Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

Is Recursive Belief Inference the Engine of Mentalizing?

Magdalena Krol¹ and Michal Krol²

¹School of Psychological Sciences, and ²Department of Economics, School of Social Sciences, University of Manchester, Manchester M13 9PL, United Kingdom

Review of Yoshida et al.

Game theorists often quote the story of Sherlock Holmes fleeing London by train in direction of Dover, and applying the following recursive reasoning: if he were to continue the journey without giving thought to his pursuers (order-one strategy), then surely Moriarty would find out and follow him to Dover. Knowing his arch-enemy well, Holmes might account for this in a second-order inference and alight at Canterbury to avoid pursuit. However, Moriarty might himself operate on the second-order of reasoning and expect Holmes to do just that, in which case the latter should progress to the third level of recursion and go to Dover. Based on his past experience with Moriarty, Holmes was able to estimate his order of sophistication, k , and optimally respond to it using a $k + 1$ order strategy (otherwise, he could continue the mind game ad infinitum without reaching a decision).

In search of the neural underpinnings of such belief-inference processes, Yoshida et al. (2010) chose an experimental task suitably underlining its potential evolutionary role, namely the conventional stag-hunt game, wherein two players independently decide to hunt a stag or a hare. If both choose the former (to cooperate), they succeed and both receive high rewards, otherwise any hare-hunter gets a low payoff and anyone in lone pursuit of the stag ends up

with nothing. Although standard game theory struggles to predict the game's outcome, the authors apply a computational model demonstrating how the theory of mind (ability to infer internal states of others) facilitates coordinating on the cooperative stag-hunting solution.

To this end, subjects were asked to play a dynamic variant of the game in which they and a computer-controlled counterpart took turns to navigate a two-dimensional grid in pursuit of a chosen prey.

This task demanded recursive inference of the counterpart's intentions, based on the history of its actions. The order of computer's strategy was varied throughout the task. An agent of type $k = 1$ assumes that the counterpart will move randomly, and hence pursues a hare—unless both players are not far (say, at most two moves) from the stag, in which case there is a chance that the counterpart will accidentally reach the prey and assist in the hunt (recall both need to go for the stag to succeed). Consequently, a $k = 2$ agent (who assumes he is dealing with a $k = 1$ type) knows that it is sufficient for him to be three moves from the stag, with the counterpart one step closer, to induce the latter to cooperate by moving toward the prey. It follows that each subsequent order of sophistication requires less proximity to the target to initiate a pursuit of the stag.

The authors evaluated two models of participants' behavior: one in which subjects followed recursive strategies at a fixed level k (as detailed above), and one in which they operated one level above their current estimate of the counterpart's k , based on the history of their counterpart's

moves in the current game. The latter theory of mind (ToM) model was designed to capture the participants' tendency to form representations of the other party's decision-making processes, as Holmes did with Moriarty. Indeed, ToM produced a better statistical fit to the behavioral data than the fixed k model, leading the authors to conclude that participants used recursive belief inference. However, the fixed k model required players to act based only on the current state of the game-grid, ignoring the history of players' moves. Such a strategy is unlikely to be used by humans, who usually consider past events when engaged in dynamic interactions, and hence it is not a demanding benchmark for model comparison. On the other hand, the ToM itself is difficult to generalize to human-to-human interactions, as it entails making judgments about the counterpart's k assuming it is fixed, rather than resulting from a similar ToM process on behalf of the other player. In other words, a ToM agent assumes other players are of the fixed k type, i.e., insufficiently refined to infer the agent's mental state the way the agent infers theirs.

Overall, the fact that the ToM model (for optimally adjusted parameters) is consistent with the observed behavior, does not mean that subjects actually use ToM, nor even any recursive inference in general. It would therefore be worth checking whether any less complicated strategy could fit the data equally well. For instance, a trivial tit-for-tat rule of moving toward the stag unless the computer did otherwise in the previous round could

Received Sept. 19, 2010; revised Oct. 12, 2010; accepted Oct. 13, 2010.

Correspondence should be addressed to Magdalena Krol, School of Psychological Sciences, Zochonis Building, University of Manchester, Oxford Road, Manchester M13 9PL, UK. E-mail: Magdalena.Krol@postgrad.manchester.ac.uk.

DOI:10.1523/JNEUROSCI.4917-10.2010

Copyright © 2010 the authors 0270-6474/10/3015711-02\$15.00/0

explain the fact that subjects cooperated more often when the computer's k increased. In other words, this tendency does not mean that subjects are necessarily affected by the counterpart's sophistication level, as Yoshida et al. (2010) conclude. Instead, the participants could merely act in response to the observable consequences of k , such as the act of following a particular type of prey, without recursive processes involved. This is important because the authors use their best-fit ToM model to compute the values of latent variables, such as the participants' k or uncertainty of their belief inference, and map those values onto the brain imaging data. Despite offering intriguing interpretational possibilities, this could be misleading if the postulated latent variables were in fact nonexistent.

Yoshida et al.'s (2010) neuroimaging results show increased activity in bilateral ventral striatum, part of the reward system, at the time of receiving payoffs for catching a prey. Interestingly, in caudal ventral striatum, this activity correlated with the latent sophistication level k (related to cooperation) obtained from the best-fit ToM model at the time of payoff. The authors argue that this particular cluster of activity reflected the rewarding feeling of a "warm glow" (Bhatt and Camerer, 2005) resulting from cooperation. This could suggest that cooperation is evolutionarily favored and reinforced by the reward system in the brain to ensure it is selected and sustained. Although consistent with other reports (Rilling et al., 2002), this result can be alternatively explained by the players' anticipation of the forthcoming reward, natural when they think the counterpart is sophisticated and cooperation-oriented, which is precisely when their own k is high. Cooperation not leading to a high-value objective might not be gratifying.

Moreover, participants' estimated k at the time of the computer's move, which the authors interpret as the level of strate-

gic thinking, correlated with activity in several regions, most notably dorsolateral prefrontal cortex (Brodmann area 9). The latter is not related exclusively to mentalizing, but is implicated in logical and probabilistic reasoning tasks (Goel and Dolan, 2004). Thus, it is plausible that the sophistication level reflected a more generic process of reasoning than a function specific to the theory of mind.

Finally, the authors found that activity in the anterior part of rostral medial prefrontal cortex (MPFC) (Brodmann area 10) correlated with the participants' estimated degree of uncertainty about the computer's k . This is crucial, because the involvement of MPFC in mentalizing (theory of mind) tasks is well established (Amodio and Frith, 2006). This could suggest that participants approached the computer agent as another person, using mentalizing to work out its intentions. Interestingly, Coricelli and Nagel (2009) found a correlation between MPFC and strategic thinking only when their participants were playing against other people, not when playing against computers. However, in their task, the computer behaved randomly and did not display any strategic behavior. Coricelli and Nagel (2009) also found an increase in MPFC activity for those participants who reasoned at higher levels of strategic sophistication, obtaining better results as a consequence. Thus, results of both Yoshida et al. (2010) and Coricelli and Nagel (2009) indicate that activity in MPFC reflects the level of engagement in the process of mentalizing. For this reason, it would be interesting to know whether Yoshida et al. (2010) found a similar correlation between the MPFC activity and players' performance in the game, in a similar vein to Bhatt and Camerer (2005) or Coricelli and Nagel (2009).

An alternative explanation is that MPFC is not necessarily specific to inferring other people's beliefs, as its activation in a task not involving human agents would suggest. The question is whether

participants use the same skill they would use to infer mental states of people to estimate the computer's strategy, or do they usually use a generic skill to solve mentalizing tasks? This question mirrors the debate among theory of mind researchers regarding the mechanism underlying this faculty. Simulation theory postulates that people infer others' internal states by using their own mind to simulate the minds of other people. In contrast, Theory theory assumes that mentalizing problems are solved simply by reasoning applied to our knowledge of other people (Carruthers and Smith, 1996). The results by Yoshida et al. (2010) support the latter. In particular, activation of areas involved in reasoning, as well as those strongly associated with mentalizing, in interactions with a computer suggest that mentalizing might be essentially a reasoning rather than simulation process. Yoshida et al. (2010) propose that recursive belief inference might be the core process behind theory of mind, i.e., the engine of mentalizing.

References

- Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci* 7:268–277.
- Bhatt M, Camerer CF (2005) Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games Econ Behav* 52:424–459.
- Carruthers, P., and Smith, P. K. (1996) *Theories of theories of mind*. Cambridge: Cambridge UP.
- Coricelli G, Nagel R (2009) Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *Proc Natl Acad Sci U S A* 106:9163–9168.
- Goel V, Dolan RJ (2004) Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition* 93:B109–B121.
- Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C (2002) A neural basis for social cooperation. *Neuron* 35:395–405.
- Yoshida W, Seymour B, Friston KJ, Dolan RJ (2010) Neural mechanisms of belief inference during cooperative games. *J Neurosci* 30:10744–10751.