What will he do next? Imaging of mentalizing in an observed iterated trust game

Christoph Mathys[1,2], Tony B. Williams[3], Lars Kasper[1], Lilian A. E. Weber[1], Ernst Fehr[3], Klaas E. Stephan[1,2]

[1] Translational Neuromodeling Unit (TNU), University of Zurich and ETH Zurich, Switzerland
[2] Wellcome Trust Centre for Neuroimaging, University College London, United Kingdom
[3] Laboratory for Social and Neural Systems Research (SNS Lab), Department of Economics, University of Zurich, Switzerland

(Combined word limit for all four sections: 4000 characters, including spaces)

Introduction

Mentalizing (i.e., putting oneself in the shoes of another) is a crucial skill for understanding interpersonal relations and has strong bearing on psychiatric disorders. In this context, trust games, which are widespread in behavioral economic research, have recently been shown to be relevant to psychiatric research [1]. However, while letting subjects engage in interactions may prove to be a valuable tool for the investigation of mentalizing and for clinical assessments, it entails logistical and conceptual difficulties: it is impractical always to need two subjects for a task, and it is unclear how to compare the behavior of two subjects who have interacted with two possibly very different partners. For these reasons, with the intention to investigate the neural mechanisms underlying mentalizing, we developed a mentalizing task that can be performed by one subject alone and is exactly the same for all subjects.

Methods

Our task was based on a trust game that involves two players, an investor and a trustee. Both receive the same initial endowment on each round. The investor may then transfer some or all of his endowment to the trustee. On the way to the trustee, the transferred amount is multiplied by three. The trustee may then repay some or all of what he has (i.e., his initial endowment plus the multiplied transferred amount) to the investor. Both players will end up with the same amount if the trustee repays twice what the investor sent him, and the more the investor sends the more the parties jointly profit. However, the interaction is already profitable to the investor as soon as the trustee sends back more than one third of what he received. In our task, subjects did not themselves participate in a trust game but observed the interactions in a game that had previously taken place between two other subjects (Fig. 1). While in an fMRI scanner, they had to predict the next decision of the trustee before that decision was revealed to them. Prediction accuracy was rewarded, and the game went over 70 rounds to allow for several buildups and breakdowns of cooperation in the previously recorded interaction. 44 healthy subjects (22 from each sex) participated. We used a recently developed hierarchical Bayesian model of learning, the hierarchical Gaussian filter (HGF) [2], to model our subjects' evolving perception of the observed interaction dynamics. This allows for subject-by-subject estimation of parameters characterizing individual differences in learning styles. Individual belief trajectories were calculated and used as regressors in the fMRI analysis (Fig. 2). Chief among them were the time-varying belief on the current cooperativeness of the trustee and the volatility of that belief. Furthermore, several questionnaire- and behavior-based psychological measures were assessed.

Results

Task performance was significantly correlated with the "openness to experience" scale of the NEO-FFI, and, in women, with the "reward responsiveness" scale of the BIS/BAS. The fMRI analysis revealed strong responses in various brain regions to very abstract inferred measures: cooperativeness of the trustee and its volatility (i.e., measures at the first and second level of hidden states in the HGF, Figs 3,4).

Conclusions

While the nature of the interaction dynamics in trust games has previously been quantified in a model-free way, we used the HGF to model our subjects' perception of the interaction dynamics in a trust game. Such a model-based fMRI analysis has the advantage of being interpretable in terms of the modeled quantities; it revealed strong responses in regions previously associated with reward, valuation, empathy, and cognitive control. Furthermore, this study introduces a task that is quick and simple, and which can crucially be performed by one subject alone while potentially retaining the anticipated advantages of assessing multiplayer interactions for clinical diagnosis and prognosis.

Figures

Fig1.png
Fig2.png
Fig3.png
Fig4.png

References (Harvard author-date style)

1. King-Casas, B., et al. (2008), 'The Rupture and Repair of Cooperation in Borderline Personality Disorder', Science, vol. 321, pp. 806-810.

2. Mathys, C., et al. (2011), 'A Bayesian foundation for individual learning under uncertainty', Frontiers in Human Neuroscience, vol. 5: 39.

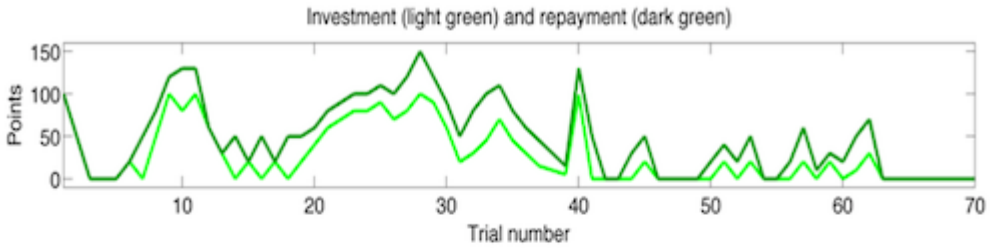

Fig. 1 | Iterated trust game prediction task. Subjects were asked to observe an iterated trust game that went over 70 rounds. They were informed that the game had actually taken place at our lab in precisely the way they would see it. They were not given any information about the two players they were observing beyond the numbers announcing the investor's investment and subsequently the trustee's repayment, sequentially for each round. Subjects knew that both players had an endowment of 100 points on each round, convertible into Swiss francs at the end of the game at a certain rate. On each trial, subjects could win 400 points by correctly predicting the trustee's repayment. They received this maximum for exactly accurate predictions, but reward fell with the squared difference between prediction and outcome, so that no reward was given for prediction errors of 20 or more points. At the end of the experiment, points were converted into Swiss francs at a rate of 0.002, resulting in a maximal total reward of 56 francs (about 59 U.S. dollars). At no point in this experiment were subjects deceived. Out of twelve subject pairs that played the iterated trust game in a pre-study, about half (seven) showed relatively uninteresting and simple-to-predict patterns of consistently high cooperation. The other five exhibited a more varied series of interactions with cooperation intermittently breaking down before picking up again. The figure above shows the course of the game chosen for use in the main study. Coaxing of the investor by the trustee is evident in many places throughout the game, for example around trials 15 and 60. Coaxing refers to a behavior where one partner in the interaction shows greater cooperation than mere reciprocation of the other partner's cooperation. In so far as coaxing leads to the establishment of trust and greater cooperation, it is beneficial to both parties. Patients with borderline personality disorder have been shown to exhibit less coaxing than healthy controls [1].

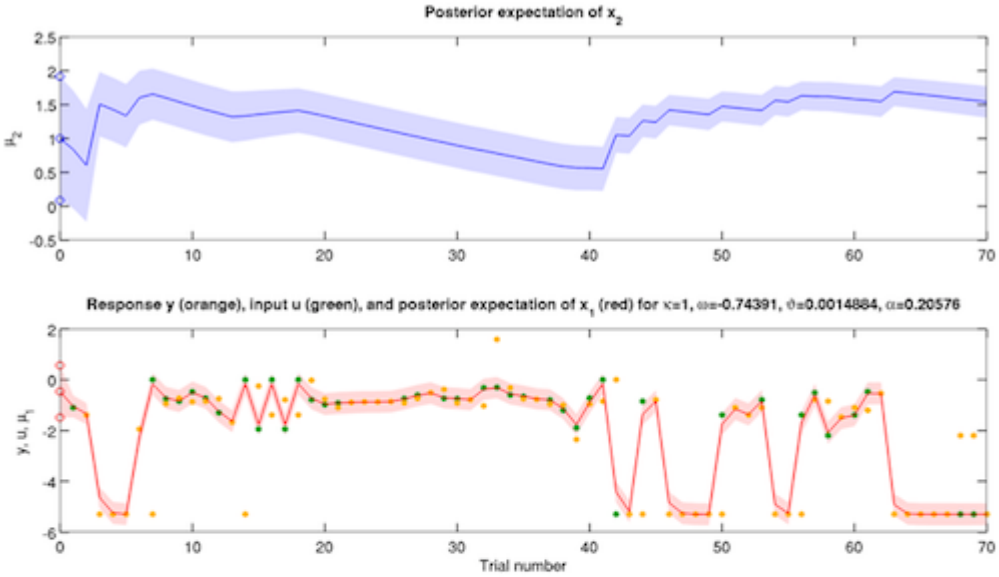

Fig. 2 | Hierarchical Gaussian filter (HGF) estimation of cooperativeness and its volatility. Trajectories for one example subject. The HGF is a hierarchical Bayesian model of learning [2] that can be used to describe how subjects track changing quantities that shape their environment. In the present application, the quantity of interest is the cooperativeness of the trustee: the subject has to track this to make good predictions about the trustee's repayment. Owing to the particular hierarchical structure of the HGF, the volatility of any quantity of interest is also tracked and can be used for inference on the neural correlates of inferring others' states of mind. Bottom: inferred cooperativeness (red line: mean; shaded area: standard deviation). Orange dots indicate predicted cooperation, green dots indicate observed cooperation. Top: volatility of inferred cooperativeness (blue line: mean; shaded area: standard deviation).

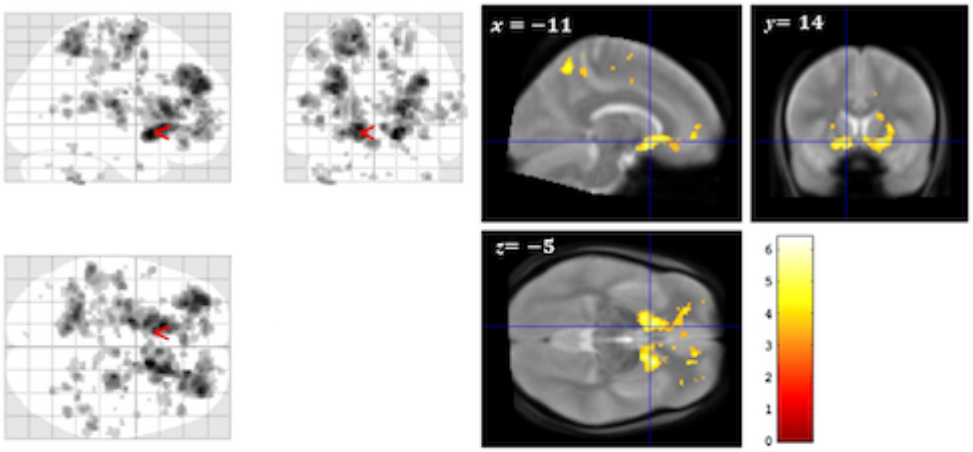

Fig. 3 | Negative correlation with inferred cooperativeness of the trustee. MRI images were acquired on a Philips 3 Tesla Achieva whole body MR with an eight-channel Philips SENSE head-coil. 40 transverse slices were measured in ascending order with a slice thickness of 2.6 mm and a gap of 0.7mm for a voxel size of 2.0×2.0×2.6 mm and a field of view of 192×192×131.3 mm. A T_2*-weighted single-shot echo-planar imaging sequence with a TR of 2500 ms, a TE of 25 ms, and a flip angle of 80° was used. fMRI data were pre-processed and analyzed using SPM8. The figure shows areas whose signal correlates negatively with the inferred cooperativeness of the trustee at a level of p<0.001 uncorrected. Significant clusters (p<0.05 FWE whole-brain corrected) are found in the ventral striatum, the dorsolateral prefrontal cortex, the orbitofrontal cortex, and left parietal cortex.

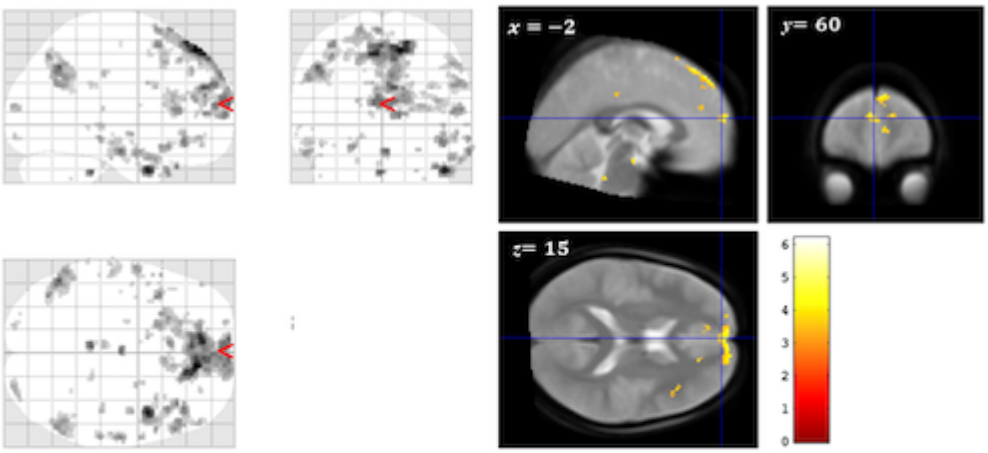

Fig. 4 | Positive correlation with volatility of inferred cooperativeness. The figure shows areas whose signal correlates positively with the volatility of the inferred cooperativeness of the trustee at a level of $p<0.001$ uncorrected. Significant clusters ($p<0.05$ FWE whole-brain corrected) are found in the medial prefrontal cortex and the left parietal cortex.