Spotlight

Neurocomputational mechanisms of adaptive mentalization in humans

Toan Nong[1,2], Jun Feng[3], and Jean-Claude Dreher[1,2,*]

'Theory of mind' (ToM) is classically investigated with 'static' inference tasks, which miss the dynamic nature of social interactions. In a recent article, Buergi, Aydogan, and colleagues combined computational modeling and neuroimaging to study the adaptive nature of mentalization (i.e., the ability to infer the continuous change of others' thoughts and intentions).

Representing other people's thoughts, feelings, intentions, what neuroscientists call mentalization or ToM, is central to social cognition. Whether mentalization is shared with nonhuman primates, how it develops in children, and whether large language models (LLMs) have ToM are topics of active debate [1–5]. Characterizing this capacity at the computational level is crucial to advancing these discussions. Yet, while decades of research have examined the neural bases of ToM, most paradigms have treated it as a static skill, often using 'false belief' tasks. By contrast, successful real-world interactions require continuous and flexible adaptation (i.e., adjusting one's inferences about others' reasoning as their strategies evolve).

In a recent paper, Buergi, Aydogan, *et al.* address this gap with the Cognitive Hierarchy Assessment (CHASE) model, a Bayesian approach that formalizes trial-by-trial updating of beliefs about an opponent's sophistication during repeated Rock–Paper–Scissors (RPS) games [6]. CHASE outperformed alternative models, capturing how individuals (N >500) adapt their recursive reasoning (called 'level-k thinking') to opponents of varying sophistication.

In an fMRI subset of participants (N = 50), belief updates derived from CHASE were associated with activity in the temporoparietal junction (TPJ), dorsomedial prefrontal cortex (dmPFC), anterior insula, and dorsolateral prefrontal cortex, which are core nodes of the 'social brain'. Crucially, multivariate decoding demonstrated that these neural patterns could predict the extent of adaptive updating both across individuals and in out-of-sample data. This finding establishes a distributed 'neural fingerprint' of adaptive mentalization, which was replicated in an independent, more demographically diverse cohort.

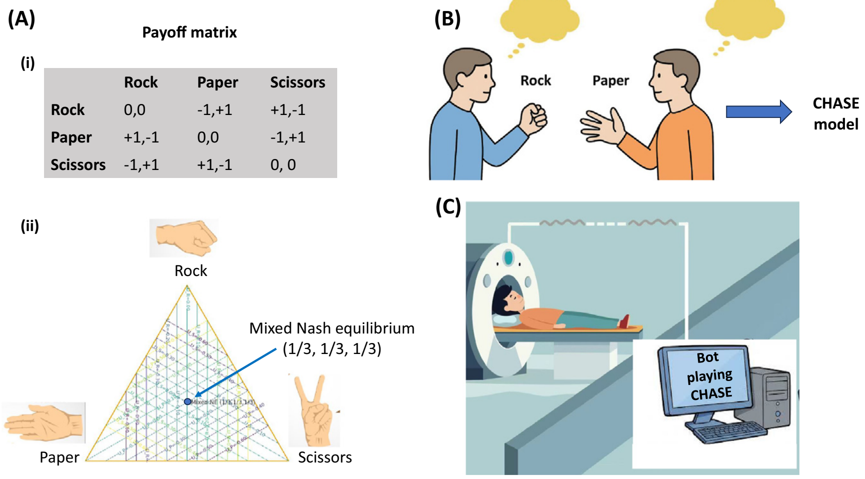
This work advances the field in three key ways. First, it reframes mentalization as a dynamic Bayesian inference process, aligning with recent decision neuroscience efforts that emphasize latent belief updating rather than static mentalization strategies [7]. Second, it reveals that TPJ activity, long associated with ToM, encodes not only belief attribution, but also the dynamic adaptation of one's mentalization depth to others' changing strategy. Third, it provides a robust neural marker with predictive validity, offering potential translational utility for assessing individual differences and clinical impairments in social cognition.

The finding that dynamic adaptation is encoded by the TPJ resonates with other recent computational accounts of social interactions. For example, when the nature of the cooperative or competitive interactions is not explicitly cued, a mixture of influence learning models describes how the brain dynamically arbitrates between cooperative and competitive intentions based on their relative reliability [7]. In this latter study, right TPJ engagement was also observed during such adaptation of one's behavior to the changing intentions of others. These results suggest that TPJ involvement in ToM reflects an even more general adaptive social inference process, whether adjusting one's mentalization depth to the opponent's strategy or adapting to others' changing intentions.

The authors' emphasis on predictive validity is particularly valuable. Using multivariate pattern analysis, both the levels of strategic sophistication and the model-inferred belief update could be decoded and predicted out-of-sample from distributed neural activity across the whole brain, but less so from the 'social brain'. This raises intriguing questions about the localization and specificity of adaptive mentalization signals. Thus, this meticulous demonstration lays a solid foundation for understanding the neurocomputational mechanisms that enable humans to flexibly adjust their mentalizing strategies in dynamic social contexts.

The RPS game provides an elegant model of adaptive mentalization. As a nontransitive game, where no single strategy dominates (rock beats scissors, scissors beat paper, and paper beats rock), it lacks a pure evolutionarily stable strategy (ESS), since any fixed choice can be exploited (Figure 1). Instead, the game has a mixed ESS, in which each option is chosen with equal probability. When players randomize their choices equally, no alternative strategy performs better on average. However, in practice, during human–human RPS games, people often deviate from this equilibrium, producing cyclic dynamics in which the frequency of one action rises and falls in turn with others [8]. For example, people might collectively move from playing more rock to more paper to more scissors, again and again.

However, in the current fMRI experiment, participants were not playing with other humans but with three bots with fixed



Trends in Cognitive Sciences

Figure 1. Schematic of the Rock–Paper–Scissors (RPS) game and experimental design. (A) (i) The payoff matrix represents wins by 1, ties by 0, and losses by –1. (ii) The probability simplex diagram for the RPS game, in which each corner corresponds to playing one pure action (Rock, Paper, or Scissors), the center point (blue dot) is the mixed Nash equilibrium (MNE), where each action is played with probability 1/3. Each set of contour lines is the expected payoff U_A(action) for Player A when she plays that pure action against an opponent mixing at the point in the simplex. (B) In several behavioral versions of the RPS game, participants (N = 456) played against human opponents, and their behavior was best explained by the Cognitive Hierarchy Assessment (CHASE) model across a wide range of game specifications. (C) In the scanner, participants (N = 50) played against three different artificial opponents based on the CHASE model with 0, 1, or 2 steps of reasoning. (A–C) generated with the help of ChatGPT.

levels of reasoning based on a simplified version of the CHASE model. This fMRI design controlled the policies of the bots and circumvented the tendency toward cyclic patterns previously observed in repeated human–human RPS games [8]. Yet, a limitation of this fMRI study is that the bots may have led participants to behave in a particular way. With repeated trials, two interacting humans could become sufficiently sophisticated to get close to the Nash equilibrium (i.e., the mixed ESS) during the final periods of the experiment, while the bots do not have this ability. Although this does not challenge the main conclusions of the authors, it questions the extent to which the behavior and brain activity identified in the current study would emerge during genuine human–human interactions, entering a cyclical pattern.

Future studies should test whether the identified neural computations generalize to group interactions. Mentalization in groups may rely on analogous mechanisms, such as simulating the average group member's mind and making predictions of others' decisions while also simulating the effects of one's own actions on the dynamics of the group [9]. Moreover, despite its strengths, the RPS may not capture the full complexity of human strategic reasoning. Developing other models of adaptive mentalization that account for more complex repeated strategic games will be useful to better characterize the neural mechanisms engaged in adaptive learning and hierarchical reasoning. For example, a recent model of adaptive mentalization used the 11–20 money request game to determine the neural computations underlying these processes [10]. This hybrid model includes three different learning processes: adaptive learning; simulated adaptive learning (i.e., simulating adaptive learning from the perspective of the opponent); and sophisticated learning (which is similar to the belief updating component in the CHASE model). It more accurately explains dmPFC activity compared with pure adaptive learning models. Moreover, it reveals a prediction error signal for the sophisticated learning process that engages the dorsolateral prefrontal cortex.

Overall, this study marks a milestone in the quest to uncover the neurocomputational bases of ToM. By integrating formal models of adaptive mentalization with neural signatures, Buergi, Aydogan *et al.* capture the essence of human social intelligence: the capacity to flexibly model, predict, and adapt to the changing minds of others.

Acknowledgments

This research benefited from the financial support of the LABEX CORTEX to J-C.D. (ANR-11-LABX-0042) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency. This work was also supported by grants from the Agence Nationale pour la Recherche to J-C.D. (ANR-24-CE37-4261-01, ANR-24-NEUC-0001 and ANR-24-EXEN-0001, which is part of the 'Plan France 2030'). J.F. received support from the Natural Science Foundation of China (72371090).

Declaration of interests

None declared by authors.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to generate three panels for the figure. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

[1]Neuroeconomics Lab, Institut des Sciences Cognitives Marc Jeannerod, Centre National de la Recherche Scientifique (CNRS), Lyon, France
[2]University Claude Bernard Lyon, Lyon 1, Villeurbanne, France
[3]School of Economics, Hefei University of Technology, Hefei, China

*Correspondence:
dreher@isc.cnrs.fr (J.-C. Dreher).

https://doi.org/10.1016/j.tics.2025.11.009

References

1. Tomasello, M. (2025) How to make artificial agents more like natural agents. *Trends Cogn. Sci.* 29, 783–786

2. Meisner, O.C. *et al.* (2025) Diverse and flexible strategies enable successful cooperation in marmoset dyads. *Curr. Biol.* 35, 4509–4521
3. Nong, T. *et al.* (2023) Computational mechanisms underlying the emergence of theory of mind in children. *PsyArXiv* Published online September 5, 2023. http://dx.doi.org/10.31234/osf.io/y876r
4. Strachan, J.W.A. *et al.* (2024) Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 8, 1285–1295
5. Akata, E. *et al.* (2025) Playing repeated games with large language models. *Nat. Hum. Behav.* 9, 1380–1390
6. Buergi, N. *et al.* A neural fingerprint of adaptive mentalization. *Nat. Neurosci.* (in press).
7. Philippe, R. *et al.* (2024) Neurocomputational mechanisms engaged in detecting cooperative and competitive intentions of others. *Nat. Commun.* 15, 3189
8. Hoffman, M. *et al.* (2015) An experimental investigation of evolutionary dynamics in the Rock-Paper-Scissors game. *Sci. Rep.* 5, 8817
9. Khalvati, K. *et al.* (2019) Modeling other minds: Bayesian inference explains human choices in group decision-making. *Sci. Adv.* 5, eaax8783
10. Feng., J. *et al.* (2025) Neural correlates of interactions between adaptive learning and hierarchical reasoning in repeated strategic games. *SSRN* Published online October 19, 2025. http://dx.doi.org/10.2139/ssrn.5625070