# Neural representations of social placebo effect

## Introduction

*"We are led to the inescapable conclusion that the history of medical treatment for the most part until relatively recently is the history of the placebo effect."*

——*Shapiro, 1959.*

Placebo effect represents the beneficial changes induced by the use of inert treatment, which could mimic the therapeutic effects of real treatment, and widely affects human well-being (Colloca et al., 2004; Enck et al., 2013). Our ancestors administered colorful treatments to cure diseases, such as the touch of royal members, spider web and worms. Although it easy to criticize these sham treatments with the benefit of hindsight, we can be hardly immune to placebo effect today (REF).

Knowing the treatment-related information (such as its efficacy, usage) is critical to induce placebo effect (Finniss et al., 2010). In clinical practice, open treatment with drug information provided by clinicians produced better effectiveness than hidden treatment in which without any information (Benedetti et al., 2003) even though the administered treatment is same. And more delineated and certain information produced stronger therapeutic efficacy (Slavenka et al., 2014; Thomas, 1987). In laboratory studies, delivering the treatment efficacy via verbal instructions has been widely used to induce placebo effect (Benedetti, Enck, Frisaldi, Schedlowski, 2014). We have already known that the necessity of treatment-related information for inducing placebo effect. But how brain encoded, represented and constructed the treatment-related information to induce placebo responses are still remained unknown.

Gaining information about the administered treatment may generate expectations of treatment outcome possibly by actively predicting the health benefits for future-oriented personal well-being. We hypothesized that the treatment-related information could be encoded and constructed in ventromedial prefrontal cortex

(vmPFC) and brain regions that were associated with believed, specific treatment effect. This hypothesis is based on two lines of evidence. On the one hand, the vmPFC is a critical brain region in self-projection (Bucker and Carroll, 2006) and imagination that via integrating knowledge to constitute a possible future episode (Benoit et al., 2014). It has been theorized as the key region to represent concepts, meaning, value, expectations related to the treatment for personal well-being in placebo studies (Roy et al., 2012; Büchel et al., 2014; Ashar et al., 2017; Geuter et al., 2017). It was also treated as an important hub to project the generated expectations to target systems (e.g., multisensory system), then induced placebo responses (e.g., pain analgesia; Bingel et al., 2006; Tracey, 2010). On the other hand, the corresponding neural activity of placebo effect were specific and highly related to the acquired treatment-related information. For example, the "pain killer" decreased neural responses in pain-sensitive regions (e.g., anterior insula, rostral anterior cingulate cortex, periaqueductal gray, thalamus, somatosensory cortex) (Wager et al., 2004; Ellingsen et al., 2013), and "anxiolytic drug" could down-regulate salience-related brain regions (e.g., dorsal anterior cingulate cortex, anterior insula) (Meyer et al., 2018). Therefore, the treatment-related information may have been encoded, represented, and further constructed in vmPFC to generate beliefs, as well as in treatment-specific brain regions.

We adopted the framework of social placebo effect (SPE; Yan et al., 2018), a latest placebo effect that can facilitate prosocial behaviors and promote social functioning, to investigate the neural representations of treatment-related information underlying placebo effect. Thus, we chose the social brain network, which was crucial for social functioning and social well-being (McCormick et al., 2018; Alcalá-López et al., 2018) as the treatment-specific brain regions.

In the single-blind, between-subjects design, participants were randomly assigned to spray+ condition and control condition. Consistent with our previous study (Yan et al., 2018), participants in spray+ condition were intranasally administered with saline (but

it was told as “oxytocin”, a neuropeptide which is highly related to prosocial behaviors). In control condition, participants were intranasally administered with saline and it was told as “saline”. We adopted and designed resting-state and information representation task as in-scanner tasks. The resting state could reflect the internal and automatic representation (Lewis et al., 2009) for prior cognitive process and recent experience (Waites et al., 2005; Hasson et al., 2009; Mildner and Tamir, 2019). Therefore, measuring the resting-state process may shed light on how people spontaneously processed the treatment-related information. Based on our hypothesis, we were interested to examine the functional connectivity within social brain network and vmPFC-based connection. Further, to capture how the treatment-related information represented and constructed, we designed information representation task, in which participants were presented with the key treatment-related sentences (i.e., sentences to describe oxytocin, *oxytocin-related* information) one by one (details see Method). This task possibly mimicked the representation process when people read the instruction of medicine in daily life, or received the verbal suggestions in placebo studies. We also included treatment-irrelevant information (sentences to describe robot, *oxytocin-irrelevant* information) as control stimuli in this task. We aimed to decipher the representation pattern of treatment-related information by using representational models (Diedrichsen et al., 2017; Kriegeskorte and Douglas, 2019; here, we adopted representational similarity analysis) which would assist to specify how information represented and organized in vmPFC and social brain network. Finally, participants finished social incentive delay task, and post-check questionnaires out of scanner (general procedure see Figure 1).

The answers to current research questions was affirmative:

(i) Spontaneous representation for treatment information has enhanced functional connectivity between vmPFC and superior frontal gyrus, as well as within social brain network in spray+ condition;

(ii) Treatment-related information was represented in vmPFC, left middle frontal, right lateral orbitofrontal cortex with more distinct and finer-grained pattern in spray+ condition.

(iii) Combing the functional connectivity strength within social brain network, neural representation of treatment information, and behavioral indices, participants in spray+ condition can be discriminated from control condition with a classification accuracy of 89%.

## Results

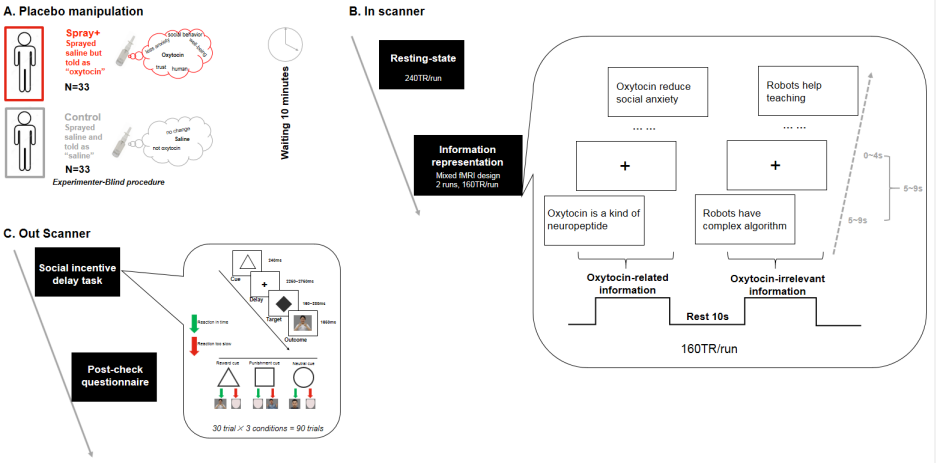



**Figure 1. Schematic of the general experimental procedure.**

(A) Experimental procedure in placebo manipulation phase. In placebo manipulation phase, participants were randomly assigned to *spray+* condition (N=33) and *control* condition (N=33). In *spray+* condition, participants learned oxytocin materials and then self-administered a nasal spray (i.e., saline spray but told as "oxytocin") under experimenter supervision. In *control* condition, the materials and procedure were same with the spray+ condition except the nasal spray was told as "saline" instead. We adopted experimenter-blind procedure (experimenters were blind to experimental hypotheses) in manipulation stage to

avoid potential experimenter bias.

(B) Experimental procedure in in-scanner phase. After 10 minutes waiting period, participants were invited to finish the resting-state, information representation task (including two categories: *Oxytocin-related* and *Oxytocin-irrelevant* information) in fMRI scanner.

(C) Experimental procedure in out-scanner phase. Then participants completed social incentive delay task, and post-check questionnaires out of scanner.

**Defining vmPFC and social brain network.**

We defined vmPFC (Figure S1) by performing a meta-analysis on previous neuroimaging studies involved "vmPFC" (based on 199 studies), implemented with the Neurosynth software (http://neurosynth.org/). To purify the vmPFC map, we set cluster-level extent threshold (k = 300 voxels) to exclude other clusters including nucleus accumbens, precuneus, hippocampus. The social brain network was defined by an elegant quantitative meta-analyses on 3972 social-affective experiments as comprising 36 distinct regions (Alcalá-López et al., 2018). Since previous placebo literature mentioned the default mode network (Ashar et al., 2016) was important for placebo effect, we also investigated the functional architectures of default mode brain network (network identification was based on the pre-defined regions and community, see Power et al., 2011).

**Spontaneous representation for treatment-related information in *spray+* condition associated with stronger functional connectivity between vmPFC and superior frontal gyrus.**

#point1-define vmPFC/define social brain network/detailed the method of seed-FC

Since the age between two treatment conditions was marginally significant ($t(64) = -1.983$, $p=0.052$), we controlled age in all analyses to exclude potentially influences induced by age difference. We used seed-based analysis to test whether vmPFC-based functional connectivity would be differed in two treatment conditions. The seed was defined by a spherical ROI (5 mm diameter) was centered on the supra‐threshold peak coordinate of defined vmPFC map (MNI coordinates, x,y,z = -2 46 -8). Interestingly, the *spray+* condition revealed stronger functional coupling strength between the vmPFC and superior frontal gyrus (SFG) (MNI coordinates, x,y,z = 20 22 60).

**Spontaneous representation for placebo information in *spray+* condition enhanced functional connectivity strength and facilitated network efficiency within social brain connectome.**

Then we examined whether the internal representation for placebo information would alert network architecture of social brain network, as well for default mode network by calculating average functional connectivity and comparing it between *spray+* and *control* condition. Interestingly, the average functional connectivity within the 36 social brain network ROIs was stronger in *spray+* than in *control* condition ($F(1,58)$[1] $= 8.291$, $p = 0.006$, $\eta_p^2 = 0.125$; after FDR correction, p=0.012. Figure 2B). But we failed to find significant difference but some trends on increasing functional connection strength within default mode network ($F(1,58)$[2] $= 3.418$, $p = 0.070$, $\eta_p^2 = 0.056$; after FDR correction, p=0.070).

#point2-why we adopted NBS/why we only considered the social brain network/the results

To further anchor the specific pairs of brain regions in which functional connectivity within social brain network was modulated by spontaneously representation, we adopted network-based statistic (NBS) approach (t-threshold: $t > 3.1$; permutation: 5000 randomizations) to identify brain regions pairs showing significant differences in functional connectivity between *spray+* and *control* condition.

The results showed a significant 7-node, 6-edge network (*spray+* - *control*; Figure 2C), and largest percentage of these links (83.33%) were with right nucleus accumbens (r-NAc). Since we only detect robust significant results on functional connectivity strength and network efficiency within social brain network, here we would not consider the other network.

#point3-why we adopted the local and global efficiency indices/the results/

More comprehensively, we conducted analysis based on graph-theory methods (Dosenbach et al., 2007; Rubinov and Sporns, 2010) to examine the network organization of whole social brain network. We mainly focused on quantifying measures of the network efficiency, which is a more biologically relevant metric to describe the information transmission within brain networks. Specifically, we calculated the global efficiency and local efficiency to elucidate the information

[1] Four participants were excluded due to excessive head movement, one participant was excluded due to his whole-brain connectivity value was outlier among all participants (>mean±3.5SD).
[2] Four participants were excluded due to excessive head movement, one participant was excluded due to his whole-brain connectivity value was outlier among all participants (>mean±3.5SD).

flows efficiency at global and local level. Higher network efficiency (globally and locally) indicated stronger information processing ability. We found that the social brain network associated with higher global efficiency in *spray+* condition ($F(1,58) = 9.981$, $p = 0.003$, $\eta_p^2 = 0.147$; after FDR correction, p=0.012), as well higher local efficiency than in *control* condition ($F(1,58) = 7.916$, $p = 0.007$, $\eta_p^2 = 0.120$; after FDR correction, p=0.014; Figure 2D & Figure 2E). For default mode network, we did not significant social placebo effect on all indices mentioned above (all $p > 0.075$; after FDR correction, $p>0.100$).

#point4-conclusion for resting part

Together, these results indicated that the spontaneous internal representation for "active" treatment information increased the functional coupling of prefrontal regions (vmPFC-SFG). Interestingly, this internal representation also allowed the stronger connectivity strength within social brain network, with the right NAc, a critical brain region for encoding motivation-related psychological processes (Schultz, 2004; Knutson and Cooper, 2005), served as the connected hub. Graph-theory based analyses suggested that this kind of spontaneous internal representation enhanced the efficiency for transmitting information within social brain connectome, which may provide the necessary prerequisites for further complex and intensive social interactions.

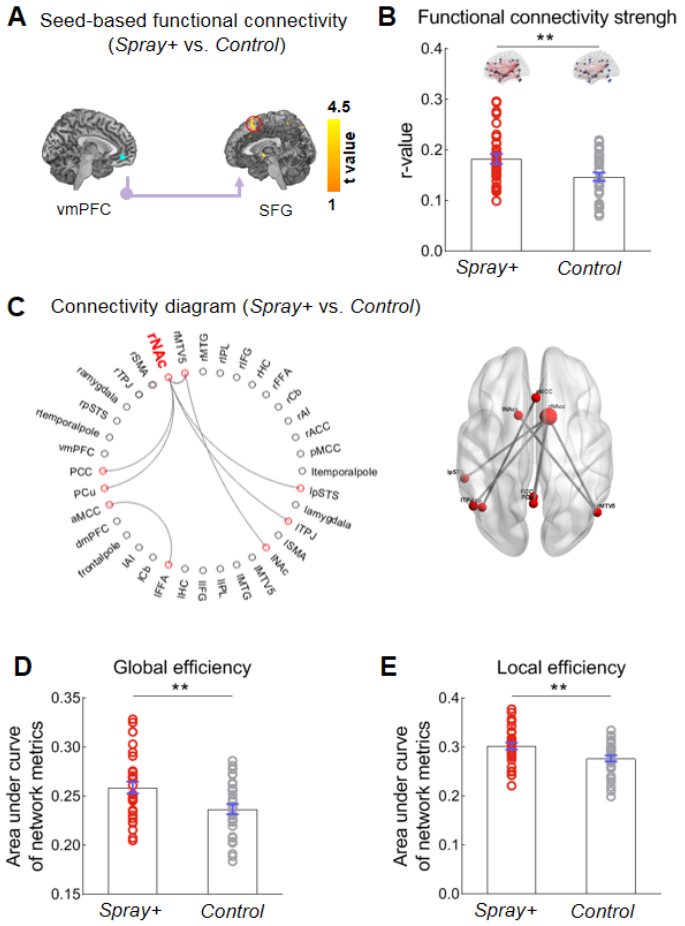



**Figure 2. Spontaneous representation for placebo information enhanced functional coupling of prefrontal regions, and reshaped social brain network with stronger connectivity strength and communication efficiency.**

(A) Results of seed-based functional connectivity. *Spray+* condition allowed stronger functional coupling between vmPFC and superior frontal gyrus. SFG, superior frontal gyrus.

(B) Social brain network's functional connectivity strength in two treatment conditions. The average functional connectivity strength within social brain network was significant higher in *spray+* condition than *control* condition.

(C) Connectivity diagram to show the internal representation (*spray+* vs. *control*) on specific

connections. Spray+ condition associated with greater functional connectivity of the right NAc (hub node) with PCC, precuneus, aMCC, left NAc, left TPJ and left STS. Red color indicated the significant nodes from network-based statistic analysis. PCC, posterior cingulate gyrus; aMCC, anterior middle cingulate gyrus; TPJ, temporoparietal junction; STS, superior temporal sulcus. (D, E) Graph theory properties within social brain network. *Spray+* condition showed higher global and local communication efficiency within social brain network. (Error bars represent standard error, $p<0.05$ *, $p < 0.01$ **).

**Stimuli and design for information representation task**

Importantly, both two groups of participants learned general knowledge about oxytocin in placebo manipulation phase, to avoid the potential influence of familiarity. Participants did not learn robot-related knowledge, since our online studies (see Figure S2, Figure S3) have shown that people were generally more familiar with robots than oxytocin, and after learning oxytocin-related materials, these two types of knowledge could reach comparable familiarity. In this task, sixty short sentences were selected as experimental stimuli and separated into two categories including "*Oxytocin-related information*" and "*Oxytocin-irrelevant information*". Sentences in *Oxytocin-related information* were made up from "oxytocin", and the high frequency words from text analysis based on oxytocin materials used in placebo manipulation phase. These sentences were designed to delineate the knowledge about oxytocin and its effect on social functioning; such as "Oxytocin increases the interpersonal interaction"). Sentences in *Oxytocin-irrelevant information* were made up from the core noun "robot", and the high frequency words from text analysis based on robot-related materials used in previous social placebo manipulation phase (Yan et al., 2018), such as "Robot can be used in hospital to help patients". Sentences in each category kept similar length and format, as well matched in valence and arousal (see SI Section 1). In scanner, participants were instructed to view each sentence from *Oxytocin-related information* and *Oxytocin-irrelevant information* then judge whether they understand or not. In out-scanner session, all participants viewed these sentences again and rated the trueness, comprehension and self-relevance for each sentence (detailed task description see Method).

**Oxytocin-related information but not Oxytocin-irrelevant information elicited higher neural activation in vmPFC in *spray+* condition.**

#introduced vmPFC, raised your hypothesis.

We aimed to empirically examine whether and how the vmPFC encoded and represented the treatment information. Placebo treatment contributed to personal well-being by appraising the benefits to oneself (Geuter et al., 2017), and vmPFC can conceptualize further states of the self (Buckner and Carroll, 2006; Hershfiled, 2011). Participants in spray+ condition received "oxytocin" and learned the information about oxytocin may imagine the oxytocin effect on themselves in the future, that is, the self-relevance of *Oxytocin-related information* may be more

evident than in control condition. Here, we hypothesized that the *Oxytocin-related information* would elicit stronger neural responses in vmPFC in spray+ condition.

#GLM results in vmPFC----treatment effect on OT and Rob separately.

We calculated the parameter estimates of signal intensity in vmPFC for each category of information. As expected, univariate analysis with Treatment (spray+ vs. control) as between-subject variable, age as covariate showed significant main effect of treatment ($F(1,60)$[3]$=9.172$, $p=0.004$, $\eta_p^2=0.112$), that spray+ condition engaged stronger neural activation for *Oxytocin-related information* (M±SD = -0.139±0.185) than control condition did (M±SD = -0.241±0.207). No significant treatment effect for *Oxytocin-irrelevant information* ($F(1,60)=0.493$, $p=0.485$, $\eta_p^2=0.008$) was found. Further, a mixed-ANCOVA with Information-type (*Oxytocin-related* vs. *Oxytocin-irrelevant*) as within-subject variable, Treatment (spray+ vs. control) as between-subject variable, and age as covariate revealed significant interaction effect of Information- type × Treatment ($F(1,60)=7.533$, $p=0.008$, $\eta_p^2=0.112$; see Figure 3A). To further confirm the results of ROI analyses, we conducted whole-brain analyses that compared spray+ and control condition in each kind of Information-type. Consistently, *Oxytocin-related information* engaged higher neural activity in vmPFC [MNI coordinates, xyz = -12 36 -6] in spray+ condition compared to control condition, while *Oxytocin-irrelevant information* did not reveal significant treatment effect in whole-brain (Figure 3B, Table S1). We also conducted a whole-brain analysis of group differences across Information-types using 2-by-2 ANOVA, and confirmed the significant interaction effect of Information-type × Treatment in superior medial prefrontal gyrus [MNI coordinates, xyz = -6 50 22].

**Finer-grained neural pattern for oxytocin-related information but not for oxytocin-irrelevant information in spray+ condition.**

#briefly introduce the motivation to do neural representation and introduce the main index.

*Oxytocin-related information* engaged higher neural activation in vmPFC in spray+ condition, this result answered *whether* the hypothesized brain region to encode the treatment-related information. To further our understanding on *how* the treatment-related2 information was encoded

[3] One participant did not have data due to technique problem, two participants were excluded due to excessive head movement.

and represented in brain, we investigate the neural representation pattern for *Oxytocin-related information* by calculating the inter-item neural similarity within all sentences in *Oxytocin-related* condition.

#results1---neural discrimination in vmPFC ROI and whole-brain; and the explanation.

The spray+ group showed less neural similarity for *Oxytocin-related information* (main effect of treatment: $F(1,60)=6.248$, $p=0.015$, $\eta_p^2=0.094$) in vmPFC. That is, neural representation of *Oxytocin-related information* was more distinct in spray+ condition than control condition (Figure 3C). Results from whole-brain analysis showed the rlOFC [MNI coordinates, xyz = 38 52 -10] and left middle frontal gyrus [MNI coordinates, xyz = -30 44 26] encoded the *Oxytocin-related information* in more distinct pattern (see Figure 3D; Table S2). We did not find treatment effect on the neural representation for *Oxytocin-irrelevant information* ($F(1,60)=0.986$, $p=0.325$, $\eta_p^2=0.016$).

#results2---hub knowledge

We then showed the neural geometric representation pattern for all sentences of *Oxytocin-related information* in significant prefrontal clusters we found (Figure 3D). To give more comprehensive and elaborated description for the representational map, we next investigated how different connection among all *Oxytocin-related information* in spray+ and control condition. Here, we aimed to find the central hub concept in spray+ condition and control condition separately by calculating degree centrality (i.e., the number of ties that a node connected) based on all participants' averaged representational similarity matrix. The sentence with biggest degree centrality value would be selected as the hub knowledge. Intriguingly, the hub information emerged in spray+ condition was "Oxytocin will influence the affective-related brain activity", while control condition was "Oxytocin is hot in research". We may conclude that the key information represented in spray+ was related to oxytocin's function, while in control condition, the hub knowledge was related to oxytocin's general knowledge.

**Greater neural discrimination pattern between *Oxytocin-related* and *Oxytocin-irrelevant* in vmPFC in spray+ condition.**

#results1-condition-based RSA

Next, we examined whether the neural representation pattern would be more differentiated

between the two types of information in spray+ condition. We first extracted the neural activation value in each information-type in each sphere (defined with a radius of 15mm) in the individual native space using a searchlight procedure. The neural pattern discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information was then computed by using 1 to minus the Pearson correlation coefficient (r) between the neural activation values of *Oxytocin-related* and *Oxytocin-irrelevant* conditions in each of the spheres in whole-brain. We extracted and looked at the neural pattern discrimination value (1-r) in pre-defined vmPFC and compared it between two treatment groups. Interestingly, we found a significant main effect of treatment ($F(1,60)=7.933$, $p=0.007$, $\eta_p^2=0.117$, controlled age, Figure 3F), that spray+ condition (M±SD = 0.732±0.142) had higher neural pattern discrimination between the two information-types than control condition did (M±SD = 0.644±0.137). Consistently, we confirmed this finding in whole-brain analyses, that the mPFC [MNI coordinates, xyz = -20 66 10] revealed more distributed neural pattern between the two types of information in spray+ than in control condition (Figure 3G, Table S3).

#results2-trial-wised RSA

Then we examined the same question with item-wised neural pattern similarity analyses. We estimated item-wised beta value for each sentence in each information-type and constructed full inter-item correlational similarity of multivoxel activity patterns (i.e., 60×60 matrix) using searchlight procedure for each individual in each treatment condition. As expected, neural pattern discrimination value in vmPFC in spray+ condition (M±SD = 0.943±0.022) was significantly larger than in control condition (M±SD = 0.925±0.026; main effect of treatment: $F(1,60)=10.480$, $p=0.002$, $\eta_p^2=0.149$; see Figure 3H), which proved again that the placebo manipulation enabled more distinct boundary for *Oxytocin-related* and *Oxytocin-irrelevant* knowledge in vmPFC in item-level. Consistently, across the whole brain, mPFC [MNI coordinates, xyz = 0 62 14] and right lateral OFC (rlOFC, [MNI coordinates, xyz = 38 52 -10]), left middle frontal gyrus (LMFG, [MNI coordinates, xyz = -30 58 16]) were found associated with more elaborated neural representation in spray+ condition (Figure 3I, Table S4).

**Greater behavioral discrimination pattern between *Oxytocin-related information* and *Oxytocin-irrelevant information* on self-relatedness in spray+ condition explained the neural discrimination in prefrontal brain regions.**

#introduce the motivation to include behavioral DM and raise the hypothesis.

To delineate the psychological underpinnings of the neural discrimination among *Oxytocin-related information* and *Oxytocin-irrelevant information*. We constructed behavioral dissimilarity matrix (behavioral DM) and then did second-order representational similarity analyses based on these matrices. The second-order representational similarity analyses can be used to compare actual neural pattern similarity (or dissimilarity) to the similarity (or dissimilarity) predicted by a theoretical model (Kriegeskorte et al., 2008; Kriegeskorte and Kievit, 2013). More similar between neural pattern and behavioral pattern, more possible that the neural pattern reflected the psychological meaning of theoretical model.

#results-behavioral DM on trueness, comprehension, self-relatedness

In present study, all participants viewed all *Oxytocin-related* and *Oxytocin-irrelevant information* again and rated the self-relatedness for each sentence in out-scanner phase. We also measured participants' rating the trueness, comprehension to construct control behavioral DMs.

All ratings were on a 9-point Likert scale: 1 = not self-related at all/not true/cannot understand, 9 = highly self-related/ very true/completely understand. We first constructed 2×2 behavioral dissimilarity matrix that characterized representational dissimilarity of each pair of information-type (one pair here, the *Oxytocin-related & Oxytocin-irrelevant information*) on the self-relatedness, trueness, comprehension rating (Figure 3J). We found behavioral discrimination pattern was more evident (i.e., smaller correlation coefficient) on self-relatedness dimension in spray+ condition ($r(32)=0.299$) than control condition ($r(31)=0.778$; significant difference, $z = -2.760$, $p=0.005$), while the behavioral dissimilarity pattern on comprehension and trueness was not so distinct between spray+ and control condition (difference between spray+ and control condition, all $p > 0.200$). More precisely, we then constructed 60×60 behavioral dissimilarity matrix that characterized behavioral representational dissimilarity of each sentence on the three kinds of ratings (Figure 3K). As expected, behavioral DM on self-relatedness showed higher discrimination pattern in spray+ condition, while behavioral DM on comprehension and trueness not.

Next, we performed a secondary whole-brain searchlight RSA (Nili et al., 2014) to identify brain regions in which the pairwise dissimilarity of neural patterns of the sixty sentences corresponded to the behavioral dissimilarity matrix of the all knowledge (sixty sentences; here, technically, we

cannot do secondary searchlight RSA based on the 2×2 behavioral DM since there was only one dissimilar value, as well as the 2×2 brain DM, just two dissimilar values cannot be conducted further correlation analyses). The whole-brain searchlight RSA that incorporated the self-relatedness dissimilar matrix revealed that the patterns of neural activity in the left middle frontal gyrus [MNI coordinates, xyz = -22 42 26] corresponded to the behavioral dissimilarity matrix (Figure 3L, Table S5). No significant clusters were found to correspond to behavioral DMs on comprehension and trueness.

These results may suggest that the greater neural discrimination for the two types of information in prefrontal brain regions in spray+ condition is more likely involved in the psychological process, in which participants simulated future episodes about the placebo effects on themselves.

All behavioral DMs on each dimension used to do whole-brain searchlight RSA were combined the behavioral DM from the two treatment conditions to avoid potential double-dipping bias (Kriegeskorte et al., 2009, NN).

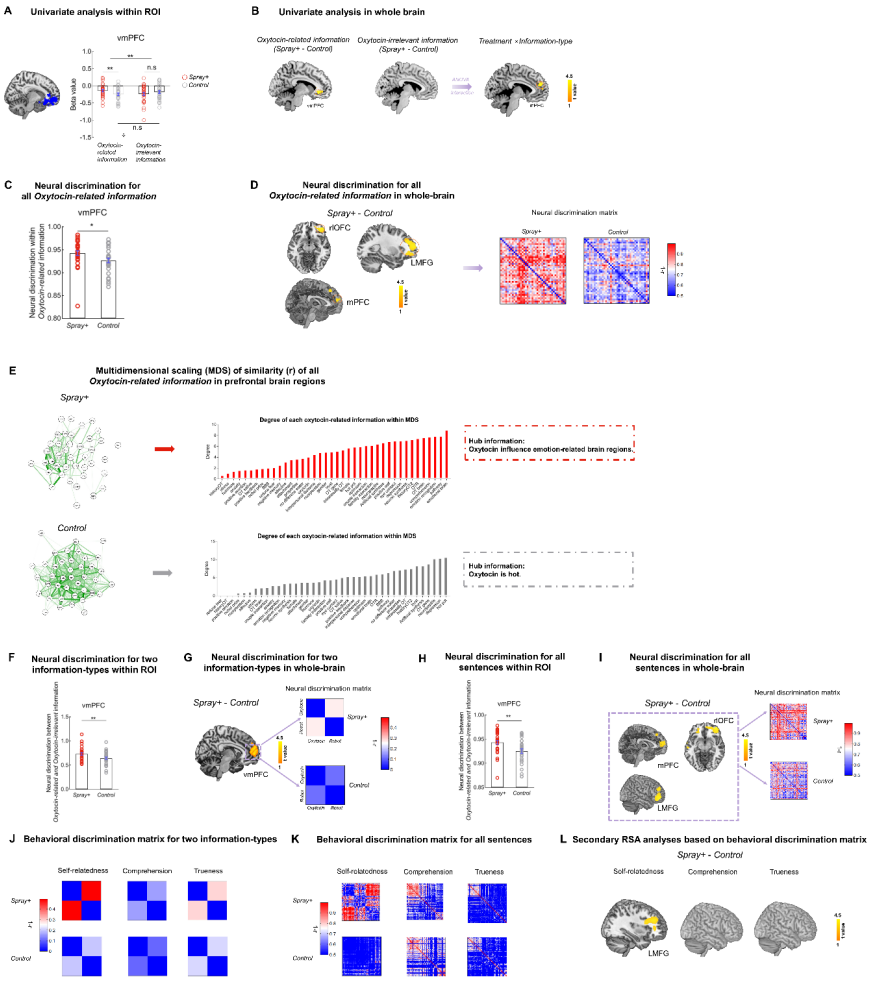
#results-behavioral DM on valence, arousal, familiarity

We also constructed behavioral DMs based on online rating including valence (*online exp1*), arousal (*online exp1*), and familiarity after learning oxytocin materials (combined data from *online exp2* & *online exp3*) and did whole-brain searchlight RSA to confirm whether the difference of information representation in two treatment conditions was related to valence, arousal, and familiarity. As predicted, there were no significant brain clusters emerged (Figure S3, Table S5). These results indicated that the distinct neural representation between *Oxytocin-related* and *Oxytocin-irrelevant information* in spray+ condition was independent of its valence, arousal and familiarity.

**Neural representation of *Oxytocin-related information* and *Oxytocin-irrelevant information* in social brain network showed similar pattern.**

To validate whether treatment-related brain regions (here, social brain network) would also show distinct neural representation pattern between *Oxytocin-related* and *Oxytocin-irrelevant information*. We extracted the neural discrimination value within *Oxytocin-related information*, as well as the discrimination value between *Oxytocin-related* and *Oxytocin-irrelevant* information (condition-based and trial-based level) within social brain network.

Similarly, the spray+ condition showed higher neural discrimination for *Oxytocin-related information* (main effect of treatment: F(1,60)=5.081, p=0.028, $\eta_p^2$ =0.078, control age, Figure 3M) in social brain network. Further, we found a marginally significant larger discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information in spray+ condition in condition-based level (F(1,60)=3.239, p=0.077, $\eta_p^2$ =0.051, control age, Figure 3N) and trial-based level (F(1,60)=5.889, p=0.018, $\eta_p^2$ =0.089, control age, Figure 3O).



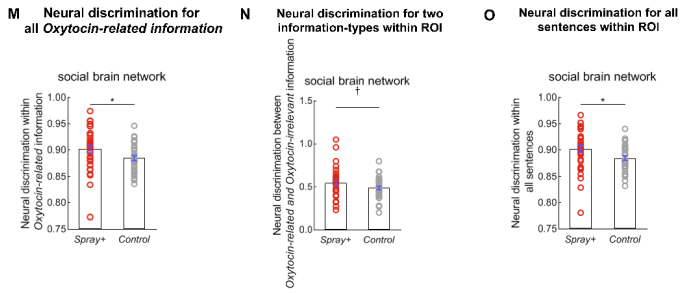



**Figure 3. vmPFC and other prefrontal brain regions encoded treatment-related information and represented the treatment-related information in elaborated and distinct pattern.**

(A) Results of region of interest for information representation task. *Oxytocin-related information* elicited higher neural response in vmPFC in spray+ condition, no treatment effect was found for *Oxytocin-irrelevant information*.

(B) Results from whole-brain univariate analysis. Consistently, *Oxytocin-related information* engaged stronger neural activation of vmPFC in spray+ condition than control condition, no significant clusters were found for *Oxytocin-irrelevant information*.

(C) Neural discrimination for all *Oxytocin-related information* in vmPFC. Spray+ condition allowed higher neural discrimination for all *Oxytocin-related information* than control condition.

(D) Neural discrimination for all *Oxytocin-related information* from whole-brain searchlight analyses. The prefrontal clusters (including rlOFC, left MFG, mPFC) encoded all *Oxytocin-related information* in more distinct way than control condition. The neural pattern discrimination matrices for all *Oxytocin-related information* in significant prefrontal clusters in each treatment were showed, obviously, *Oxytocin-related information* associated with more distributed neural pattern in spray+ condition.

(E) Multidimensional scaling of neural pattern discrimination for *Oxytocin-related* and *Oxytocin-irrelevant* information in each treatment. Multidimensional scaling of neural pattern discrimination was the averaged discrimination matrix based on all subjects in specific treatment condition. Further, we calculated the degree centrality to find which knowledge would be the hub knowledge in spray+ and control condition separately. The degree centrality

associated with each sentence was presented with ascending order. The Hub knowledge in each treatment condition was presented in the right panel.

(F) Condition-level pattern discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information in vmPFC in two treatment conditions. Spray+ condition associated with higher neural discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* in vmPFC, but control condition failed to differentiate these two types of information.

(G) Condition-level pattern discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* from whole-brain searchlight analyses. Results of whole-brain RSA were similar and consistent with ROI-based findings, the vmPFC had been identified to encode distinct neural representation between two types of information.

(H) Trial-wised pattern discrimination (among sixty sentences) between *Oxytocin-related* and *Oxytocin-irrelevant* information in two treatment conditions. Spray+ condition had larger trial-wised pattern discrimination than control condition did.

(I) Trial-wised pattern discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information from whole-brain searchlight analyses. The mPFC, right lOFC, and LMFG showed more discriminated neural pattern between two types of information in spray+ condition. The neural discrimination pattern in these significant clusters in each treatment were visualized in the right panel. LMFG, left middle frontal gyrus; rlOFC, right lateral orbitofrontal cortex.

(J) Behavioral dissimilarity matrices between *Oxytocin-related* and *Oxytocin-irrelevant* information from out-scanner rating on self-relatedness, trueness, comprehension. The self-relatedness dissimilar matrix showed significant distinct pattern in spray+ condition than control condition. But no significant differences were found on trueness and comprehension matrices.

(K) Trial-wised behavioral dissimilarity matrices (among sixty sentences) from out-scanner rating on self-relatedness, trueness, comprehension. Consistently, self-relatedness dissimilar matrix showed more distinct pattern in spray+ condition, while trueness and comprehension dissimilarity matrices not.

(L) Whole-brain results from second-order searchlight analyses based on the trial-wised behavioral dissimilarity matrices on self-relatedness, trueness, comprehension. Only the self-

relatedness behavioral dissimilarity matrix associated with significant brain clusters in LMFG in spray+ condition.

(M) Neural discrimination for all *Oxytocin-related* information in social brain network. Spray+ condition allowed higher neural discrimination for all *Oxytocin-related* information.

(N) Condition-level pattern discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information in social brain network in two treatment conditions. Spray+ condition associated with marginally higher neural discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information in social brain network.

(O) Trial-wised neural discrimination between *Oxytocin-related* and *Oxytocin-irrelevant* information in social brain network in two treatment conditions. Spray+ condition had larger item-wised neural pattern discrimination than control condition did.

(Error bars represent standard error, $p<0.08$†, $p<0.05$ *, $p < 0.01$ **, n.s, not significant).

**Shared neural representation for *Oxytocin-related* information in prefrontal brain regions.**

It has been known that the participants in spray+ condition showed more distinct and finer-grained neural representation for *Oxytocin-related* information in prefrontal brain regions. Does the neural representational pattern for the *Oxytocin-related* information be shared across all participants? Or each subject had its own representation pattern to make appraisal the placebo information? To address these questions, we calculated across-brain neural representation similarity for all *Oxytocin-related* information. Consistent with prior studies using between-subjects pattern analysis (Shinkareva et al., 2011; Chen et al., 2016), for each participant, the neural pattern of all sentences in *Oxytocin-related* condition was compared (i.e., Pearson correlation analysis) to the neural pattern of all same information in the remaining participants in each given sphere within the significant brain clusters from the neural representation for *Oxytocin-related* information (i.e., clusters in Figure 3I). This process was repeated reiteratively until every participant had been compared to the remaining averaged maps.

We identified the brain regions encoded the significant across-brain representation similarity in spray+ were the left middle frontal gyrus [MNI coordinate, xyz = -18/48/18] and right middle frontal gyrus [MNI coordinate, xyz =34/40/26] (Figure 4). These brain regions also involved in the *Oxytocin-related* information representation process for single participants. Thus, we may conclude that the participants shared the neural representation for treatment-related information in spray+ condition.

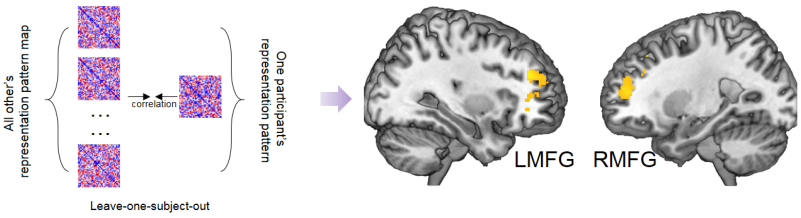



Figure 4. **Participants in spray+ condition had shared neural representation in prefrontal regions for *Oxytocin-related* information.** The LMFG and RMFG showed significant between-subjects representation similarity for all *Oxytocin-related* information. LMFG, left middle frontal gyrus; RMFG, right middle frontal gyrus.

**Social placebo effect on increasing motivation to seek social reward and avoid social punishment**

We adopted social incentive delay (SID) task (Spreckelmeyer et al., 2009; see Method) as the out-scanner behavioral measurement to investigate the placebo effectiveness on social reward. Participants had to press the button as quickly as possible to get potential reward or avoid punishment before a target symbol (a black diamond) disappeared on screen in a given time. Three kinds of cues ("triangle", "square", "circle") which were preceded by the rhombus trigger indicated the potential outcome were *happy face-oval shape* (reward trials), *angry face-oval shape* (punishment trials), *neutral face-oval shape* separately (neutral trials) (all stimuli were differed in each trial except the oval shape, stimuli were adopted from passive viewing social stimuli task). Faster reaction time or higher hit rate indicates higher motivation to get reward and avoid punishment (Knutson et al., 2001; Delmonte et al., 2012). Treatment effects were significant for hit rate in reward ($F(1,63)=4.094$, $p=0.047$, $\eta_p^2=0.063$) and punishment condition ($F(1,63)=8.037$, $p=0.006$, $\eta_p^2=0.116$), that the spray+ condition showed higher hit rate than control condition (Figure 5A), no significant results were found for neutral trials ($F(1,63)=1.481$, $p=0.228$, $\eta_p^2=0.024$). We did not observe the treatment effect on response time (all $p > 0.200$). And the time for participants to make response was no difference between two treatment groups ($F(1,63) = 0.045$, $p=0.833$, $\eta_p^2=0.001$).

**Multi-modal features served as biomarkers for classifying participants in spray+ condition from control condition.**

We conducted classification analysis to examine whether the spontaneous representation, stimuli-based representation of treatment-related information, and the placebo responses on social reward would be the critical biomarkers of social placebo effect. To avoid potential double dipping (Kriegeskorte et al., 2009), neural data (pattern discrimination and activation) entering the classification analysis were from the well-defined ROIs (i.e., social brain network, vmPFC from neurosynth). We extracted these features from each task and entered as features (all features' name see Table S6) in the classification analysis using supervised machine learning method, linear support vector machine algorithm (http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm/) with a standard 10-fold cross-validation.

The multi-modal features yield a classification accuracy of 89.47%. This classifier's sensitivity and specificity to spray+ were 86.21% and 92.86% .The classification results may further confirm that the neural account of placebo effect on social behaviors could be attributed to the enhanced connection strength within social brain network, finer-grained and distinct neural representation for placebo knowledge in vmPFC and increased sensitivity to social reward (Figure 5B).

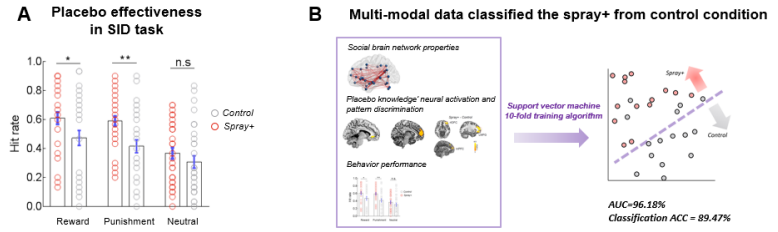



**Figure 5. Placebo effectiveness on behavior performance and classification results from multi-modal data.**

(A) Placebo effect on social incentive delay task. Spray+ increased the hit rate for seeking social reward avoiding social punishment.

(B) Multi-modal data classification results. Support vector machine algorithm with the social brain network properties, neural pattern results in vmPFC from information representation task, and behavioral placebo responses classified spray+ condition from control condition with accuracy of 89.47%. (Error bars represent standard error, $p<0.05$ *, $p < 0.01$ **, n.s, not significant).

## Discussion

# hard writing, para01- results summary, yeah!

Placebo effect is intriguing therapeutic-like effect which can be induced and modulated by treatment information. The "placebo", originally, is Latin for "*I shall be pleasing*". In current study, the neural representation of treatment information was exactly associated with brain regions (vmPFC, OFC, DLPFC) in which highly related to "*I*", as well the treatment-specific brain network to target another key word "*be pleasing*". The distinct neural representation of treatment information in future-

oriented, self-related, and treatment-related brain regions served as the important biomarkers for identifying placebo effect.

# hard writing, para02- resting, what does the enhanced MPFC-SFG mean?

Interregional communication in resting-state reflected the brain internal representation for recent experiences (Staresina et al., 2013). For example, participants would show stronger *fusiform face area-lateral occipital gyrus* (FFA-LO) coupling in resting scan after Face-Object encoding compared to baseline resting state (Tambini et al., 2010). Functional coupling between the bilateral motor cortices increased in resting scan following a motor task (Tung et al., 2013, NI). (talking about how recent experience would shape the spontaneous brain activity, as the background before introducing results, done) In current study, the spontaneous neural representation for recent experience that receiving treatment to enhance social functioning allowed increased coupling between vmPFC and superior frontal gyrus (SFG), and stronger functional connectivity strength within social brain network. (summary for resting-state results, done) The functional connectivity between vmPFC and superior frontal gyrus had been found to be related to goal-directed behaviors (Chase and Phillips, 2015), and the decoupling of these two brain regions indicated the social deficits in bipolar disorders (Marchand et al., 2014). Moreover, the vmPFC was located in anterior default mode network (Damoiseaux et al., 2008, CC) which were frequently activated in reflecting present self, while the superior frontal gyrus was the important component of posterior default mode network which had been identified when people were thinking about future self (D'Argembeau et al., 2010, SN). (details for explaining the FC between vmPFC and SFG, done) The enhanced functional coupling between vmPFC and SFG may suggest the potential psychological process that projecting current self into future social context (results implication#1, done)

# hard writing, para03- resting, what does the strengthened SBN mean?

The social brain network can be an overarching framework to understand the social cognition and capacities (Alcalá-López et al., 2018). Brain regions within the social

brain network have consistent neural activity increases during social cognitive process, such as social evaluation (McCormick et al., 2018) and social interaction (Schmälzle et al., 2017). The social brain network was crucial for social functioning and social well-being. Impaired activation or functional connectivity within social brain network has been identified to be correlated with disrupted social functioning in frontotemporal dementias (Marshall et al., 2019) and autism spectrum disorder (ASD; Anteraper et al., 2019). Dysmaturation of the social brain network has been found in newborns with family history of autism spectrum disorder (Ciarrusta et al., 2019). (details for introducing social brain network, done) We identified that the internal representation for believed treatment information had increased the functional coupling strength within social brain network, more specifically, we found the main connected hub--right NAc, showed stronger connection with brain regions within mentalizing network (i.e., PCC, Precuneus, left TPJ) after receiving placebo manipulation. Both the reward circuitry and mentalizing network had been found as the two key determinants underlying social bonding behaviors (Atzil et al., 2017). These evidence may suggest that participants in resting scan were not passively taking rest, but possibly showed active cognitive process such as simulating future episodes, forming the motivation to represent others' minds to prepare for future social behaviors (results implication#2, done).

# hard writing, para04- Why is the vmPFC to encode treatment information?

In the information representation task, concrete treatment information was mainly found to be encoded in vmPFC, in unique and fine-grained way. Moreover, the elaborated neural representation for treatment information in vmPFC and other prefrontal brain regions was tightly associated with the self-related psychological process. And the encoding and representation pattern was also common for all participants who received the "active" treatment. (results summary, done) The vmPFC is implicated in self-representation (Kelly et al., 2001; Sui et al., 2012), self-projection (i.e., imaging and projecting self into future scenarios; Buckner and Carroll, 2006) and tracking the value of outcomes after integrating information about long-term

goals (e.g., health value) (Hare et al., 2009, science). Moreover, studies of placebo analgesia reveal reliable activation increases in the vmPFC (Geuter et al., 2017), and the vmPFC activity was correlated with the strength of expectations of analgesia (Wager et al., 2011). Together, the vmPFC is more likely to involve in making meaning for personal well-being and future prospects, which would further drive psychological and physiological affective responses (Roy et al., 2012). (introduce vmPFC's function which was highly related current findings, done) We found that the vmPFC was crucial for treatment concepts encoding may suggest that participants who received believed active treatment may evaluate the treatment as beneficial, valuable thing for themselves, and integrate the treatment information for imaging future health beneficial situations to guide posterior behaviors. (implication#1 for PE process, done)

# hard writing, para05- vmPFC, what does elaborated neural representation mean?

The representational similarity analysis showed that in spray+ condition, the boundary between treatment-related information and treatment-irrelevant information in vmPFC was more clear both in category-level and item-level, and the spray+ condition allowed more elaborated neural representation pattern for all treatment-related items (i.e., all oxytocin-related information) (more specific RSA results summary, done) Generally, there have been some studies (Stolier and Freeman, 2016; Feng et al., 2018; Freeman et al., 2018) revealed the elaborated representation pattern (i.e., dissimilar representation) associated with distinct knowledge (e.g., male vs. female; self vs. stranger; black face vs. white face), that is, with less uncertainty to identify and recognize specific stimuli among others. The vmPFC was found to encode distinct neural representation between self and others (no self), at the same time, represented different dimensions of self-image in elaborated pattern while represented the dimensions in blurring way for others (Feng et al., 2018; Thornton et al., 2019). The between-category (i.e., treatment-related vs. treatment-irrelevant information) and within-treatment information fine-grained neural representation for treatment concepts in vmPFC and other prefrontal brain regions may suggest again that

participants represent the self in the context of future treatment benefits.
(implication#2 for PE process, done)

# hard writing, para06- inferring the potential PE process.

Although the current results only unraveled the treatment information representation pattern under social placebo framework, we speculated that other types of placebo effects (such as pain analgesia, reducing negative affect) may share the common principle to represent treatment situation. In general, we speculated the process to induce placebo effect could be: people rapidly integrate treatment information into vmPFC (as well as OFC, DLPFC) and treatment-related brain regions to form a high precision, elaborated constructed self-related internal model (i.e., generating expectations of the beneficial treatment effects on self), and further govern placebo responses.

# hard writing, para07- conclusion

In conclusion, the current study provides a novel perspective to understand the neural mechanism of placebo effect. We unravel the neural representation of treatment information, the important element to induce placebo effect, was located in self-related, value-related brain regions (vmPFC, OFC, DLPFC), as well as treatment-effect related brain regions in distinct, unique way. Out findings may unpack the critical neural process unerlying placebo responses and explain why giving treatment information would enable to induce placebo effects.

## Method

***Ethic approval***. The experimental procedures of all experiments were in line with the standards set by the Declaration of Helsinki and were approved by the local Research Ethics Committee of the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University. Participants provided written informed consent after the experimental procedure had been fully explained and were reminded of their right to withdraw at any time during the study.

***Sample estimation for fMRI study***. Prior to data collection, we conducted sample size estimation using G*Power3.1 (Faul et al., 2009) to determine the number of participants sufficient to detect a reliable effect. Based on the estimated effect size from our recent study examining placebo effect on social behaviors (Cohen d'= 0.774, Yan et al., 2018), 62 participants were needed to detect a significant effect ($\alpha = 0.05$, $\beta = 0.85$, two sample T test), We planned to recruit 66 participants (assuming 5% participants would be removed from the fMRI data analysis due to excessive head movement).

***Participants***. We recruited 569 participants as paid volunteers in all. Specifically, we recruited 66 males in fMRI experiments (age range: 18-31 years, mean age ± SD = 22.106 ± 2.856 years) as paid volunteers and randomly assigned to spray+ condition (N=33, mean age ± SD = 21.424 ± 2.264 years) or control condition (N=33, mean age ± SD = 22.788 ± 3.238 years). All participants in our current fMRI studies had normal or corrected-to-normal vision, and reported no history of neurological, endocrine or psychiatric disorders, who majored in psychology in college or recently participated in any other drug study were not recruited. And all participants were instructed to refrain from smoking or drinking (except water) for 2 h before the experiment. Consistent with previous empirical study of placebo effects on social behaviors (Yan et al., 2018), we recruited only male participants. The age between two treatment conditions (spray+ vs. control) reached marginally significant ($t(64) = -1.983$, $p=0.052$), so we controlled age in all analyses to exclude potentially confounding

influences induced by age difference. Participants in two treatment conditions were matched in relevant personality and mood-related traits (see Table S8).

To validate the experimental design for placebo manipulation phase, and stimuli used for information representation task and social incentive delay task, we recruited another 503 participants as paid volunteers for online experiments as follows: n=202 for online exp1 (145 females; age range: 18-56 years, mean age ± SD = 23.663 ± 4.925 years), n=72 for online exp2 (45 females; age range: 18-50 years, mean age ± SD = 23.736 ± 4.907 years), n=90 for online exp3 (63 females; age range: 18-48 years, mean age ± SD = 22.366 ± 3.657 years), and n=139 for online exp4 (69 females; age range: 18-45 years, mean age ± SD = 25.394 ± 3.818 years). Details for all online experiments see below.

***Online exp1.*** Participants were instructed to complete the rating task for valence and arousal for each sentence. Valence rating was on a 9-point Likert scale: -4 = very negative, 4 = very positive; arousal rating was scaled from 1 (not at all) to 9 (very high). We also measured the initial familiarity for oxytocin and robot at the beginning, participants were asked to report: 1) "How are you familiar with oxytocin?" (11-point Likert scale: 1 = not familiar at all, 9 = very familiar); 2) "How are you familiar with robot?" (11-point Likert scale: 1 = not familiar at all, 9 = very familiar).

***Online exp2.*** We conducted online exp2 in which participants learned oxytocin materials first and then rated the familiarity from 1 (not at all) to 9 (very familiar) for each sentence. We measured initial familiarity of oxytocin and robot as *online exp1* did.

***Online exp3.*** Participants were firstly instructed to rate the familiarity for the sixty sentences and then learned oxytocin materials, finally they rated the familiarity for all sentences again.

***Online exp4.*** *Online exp4* was conducted to examine the valence and arousal for the stimuli used in social incentive delay task for each condition (i.e., reward, punishment, neutral). A total of 130 participants actually completed the survey (63 females; age range: 18-45 years, mean age ± SD = 25.684 ± 4.092 years). Valence rating was on a 19-point Likert scale: -9 = very negative, 0 = neutral, 9 = very

positive; arousal rating was scaled from 0 (not at all) to 9 (very high). Valence in the three conditions was reward > neutral > punishment, statistics see Table S10. Moreover, valence rating for stimuli in reward condition was significantly larger than zero ($t(129)=30.035$, $p<0.001$), valence rating for punishment stimuli was significantly smaller than zero ($t(129)=-11.802$, $p<0.001$), and valence in neutral condition was no different with zero ($t(129)=1.754$, $p=0.082$). Arousal among the three conditions was reward > punishment > neutral, and stimuli in all conditions had higher arousal rating than medium level (all $p < 0.001$; one-sample T test, test value is 5) see Table S10. Stimuli and original data are available at https://github.com/psywalkeryanxy.

***Participants exclusion for formal experiments***. Four subjects in resting-state task (in spray+ condition, $N_{excluded}=2$; in control condition, $N_{excluded}=2$) were excluded from the subsequent analyses due to significant head movement (above 2.5 mm or 2.5° in any directions), and one participant was excluded due to his whole-brain functional connectivity value was outlier across all participants (>mean±3.5SD). Therefore, 61 participants were ($N_{spray+}=31$; $N_{control}=30$) included in the final data analysis. One subject in information representation task had incomplete functional images due to technique problem, and one participant were excluded owing to excessive head movement during scanning (above 2.5 mm or 2.5° in any directions), and one participant quit the experiment after resting-state session. Thus the final subject number for information representation task was 63 ($N_{spray+}=32$; $N_{control}=31$). For out-scanner phase, two participants did not complete the social incentive delay task. The final sample for this task was 64 participants ($N_{spray+}=32$; $N_{control}=32$).

***General procedure***. Participants first completed the Positive and Negative Affect Scale (PANAS, Clark and Tellegen, 1988), State Anxiety Inventory (SAI, Spielberger, 1985) which were used to measure their current mood. The participants were self-administered a nasal spray (saline spray; it was told as “oxytocin” in spray+ condition, told as “saline” in control condition) under experimenter supervision in placebo manipulation phase. After 10 min, participants were invited to the resting-

state, information representation task in fMRI scanner, and were introduced to complete social incentive delay task out of scanner (task illustration see Figure 1). Finally, participants filled out a survey for mood measurement again (PANAS and SA, results see Table S7), and personality questionnaires (results see Table S8), post-check questionnaire (results see Table S9).

***Experimenter-Blind procedure***. To avoid potential experimenter bias, experimenters of the placebo manipulations were blind to experimental hypotheses. Moreover, placebo manipulation and experimental tasks were conducted separately by different experimenters.

***Placebo manipulation phase***. In the spray+ condition, participants learned oxytocin materials on a self-paced basis and then were intranasally administered with saline (but it was told as "oxytocin"). The spray was administered to each participant three times, and each administration consisted of one inhalation into each nostril. Participants took a rest (they were told it was a time period waiting for treatment to produce effects) for 10 minutes and then performed the experimental tasks. In control condition, the materials and procedure were same with the spray+ condition except the nasal spray was told as "saline" instead. Oxytocin materials used in current experiments were adopted from previous social placebo study (Yan et al., 2018). Experimenters of the placebo manipulation phase were blind to experimental hypotheses.

***Resting state scan.*** In resting state session, participants with their eyes open and were instructed to attend to a black fixation cross centrally presented on a grey projection screen for 8 minutes (240TR).

***Experimental procedure for information representation task.*** The information representation task employed a mixed block and even-related fMRI design (Donaldson, 2004). There were two functional scans lasting 320s (160TR) in each session. Each functional scan consisted of six blocks (two blocks of each sentences category; (*i*) oxytocin-knowledge about its general knowledge; (*ii*) oxytocin-knowledge about its function; (*iii*) robot-knowledge) of 5 sentences from the same category. In each block, the sentence (from the same category) was presented for 5-9s

删除了： ***Stimuli for knowledge representation task.*** For the knowledge representation task, sixty short sentences were selected as experimental stimuli and separated into two categories including "*Oxytocin-knowledge*" (knowledge about oxytocin; these sentences were designed to delineate the knowledge about oxytocin and its effect on social functioning; such as "Oxytocin increases the interpersonal interaction"), and "*Oxytocin-irrelevant knowledge*". Here we included oxytocin-irrelevant knowledge as control stimuli to elucidate whether the representation for these information would be differed between the two treatments. We chose the robot-related knowledge (referred as *Robot-knowledge*) as oxytocin-irrelevant control materials based on previous social placebo study (Yan et al., 2018). Sentences in *Robot-knowledge* were made up from the core noun "robot", and the high frequency words from text analysis based on robot-related materials used in previous social placebo manipulation phase (Yan et al., 2018), such as "Robot can be used in hospital to help patients". Similarly, sentences in *Oxytocin-knowledge* were made up from "oxytocin", and the high frequency words from text analysis based on oxytocin materials used in placebo manipulation phase. Sentences in each category kept similar length and format, as well matched in valence and arousal (results see SI Section 1). ↵

(average duration = 7s), then followed by a fixation cross. The time duration for fixation cross was determined by the duration of the sentence and participants' response time. For example, if the presentation time for one sentence was 8s, and participants made response in 6s, then the fixation across was presented for 2s. The order of blocks and sentences were designed to present in pseudo-random order and were applied to all participants.

In each trial, participants were asked to make understand/not understand responses to each sentence by pressing one of the two buttons with the right index or middle finger. We combined the two sub-categories (i.e., general knowledge and function of oxytocin) of oxytocin knowledge (referred as *Oxytocin-related information*) due to the factor analyses based on online ratings failed to classify them into two distinct categories (SI, TableS?).

The self-reported comprehension rating in scanner for *oxytocin-irrelevant* information (*Robot-knowledge*) was no difference in two treatment conditions ($F(1,60)=0.744$, $p=0.392$, $\eta_p^2=0.012$), but showed difference for *Oxytocin-related information* (spray+ > control; $F(1,60) = 5.926$, $p=0.018$, $\eta_p^2=0.090$). Suggesting that the placebo manipulation was successful, that participants who believed they sprayed "oxytocin" had more clear and deep comprehension.

In out-scanner phase, participants were asked to complete a post-rating task for information representation task, in which participants were asked to report for each sentence: (*i*) "How true do you think this sentence is?" (on a 9-point Likert scale: 1 = not interesting at all, 9 = extremely interesting); (*ii*) "How much do you understand this sentence?" (1 = not understand at all, 9 = understand extremely well); and (*iii*) "How much self-relatedness do you feel for this sentence?" (1 = not relate to myself at all, 9 = extremely relate to myself). These three kinds of questions corresponded to the trueness rating, comprehension rating, and self-relatedness rating.

***Social incentive delay task.*** #general task description. The current study adopted revised social incentive delay task (Spreckelmeyer et al., 2009) which consisted of 90 trials. Each trial started with the appearance of one of the three cues on the screen for 240ms then followed by a fixation cross for 2250-2750ms and the target symbol for

160-280ms (depends on participants' performance in practice trials), finally presented with the outcome (duration for presenting outcome: 1650ms) according to participants' responses. #details about the task. Cues preceding target signaled potential reward (denoted by black triangle) or potential punishment (denoted by black square) or neutral stimuli (denoted by black circle). The time for presenting target symbol was adjusted by individual reaction time in practice session to maintain suitable task difficulty. If participant's hit rate was less than 50%, the target time would equal to his averaged reaction time across all trials in practice session (less than 800ms); otherwise, the time would be decreased by 15% of the practice response time (i.e., practice response time×85%) but was restricted to a range from 160ms to 280ms. Participants would gain social reward (reward condition) or avoid social punishment (punishment condition) or see the neutral outcome (neutral condition) upon successful reaction in time before a cued target symbol (a black diamond) disappeared on the screen. There were 30 trials in each condition and the presentation order for each trial was randomly set across all participants.

**Data analysis**

***Image acquisition***. Functional brain images were acquired using a 3-Tesla Siemens Trio scanner at the Beijing Normal University. Blood oxygen level-dependent (BOLD) gradient echo planar images (EPIs) were obtained using a 12-channel head coil [64 × 64 × 37 matrix with 3.44 × 3.44 × 5 mm spatial resolution, repetition time (TR) = 2000ms, echo time (TE) = 30ms, flip angle = 90°, field of view (FOV)=24 × 24 cm]. A high-resolution T1-weighted structural image (256 × 256 × 144 matrix with a spatial resolution of 1 × 1 × 1.33 mm, TR = 2530ms, TE = 3.37ms, inversion time (TI) = 1100ms, flip angle = 7°) was subsequently acquired.

***Data preprocessing for resting state.*** Preprocessing for resting-state data consisted of standard procedure, by using statistical parametric mapping (SPM12, Wellcome Trust Centre for Neuroimaging, London) and the Data Processing Assistance for Resting-State fMRI (DPARSF V4.4; Yan and Zang, 2010). The first 10 volumes of the functional images were discarded to avoid initial steady-state problems, followed by

preprocessing procedure including slice timing, motion correction, nuisance signals regression (including 24 head motion parameters (Friston et al., 1996) and five CompCorr signals (Behzadi et al., 2007)). Then the corrected and regressed functional images were normalized to the Montreal Neurological Institute (MNI) space and were resampled to 2 mm isotropic voxels, followed by spatial smoothing with a 4 mm full-width at half-maximum Gaussian kernel. Then the time series of BOLD signal were filtered with band-pass filtering (0.01-0.15 Hz) which would allow comparison of both resting-state and task data (Sun et al., 2004; Hearne et al., 2017).

***Functional connectivity estimation in resting state fMRI.*** We aimed to examine the differences in functional connectivity strength within social brain network between spray+ and control condition by conducting functional connectivity analyses. We adopted the well-defined social brain network (36 distinct regions) from a meta-analysis of neuroimaging studies in social neuroscience (Alcalá-López et al., 2017). After preprocessing, time-series were extracted from each ROI and then did Pearson correlation with each pair of time series within the social brain atlas. The resulting correlation coefficient (*r-value*) in each node yielded a 36 × 36 connectivity matrix for each participant. The mean functional connection strength was calculated by averaging the r-values within the matrix.

***Graph theory analysis.*** To ensure and increase biological interpretability, all graphic analyses were based on an established and extensively validated functional parcellation system (Power et al., 2014). We investigate the graph properties within the social brain network. The functional matrix was thresholded (to exclude the bias of specific density thresholds, threshold values ranged from 0.05 to 0.50 with an increment of 0.05) to yield binary networks. We investigated and calculated the network efficiency (global efficiency and local efficiency) by the tools implemented in the Brain Connectivity Toolbox (Rubinov and Sporns, 2010). To provide a systematic and stable network assessment, we calculated the network metric in a threshold-independent way by using the area under the curve (AUC, i.e., the integral) for further statistic evaluation. We also constructed weighted network, the results were similar with binary ones (see Figure S4).

Additionally, we extracted default-mode network, salience network from the power 264 nodes and constructed functional connectivity matrix within each network for each participant as control analyses.

***Data preprocessing for information representation task.***

Preprocessing for this task was performed using SPM12. The functional images were spatially realigned to the first volume to correct for head motion, and then corrected for slice acquisition timing. Subsequently, functional images were coregistered to each participant's gray matter image segmented from corresponding high-resolution T1-weighted image, then spatially normalized to the Montreal Neurological Institute (MNI) coordinate system and resampled into 2-mm isotropic voxels. Finally, spatially smoothed with an isotropic 6mm FWHM Gaussian kernel.

***Univariate analysis for information representation task.*** After preprocessing, we constructed first-level GLM that included two regressors that modeled "*Oxytocin-related* information" trials and "*Oxytocin-irrelevant* information" trials. All regressors were modeled with event durations of response time and were convolved with a double-gamma hemodynamic response function at the first level. In addition, six rigid-body realignment parameters (three translations and three rotations) of each participant were included as nuisance covariates to regress out effects related to head movement-related signal change. A high-pass temporal filter with a cut-off period of 256s was applied to remove high frequency noise. The resulting GLM was corrected for temporal autocorrelations using a first-order autoregressive model (AR(1)). In order to increase signal-to-noise ratio (SNR), global scaling was used to minimize the global intensity changes for each image. With this GLM, individual maps of parameter estimates were generated for three contrasts of interests: (1) *Oxytocin-related* information induced brain activation; (2) *Oxytocin-irrelevant* information induced brain activation; and (3) the difference between *Oxytocin-related* information and *Oxytocin-irrelevant* information in brain activation.

***Condition-based neural representation similarity analysis (RSA) for information representation task.*** We aimed to investigate whether the placebo treatment would differentiate the *Oxytocin-related information* and *Oxytocin-irrelevant information*. We conducted representational similarity analysis (RSA, Kriegeskorte et al., 2008; Nili et al., 2014) between *Oxytocin-related information* and *Oxytocin-irrelevant information*. The smaller neural similarity between *Oxytocin-related information* and *Oxytocin-irrelevant information* indicated higher neural discrimination (measured as 1 minus the neural similarity) between these two types of information in human brain. The preprocessing for neural pattern similarity analysis only included realignment of the functional images to the first volume, slice acquisition timing and coregistered to the T1 structural volume, thus the functional volumes remained unsmoothed and in their native space. For each voxel in the individual native brain image, a sphere with a radius of 15 mm was defined. The neural similarity between *Oxytocin-related information* and *Oxytocin-irrelevant information* was estimated by computing the pair-wise Pearson correlation for *Oxytocin-related* and *Oxytocin-irrelevant* condition using the multivoxel neural response value (i.e., beta value) for each category within the given sphere (Charest et al., 2014). Then the Fisher-transformed similarity value was assigned to the center voxel of the sphere.

***Trial-wised neural representation similarity analysis (RSA) for information representation task.*** We conducted trial-wised representational similarity analysis to precisely capture how the brain represent the knowledge in item-level. The processing procedure was same as condition-based RSA did. Then we conducted single-trial parameter estimates analysis via least squares separate (LSS, Mumford et al., 2012) to estimate the beta value for each sentence. Beta estimates of each trial (i.e., each sentence per category) was estimated in a separate GLM. The first regressor in the GLM was modeled as the regressor of interest, and the second regressor was the remaining trials. In all, we constructed 60 GLMs for 60 sentences. As in consequence, each category would have a beta-series vector (e.g., forty beta values in *Oxytocin-related information* and twenty beta values in *Oxytocin-irrelevant information*) in each voxel. Then for each voxel, a sphere with a radius of 15 mm was defined in the individual native brain space. The neural similarity among all sentences within the defined sphere was computed by the pair-wise Pearson correlation for each pair of sentences using the beta value from each sentence, and then averaged the Fisher-

transformed correlation value across all sentences or oxytocin sentences and assigned to the center voxel of the sphere. In the current study, we first calculated the pattern similarity within *Oxytocin-related information* and then computed the pattern similarity within sixty sentences.

***Trial-wised second-order representation similarity analysis (RSA) for information representation task.***

We compared the neural pattern dissimilarity model (i.e., neural DM) among all knowledge with the behavioral dissimilarity models (i.e., behavioral DMs including trueness, comprehension, self-relatedness, valence, arousal and familiarity) in each voxel of the brain using the searchlight procedure (Kriegeskorte et al., 2006). The neural DM was constructed by 1 minus the correlation coefficient between the multivoxel pattern vectors of each sentence pair. Next, the Spearman rank correlation between the neural DM and behavioral DMs were computed and assigned to the central voxel of the defined sphere. As such, the searchlight procedure produced Spearman ρ values on each voxel for each participant, which were then Fisher-transformation for further statistical analyses. Higher ρ values indicated more similar between the neural pattern and behavioral models

***Trial-wised across-brains neural representation similarity analysis for information representation task.***

The preprocessing procedure for between-subjects RSA echoed preceding method for trial-wised representational similarity analysis except transforming brain data to a common space (resampled to 2×2×2 $mm^3$ voxels) before doing whole-brain searchlight analysis. For each participant, the neural pattern similarity for each sentence in *Oxytocin-related information* was compared to the neural pattern similarity of the same sentence in the remaining participants in each given sphere (radius was 15mm). This process was repeated until every participant had been compared to the remaining averaged maps. The preceding analyses resulted in a single brain map with similarity value mapped in each voxel per participant.

***Group level analyses for all representation similarity analyses.***

For all representation similarity analyses, the correlation coefficients in each voxel were Fisher-z transformed prior to statistical tests. The searchlight procedure was conducted in the native space for every participant, and the resulting z maps were then normalized to standard space (resampled to $2 \times 2 \times 2$ $mm^3$ voxels), finally spatially smoothed with an isotropic 6mm FWHM Gaussian kernel and entered into group analyses.

Statistical significance was determined at the group level using a random effects analysis. Significant clusters from group level analyses were determined using a height threshold of $P < 0.001$ and an extent threshold of $P < 0.05$ with cluster-based FWE correction. We also applied voxel-wise inference using the FWE-corrected threshold of $P = 0.05$ on the whole-brain analysis, given recent concern over cluster-wise inferences. For condition-based representational similarity analysis between *Oxytocin-related* and *Oxytocin-irrelevant* information, the peak voxels in medial prefrontal gyrus survived voxel-wise FWE correction ($P = 0.007$). And peak voxels in rlOFC from the trial-wised representational similarity analysis for all sentences survived in voxel-wise FWE correction ($P = 0.021$).

***Classification analysis.***

We conducted classification analyses to examine whether the neural activity and behavioral social placebo responses would accurately classify spray+ treatment from control treatment. Specifically, the neural features included the social brain network functional connectivity strength, network efficiency indices, and the neural discrimination value (1- neural similarity) of brain regions which were found in information representation task, the behavioral features were the hit-rates in all conditions in social incentive delay task. The classification analysis using supervised machine learning method, linear support vector machine algorithm (http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm/) with a standard 10-fold cross-validation. We randomly divided the data into 10 subgroups, used the trained classifier from the nine subgroups to predict the performance on the left one subgroup, and repeated the

procedure for 10 times.

***Data and Code Availability***

Data and analysis scripts for this paper can be found at:

Open Science Framework: (link)

Github:

Clinical trials: (link)