Predicting the other in cooperative interactions

Alan G. Sanfey[1,2], Claudia Civai[1], Peter Vavra[1,2]

[1] Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Kapittelweg 29, 6525 EN Nijmegen, the Netherlands

[2] Behavioral Science Institute, Radboud University Nijmegen, Postbus 9104, 6500 HE Nijmegen, the Netherlands

Corresponding author: Sanfey, A.G. (alan.sanfey@donders.ru.nl)

Abstract

Recent research has shown that a collection of neurons in dorsal anterior cingulate cortex of Rhesus monkeys may specifically encode the choice selection of an interaction partner. This raises interesting and important questions as to the nature of theory of mind processes in social interactive decision-making, with potential societal implications.

One notable aspect of human decision-making is the ubiquity of our cooperative interactions [1], both with specific others or with societal institutions more broadly. We generally return the favors of others - we help a friend move in the expectation of future help in return. We also cooperate on a larger scale, for example we pay our taxes when we could potentially avoid doing so. Many of these social choices are risky, that is, we are often unsure if our positive acts will indeed be reciprocated in the future, and a key component of our decisions to cooperate is to what degree we can predict that our partner in the exchange will be willing to commit to cooperation. Therefore one extremely important aspect of understanding the motivations and mechanisms underlying these important choices is how we represent the likely decisions of others.

In a recent compelling paper, Haroush & Williams [2] outline the case for a grouping of neurons in the primate brain that appear to specifically encode the choice selection of an interaction partner. These neurons, in the dorsal anterior cingulate cortex of Rhesus monkeys, were observed using single-unit recording while the monkeys played a variant of the oft-studied iterative Prisoner's Dilemma game, where players must decide to either cooperate with a partner for a potential joint positive gain, or defect to guarantee themselves a payoff at the expense of their partner. Using these signals, the monkey's own choice could be correctly predicted on over 65% of rounds. However, using the same signals they were also able to predict the *other,* physically present, monkey's unobserved choices with even higher accuracy, namely 79%. In other words, these dACC

neurons encoded information enabling the monkey to, at least in principle, predict the other's future behavior with high accuracy.

The concept of Theory of Mind (ToM) refers to the ability to understand and predict the behavior of others, and by isolating cells that appear to represent the as yet unknown intentions of a game partner, these data support the idea that ToM is a fundamental and specific process, and raise intriguing questions on three distinct levels. On the computational level: under which circumstances these neurons get engaged, and how this impacts strategic decisions. On the neural level: how do these “other”-encoding dACC neurons fit into a larger ToM network that implements these computations. Finally, on a socio-behavioral level: to what extent the physical presence of others modulates the perception of social context.

Firstly, these results shed light on the circumstances under which ToM is engaged. In a control experiment, when the first monkey defected and this choice was explicitly shown to the second monkey, the latter defected in turn on over 90% of the rounds; that is, the second monkey successfully avoided exploitation. Notably however, when the monkeys made their choice without directly observing the decision of the other, they cooperated substantially more often. Given that the neural predictions were very accurate, and so presumably should not lead to different decisions than observation, what underlies this difference in cooperation rates? Does revealing one's intentions explicitly change how the ToM network of others is engaged, thus altering the tendency

to cooperate by impacting the certainty of beliefs about the other's behavior? It may require a revision of current models of strategic behavior to account for these different levels of cooperation. Recent computational models based on human experiments with similar two-player games suggest that people adapt their decisions based on how they expect others will behave [3,4] as well as how they believe others expect them to behave [5]. One possible extension based on these results could therefore be to explicitly model how certain we are about such beliefs and how this (un)certainty affects our choices.

In terms of the broader neural basis of ToM, in humans this network encompasses posterior areas such as temporo-parietal junction (TPJ) and superior temporal sulcus (STS), as well as anterior areas such as medial prefrontal cortex (MPFC) and dACC, with each connection appearing to subserve a specific function [6]. The present paper offers useful and novel insights as to the neuronal specificity of the dACC: disrupting this area via electrical stimulation seemed to specifically affect how the last interaction was taken into account. Without disruption, monkeys cooperated more on rounds following mutual cooperation; with disruption, monkeys cooperated substantially less often after rounds where the partner had cooperated, appearing not to take that information into account. These findings suggest a specific contribution of the dACC within the ToM network: processing information about past events. Future studies could assess whether this area is engaged in the retrieval of the other's behavior, the assessment of the other's actions, or perhaps the integration of all of this information. Further, it would be

useful, if technically feasible, to consider simultaneously recording from other regions of this network in order to shed light on the neuronal functionality of the network as a whole.

Finally, the findings also relate to the physical presence of another individual, suggesting the existence of a specific social context sensitivity, even at the neural level: Playing with monkeys who were in another room yielded a reduced cooperation rate, to one indistinguishable from when playing with a computer partner. Importantly, the number of neurons encoding the other's unobserved choices also decreased in this context, suggesting a role for physical presence in social preference. Results from a variety of studies demonstrate that humans perceive agency and intentions even in the absence of the physical presence of the other. For example, just knowing that the partner in a Prisoner's Dilemma is human increases cooperation, as well as the ToM's network associated activation [7]. However, it has been shown that an increase in physical distance does lead to a decrease in cooperation [8,9], and thus it could indeed be the case that as one's game partner becomes physically more distant a reduction in the neural encoding of social context, and, consequently, a reduction in prosocial behavior is observed. Understanding the extent of this neuronal specificity for others' physical presence in areas associated with complex decision-making, such as dACC and MPFC, would be a useful direction for future research. Given the increasing lack of face-to-face interaction in modern society, it may be fruitful to consider ways of boosting direct

contact in situations where social bonds are desirable, such as social media or internet banking, demonstrating the use of these approaches in assisting with policy designs.

References

1 Clutton-Brock, T. (2009) Cooperation between non-kin in animal societies. *Nature* 462, 51–7

2 Haroush, K and Williams, Z.M. (2015) Neuronal prediction of opponent's behavior during cooperative social interchange in primates. *Cell* 160, 1-13

3 Chang, L. J., and Sanfey, A. G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience* 8, 277–84

4 Xiang, T., et al. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience* 33, 1099–108

5 Chang, L. J., et al. (2011). Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 70, 560–72

6 Abu-Akel, A., and Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* 49, 2971-2984

7 Rilling, J. K., et al. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage* 22, 1694-1703

8 Sensenig, J., et al. (1972). Cooperation in the prisoner's dilemma as a function of interpersonal distance. *Psychon Sci* 2, 105-106

9 Bradner, E., and Mark, G. (2002, November). Why distance matters: effects on cooperation, persuasion and deception. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work* (pp. 226-235). ACM