Is the boss watching?

Amit Etkin

A combination of computational modeling, neuroimaging and a causal manipulation of brain activity in humans reveals how the brain represents beliefs about how our choices will affect those of others we interact with.

As long as there have been bosses, there have been employees wondering whether the boss will check up on them and thus whether they should work or slack off. Neuroeconomists have, in turn, codified this intricate social dance in behavioral tasks and computational models. In the task used by Hill *et al.*[1] in this issue of *Nature Neuroscience*, participants play the role either of boss or employee. In each trial, the boss decides whether to check on the employee or not, and the employee decides whether to work or to shirk their duties. Employees win money when they correctly predict the actions of the boss (slacking when the boss does not check, working when they check), whereas bosses win money when catching an employee slacking or when an employee works without them checking. Prior computational modeling using this task[2] considered several scenarios. These ranged from decisions being driven purely by recently rewarded choices through more complicated models that consider the influence a player believes their choice will have on their opponents' behavior. It turns out that a model that accounts for this influence signal fits the data best[2], thereby opening the door to understanding the neurobiological basis of this behavior and the computations underlying it. So how is the influence signal calculated in the brain, and how does it drive decisions?

By combining computational modeling of participating employees' behavior, recording neural activity with functional magnetic resonance imaging (fMRI) and manipulating brain function using transcranial magnetic stimulation (TMS), Hill *et al.*[1] demonstrate a causal role for the right temporoparietal junction (rTPJ) in computing this key influence signal. However, to understand the significance of this work, let us first examine what computational modeling and fMRI, without TMS, can and cannot tell us.

Amit Etkin is in the Department of Psychiatry and Behavioral Sciences and the Stanford Neurosciences Institute, Stanford University, Stanford, California, USA, and with the Veterans Affairs Palo Alto Healthcare System and the Sierra Pacific Mental Illness, Research, Education, and Clinical Center (MIRECC), Palo Alto, California, USA.
e-mail: amitetkin@stanford.edu

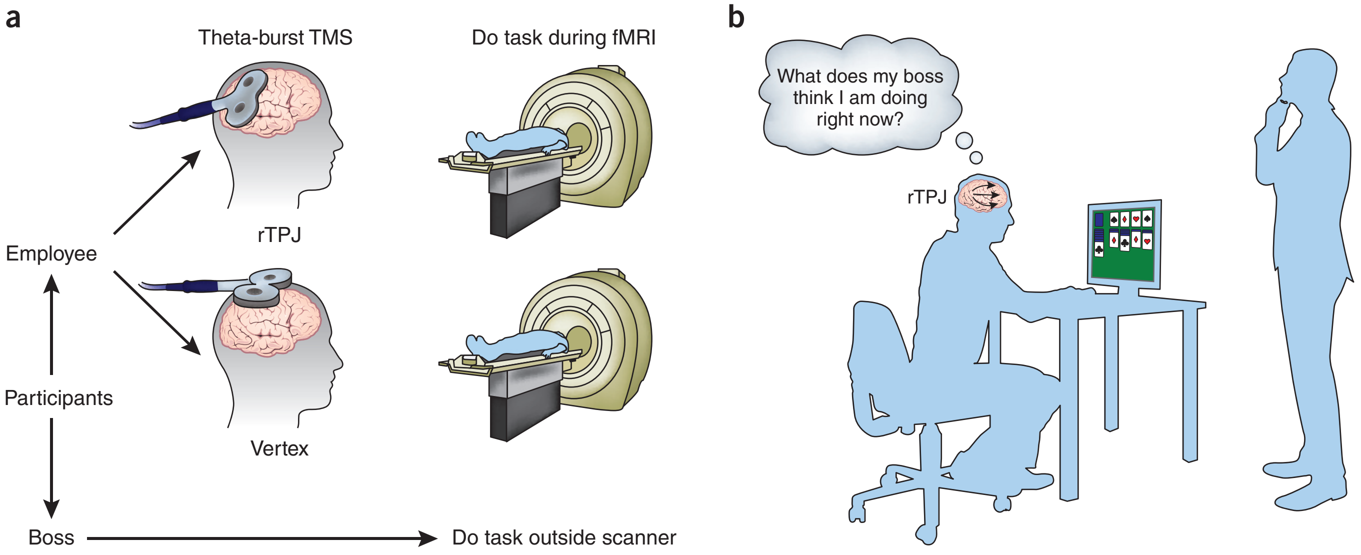
First, when people refer to computational models in decision-making, it is important to appreciate that these are typically models of presumed latent mental computations meant to explain choice behavior. Most often, such models result in values for these latent parameters for each trial. Second, computational models are in turn typically related to brain activity by seeing the degree to which trial-to-trial changes in latent model parameters explain trial-to-trial changes in fMRI signal across the brain[3,4]. While by this point we are used to seeing decision-making studies successfully employ this strategy, it is by no means guaranteed that a presumed latent mental computation, as codified in a model, necessarily has an fMRI-measureable neural correlate. Moreover, even detection of an fMRI model-correlate in some brain region does not itself indicate anything about a causal relationship between activity in that region and the behavior being modeled. Indeed, this is the great weakness of neuroimaging research in humans: namely, that simply building stronger correlations through sophisticated modeling and data analysis still does not win us causal mechanistic insights[5]. This is where TMS comes in and where the approach of Hill *et al.*[1] really stands out.

TMS is a form of focal noninvasive brain stimulation. Neurostimulation is achieved by submillisecond magnetic pulses, which cross the scalp and skull unimpeded and depolarize populations of cortical neurons underlying the coil[6]. Repetitive stimulation with TMS, in the form of particular temporal patterns, can result in plasticity that either increases or decreases regional excitability and that outlasts the stimulation itself. Hill *et al.*[1] used one such pattern, continuous theta-burst stimulation, which has been shown in motor cortex to decrease regional excitability[7]. As prior computational modeling-based fMRI analysis of their task found that rTPJ activity tracked trial-to-trial fluctuations in the influence signal[2], they compared the consequences of theta-burst modulation of the rTPJ to a similar manipulation of a control region (stimulation over the vertex of the head; **Fig. 1a**). Critically, what use of TMS brought them, in the context of computational modeling of behavior and recording of fMRI, is the ability to make strong claims about the causal contribution of the rTPJ to behavior in the task.

Hill *et al.*[1] verified that, as expected, theta-burst modulation of the rTPJ blunted its ability to track the computational model-calculated influence signal. Moreover, the magnitude of the influence parameter itself was reduced in individuals in whom the rTPJ had been inhibited, which resulted in a lower likelihood of these individuals switching choices across successive trials. The reasoning here is that if rTPJ inhibition prevented the subjects playing employees from accurately estimating how their choices would affect the employer's behavior, they would have an impoverished representation of the likely strategy an employer might use to outsmart them. As a consequence, they would be more likely to repeat the same choice they had just made. By contrast, employees with intact rTPJ influence signals would contemplate the employer's strategy and switch choices more often to keep the employer on his or her toes.

This, however, is not where the story ended. By virtue of having causally manipulated activity in the rTPJ, the authors could examine the consequence of doing so on activity in downstream regions that have also been implicated in influence calculation or in other aspects of decision-making (**Fig. 1b**). Based on prior results implicating the dorsomedial and ventromedial prefrontal cortices as important partners for the rTPJ in this task and in decision-making more broadly[2], Hill *et al.*[1] examined their connectivity. They found that rTPJ inhibition blunted its relationship with both the dorsomedial and ventromedial prefrontal cortices. Though neuroimaging alone cannot demonstrate the directionality of causal relationships, the fact that only rTPJ activity was manipulated with TMS suggests that information passes from rTPJ to the dorsomedial and ventromedial prefrontal cortices in the task. Thus, beyond expanding our knowledge of the connectivity-based mechanisms of decision-making in this specific task, these findings further illustrate the kind of insights only possible by combining causal intervention with neuroimaging.

From the perspective of understanding what the rTPJ does during decision-making more generally, these findings converge with recent



Kim Caesar/Springer Nature

Figure 1 Experimental task and key findings. (**a**) Participants play either the role of the employee or that of the boss in a 'shirk or work' task. Employees receive theta-burst TMS to inhibit the rTPJ or at a vertex control site, and then they perform the task in the fMRI scanner. Bosses perform the task at the same time but not in the scanner. (**b**) The rTPJ causally underpins computation of the influence an employee's choices have on the behavior of their bosses. This involves connectivity of the rTPJ with dorsomedial and ventromedial prefrontal regions.

work examining the behavioral consequences of the same rTPJ theta-burst manipulation on attentional reorienting and theory-of-mind tasks[8]. In that study, researchers found that inhibiting the rTPJ leads to impairments in the reorientation of attention after invalid cues in a cognitive task and higher error rates in response to false beliefs in a theory-of-mind task. Thus, in the context of the social interactions in the employer-versus-employee task used by Hill *et al.*[1], the rTPJ inhibition-driven impairments in switching between choices (which would help employees outsmart employers) may reflect a more general role for this region in exploration of alternative strategies in situations wherein adopting another strategy would be beneficial. The current findings also emphasize that the rTPJ does not do this alone but rather drives activity in medial prefrontal regions in support of optimal decision-making. Indeed, the greater the blunting of rTPJ–ventromedial prefrontal connectivity by TMS, the less subjects accounted for how their choices influenced others into their own decisions.

These findings also elegantly illustrate a much-needed shift in approach that will become increasingly important in human neuroscience: a shift toward a more direct investigation of causal mechanisms in the context of sophisticated computational accounts of behavior. No amount of correlational computational modeling could provide the kind of strong conclusions made possible by directly manipulating regional brain function. Conversely, simplistic accounts of the behavioral consequences of experimental manipulations of brain activity, even if causal, may nonetheless not inform an understanding of information coding nor arbitrate between competing theoretical models. Neuroimaging as a field has matured to the point where these are now the critical questions to answer. Moreover, the present study provides a new perspective on neuroimaging studies of psychiatric conditions, wherein abnormalities in the activation of the rTPJ have been noted in individuals with schizophrenia[9] and autism[10]. Discovery of a specific causal role for the rTPJ in influence calculation during social decision-making sets up testable hypotheses about the nature of abnormal neural computations in such individuals. Arguably, the combination of computational modeling, neuroimaging and TMS described by Hill and colleagues will also accelerate convergent 'computational psychiatry' efforts.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Hill, C.A. *et al. Nat. Neurosci.* **20**, 1142–1149 (2017).
2. Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. *Proc. Natl. Acad. Sci. USA* **105**, 6741–6746 (2008).
3. Behrens, T.E., Hunt, L.T. & Rushworth, M.F. *Science* **324**, 1160–1164 (2009).
4. Gläscher, J.P. & O'Doherty, J.P. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 501–510 (2010).
5. Poldrack, R.A. & Farah, M.J. *Nature* **526**, 371–379 (2015).
6. Hallett, M. *Neuron* **55**, 187–199 (2007).
7. Chung, S.W., Hill, A.T., Rogasch, N.C., Hoy, K.E. & Fitzgerald, P.B. *Neurosci. Biobehav. Rev.* **63**, 43–64 (2016).
8. Krall, S.C. *et al. Hum. Brain Mapp.* **37**, 796–807 (2016).
9. Kronbichler, L., Tschernegg, M., Martin, A.I., Schurz, M. & Kronbichler, M. *Schizophr. Bull.* https://doi.org/10.1093/schbul/sbx073 (2017).
10. Pantelis, P.C., Byrge, L., Tyszka, J.M., Adolphs, R. & Kennedy, D.P. *Soc. Cogn. Affect. Neurosci.* **10**, 1348–1356 (2015).