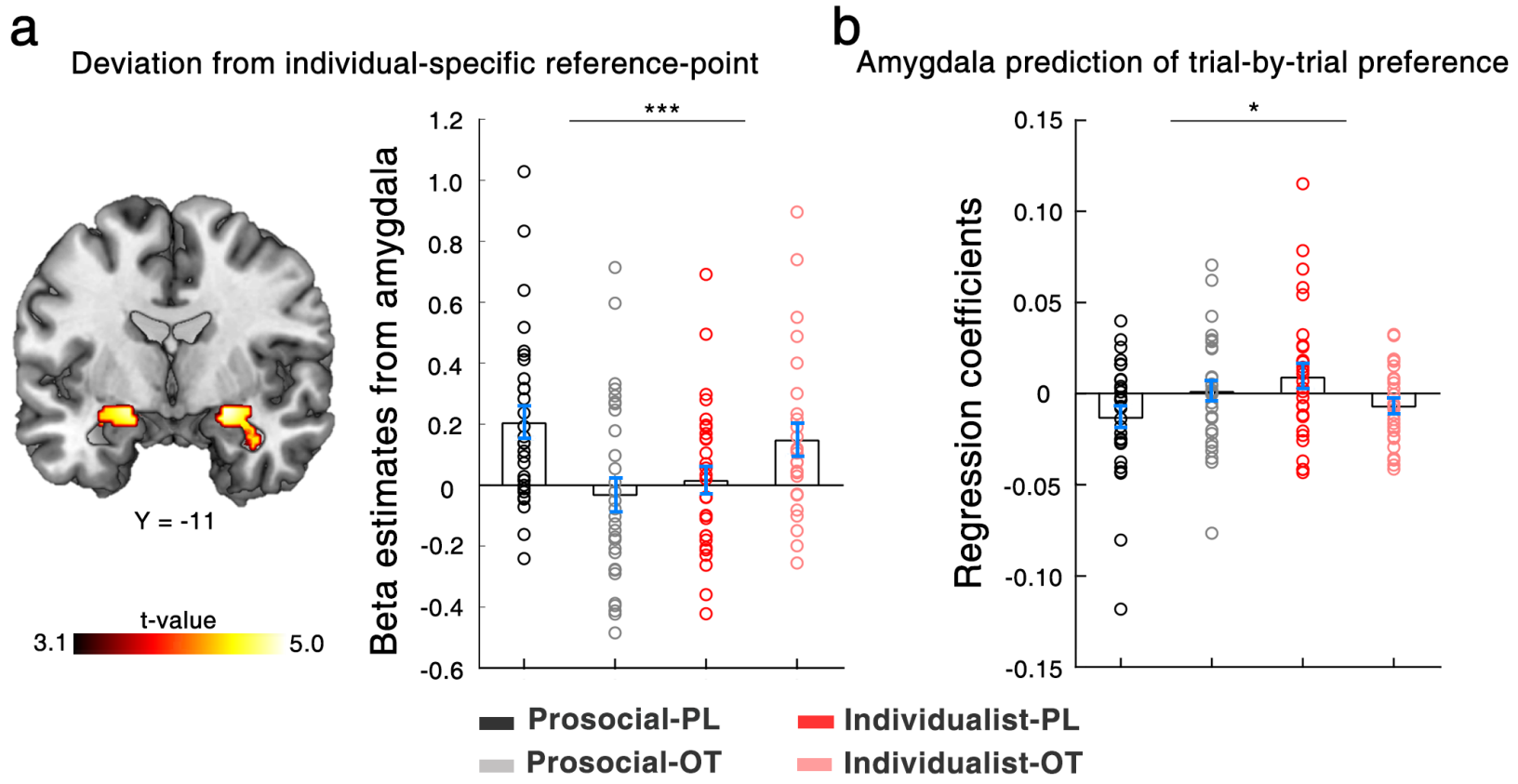
59. Andrade, A., Paradis, A.-L., Rouquette, S. & Poline, J.-B. Ambiguous results in functional neuroimaging data analysis due to covariate correlation. *Neuroimage* **10**, 483-486 (1999).

60. Wilson, R.C. & Niv, Y. Is model fitting necessary for model-based fMRI? *PLoS. Comput. Biol.* **11**, e1004237 (2015).


a
Self-other allocation
$Other
90°
(+4, +3)
180°
0°
$Self
θ
-90°
self
+4
+
other
+3
3 s
Preference rating
(1 to 4)
+
1-5 s
self
+3
+
other
-4
3 s
b
Preference
(z-scored)
Monetary outcome for other ($Other)
Monetary outcome for the self ($Self)
c
cosθ
β1cosθ+β2sinθ=Rcos(θ-φ)
sinθ
Preference (z-scored)
-90°
0°
φ
90°
180°
β2
φ
β1
90°
180°
0°
-90°


a
fMRI during-scan
***
Social reference point (φ)
60
40
20
0
-20
-40
-60
n.s
*
b
Online-replication
***
Social reference point (φ)
80
60
40
20
0
-20
-40
-60
c
Online-replication
r = 0.55, p < 0.001
SVO scores
70
50
30
10
-10
-30
-50
-20
-10
0
10
20
30
40
50
social reference point (φ)
Prosocial-PL
Prosocial-OT
Individualist-PL
Individualist-OT
d
fMRI post-scan
***
Social reference point (φ)
80
60
40
20
0
-20
-40
-60
n.s
*
e
Oxytocin-replication
***
Social reference point (φ)
60
40
20
0
-20
-40
-60
n.s
*
f
Oxytocin-replication
r = -0.23, p = 0.041
SVO scores
60
40
20
0
-20
-40
-60
-80
-20
0
20
40
60
social reference point (φ)


a
Deviation from individual-specific reference-point
Y = -10
t-value
3.1
5.0
Beta estimates of amygdala
1.2
0.9
0.6
0.3
0.0
-0.3
-0.6
**
Prosocial
Individualist
b
Deviation from individual-specific reference-point
Z = -14
Beta estimates of lOFC
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
*


a
Deviation from individual-specific reference-point
Y = -11
t-value
3.1
5.0
Beta estimates from amygdala
1.2
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
***
b
Amygdala prediction of trial-by-trial preference
Regression coefficients
0.15
0.10
0.05
0
-0.05
-0.10
-0.15
*
Prosocial-PL
Prosocial-OT
Individualist-PL
Individualist-OT