Neurocomputational approaches to social behavior

Konovalov, Arkady ; Hu, Jie ; Ruff, Christian C

DOI: https://doi.org/10.1016/j.copsyc.2018.04.009

Neurocomputational Approaches to Social Behavior

Arkady Konovalov*, Jie Hu*, Christian C. Ruff*

Laboratory for Social and Neural Systems Research (SNS-Lab)
Department of Economics
University of Zurich
*All authors contributed equally to this work

Abstract

Social decision-making is increasingly studied with neurocomputational modelling. Here we discuss how this approach allows researchers to better understand and predict behavior in social settings. Using examples from the study of resource distributions and social learning, we illustrate how this methodology provides a flexible way to quantify social values and beliefs, identify specific motives and cognitive processes underlying social choice and learning, and arbitrate between competing theories of social behavior. We also critically discuss open questions and potential problems associated with this methodology.

Introduction

How humans behave in social settings is of central interest to the behavioral sciences. Classic theories propose that social behavior can be predicted based on either motives of the agent (e.g., [1]) or characteristics of the specific social situation (e.g., [2]). While these theories have been invaluable for structuring research, many of them have been purely qualitative. This has made it hard to predict the strength of corresponding behavioral effects and to determine how factors captured by different theories interact in a given setting.

One way to overcome these problems is to employ neurocomputational models of affective and cognitive processes underlying social behavior. Most of these models were developed to capture decision-making, learning, and motivation in non-social settings, and some of them even originated in academic disciplines not focused on behavior (e.g., computer science). However, adaptations of these models allow researchers to formally integrate both personal and situational characteristics into a model of the dynamic control of social behavior.

Here we outline some of the advantages (and potential shortcomings) of this approach, by discussing modelling studies of two aspects of social behavior: Resource allocation and social learning. We briefly introduce the models employed in these two domains and discuss example studies addressing questions that would be hard to study with purely qualitative approaches. We close by summarizing potential shortcomings and open questions in this field of study.

Modelling resource allocations

An important problem for families, groups, organizations, and societies is the question how resources and costs (time, money, attention, effort, etc.) should be distributed among its members. How humans determine these allocations has puzzled researchers for a long time, leading to different proposals about the motives (e.g., concerns about fairness, efficiency, or self-interest [3]) or context factors (e.g. others' behavior or social distance [4]) that may guide these choices. However, which motive or factor may be most important has been debated for decades [5,6]; it is furthermore unclear how these motives and factors interact in given contexts. This question is difficult to study with comparisons of discrete experimental conditions, which usually only vary one factor or motive at a time [7–9].

To overcome this limitation, researchers have begun to employ computational models to investigate these questions in one overarching framework. The main class of models used for this purpose originated in behavioral economics and assumes that humans select actions to maximize their projected utility (often referred to as "value" in psychology/neuroscience models; [10]). Using this approach allows researchers to quantify how strongly different motives and context factors influence behavior, by specifying a model of how the agent should transform relevant experimental variables (e.g., payoffs for different people) into expected utilities/values of choice options (e.g., accepting/rejecting the distribution). Fitting this model to observed choices with well-established computational methods (reviewed e.g. in [11]) makes it possible to infer the strength of the corresponding motive. For example, a popular type of such models quantifies inequity aversion, the preference for maximizing fairness (i.e., minimizing inequity) in the resource distribution, by including the payoff difference between two participants as a component into the utility function ([12], see **Figure 1A**).

Using these models has allowed researchers to make progress in understanding determinants of resource distribution behavior in several ways. First, the models' quantitative predictions allow conclusions that go far beyond the statement that people "care about fairness". For instance, studies using computational models can not only capture inter-individual differences in resource-distribution choices and the underlying brain structures but can also predict how a given individual will respond to changes in the cost of altruistic behaviors [13]. Moreover, the models' quantitative predictions make it possible to identify brain activity correlating with each trial's model-inferred motivational tendencies, rather than just observed choices. Any correspondence between predicted motives and neurophysiological processes thus empirically supports the model assumptions (see also Figure 2). It would be hard to derive such predictions using purely qualitative theories about the role of fairness concerns for resource distribution behavior.

Second, the fact that different motives can simultaneously impact on perceived utility allows researchers to clarify the relationship between different internal motives underlying distribution behaviors. For example, apart from concern for inequity/fairness, people also care about efficiency of resource distributions (the total welfare gained across all group members by any choice) [5]. By integrating these two components in a single model, studies have shown that concerns about inequity versus

efficiency reflect at least partially distinct, parallel motivational tendencies: Neural activity in different brain structures was found to correlate with each distribution's inequity (insula), efficiency (putamen), or model-derived utility (caudate /septal subgenual region) [14]). Moreover, pharmacological manipulation of the neurotransmitter dopamine causally modulated individuals' inequity aversion without influencing concerns about others' payoff [15], further emphasizing the distinct neurocomputational implementation of both competing motives.

Third, use of neurocomputational models allows researchers to test how different contexts influence the expression of individual motives and the resulting social behavior [16]. For instance, people tend to be more generous toward close others than to distant strangers [17]. Use of a social discounting model made it possible to characterize and predict how any given social distance influences the value placed on the other's payoff in the utility function [18]. The plausibility of this model was underlined by findings that the discounted utility curve correlated with neural activity in one specific brain area (the temporoparietal junction TPJ; [19]), and that brain-stimulation-disruption of neural activity in this brain area systematically changed the influence of social distance on resource-sharing [20].

Fourth, the detailed predictions of competing models make it possible to formally select the model best able to explain behavior and correlated physiological processes. For instance, unequal distributions not only elicit motivational tendencies modelled with utility functions but also impact on beliefs about how other people should act in these contexts [16,21]. A recent study modelled these belief changes in a Bayesian framework (**Figure 1B**) and could predict how participants adapted their expectations about others' distribution offers in different contexts [21]. This belief model was better at accounting for observed behavior than utility-based inequity-aversion models, suggesting that under certain circumstances, expectations about others' behavior, rather than concerns for inequity, can dominate distribution behaviors [22].

Last but not least, use of neurocomputational models makes it possible to identify the degree to which apparently fair resource distribution choices result from genuine fairness concerns or rather from noisy, unsystematic behavior. This was made possible by the use of drift diffusion models (DDMs), a class of models that stems from cognitive psychology and characterizes choices as a sequential sampling process that

accumulates noisy evidence until a criterion is reached ([23], see **Figure 1C**). This model allowed researchers to identify for each individual the degree to which distribution choices reflect genuine concerns for fairness or inconsistent application of general decision criteria [24]. That both social motives and general decision processes/criteria determine social decisions is also evident from demonstrations that DDMs fitted to non-social choices (i.e., food choices) allows researchers to make accurate out-of-sample predictions of social resource distribution choices [25] .

Learning

Many social motives or beliefs are not static predispositions but can change with experience [26]. This has motivated development of models that dynamically update utility/values and beliefs, thus allowing prediction of an individual's future actions based on her previous experience. The most popular class of these models characterizes these changes as a dynamic reinforcement-learning (RL) process in which values or beliefs are updated by so-called prediction errors (see **Figure 1D**). These dynamically-changing representations form the basis for choice in similar ways as for static models (e.g., via DDMs, see **Figure 1C**). The corresponding models provide a good fit for different types of learning [27] and have been neuro-biologically validated: Dopamine neurons and their neural projections in the ventral striatum (VS) encode prediction errors, with increased firing rates for unexpected rewards and suppressed activity when expected rewards are omitted [28]. This modelling framework makes it possible to capture trial-by-trial learning about purely social contexts, such as others' experiences or actions. Compared to simple comparison of two conditions (e.g. reward/no-reward or social/non-social context) [29], studies employing these models can provide more mechanistic accounts of how experience gradually changes social motivations (**Figure 2**). This brings several advantages.

For example, use of these models can illuminate whether people learn in a similar or distinct manner from personal and social experience. Recent studies suggest the former, since modelled reward prediction errors (RPEs) correlate with activity in similar brain structures when individuals either experience rewards themselves or merely observe others receiving these rewards [30,31]. However, prediction errors about other's actions (rather than reward experiences) are reflected in activity of

distinct brain areas [32]. This suggests both overlapping and distinct mechanisms involved in learning from personal experience or social observation.

Computational modelling also helps researchers to study the origins of conformity [33], which is traditionally thought to reflect the motivation to align one's behavior with a group. Computational studies now suggest that conformity may rather reflect optimal learning from social signals that resolve informational uncertainty [34,35]. Corresponding models can predict quantitatively how choices are shaped by various factors (e.g., group size, strength of preferences, or private information) [35]. Moreover, the model-predicted informational conformity involves neural structures overlapping with those underlying learning from own experiences [36,37]. However, other studies show that individual risk preferences and the correlated neural activity can change during observation of other people's risky choices [38], suggesting that some types of conformity can also reflect changes in motivational tendencies rather than the aim to resolve informational uncertainty.

Like their static counterparts described before, learning models can identify the parallel use and possible dominance of different learning processes. For example, strategic interactions as investigated by repeated games (e.g., "rock-paper-scissors") are in principle optimally solved by randomizing over all three choices [39]. However, empirical choices often exhibit sequential contingencies that can be exploited by an agent learning the opponent's choice tendencies. This learning can focus on the opponent's choice history (first-order beliefs, "she tends to play rock, so I'll play paper") but also on one's own choices (second-order beliefs; "I just played paper, so she might think I will play it again, so I will play rock now"). These two types of beliefs are updated in parallel in the influence learning model [40] (**Figure 1E**). Confirming this model, studies show that both types of prediction errors are encoded simultaneously by different neural structures (first-order beliefs in the dorsal anterior cingulate cortex (dACC) [41] and second-order beliefs in the temporoparietal junction (TPJ)). Importantly, brain-stimulation-disruption of TPJ activity selectively reduces individuals' ability to employ second-order beliefs during competitive interactions, providing evidence that this model-predicted computation is indeed necessary to control strategic behaviour [42] (**Figure 2**).

Finally, modelling studies are shedding light on the question of how social and personal motivations may be integrated in the control of behavior. In studies of

consensus decision making, model-predicted learning of personal preferences or other's opinions takes place in distinct neural structures (ventro-medial prefrontal cortex (vmPFC) versus TPJ), but both types of signals are integrated in the dorsal ACC [43]. Another study showed that while model-predicted estimates of participant's own performance in a social game were always encoded in the ACC, their estimates of their own and others' abilities were merged in the dmPFC, taking into account whether the social context demanded competition or cooperation [44]. Thus, computational modelling allowed researchers to show directly that identical social information can be integrated with personal information in different ways depending on social context.

Implications and open questions

We have illustrated how neurocomputational models can benefit the study of social behavior, by discussing example studies of resource distributions and social learning. Space constraints precluded us from discussing other interesting studies using this approach to investigate, e.g., moral decision making [45–47], honesty [48], or reciprocity [49,50]. Research in these domains also clearly benefits from the use of neurocomputational models; however, note that the specific models reviewed here may not transfer seamlessly to all types of social behavior so that distinct computational approaches may have to be developed. While the preceding sections have emphasized some advantages of using neurocomputational models to study social behavior, we close by highlighting potential problems associated with this approach.

The use of different models to capture either values, beliefs, or choice processes carries the danger that the choice of model (rather than the behavior in itself) determines the focus of investigation. Studies of repeated social interactions typically characterize behaviors as learning/prediction problems, whereas studies of resource distribution usually focus on motivations. Since the models (and tasks) used to capture belief learning or preferences differ, at least some of the differences in processes thought to drive these behaviors may reflect use of different computational frameworks. Further progress of the field may require more unified approaches to modelling different types of behaviors.

This problem can be illustrated by studies of trust, which has mainly been characterized as a learning problem. However, since trusting others makes oneself vulnerable to others' exploitation, people have to resolve conflicts between potential

profit and at least three other concerns: loss aversion, inequity aversion, and betrayal aversion [51]. Few studies have examined the neurocomputational mechanisms underlying these kinds of aversions in (dis)trust decisions, which could be studied with mixture models assigning weights to these different concerns [52].

In closing, we stress that while neurocomputational models can clearly advance understanding and prediction of social behavior, they only provide a restricted view of the underlying motives, cognitions, and processes. Moreover, some models may fit behavior and brain activity largely because of their general flexibility to fit many different patterns of data. Empirical studies should therefore strive to provide evidence that latent parameters of models actually reflect processes that can be selectively changed by experimental interventions [42]. Ultimately, the best models may be those that can structure research about what drives social behavior, pretty much like classic theories, but now with a more quantitative and mechanistic focus. ”Essentially, all models are wrong, but some are useful” [53].

References

[1] C.D. Batson, B.D. Duncan, P. Ackerman, T. Buckley, K. Birch, Is empathic emotion a source of altruistic motivation?, J. Pers. Soc. Psychol. 40 (1981) 290–302. doi:10.1037/0022-3514.40.2.290.

[2] L. Festinger, A theory of social comparison processes, Hum. Relat. 7 (1954) 117–140.

[3] C.F. Camerer, G. Loewenstein, Information, fairness, and efficiency in bargaining, In Psychological Perspectives on Justice Theory and Applications. Edited by B.A. Mellers, J. Baron. Cambridge University Press (1993) 155-180.

[4] M.J.J. Handgraaf, E. Van Dijk, D. De Cremer, Social utility in ultimatum bargaining, Soc. Justice Res. 16 (2003) 263–283. doi:10.1023/A:1025940829543.

[5] D. Engelmann, M. Strobel, Inequality aversion, efficiency, and maximin preferences in simple distribution experiments, Am. Econ. Rev. 94 (2004) 857–869.

[6] E. Fehr, M. Naef, K.M. Schmidt, Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment, Am. Econ. Rev. 96 (2006) 1912–1917.

[7] E. Xiao, D. Houser, Emotion expression in human punishment behavior., Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 7398–7401. doi:10.1073/pnas.0502399102.

[8] A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom, J.D. Cohen, The neural basis of economic decision-making in the ultimatum game, Science. 300 (2003) 1755–1758. doi:10.1126/science.1082976.

[9] M.J. Crockett, A. Apergis-Schoute, B. Herrmann, M. Lieberman, U. Muller, T.W. Robbins, L. Clark, Serotonin modulates striatal responses to fairness and retaliation in humans, J. Neurosci. 33 (2013) 3505–3513. doi:10.1523/JNEUROSCI.2761-12.2013.

[10] E. Fehr, I. Krajbich, Social preferences and the brain, In Neuroeconomics: Decision making and the brain (Second Ed.), Edited by P.W. Glimcher, E. Fehr. Elsevier, (2014) 193–218.

[11] J. Annis, T.J. Palmeri, Bayesian statistical approaches to evaluating cognitive models,

Wiley Interdiscip. Rev. Cogn. Sci. 9 (2018) e1458. doi: 10.1002/wcs.1458.
[12] E. Fehr, K.M. Schmidt, A theory of fairness, competition, and cooperation, Q. J. Econ. 114 (1999) 817–868.
[13] Y. Morishima, D. Schunk, A. Bruhin, C.C. Ruff, E. Fehr, Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism, Neuron. 75 (2012) 73–79. doi:10.1016/j.neuron.2012.05.021
[14] M. Hsu, C. Anen, S.R. Quartz, The right and the good: Distributive justice and neural encoding of equity and efficiency, Science. 320 (2008) 1092–1095. doi:10.1126/science.1153651.
[15] I. Saez, L. Zhu, E. Set, A. Kayser, M. Hsu, Dopamine modulates egalitarian behavior in humans, Curr. Biol. 25 (2015) 912–919. doi:10.1016/j.cub.2015.01.071.
This study shows with pharmacological manipulation that dopamine selectively modulates preferences for fairness, rather than influencing concerns for other's interests or general altruistic preferences.

[16] N.D. Wright, M. Symmonds, S.M. Fleming, R.J. Dolan, Neural segregation of objective and contextual aspects of fairness., J. Neurosci. 31 (2011) 5244–5252. doi:10.1523/JNEUROSCI.3138-10.2011.
[17] K.M. Vekaria, K.M. Brethel-Haurwitz, E.M. Cardinale, S.A. Stoycos, A.A. Marsh, Social discounting and distance perceptions in costly altruism, Nat. Hum. Behav. 1 (2017) 0100. doi:10.1038/s41562-017-0100.
[18] B. Jones, H. Rachlin, Social discounting, Psychol. Sci. 17 (2006) 283–286.
[19] T. Strombach, B. Weber, Z. Hangebrauk, P. Kenning, I.I. Karipidis, P.N. Tobler, T. Kalenscher, Social discounting involves modulation of neural value signals by temporoparietal junction, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 1619–1624. doi:10.1073/pnas.1414715112.
[20] A. Soutschek, C.C. Ruff, T. Strombach, T. Kalenscher, P.N. Tobler, Brain stimulation reveals crucial role of overcoming self-centeredness in self-control, Sci. Adv. 2 (2016) e1600992. doi:10.1126/sciadv.1600992.
[21] T. Xiang, T. Lohrenz, P.R. Montague, Computational substrates of norms and their violations during social exchange., J. Neurosci. 33 (2013) 1099–1108. doi:10.1523/JNEUROSCI.1642-12.2013.
[22] L.J. Chang, A.G. Sanfey, Great expectations: Neural computations underlying the use of social norms in decision-making, Soc. Cogn. Affect. Neurosci. 8 (2013) 277–284. doi:10.1093/scan/nsr094.
[23] R. Ratcliff, G. McKoon, The diffusion decision model: Theory and data for two-choice decision tasks, Neural Comput. 20 (2008) 873–922.
[24] C.A. Hutcherson, B. Bushong, A. Rangel, A neurocomputational model of altruistic choice and its implications, Neuron. 87 (2015) 451–462. doi:10.1016/j.neuron.2015.06.031.
Using a multi-attribute drift-diffusion model (DDM), this study identifies distinct contributions of preferences versus general decision noise to altruistic choices and provides neural evidence that right TPJ represents of others' interests.

[25] I. Krajbich, T. Hare, B. Bartling, Y. Morishima, E. Fehr, A common mechanism underlying food choice and social decisions, PLoS Comput. Biol. 11 (2015) e1004371. doi:10.1371/journal.pcbi.1004371.
Using a sequential sampling model (SSM), the authors show that the speed and identity of social choices (e.g., about resource distribution) can be predicted by model parameters derived from data about non-social decisions (i.e., food

choice), suggesting that decision making across these two domains may share common processes.

[26] T.E.J. Behrens, L.T. Hunt, M.F.S. Rushworth, The Computation of Social Behavior, Science. 324 (2009) 1160–1164. doi:10.1126/science.1169694.
[27] N.D. Daw, K. Doya, The computational neurobiology of learning and reward, Curr. Opin. Neurobiol. 16 (2006) 199–204. doi:10.1016/j.conb.2006.03.006.
[28] W. Schultz, A Neural Substrate of Prediction and Reward, Science. 275 (1997) 1593–1599. doi:10.1126/science.275.5306.1593.
[29] M. Stallen, A.G. Sanfey, Cooperation in the brain: neuroscientific contributions to theory and policy, Curr. Opin. Behav. Sci. 3 (2015) 117–121. doi:10.1016/j.cobeha.2015.03.003.
[30] S. Collette, W.M. Pauli, P. Bossaerts, J. O'Doherty, Neural computations underlying inverse reinforcement learning in the human brain, ELife. 6 (2017).
[31] M.R. Hill, E.D. Boorman, I. Fried, Observational learning computations in neurons of the human anterior cingulate cortex, Nat. Commun. 7 (2016) 12722. doi:10.1038/ncomms12722.
[32] S. Suzuki, N. Harasawa, K. Ueno, J.L. Gardner, N. Ichinohe, M. Haruno, K. Cheng, H. Nakahara, Learning to Simulate Others' Decisions, Neuron. 74 (2012) 1125–1137. doi:10.1016/j.neuron.2012.04.030.
[33] S.E. Asch, Studies of independence and conformity: I. A minority of one against a unanimous majority., Psychol. Monogr. Gen. Appl. 70 (1956) 1.
[34] U. Toelch, D.R. Bach, R.J. Dolan, The neural underpinnings of an optimal exploitation of social information under uncertainty, Soc. Cogn. Affect. Neurosci. 9 (2014) 1746–1753. doi:10.1093/scan/nst173.
[35] R.E. Huber, V. Klucharev, J. Rieskamp, Neural correlates of informational cascades: brain mechanisms of social influence on belief updating, Soc. Cogn. Affect. Neurosci. 10 (2015) 589–597. doi:10.1093/scan/nsu090.
This paper shows that a neurocomputational model of Bayesian learning can identify neural underpinnings of social belief updating and biases towards private information.

[36] C.J. Charpentier, C. Moutsiana, N. Garrett, T. Sharot, The Brain's Temporal Dynamics from a Collective Decision to Individual Action, J. Neurosci. 34 (2014) 5816–5823. doi:10.1523/JNEUROSCI.4107-13.2014.
[37] V. Klucharev, K. Hytönen, M. Rijpkema, A. Smidts, G. Fernández, Reinforcement Learning Signal Predicts Social Conformity, Neuron. 61 (2009) 140–151. doi:10.1016/j.neuron.2008.11.027.
[38] D. Chung, G.I. Christopoulos, B. King-Casas, S.B. Ball, P.H. Chiu, Social signals of safety and risk confer utility and have asymmetric effects on observers' choices, Nat. Neurosci. 18 (2015) 912–916. doi:10.1038/nn.4022.
[39] T. Chmura, S.J. Goerg, R. Selten, Learning in experimental games, Games Econ. Behav. 76 (2012) 44–73. doi:10.1016/j.geb.2012.06.007.
[40] A.N. Hampton, P. Bossaerts, J.P. O'Doherty, Neural correlates of mentalizing-related computations during strategic interactions in humans, Proc. Natl. Acad. Sci. 105 (2008) 6741–6746.
[41] L. Zhu, K.E. Mathewson, M. Hsu, Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning, Proc. Natl. Acad. Sci. 109 (2012) 1419–1424. doi:10.1073/pnas.1116783109.

[42] C.A. Hill, S. Suzuki, R. Polania, M. Moisa, J.P. O'Doherty, C.C. Ruff, A causal account of the brain network computations underlying strategic social behavior, Nat. Neurosci. 20 (2017) 1142–1149. doi:10.1038/nn.4602.
This paper utilizes transcranial magnetic stimulation (TMS) and a learning model to demonstrate a causal link between mentalizing computations in the TPJ and formation of second-order beliefs about the opponent during competitive interactions.

[43] S. Suzuki, R. Adachi, S. Dunne, P. Bossaerts, J.P. O'Doherty, Neural Mechanisms Underlying Human Consensus Decision-Making, Neuron. 86 (2015) 591–602. doi:10.1016/j.neuron.2015.03.019.
This paper uses a computational model to show how individuals integrate their own preferences with information about the social consensus.

[44] M.K. Wittmann, N. Kolling, N.S. Faber, J. Scholl, N. Nelissen, M.F.S. Rushworth, Self-Other Mergence in the Frontal Cortex during Cooperation and Competition, Neuron. 91 (2016) 482–493. doi:10.1016/j.neuron.2016.06.022.
This study investigates how the brain tracks and aggregates performances of the individual and others, suggesting that these representations are interdependent.

[45] M.J. Crockett, Z. Kurth-Nelson, J.Z. Siegel, P. Dayan, R.J. Dolan, Harm to others outweighs harm to self in moral decision making, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 17320–17325. doi:10.1073/pnas.1408988111.
[46] M.J. Crockett, J.Z. Siegel, Z. Kurth-Nelson, O.T. Ousdal, G. Story, C. Frieband, J.M. Grosse-Rueskamp, P. Dayan, R.J. Dolan, Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making, Curr. Biol. 25 (2015) 1852–1859. doi:10.1016/j.cub.2015.05.021.
[47] M.J. Crockett, J.Z. Siegel, Z. Kurth-Nelson, P. Dayan, R.J. Dolan, Moral transgressions corrupt neural representations of value, Nat. Neurosci. 20 (2017) 879–885. doi:10.1038/nn.4557.
[48] L. Zhu, A.C. Jenkins, E. Set, D. Scabini, R.T. Knight, P.H. Chiu, B. King-Casas, M. Hsu, Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest, 17 (2014) 1319–1321.
[49] D.S. Fareri, L.J. Chang, M.R. Delgado, Computational substrates of social value in interpersonal collaboration, J. Neurosci. 35 (2015) 8170–8180. doi:10.1523/JNEUROSCI.4775-14.2015.
[50] D.S. Fareri, L.J. Chang, M.R. Delgado, Effects of direct social experience on trust decisions and neural reward circuitry, Front. Neurosci. 6 (2012) 148. doi:10.3389/fnins.2012.00148.
[51] I. Bohnet, R. Zeckhauser, Trust, risk and betrayal, J. Econ. Behav. Organ. 55 (2004) 467–484. doi:10.1016/j.jebo.2003.11.004.
[52] G. Nave, C. Camerer, M. McCullough, Does oxytocin increase trust in humans? A critical review of research, Perspect. Psychol. Sci. 10 (2015) 772–789. doi:10.1177/1745691615600138.
[53] G.E.P. Box, N.R. Draper, Empirical model-building and response surfaces, John Wiley & Sons, 1987.
[54] R.S. Sutton, A.G. Barto, Reinforcement learning: an introduction, MIT Press, Cambridge, Mass, 1998.

Figure 1. Computational models of behavior

A. Inequity aversion model

$$U(x_i, x_j) = x_i - \frac{\alpha}{n-1}\sum_{j=1..n}\max(x_j - x_i, 0) - \frac{\beta}{n-1}\sum_{j=1..n}\max(x_i - x_j, 0)$$

This model determines the utility $U(x_i, x_j)$ of a resource distribution characterized by the agent's payoff x_i and the payoff x_j to n other people. Inequity aversion is captured by subtracting the weighted payoff difference $(|x_i - x_j|)$ from the personal payoff x_i [12]. The strength of this concern is captured by the weighting parameters α and β, as expressed in disadvantageous $(x_j > x_i)$ and advantageous $(x_j < x_i)$ conditions, respectively.

B. Bayesian observer model

This model allows researchers to examine how people form beliefs about other people's distribution offers. The prior belief about the distribution of offers is a Gaussian with mean μ and variance σ^2:

$$p(u) = p(u|\mu, \sigma^2)p(\mu, \sigma^2)$$

When the participant observes an offer x_t at trial t, she will update this belief:

$$p(u_t|x_t) = \frac{p(x_t|u_t)p(u_{t-1})}{p(x_t)}$$

This updated belief becomes the new prior. With this model, researchers can, for example, derive the expected offer (norm) at trial t: $E[u_t] = u_t$, or the deviation of a given offer from expectations: $\delta_t = x_t - E[u_{t-1}] = x_t - u_{t-1}$ [21].

C. Drift diffusion model

$$dy(t) = v(\Delta u) \cdot dt + \sigma \cdot dW$$

Drift diffusion models formalize the decision process itself as a sequential sampling process that accumulates noisy evidence for one choice option over the other (e.g., difference in utilities between two options) until a criterion is reached [23]. $y(t)$ is the amount of accumulated evidence at time t, Δu is the difference in utilities between two options, v is the parameter quantifying the efficiency of evidence accumulation (drift rate), and σ is the Gaussian noise parameter of the Wiener process dW. Additionally, DDMs also include parameters capturing bias in the choice process, decision threshold, and non-decision time.

D. Reinforcement learning (RL) model

The simplest and most popular RL model is the Rescorla-Wagner rule [54]. This model assumes that, at time (or trial) t, the brain computes values for available actions (Q_t) and updates these values as follows:

$$Q_{t+1} = Q_t + \alpha \cdot \delta_t,$$

where α is the learning rate, and δ_t is the prediction error, the difference between the actual reward received at time t (r_t) and the expected reward, which is the stored action value:

$$\delta_t = r_t - Q_t.$$

E. Influence learning model

This model updates the individual's future belief p_{t+1} about what the opponent will choose using the past belief p_t and two types of weighted prediction errors:

$$p^*_{t+1} = p^*_t + \eta(P_t - p^*_t) + \kappa(Q_t - q^{**}_t),$$

$\eta(P_t - p^*_t)$ represents *fictitious play*, i.e., learning of the opponent's choice history ("first-order

beliefs"), with η being a learning parameter scaling the prediction error. $\kappa(Q_t - q_t^{**})$ represents learning how the player's own actions *influence* the opponent's choices ("second-order beliefs"; again, κ reflects a separate learning rate)[40]. This model allows separation and quantification of two competing motivations that can influence strategic choice.

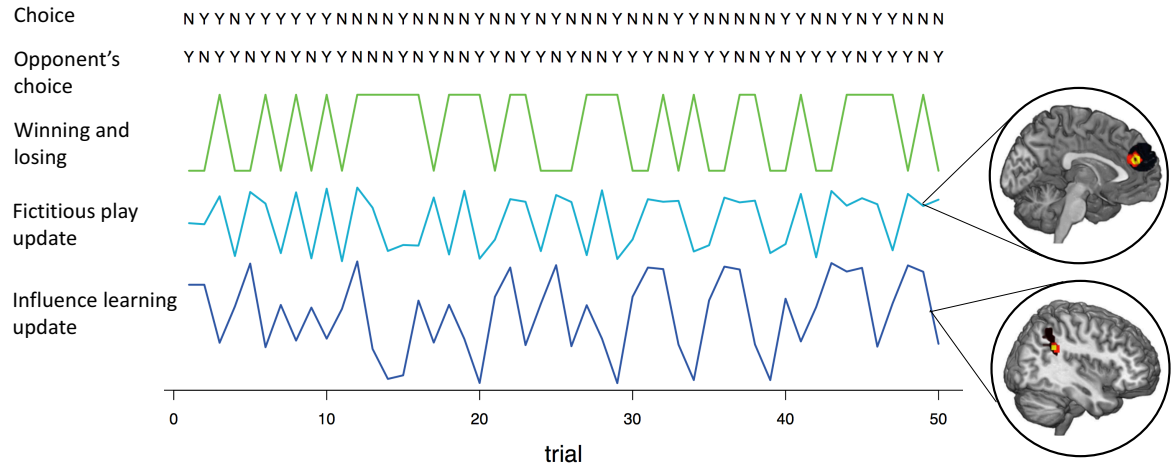

Figure 2. An illustration of how neurocomputational modelling allows identification and separation of different cognitive processes underlying learning in a strategic game. The subject plays the role of an employee choosing between "working" (Y) and "shirking" (N), while the opponent outside of the scanner is choosing to "inspect" (Y) or "not inspect" (N). Subject and opponent have opposing aims, since the subject wins when both players' choices match (Y/Y or N/N) but the opponent wins when they do not match (Y/N or N/Y). For the choices observed on each trial (given in the top two lines of the figure), behavior and brain activity can be regressed on simply the observed outcomes (winning or losing, green line depicts the corresponding regressor). This would allow only limited inference on cognitive processes contributing to the choice. However, fitting the computational learning model described in Figure 1E allows prediction of how participants should use the outcomes to update their predictions of what the opponents will play based on two types of beliefs: First-order beliefs about the opponent's choice tendencies ("fictitious play") or second-order beliefs about how their choices influence the opponent ("influence learning"). Evidence for these types of learning is given by the fact that both predicted updates correlate with activity in two different regions in the brain (dorsal ACC and TPJ) [42], and that brain-stimulation-disruption of TPJ activity reduces the participants' ability to engage in influence learning [42] .