Inappropriate use of sophisticated mentalisation produces paranoia: a new formal theory

Nitay Alon[1,2], Peter Dayan[1,3], Vaughan Bell[4], Michael Moutoussis[5,6], and Joseph Barnby[7]

[1]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[2]Department of Computer Science, Hebrew University of Jerusalem, Jerusalem, Israel
[3]Department of Computer Science, University of Tübingen, University of Tübingen, Tübingen, Germany
[4]Clinical, Educational, and Health Psychology, University College London, UK
[5]Wellcome Centre for Human Neuroimaging, University College London, London, UK
[6]Max-Planck – UCL Centre for Computational Psychiatry and Ageing, University College London, London, UK
[7]Department of Psychology, Royal Holloway, University of London, London, UK

March 15, 2023

The ability to ascribe beliefs, desires, and intentions to others is known as Theory of Mind (ToM). Moreover, one can envision others as thinking about its beliefs in a nested way. This enables humans to navigate complex social situations and to achieve joint goals. Depending on the social demands, different depths of mentalisation (DoM) are required for successful social outcomes, ranging from shallow to hierarchical and recursive.

Inaccurate mentalisation has been suggested as an explanatory mechanism in a number of psychiatric disorders, such as personality disorder and psychosis. For example, instead of taking the actions of others at face value, a self may ascribe hidden motives to others, viewing them as competitive agents engaged in complex and clandestine strategic planning.

Here, we present a computational model based on hierarchical ToM to formalise the notion that aspects of paranoia - the over-ascription of self-focused intent-to-harm from others - might be explained as the over-attribution of competitive intentions and strategic sophistication when faced with ambiguous, noisy, or prosocial behaviour. We present a variant on an Ultimatum game, called the External Agency Test, in which a self (the receiver) negotiates with an other (the proposer) for rewards by accepting or rejecting monetary offers. We find that when selves inappropriately attribute incorrect sophisticated mental states to others, their performance in the task is impaired, and confidence in their incorrect inferences increase, and this is will be more pronounced in those with higher trait paranoia.