Computational Social Cognition: Approaches and Challenges

Ismail Guennouni (ismail.guennouni@zi-mannheim.de) Z.I Mannheim and University of Heidelberg

Joseph M Barnby (joseph.barnby@rhul.ac.uk) Royal Holloway, University of London

Julian Jara-Ettinger (julian.jara-ettinger@yale.edu) Yale University

Rebecca Saxe (saxe@mit.edu) MIT

Maarten Speekenbrink (m.speekenbrink@ucl.ac.uk) University College London

Keywords: Computational Models; Social Cognition; Theory of Mind

Introduction

Predicting the actions and reactions of others is crucial to successful social interaction. When deciding whether to bluff in a game of poker, we consider the chances that the other players will fold or continue to play and unmask our bluff. When deciding whether to tell our boss that their plans are likely to have adverse effects, we consider a range of reactions, from being grateful for our honesty to being dismissed out of spite. Such predictions are highly uncertain and complex, not least because the other's (re)actions usually result from them making equally complex and uncertain inferences about us. Nevertheless, we are often remarkably successful – although sometimes utterly wrong – in our social inferences. How do we explain these successes and failures?

Theory of Mind (ToM) or mentalizing, referring to the ability to represent another's latent mental states (e.g. thoughts, beliefs, motivations, and emotions) is a central concept in theories of social inference (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Barnby, Bellucci, et al., 2023; Radkani & Saxe, 2023). That predictions of another's actions derive from inferences of such latent states is a vague –yet meaningful – statement. How are these inferences made? And what are the benefits of representing others in terms of their beliefs, motivations, and emotions, beyond predicting their (re)actions?

There is a growing consensus that computational models are required to reduce this ambiguity and improve the falsifiability of theories of social interaction (FeldmanHall & Nassar, 2021). These models should be both mathematically tractable and psychologically plausible, offering insights that are scientifically robust and socially relevant. Given the complexities of social interactions, it is unlikely that models which work well in non-social domains (e.g. standard reinforcement learning models) can be called upon without modification.

This symposium brings together researchers who have taken up the challenge of computational social cognition with a variety of approaches, including generative Bayesian models (Barnby, Dayan, & Bell, 2023), inverse reinforcement learning (Jara-Ettinger, 2019), unsupervised clustering techniques (Guennouni & Speekenbrink, 2022), and predictive coding (Koster-Hale & Saxe, 2013). The symposium aims to identify the unique challenges that face computational social cognition, how these are met by the different approaches, and what remains to be addressed in the future.

Contributors

Joseph Barnby is a computational and cognitive neuroscientist, and Assistant Professor at Royal Holloway, University of London. He received his PhD in Cognitive Neuroscience from King's College London. His lab – the Social Computation and Representation Lab – is interested in the brain basis of social interaction, using and developing computational models imbued with Theory of Mind to provide novel theoretical frameworks and empirical evidence to inform the aetiology of psychiatric disorder, and to develop more dynamic artificial systems.

Ismail Guennouni is a computational cognitive scientist and Postdoctoral Research Fellow at the AI Health Innovation Cluster. He received his PhD in Psychology from UCL with a focus on experimental and computations approaches to strategic social interaction. He is interested in exploring how cognitive interventions, combined with computational modelling of behaviour can help address social dysfunction inherent in many mental health disorders.

Julian Jara-Ettinger is an associate professor of psychology at Yale University, with affiliations to the Computer Science department, the Cognitive Science program, and the Wu Tsai institute. Julian received his PhD in Cognitive Science at MIT. At Yale, Julian's research group–the computational social cognition lab–aims to characterize the representations and computations that support human social cognition, understand how they emerge and develop, and use them to build more human-like machine social intelligence.

Rebecca Saxe is a social cognitive neuroscientist, John W Jarve (1978) Professor of Cognitive Neuroscience, and Associate Dean of Science at MIT. She received her PhD in Cognitive Science at MIT, and was a junior fellow in the Harvard Society of Fellows. Her lab studies the development and neural basis of human cognition, focusing on social cognition.

Maarten Speekenbrink is Professor of Mathematical Psychology at UCL. His research combines computational models and behavioural experiments to identify core elements of human learning and decision making. In recent work, he focuses on how these processes operate in social interactions with other agents.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Walking a mile in their shoes - Refining the neural and computational foundations of social representation

Joseph Barnby

Strategic reasoning is essential to avoid deception. Too much vigilance can however lead to false beliefs about a partner's intended harm. This talk will focus on recent work developing mathematical models of how humans build recursive maps of their social partners for strategic interaction, testing which neurochemical and social factors cause this ability to go awry, and how this may explain psychopathological symptoms. Making small mis-calibrated changes to the way in which artificial agents interact causes social interaction to break down and can account for a several psychopathological symptoms observed in the clinic. It also identifies the necessity of calibrating artificial systems to their users to ensure ingenuous behaviour is not mistaken as threat.

Hidden Markov models to capture the dynamics of strategic interaction and build adaptive human-like agents

Ismail Guennouni

In social environments, the outcomes of our actions are influenced by the behavior of others, necessitating mutual adaptation. Static opponent models are inadequate when dealing with players who are also learning and modelling our strategies. In this talk, I will introduce a framework based on hidden Markov models, tailored to capture the evolving dynamics of interactions in repeated economic games. I will show how this approach effectively characterizes player behavior through analysis of dyadic human exchange data. I will then explore how these HMM models facilitate the development of adaptive artificial agents that can emulate human behavior in economic games, whilst offering a higher degree of experimental control. Additionally, I will demonstrate the application of these agents in intervention studies aimed at bolstering cooperative behavior in such games.

Social representations as probabilistic programs

Julian Jara-Ettinger

Virtually all areas of uniquely-human intelligence, from moral reasoning to language understanding, rely on social cognition. Characterizing its computational structure is therefore a central challenge in cognitive science and critical towards engineering more human-like AI. In this talk I will present a computational framework of social cognition, where agent behavior is represented hierarchically through a combination of symbolic propositional programs, with internal non-symbolic continuous representations of mental state contents. I show experimental evidence that this framework captures how people make sense of behavior, including cases with complex reward structures that are not adequately captured by previous approaches.

Using inverse planning to learn from and communicate with social actions

Rebecca Saxe

Any social action has multiple possible explanations. For example, consider punishment. If an authority chooses to punish a norm violation, one possible explanation is that the norm violation is morally wrong, and the authority is impartial. Another possible explanation is that the norm violation is relatively innocuous, and the authority is biased against the target. I will present a formal cognitive model of how people accommodate the ambiguity of social actions, and rationally jointly update beliefs about the situation (i.e. the norm violation) and the actor (i.e. the authority's motives), depending on their priors. This model predicts when the beliefs of different people observing the same social interactions will rationally diverge, and fits human inferences across three studies (N=1260). The model of observers' inferences can be further embedded, recursively, in a model of planning, to explain how people anticipate the reputational consequences of their decisions and plan communicative social actions.

Panel discussion

Maarten Speekenbrink

The talks will be followed by a panel discussion, introduced by Maarten Speekenbrink. He will aim to integrate insights from the four talks, and highlight future challenges and directions for the field. These will then be discussed by the panel of speakers.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Barnby, J. M., Bellucci, G., Alon, N., Schilbach, L., Bell, V., Frith, C., & Dayan, P. (2023). Beyond theory of mind: A formal framework for social inference and representation. *PsyArXiv*. http://doi.org/10.31234/osf.io/cmgu7

Barnby, J. M., Dayan, P., & Bell, V. (2023). Formalising social representation to explain psychiatric symptoms. *Trends in Cognitive Sciences*, *27*, 317–332.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057.

Guennouni, I., & Speekenbrink, M. (2022). Transfer of learned opponent models in zero sum games. *Computational Brain & Behavior*, *5*(3), 326–342.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron*, *79*(5), 836–848.

Radkani, S., & Saxe, R. (2023). What people learn from punishment: Joint inference of wrongness and punisher's. In *Proceedings of the annual meeting of the cognitive science society, 45 (45)*.