bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title: A network supporting social influences in human decision-making

Authors: Lei Zhang[1,2]*, Jan Gläscher[1]*

Affiliations:

[1]Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany.

[2]Neuropsychopharmacology and Biopsychology Unit, Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, 1010 Vienna, Austria

*Correspondence to: lei.zhang@uke.de or glaescher@uke.de.

Short title: (40 characters): Neurocomputational model of social influence

Abstract:

Most human decision-making takes place in a social context, which could influence individual decisions. Using a novel probabilistic reversal learning task in a social setup, we observed opposite effects of social influence on choice and confidence. People succumb to the group when confronted with dissenting information whereas increase their confidence when observing confirming information. Using computational modeling and functional neuroimaging of goal-directed learning, we were able to separate normative influence leading to changes in choice behavior and informational influence resulting in changes in value computations, and identified their unique

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

neural representations. Subsequent valuation was accommodated by both reward prediction error and social prediction error. Moreover, we established an interaction of two brain networks related to processing reward and social information, which modulates behavioral adjustment.

One Sentence Summary:

Social influence modulates goal-directed learning and the interaction between the brain's reward hub and social hub.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Main Text:

Most of our everyday decisions are made in a social context. This affects both big and small decisions alike: we care about what our family and friends think of which major we choose in college, and we also monitor other peoples' choice at the lunch counter in order to obtain some guidance for our own menu selection. Behavioral studies have examined social influence as expressed by conformity (*1*) and have classified two major sources of social influence: normative and informational influence (*2–4*). Normative influence leads to public compliance, but individuals may maintain private beliefs; whereas informational influence hypothesizes that social information is integrated into the own valuation process. Neuroscience studies have recently attempted to assess the neurobiological underpinnings of both two types of influence (*5–8*). However, results are inconclusive, and more importantly, few of them have addressed the neurocomputational interaction between normative and informational influence in conjunction with individuals' own valuation processes. This is largely due to the challenge that most studies (*5–8*) relied on subjective judgment tasks where no feedback was given, which hindered the investigation of private belief, and due to a lack of a comprehensive computational model that quantifies and isolates latent determinants relevant to behavioral change.

Here we seek to establish such a comprehensive account of social influence in decision-making at the behavioral, computational, and neurobiological level including distinct, yet interacting brain regions instantiating social decision-making in humans. We ask whether social influence has a distinct neurocomputational representation, and how it is integrated with an individual's own value computation. Our model recapitulates crucial decision variables associated with behavioral

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

adjustments, allowing us to directly probe the network of interacting regions. We hypothesize that normative influence has its basis in mentalizing processes encoded in the right temporal parietal junction (rTPJ; *9–11*). Besides, we hypothesize that informational influence involves modulation of social learning signals by the anterior cingulate cortex (ACC; *12*, *13*). In addition, we anticipate that an individual's own valuation is computed via direct reinforcement learning (RL; *14*) encoded in the ventromedial pre-frontal cortex (vmPFC; *15*). We also propose an interaction of two brain networks related to processing social information (e.g., rTPJ) and to reward information (e.g., striatum), whose coupling is modulated by behavioral adjustment (*16*).

We test these hypotheses by employing a novel paradigm that allows multiple players to interact with each other in real-time while engaging in a probabilistic reversal learning task (PRL; *17*) using functional magnetic resonance imaging (fMRI). The task design of our multi-phase paradigm (fig. S1) enables us to tease apart every crucial behavior under social influence. Participants in groups of five (one in the MRI scanner) began each trial with their initial choice between two abstract fractals with complementary reward probabilities, followed by their first post-decision wager (an incentivized confidence rating; *18*). After sequentially uncovering their peers' first decisions in order of their subjective preference, participants had the opportunity to adjust both their choice and bet. The final choice and bet were then multiplied to determine the outcome on that trial (Fig. 1A). It is worth noting that participants' actual choices were communicated via intranet connection to every other participant, thus increasing the ecological validity of the task. These dynamically evolving group decisions allowed us to parametrically test the effect of group consensus, although participants were aware that outcomes were only dependent on their own choice and not that of the group, which prevented cooperative and competitive motives.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PRL requires participants to learn and continuously update action-outcome associations, thus creating enough uncertainty such that group decisions are likely to influence the choice and bet in the 2nd decision (i.e., inferring normative influence), and examine whether the others' learning behavior at the end of the trial was integrated into their own learning (i.e., implying informational influence).

Model-free analyses showed that 185 healthy participants indeed altered both their first choice and bet after observing the group decision, but in the opposite direction. Both second choices and bets were modulated by a significant interaction between the relative direction of the group (with vs against the participant's 1st choice) and the group consensus (2:2, 3:1, 4:0, view of each participant, fig. S2). Participants showed an increasing trend to switch their choice toward the group decision when faced with more dissenting social information, whereas, they were more likely to persist when observing agreement with the group (direction x consensus interaction, $F_{1,574} = 55.82$, $P < 0.001$) (Fig. 1B and table S1). Conversely, participants tended to increase their bets as a function of group consensus when observing confirming opinions, but sustained their bets when being contradicted by the group ($F_{1,734} = 4.67$, $P < 0.05$) (Fig. 1C and table S2). We further verified the benefit of behavior adjustment: social information facilitated learning (fig. S3, table S3, table S4), demonstrating that the adjustment is not a result of perceptual salience. Next, we sought to uncover what computations underlay these behavioral adjustments.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Using computational cognitive modeling, we aimed to formally quantify the latent mechanisms by dissociating the two types of social influence at the computational level, and particularly, by unraveling how informational influence was incorporated into one's own learning process. In addition to existing RL models on social influence (*19, 20*), our model accommodates multiple players, and is able to simultaneously estimate participants' two choices and two bets under the hierarchical Bayesian analysis workflow (*21*). Our efforts to construct these models were guided by two design principles: (1) separating of the individual's own value (V_{self}) and the vicarious value of others (V_{other}), which were combined into a choice value for the 1st choice (V_{combined}) using linear weighting, and (2) separating instantaneous normative social influence on the second choice and social learning from observing the performance of other players. We modeled the second choice as a function of two counteracting influences: (1) the group dissension ($\text{N}_{\text{against}}$) representing the instantaneous influence and (2) the difference between the participants' action values in the 1st choice ($V^{\text{chosen}} – V^{\text{unchosen}}$) representing the distinctiveness of the current value estimates. When this value difference is large, participants are less likely to succumb to social influence from dissenting information. When all outcomes were delivered at the end of the trial, both own and vicarious value were updated on a trial-by-trial basis: V_{self} was updated with a reward prediction error (RPE; *22*); meanwhile, V_{other} was updated through tracking a preference-weighted discounted reward history of all four other group members (Fig. 1D). By validating the necessity of the social learning aspect and ruling out four other competing observational learning processes (fig. S4 and table S5), we demonstrated that the above model provided the best out-of-sample predictive power, and its posterior prediction indeed captured behavioral findings of our model-free analyses (Fig. 1B and 1C, *23*).

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We then sought to establish the functional association between model parameters and the model-free behaviors. Parameter results (fig. S5 and S6) hinted that the extent to which participants learned from themselves and from the others was on average comparable, suggesting that an integrated value from one's direct learning and informational influence was computed to guide future decisions. Furthermore, parameters related to normative influence were well-capable of predicting the individual difference of participant's behavioral adjustment (Fig. 1E and 1F). Once we had uncovered the latent elements of the decision processes under social influence, we were then able to test how they were computed and implemented at the neural level using model-based fMRI (*24*).

The first part of our imaging analyses focused on how distinctive decision variables (table S6) were parametrically modulated in the brain. Our model distinguished between two value signals and suggested that an integrated signal was associated with participants' initial action and bet. Consequently, we now aimed to test the hypothesis that distinct and dissociable brain regions were recruited to implement these computational signals. Indeed, using a double-dissociation approach, we found that the chosen V_{self} was exclusively encoded in the vmPFC and the chosen V_{other} was exclusively mediated by the ACC gyrus (ACCg) (Fig. 2 and table S7). In addition, the medial prefrontal cortex (mPFC) was functionally coupled with both vmPFC and ACCg (fig. S11 and table S10), suggesting a neural encoding for the integrated value signal. Besides, the RPE was firmly associated with activities in the nucleus accumbens (NAcc), a region that is well-studied in the literature (*22*). However, to qualify as a region encoding an RPE signal, activities in the NAcc ought to covary positively with the actual outcome (i.e., reward) and negatively with the

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expectation (i.e., value) (fig. S8 and table S7). This property thus provides a common framework to test the neural correlates of any error-like signal (*12*).

We next turned to disentangle the neural substrates of the instantaneous social influence and the subsequent behavioral adjustment. As we have validated enhanced learning using such social information (fig. S3), we reasoned that participants might process other co-players' intentions relative to their own first decision to make subsequent adjustments. Based on this reasoning, we assessed the parametric modulation of preference-weighted $N_{against}$, and found that activity in rTPJ and other regions was positively correlated with the dissenting social information (fig. S9 and table S8). In addition, the resulting choice adjustment (switch > stay) covaried with activity in bilateral dorsolateral prefrontal cortex (dlPFC) (*17*, *25*), commonly associated with executive control and behavioral flexibility. In contrast, the vmPFC was more active during stay > switch trials (*17*), reminiscent of its representation of one's own valuation (fig. S10 and table S9). In summary, our model-based fMRI analysis (*24*) revealed two distinct brain networks representing social information and reward and value processing.

In a next step, we sought to establish how these network nodes are functionally coupled to bring about socially-induced behavioral change and to uncover additional latent computational signals that would otherwise be undetectable by conventional general linear models. Using a psycho-physiological interaction (PPI, *26*) seeded in rTPJ, we investigated how behavioral change at the 2nd decision modulated the functional coupling between the social information represented in rTPJ and other brain regions. This analysis identified the left putamen (lPut) (Fig. 3A, 3B and table S10). Closer investigations into the computational role of lPut revealed that it did not correlate

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with the two components of an RPE (fig. S12), but rather positively with the actual agreement (approximated by 1-$N_{against}$) and negatively with the expected agreement (approximated by the value difference V^{chosen} - $V^{unchosen}$, as individuals who maintain a larger value difference may expect more agreement), effectively encoding a social prediction error (SPE, Fig. 3C). Taken together, these analyses demonstrate that functional coupling between neural representations of social information and an SPE is enhanced, when this social information is leading to a behavioral change.

In a last step, using a physio-physiological interaction (*26*) we investigated how the neural representation of switching at 2nd decision in the left dlPFC modulated the functional coupling of rTPJ and other brain regions. This analysis revealed that activity in rTPJ positively modulated the coupling between vmPFC and ACC, which strikingly overlapped with the regions that represented the two value signals (Fig. 3D-F, table S10). Therefore, it seems that the interplay of neural representations of social information and the propensity for behavioral change leads to the updating of both values signals obtained via both direct learning and observational learning.

Social influence is a powerful modulator of individual choices (*27*). In a comprehensive neurocomputational approach to social decision-making we were not only able to identify a network of brain regions that represents and integrates social information of others, but also characterize the computational role of each node in this network in detail (Fig. 4), suggesting the following process model: one's own decision is guided by a combination of value signals from direct learning (V_{self}) represented in vmPFC (Fig. 2C; *15*) and from observational learning (V_{other}) represented in a section of ACC (Fig. 2B) that is also closely related to estimates of the volatility

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of others' choices (*12*) and to error detection and response conflict resolution (*28*, *29*). The decisions of others are encoded with respect to the own choice in rTPJ (fig. S9), an area linked, but not limited to representations of social information and agents in a variety of tasks (*10–12*, *30*). In fact, rTPJ is also related to Theory of Mind (*9*) and other integrative computations such as multisensory integration (*31*) and attentional processing (*32*). Dissenting social information gives rise to a novel social prediction error (actual and expected agreement with group decision) represented in lPut (Fig. 3A, 3C and fig. S12), unlike the more medial NAcc, which exhibits the neural signature of a classic RPE (fig. S8; *22*). The interplay of lPut and rTPJ affects behavioral change toward the group decision and in combination with its neural representation of choice switching in left dlPFC (Fig. 3D). This triggers the update of direct learning in vmPFC (V_{self}) and observational learning in ACC (V_{other}), thus closing the loop of decision-related computations in social contexts.

In summary, we show behavioral and computational evidence that normative social influence alters individuals' action and confidence, and informational social influence is incorporated into their own valuation processes. Moreover, we found a network of distinct, yet interacting brain regions substantiating specific computational variables. Such a network is in a prime position to process decisions of the sorts mentioned in the beginning, where – as in the example of a lunch order – we have to balance our own experienced-based reward expectations with the expectations of congruency with others and use the resulting error signals to flexibly adapt our choice behavior in social contexts.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

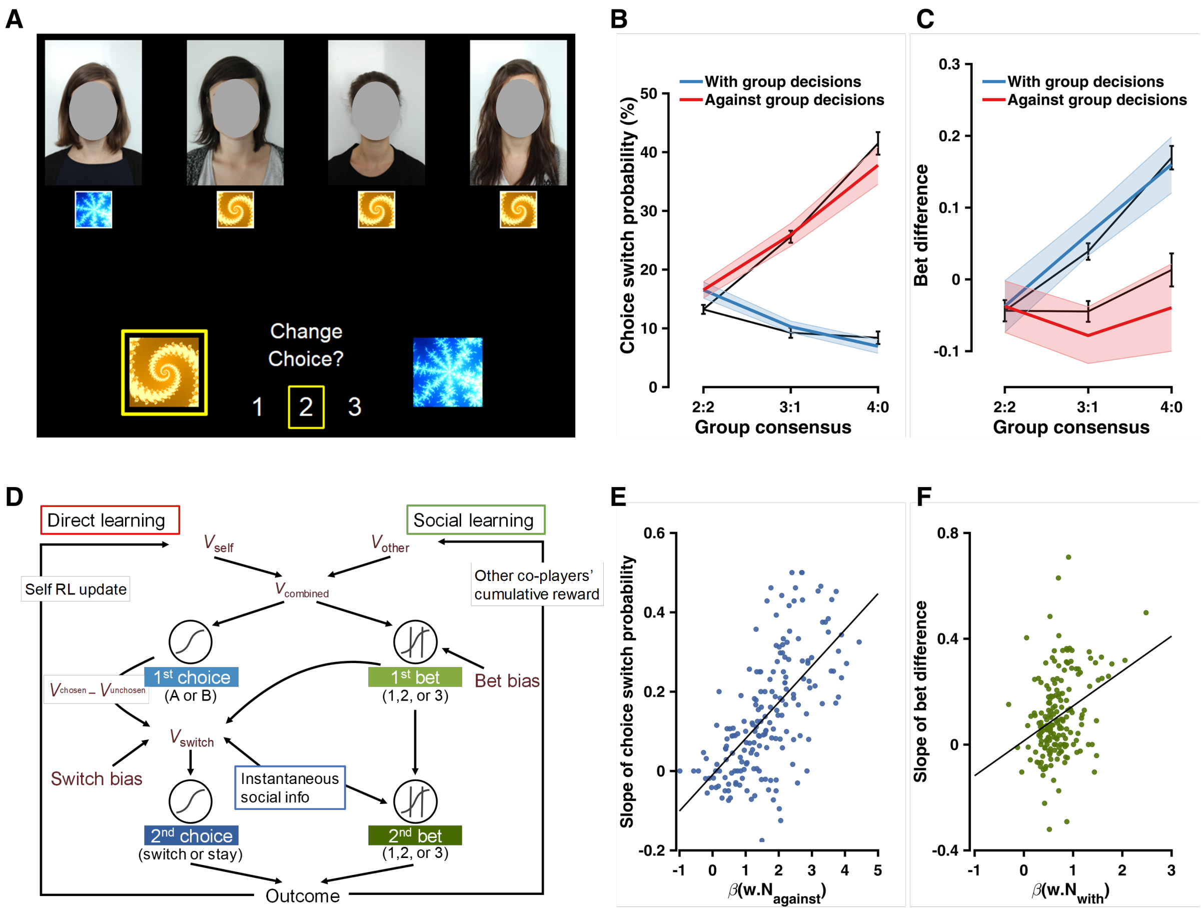

Fig. 1. Experimental task, behavioral results, and computational model. (**A**) Excerpt of the experimental task display (2nd choice). For a full trial display, see fig. S1. (**B**) Switch probability at 2nd choice and **(C)** bet difference (2nd bet – 1st bet) as a function of the majority of the group's 1st decision (with, against) and the group consensus. All black lines indicate actual data (± within-subject SEM). Shaded error bars represent the 95% highest density interval (HDI) of the mean effect computed from the winning model's entire posterior samples (posterior predictive check). **(D)** Schematic of the computational model (see main text and supplement for details). **(E)** Relationship between contradicting social information (preference-weighted $N_{against}$) and the susceptibility to social influence (slope of switch probability). **(F)** Relationship between confirming social information (preference-weighted N_{with}) and the bet difference.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

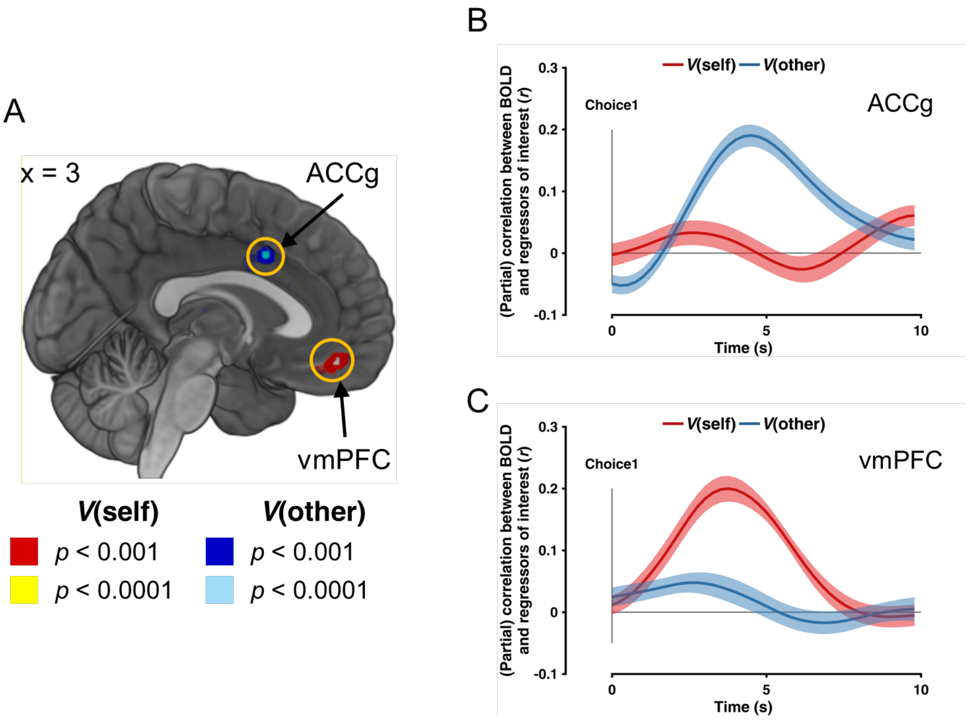

Fig. 2. Neural substrates of dissociable value signals. (**A**) The neural representation of $V_{\text{self}}^{\text{chn}}$ and $V_{\text{other}}^{\text{chn}}$ are encoded in the vmPFC (red/yellow) and the ACCg (blue/light blue), respectively. (**B**) and (**C**) Time series estimates (*12*) demonstrate a double dissociation of the neural signatures of the value signals. (**B**) The vmPFC is positively correlated with $V_{\text{self}}^{\text{chn}}$, but not with $V_{\text{other}}^{\text{chn}}$, whereas (**C**) the ACCg is positively correlated with $V_{\text{other}}^{\text{chn}}$, but not with $V_{\text{self}}^{\text{chn}}$.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

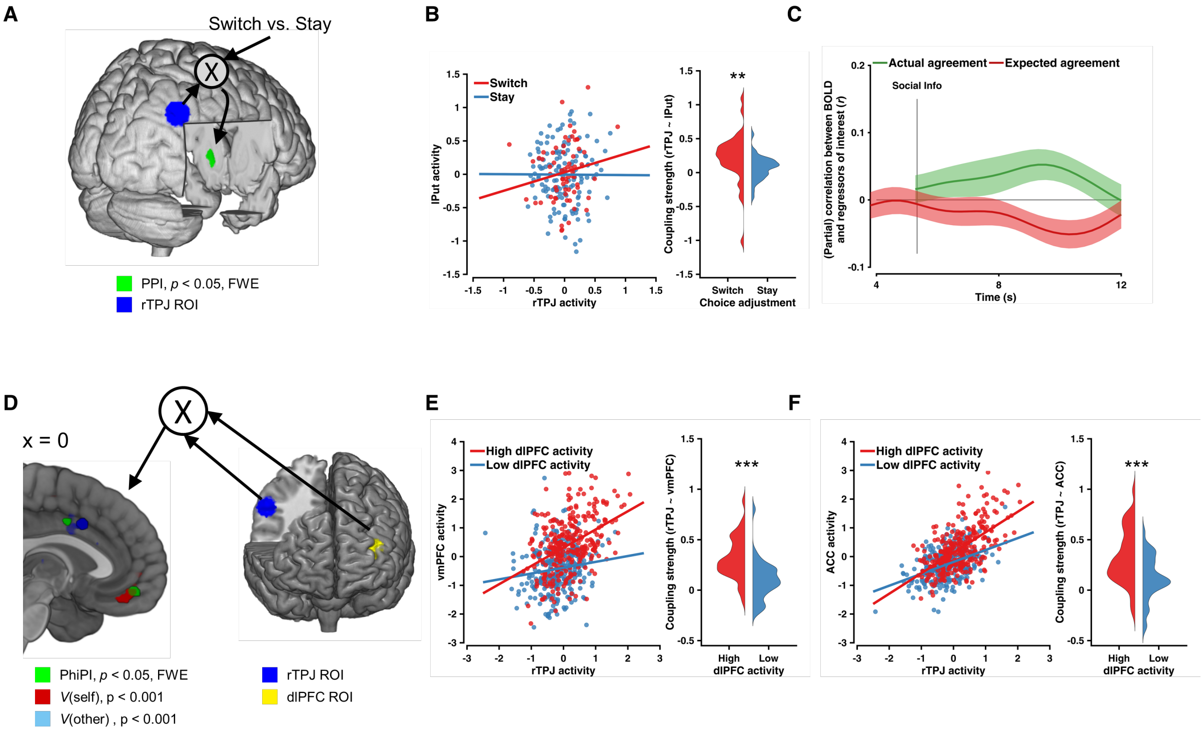

Fig. 3. Functional connectivity between reward-related regions and social-related regions. (**A**) The functional connectivity between the left putamen (green) and the seed region rTPJ (blue) is modulated by the choice adjustment (switch vs. stay). (**B**) Correlation of activity in seed and target region for both switch and stay trials in an example subject and histogram of coupling strength across all participants for switch and stay trials. **(C)** The BOLD time series in the left putamen (PPI target) exhibits a social prediction error (positive correlation with the actual agreement, and negative correlation with the expected agreement). (mean effect across participants ± SEM. (**D**) Two seed regions, the rTPJ (blue), which responds to the social information, and the left dlPFC (yellow), which encodes the choice adjustment, elicit connectivity activations in the vmPFC and the pMFC (both in green), which partially overlap with the latent value signals (i.e., $V_{\mathrm{self}}^{\mathrm{chn}}$; red; and $V_{\mathrm{other}}^{\mathrm{chn}}$; blue), as in Fig. 2A. **(E)** and **(F)** Correlation plots of seed and target regions for both high and low dlPFC activity in an example subject and histograms of seed-target coupling strengths across all participants for high and low dlPFC activity.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

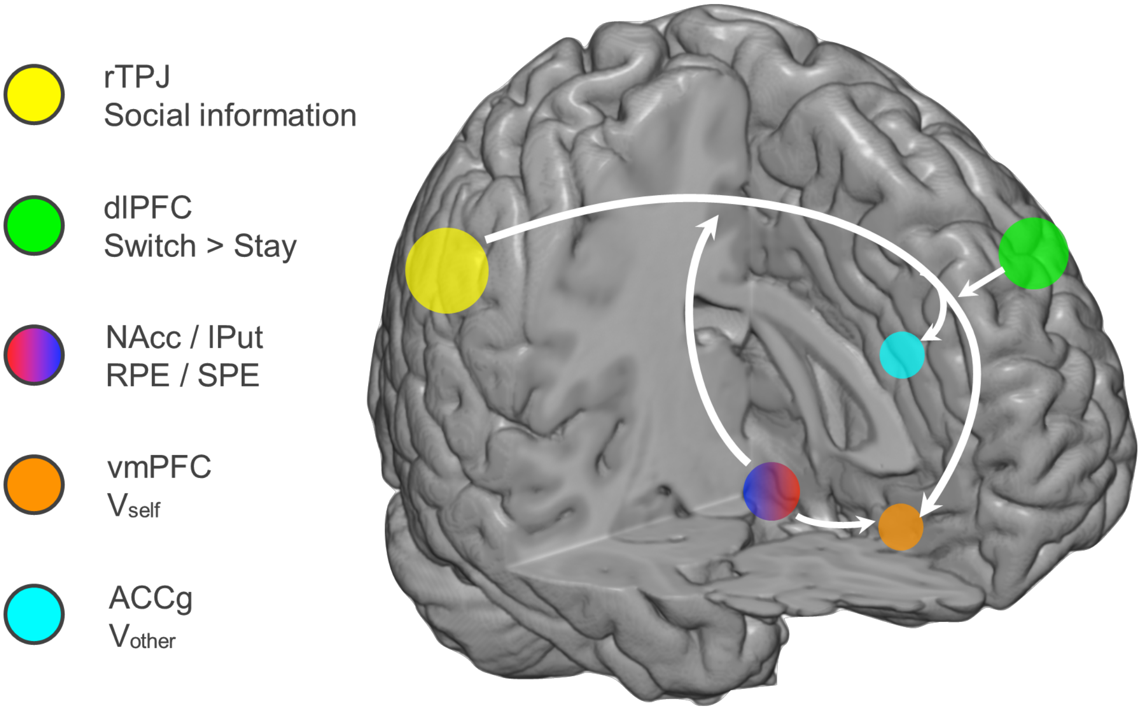

Fig. 4. Schematic of the of the network supporting social influence in decision-making as uncovered in this study (for details see main text)**.** 5

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References and Notes:

1. S. E. Asch, Psychol. *Monogr. Gen. Appl.* **70**, 1–70 (1956).
2. R. B. Cialdini, N. J. Goldstein, *Annu. Rev. Psychol.* **55**, 591–621 (2004).
3. U. Toelch, R. J. Dolan, *Trends Cogn. Sci.* **19**, 579–589 (2015).
4. E. Fehr, I. Schurtenberger, *Nat. Hum. Behav.* **2**, 458–468 (2018).
5. V. Klucharev, K. Hytönen, M. Rijpkema, A. Smidts, G. Fernández, *Neuron*. **61**, 140–51 (2009).
6. D. K. Campbell-Meiklejohn, D. R. Bach, A. Roepstorff, R. J. Dolan, C. D. Frith, *Curr. Biol.* **20**, 1165–1170 (2010).
7. J. Zaki, J. Schirmer, J. P. Mitchell, *Psychol. Sci. a J. Am. Psychol. Soc. / APS*. **22**, 894–900 (2011).
8. S. A. Park, S. Goïame, D. A. O'Connor, J.-C. Dreher, *PLoS Biol.* **15**, e2001958 (2017).
9. C. D. Frith, U. Frith, *Science*. **286**, 1692–1695 (1999).
10. R. Saxe, N. Kanwisher, *Neuroimage*. **19**, 1835–1842 (2003).
11. A. N. Hampton, P. Bossaerts, J. P. O'Doherty, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 6741–6746 (2008).
12. T. E. J. Behrens, L. T. Hunt, M. W. Woolrich, M. F. S. Rushworth, *Nature*. **456**, 245–249 (2008).
13. S. Suzuki *et al.*, *Neuron*. **74**, 1125–1137 (2012).
14. R. S. Sutton, A. G. Barto, *Introduction to reinforcement learning* (MIT press Cambridge, 1998)
15. O. Bartra, J. T. McGuire, J. W. Kable, *Neuroimage*. **76**, 412–427 (2013).
16. T. A. Hare, C. F. Camerer, D. T. Knoepfle, J. P. O'Doherty, a. Rangel, *J. Neurosci.* **30**, 583–590 (2010).
17. J. Gläscher, A. N. Hampton, J. P. O'Doherty, *Cereb. Cortex*. **19**, 483–495 (2009).
18. N. Persaud, P. McLeod, A. Cowey, *Nat. Neurosci.* **10**, 257–261 (2007).
19. G. Biele, J. Rieskamp, L. K. Krugel, H. R. Heekeren, *PLoS Biol.* **9** (2011).
20. A. O. Diaconescu *et al.*, *PLoS Comput. Biol.* **10**, e1003810 (2014).
21. B. Carpenter *et al.*, *J. Stat. Softw.* **76** (2017)
22. W. Schultz, P. Dayan, P. R. Montague, *Science.* **275**, 1593–1599 (1997).
23. Materials and methods are available as supplementary materials online.
24. J. D. Cohen *et al.*, *Nat. Neurosci.* **20**, 304–313 (2017).
25. C. J. Burke, P. N. Tobler, M. Baddeley, W. Schultz, *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14431–14436 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. K. J. Friston *et al.*, *Neuroimage*. **6**, 218–229 (1997).
27. C. C. Ruff, E. Fehr, *Nat. Rev. Neurosci.* **15**, 549–562 (2014).
28. C. S. Carter *et al*, *Science.* **280**, 747–749 (1998).
29. M. M. Botvinick, *Cogn. Affect. Behav. Neurosci.* **7**, 356–366 (2007).
30. S. Suzuki, R. Adachi, S. Dunne, P. Bossaerts, J. P. O'Doherty, *Neuron*. **86**, 591–602 (2015).
31. M. Tsakiris, L. Carpenter, D. James, A. Fotopoulou, *Exp. Brain Res.* **204**, 343–352 (2010).
32. M. Corbetta, G. L. Shulman, *Nat. Rev. Neurosci.* **3**, 201–215 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments:

We thank N. Daw for help in developing the computational model, A. Bert, K. Weisel, J. Spilcke-Liss, J. Majewski, and all radiographers for help with data acquisition; and xx for helpful feedback on earlier versions of the manuscript. **Funding:** J. G. was supported by the Bernstein Award for Computational Neuroscience (BMBF 01GQ1006), the Collaborative Research Center "Cross-modal learning" (DFG TRR 169), and the Collaborative Research in Computational Neuroscience (CRCNS) grant (BMBF 01GQ1603). L. Z. was supported by the International Research Training Groups "CINACS" (DFG GRK 1247), and the Research Promotion Fund (FFM) for young scientists of the University Medical Center Hamburg-Eppendorf. **Author contributions:** JG. conceived the research idea. LZ and JG designed experiments. LZ acquired data. JG and LZ designed computational model, performed analyses and wrote the paper. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** Data and analyses scripts are available upon request to the authors.

bioRxiv preprint doi: https://doi.org/10.1101/551614; this version posted February 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Materials:

Materials and Methods

Supplementary Text

Figures S1-S12

Tables S1-S10

References and Notes (33-62)