Published in final edited form as:
*Brain Imaging Behav*. 2009 March 1; 3(1): 24–37. doi:10.1007/s11682-008-9047-y.

# Brain Activity Dissociates Mentalization from Motivation During an Interpersonal Competitive Game

**Michal Assaf**
Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, 200 Retreat Ave., Hartford, CT 06106, USA

Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

**Itamar Kahn**
Howard Hughes Medical Institute at Harvard University, Cambridge, MA, USA

**Godfrey D. Pearlson**
Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, 200 Retreat Ave., Hartford, CT 06106, USA

Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

**Matthew R. Johnson**
Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, 200 Retreat Ave., Hartford, CT 06106, USA

Interdepartmental Neuroscience Program, Yale University, New Haven, CT, USA

**Yehezkel Yeshurun**
School of Computer Science, Tel Aviv University, Tel Aviv, Israel

**Vince D. Calhoun**
Olin Neuropsychiatry Research Center, Institute of Living, Hartford Hospital, 200 Retreat Ave., Hartford, CT 06106, USA

Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA

The Mind Research Network, Albuquerque, NM, USA

Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM, USA

**Talma Hendler**
Functional Brain Research Center, Wohl Institute for Advanced Imaging, Tel Aviv Sourasky Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

## Abstract

Studies demonstrating selective brain networks subserving motivation and mentalization (i.e. attributing states of mind to others) during social interactions have not investigated their mutual independence. We report the results of two fMRI studies using a competitive game requiring players to use implicit 'on-line' mentalization simultaneously with motivational processes of gains and losses in playing against a human or a computer opponent. We delineate a network, consisting of bilateral temporoparietal junction, temporal pole (TP), medial prefrontal cortex (MPFC) and right fusiform



massaf@harthosp.org.

gyrus, which is sensitive to the opponent's response (challenging>not challenging the player) and opponent type (human>computer). This network is similar to a known explicit 'off-line' mentalization circuit, suggesting its additional involvement in implicit 'on-line' mentalization, a process more applicable to real-life social interactions. Importantly, only MPFC and TP were selective to mentalization compared to motivation, highlighting their specific operation in attributing states of mind to others during social interactions.

### Keywords

Theory of mind; Reward; Medial prefrontal cortex; Temporoparietal junction; Temporal pole

## Introduction

Human interpersonal relationships involve effectively processing our own and others' states of mind (including desires, goals and beliefs). These cognitive processes are often referred to as motivation and mentalization (also known as 'Theory of Mind'), respectively. As an example of the interplay between mentalization and motivation in a social interaction one can think about a common situation where an employee is upset with his/her boss. The initial desire of the employee might be to burst into the boss's office to confront him/her (i.e. self-motivation); however, most employees will stop and think about the boss's potential emotional reaction and action (i.e. "my boss will be upset and fire me"; thinking about the boss's potential state of mind refers to the process of mentalization) and this prospective outcome might prevent them from action (i.e. motivation to avoid punishment). Motivation and mentalization may engage selective brain networks as supported by lesion and neuroimaging studies (Dolan 2002; Phan et al. 2002; Adolphs 2003; Frith and Frith 2003; Ressler 2004). However, while these processes likely occur simultaneously in most social interactions, no prior study has clearly differentiated between the neural networks underlying each.

The specialized mentalization-related network likely involves temporoparietal junction (TPJ), temporal pole (TP), and medial prefrontal cortex (MPFC) (for review see Frith and Frith 2003). Most previous studies used explicit 'off-line' mentalization tasks requiring participants to retrospectively attribute states of mind to others after explicitly being asked to think about the other person. However, in real-life, mentalization is typically an implicit 'on-line' process requiring ongoing prediction of other peoples' behavior without explicit instructions to do so (hence, implicit). The few studies exploring implicit 'on-line' mentalization using interactive games consistently show activation only in MPFC (McCabe et al. 2001; Gallagher et al. 2002; Rilling et al. 2004). This raises the possibility that 'on-line' mentalization activates different network than 'off-line' mentalization. In addition, interactive paradigms, unlike 'off-line' mentalization paradigms, also engage reward- and/or punishment-related motivational processes (i.e. appraising the value to the self of rewards and/or punishments, real or prospective, that are potential consequences of specific social situations); however, these processes have not been explored separately. The importance of this distinction is evident in the MPFC, which is implicated in both mentalization and reward processing associated with gains (e.g. Delgado et al. 2000; Pochon et al. 2002; Knutson et al. 2003). Thus, it remains to be demonstrated that its activation during interactive games is mentalization-selective and not motivation-related. The current study aimed to characterize the selective brain networks of motivation and mentalization during an interactive, two-person, on-line game.

We performed two functional MRI (fMRI) experiments as subjects played a competitive computerized Domino game, to characterize the neural circuit subserving implicit 'on-line' mentalization independently from motivation processes related to gains and losses (Fig. 1a, Kahn et al. 2002). In experiment I, players were told that their opponent was the human

experimenter, and thus believed that they participated in an interpersonal competitive game. Consequently, players were simultaneously engaged in both *mentalization* of their opponent's behavior (trying to predict the opponent's next move based on his/her last moves) and processing *motivational* aspects related to losses and gains (determined by their own moves and the opponent's responses). These processes mostly occurred in a well-defined interval after the opponent's response was revealed to the player (the Response to Outcome interval, Fig. 1a). This new information about the opponent was used by the players to update their representation of the opponent's state of mind (i.e. mentalization) and it determined the outcome of a specific round of the game (gain/loss, i.e. motivation). We hypothesized that by taking into account the motivational processes and their neural correlates we could delineate a selective 'on-line' mentalization network. To further examine the selectivity of the proposed mentalization network to processing the actions of a human agent, during a second experiment players were told they were playing against either a human or a computer opponent. We predicted that brain regions selective for 'on-line' mentalization would be more active in the former case.

## Methods

### Subjects

**Experiment I—**Nineteen healthy subjects (2 left-handed, 10 males), ages 21 to 56 years (mean ±SD: 32.3±10.4) with estimated full scale IQ (estimated by the National Adult Reading Test-Revised) of 94 to 122 (112±7) participated in this experiment. Participants had no current or past history of any psychiatric Axis I diagnosis, assessed by the Structured Clinical Interviewed for DSM-IV Axis I Disorders (SCID; First et al. 2002), no major physical illness, current or past history of neurological disease or substance abuse, no history of head trauma causing loss of consciousness, assessed by detailed interview, and negative urine screening for recreational drugs.

**Experiment II—**Twenty-five healthy participants were recruited for this experiment (none of whom participated in experiment I); however, data for eighteen participants (2 left handed, 9 males), ages 19 to 55 years (mean±SD: 28.8±12.4) with estimated full-scale IQ score (available for 14 subjects only) from 99 to 120 (111±6) is reported. Four subjects were excluded from analysis due to excessive head movements (>6 mm), and three because they scored all mentalization statements on a Likert scale (e.g. "I usually chose which chip to play next according to my opponent's last moves" and "When my opponent didn't challenge me I tended to pick a matching chip in my next step") identically for both human and computer *and* indicated that they did not differentiate between the two in terms of their strategy (i.e. they replied 'no' to the question "Did you have different strategies for the games you played against the experimenter and for the games you played against the computer?"). Inclusion criteria were identical to experiment I. There were no significant differences between participants in the two experiments on age ($t(35)=0.9$, n.s.), gender ($\chi(1)=0.03$, n.s.), race ($\chi(1)=0.67$, n.s.) and full-scale IQ ($t(30)=0.3$, n.s.). Participants in both experiments provided written informed consent, approved by the Hartford Hospital Institutional Review Board.

**The domino game paradigm:** The paradigm, described in Fig. 1a, is a two-player competitive computerized game. The scanned subject is the player while a computer randomly generates the opponent's responses. However, to test the uniqueness of human-related mentalization, players can be told that they are either playing against a human (Experiment I and two runs of Experiment II) or a computer opponent (two runs of Experiment II). Thus, from their perspective, during the human-opponent runs they are playing in an interpersonal competitive context. The game is composed of a pool of 28 domino-like chips; at the beginning of each game, 12 random domino chips are assigned to the player (shown face up on the screen), four

undisclosed chips are assigned to a bank and an opponent domino chip (constant throughout the game) shows face-up on the top-corner of the board. The 11 remaining chips from the overall pool are not used in the specific game. Each assigned chip can either match the opponent's chip (have one of the opponent chip's numbers) or not. The player's goal is to dispose of all assigned chips before the game ends (4 min). In the game context, matching chips are considered 'safe' moves and non-matching chips are considered 'risky' moves or 'bluffs', since they are associated with gains and losses of chips, respectively. It is only possible to win and to collect the resulting monetary bonus ($10/game) by occasionally bluffing (i.e. playing a non-matching chip).

During each round of the game, the player decides which chip to play next, places it face down adjacent to the opponent's chip and awaits the opponent's response. The opponent can either challenge the player by asking him/her to reveal the chosen chip or not challenge, allowing him/her to move on to the next round. Each round progresses according to the following commands, presented to the player both visually and aurally (Fig. 1a): (a) 'Choose' instructs the subject to mentally select a chip to be played next. The player then decides to either pick a matching or a non-matching chip; (b) 'Ready' instructs the subject to move a cursor (using his/her dominant hand) to the chosen chip. These first two steps each last 4 s; (c) 'Go' instructs the player to put the chosen chip face down next to the opponent's chip as quickly as possible. The player then awaits the opponent's response (either 3.4, 5.4 or 7.4 s) of either (d) 'Show' or 'No-Show'. The former command exposes the player's selected chip (revealing whether they played safe or bluffed), while the latter leaves it unexposed. The next round of the game starts by presenting the 'Choose' command again following a 5.4±2 second pause.

Based on the players' choice and opponent's response there are four possible consequences per game round: (1) Show of Match chip: the choice of a matching chip is exposed and the players rewarded by disposal of the selected chip and one additional random chip from the game board. At the end of these trials players dispose of two chips (*overt gain*); (2) Show of Non-Match chip: the players' choice of non-matching chip is exposed, and they are punished by receiving back the selected chip plus two additional chips (from the bank or previously played chips, thus not chosen by the player), for a total of three chips (i.e. *overt loss*); (3) No-Show of Non-Match chip: a choice of a non-matching chip remains unexposed, and only the selected chip is disposed of, so the players get away with a non-matching choice (*relative gain*); and (4) No-Show of Match chip: the choice of a matching chip is not exposed and only the selected matching chip is disposed of, so the players are relatively punished as they could have disposed of another chip (*relative loss*). For analysis purposes, it was assumed that the quality of the player's mentalization can be modified by the different opponent's responses ('Show' vs 'No-Show') the player needs to process. On the other hand, the motivation factor is tested in terms of the values of the outcome (i.e. gains vs. losses) (see Fig. 1b and c).

Rounds continue until the players win (by getting rid of all of their chips) or lose (when either 240 s have passed or they receive all the chips from the bank and the board back, and there are no more chips they can get).

During *experiment I*, participants played Domino games over three scan runs of 15 min each for an average of 12.5 games (due to technical problems, for two subjects only two runs were used). Participants were told they were playing against the experimenter, whom they met prior to the scan, outside the scanner. The experimenter talked to the participant after each run making competitive comments about the games just played (such as "you really got me this time…"). To ensure that players were engaged in the game and believed that winning was possible, if they did not win during the first run, the first game of the second run was not automated and the experimenter "threw" the game, ensuring that the player won. Five of the

19 players played a non-automated game; these games were excluded from the analysis. Games shorter than one minute (5.65% of all games) were not analyzed.

During *experiment II*, participants played Domino games over four scan runs of 15 min each for an average of 16.7 games. Participants were told that they were playing against the experimenter for two runs and against the computer for two (on average: 8.3 games against human and 8.4 games against computer). Players were told that the computer's moves were random. The order of the human and computer opponent runs was counterbalanced across subjects. Participants were told who they are playing against via headphones right before each run had begun. In this group only one player played a non-automated game, which was excluded from the analysis. Again, games shorter than one minute (5.32% of all games) were not analyzed.

For both experiments, participants practiced the game outside the scanner prior to scanning. Scanning began when the experimenter was convinced that participants understood the game's rules. A thorough debriefing was carried out immediately after scanning during which participants were asked about their emotions and strategies while playing the Domino paradigm using open-ended questions and a Likert scale questionnaire where participants rated their responses on a scale of 1(least) to 5 (most) agreement to statements (see examples above and in the “Results” section).

**Behavioral data analysis:** Likert scale scores were analyzed using one sample *t*-test against the score 3 (middle score). For experiment II, an additional paired *t*-test was calculated for each statement to test for differences between human- and computer-opponents.

To characterize player's choices during the game, a Non-Match Index was defined as the ratio between the number of times a player chose a non-matching chip to the total number of choices. This index represents an unbiased choice when equal to 0.5 (exactly half of the choices were non-matching choices), a biased choice for matching chips when smaller than 0.5 or for non-matching chips when greater than 0.5.

**Functional MRI acquisition:** Blood oxygenation level dependent (BOLD) data were collected with a T2*-weighted echo planar imaging (EPI) sequence (TR/TE=1,860/27 msec, Flip angle=70°, Field of view=22 cm with a 64×64 acquisition matrix) using a Siemens Allegra 3 Tesla scanner. Thirty-six contiguous axial functional slices of 3 mm thickness with 1 mm gap were acquired, yielding 3.4×3.4×4.0 mm voxels. Overall, 492 images were acquired during each run, including six ‘dummy’ images at the beginning to allow global image intensity to reach equilibrium, which were excluded from data analysis.

**Functional data analysis Experiments I and II—**Imaging data were analyzed using SPM2 (Wellcome Department of Cognitive Neurology, London, UK). Each individual's data set was realigned to the first ‘non-dummy’ image using the INRIAlign toolbox (A. Roche, INRIA Sophia Antipolis, EPIDAURE Group), spatially normalized to the Montreal Neurological Institute space (Friston et al. 1995) and spatially smoothed with a 9 mm isotropic (FWHM) Gaussian kernel.

We defined four intervals of interest to use for fMRI analyses (Fig. 1a): The Decision-making interval was defined from the ‘choose’ command onset to the ‘ready’ onset during which players were instructed to decide on their next move without being able to move on the board. The Ready interval was defined from the onset of ‘ready’ to the onset of ‘go’. These first two intervals lasted 4 s each. The third interval, Anticipation of Outcome, started after the selected chip was placed face down beside the opponent's chip and ended with the opponent's response. This interval was sorted according to the player's choice of matching or non-matching chips.

The fourth interval, Response to Outcome, started after the opponent's response and ended with the next 'choose' onset. Trials were sorted according to the player's choice and the opponent's response to derive the four conditions described above (Show Match, Show Non-Match, No-Show Match and No-Show Non-Match). The third and fourth intervals were jittered, randomly lasting 3.4, 5.4 or 7.4 s each (Dale and Buckner 1997).

**Experiment I—**For every subject, a general linear model (GLM) was estimated with SPM using the 'Ready', 'Anticipation of Outcome' and 'Response to Outcome' intervals as regressors while the 'Decision-making' interval was not modeled, and hence was used as 'baseline' (Kahn et al. 2002). High-pass filter with a cut-off of 128 s was applied to correct for EPI signal low frequency drift. Next, individual statistical parametric maps of the four Response to Outcome interval conditions were calculated (results related to the Anticipation of Outcome interval are described in Kahn et al. 2002). These maps were entered into a random effects group analysis using a 2×2 (Mentalization: show/no-show × Motivation: gain/loss) ANOVA (Fig. 1b). As mentioned earlier, we consider both 'Show' and 'No-Show' conditions to entail a mentalization related parameter. We thus expected regions related to mentalization to show differential activation along this parameter (Fig. 1b and c). To further evaluate the magnitude of the results, contrast values of individual subjects' time course were calculated per condition at the points of maximum group activations, and these values were entered into an ANOVA analysis using SPSS™ (SPSS Inc., Chicago, IL).

**Experiment II—**The same individual GLMs were defined with separate regressors for Human- and Computer Opponents and individual statistical parametric maps of the four Response to Outcome interval conditions were calculated separately for the two different opponents (a high-pass filter was applied as described for experiment I). These maps were masked with the group activation map of the main effect of Mentalization from experiment I (Fig. 3a) before they were entered into second level group ANOVAs (hence all group results presented were based on mask). This allowed us to perform a region of interest (ROI) analysis within the regions identified in experiment I as related to mentalization. Two random-effects 2×2 ANOVAs maps were calculated for experiment II. First, as for the experiment I, the masked-individual statistical maps of human opponent only were entered into a random effects group analysis of 2×2 (Mentalization × Motivation) ANOVA to replicate the first study's results. Second, a 2×2 (Opponent Type: Human/Computer × Mentalization) ANOVA was calculated. We then masked (using an inclusive mask) the contrast Human Mentalization (Show > No-Show; note that this is equivalent to the main effect of Mentalization in the first ANOVA) with the contrast of Human>Computer (main effect of Opponent Type) to identify regions common to both contrasts. As for the first experiment, contrast values of individual subjects' statistical maps were extracted at the points of maximum group activations for both contrasts (if coordinates of the maximal activation within the same cluster were different), and these values were entered into an ANOVA analysis.

For both studies, for the purpose of additional anatomic precision, group contrasts were overlaid on a surface-based representation of the MNI canonical brain using the SPM SurfRend toolbox (I. Kahn: http://spmsurfrend.sourceforge.net). The surfaces were then rendered using FreeSurfer (CorTechs Labs, Inc., Charlestown, MA). Reported coordinates were converted to Talairach space (Talairach and Tournoux 1988) using Matthew Brett's scripts (MRC Cognition and Brain Sciences Unit, Cambridge, England; http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach).

# Results

## Experiment I: Playing against a human opponent

**Behavioral data—**Analyses of the Likert scale responses showed that subjects were engaged in the game and winning was important to them, e.g. for the statement "I wanted to win the game very much" the mean score ±SD was 4.30±0.95 ($t(18)=6.00$, $p<0.0001$). We examined players' strategy by asking if the players selected their chips intentionally or randomly. The score for "I played the chip I thought was best for each particular game" was 4.68±0.75 ($t(18)=9.80$, $p<0.0001$). Lastly, we checked whether subjects were involved in mentalization during the game. Because the use of an older, shorter, debriefing form, only 13 of the 19 players were asked to respond to the mentalization statements. Ten of the 13 subjects rated these statements as four or five (e.g. "I usually chose the next chip to play according to my opponent's last moves": 3.85±0.69, $t(12)=4.43$, $p=0.001$ and "I took my opponent's last moves into account before deciding which chip to play next":3.70±0.95, $t(12)=2.64$, $p=0.02$, respectively). In response to an open-ended question about the influence of the opponent's moves on the players' decisions during the games, all 19 participants indicated they were *trying to predict their opponent's next move* and to act accordingly. These results indicate that all participants were actively involved in mentalization processes during the games.

To evaluate the motivational aspect of the game (related to gains and losses) we assessed the players' appraisal of the various outcomes. In response to open questions, players reported perceiving each challenge of a safe move as a gain and each challenged bluff as a loss. Similarly, non-challenged bluffs and safe moves were perceived as gains and losses, respectively, although no actual gain or loss was given. However, on the Likert scale the response to overt gains and relative gains ("I felt glad when…") were significant; while the reactions to overt loss ("I felt angry when…") were not (gain: 4.42±0.96, relative gain: 4.30±1.25; loss: 2.58 ±1.4, relative loss: 2.15±1.2). Importantly, results from paired *t*-tests showed no significant differences between overt and relative gains or overt and relative losses ($t(18)=0.26$ and $t(18)=1.45$, $p>0.1$, respectively). These results suggest that the primary game motivator was gains rather than losses and that participants did not find overt gains and losses to be more emotionally salient than the relative events.

An analysis of participants' choices while playing the Domino game in the scanner showed that subjects chose equally between playing a non-matching chip (risk taking and 'bluffing' their opponent) and playing a matching chip (playing safe). The average player's Non-Match index was 0.5±0.1. A one-way ANOVA of this index by time (4 min) revealed a significant main effect of time ($F(1,3)=16.5$, $p<0.0001$) such that players chose to 'bluff' their opponent more towards the end of a game than at the beginning (Fig. 2, orange line).

**Functional brain data—**To examine the network involved in implicit 'on-line' mentalization regardless of the outcome, we calculated the main effect of Mentalization (i.e. Opponent's Response of 'Show' vs. 'No-Show') with a whole-brain analysis. Table 1 and Fig. 3a present the resulting SPM analysis for peak activation corrected for multiple comparisons using a false discovery rate (FDR) with significance level represented by $q_{FDR}$ (Genovese et al. 2002). The resulting network ($q_{FDR}<0.005$) composed of bilateral temporal poles (TP); temporoparietal junctions (TPJ), including superior temporal sulcus and middle temporal gyrus (STS and MTG); medial prefrontal cortex (MPFC), including the paracingulate cortex; midbrain and cuneus; right ventrolateral prefrontal cortex (VLPFC); fusiform gyrus (FG); and postcentral gyrus.

To assess activation related to the value of outcome (i.e. gains vs. losses) within these regions independent of mentalization processes, we examined the main effect of Motivation (Fig. 1b). We first masked this analysis with the map of the Mentalization main effect, such that the

regions sensitive to the latter were used as ROIs for the former. No effect of motivation was found in the regions that showed a significant main effect of Mentalization at the same threshold ($q_{FDR}<0.005$). However, at a lower threshold ($q_{FDR}<0.05$), bilateral TPJ and right FG and VLPFC showed a significant Motivation effect (Left TPJ: $F(1,18)=5.99$, $p=0.025$; Right TPJ: $F(1,18)=9.28$, $p=0.007$; FG: $F(1,18)=6.40$, $p=0.021$; VLPFC: $F(1,18)=17.38$, $p=0.001$; see Fig. 3), such that activation during gains exceeded losses.

Importantly, a follow-up whole-brain analysis showed main effect of Motivation ($q_{FDR}<0.005$) in bilateral nucleus accumbens (NAcc); orbitofrontal cortex (OFC); postcentral and middle frontal gyri; precuneus; and cerebellum (Table 1 and Fig. 3b). These regions were activated more during (overt and relative) gains than (overt and relative) losses. Figure 3b shows 3D views of activation maps for the mentalization and motivation effects, including overlapping regions. It is evident that only the TP and the MPFC showed an effect for the Mentalization that was independent of the Motivation effect, while NAcc and OFC only demonstrated a unique effect for Motivation. These results reflect a regional dissociation for processing mentalization and motivation.

Finally, no brain region showed an interaction between mentalization and motivation effects (even at $p<0.05$, uncorrected). It is important to note that the interaction in this ANOVA is related to the player's choice in the current move, since we are looking at non-match vs. match choices ((No-Show No-Match + Show No-Match) > (Show Match + No-Show Match)). These results imply that after learning what their opponent's response is, players are no longer concerned about their choice per se, but about the outcome (i.e. gain vs. loss) as determined by their choice and the opponent's response and about the opponent's state of mind as represented by his/her response (i.e. mentalization).

## Experiment II: Playing against human and computer

**Behavioral data—**Post-scan debriefing data showed that players were engaged in the games and similarly excited to win when playing both against human and computer opponents. Specifically, the score for “I wanted to win the game very much” was 4.72±0.46, $t(17)=15.85$, $p<0.0001$ for Human Opponent and 4.83±0.38, $t(17)=20.28$, $p<0.0001$ for Computer Opponent. Paired *t*-test was not significant ($p=0.16$). In addition, players selected their chips similarly (intentionally and not randomly) in the two conditions (e.g. “I played the chip I thought was best for each particular game”: Human: 4.28±0.83, $t(17)=6.50$, $p<0.0001$; Computer: 4.33±0.90, $t(17)=6.23$, $p<0.0001$; paired *t*-test was not significant: $p=0.72$).

Participants reported that they considered their opponent's moves when playing against both opponents (“I usually chose the next chip to play according to my opponent's last moves”: Human: 3.61±1.04, $t(17)=2.50$, $p=0.02$; Computer: 3.56±1.29, $t(17)=1.82$, $p=0.08$; and “I took my opponent's last moves into account before deciding which chip to play next”: Human: 4.00 ±0.91, $t(17)=4.67$, $p<0.0001$; Computer: 3.89±1.02, $t(17)=3.69$, $p=0.002$; paired *t*-tests not significant: $p=0.85$, 0.63, respectively).

Analyzing participants' in-scanner game playing choices revealed no differences between Human and Computer opponents. On average, with both opponents, players chose equally between risky and safe moves (Non-Match index was 0.49±0.1 for both opponents). A 2×4 ANOVA (Opponent Type: Human/Computer × Time: 4 min) showed no significant main effect of Opponent Type ($F(1,17)<1$) but a significant main effect of Time ($F(1,17)=3.31$, $p=0.049$) for this index, similar to Experiment I (see Fig. 2). In addition, analyses of numbers of games won and lost to both opponents during the two runs of each revealed no significant differences between the two opponents and two runs (Wins: $F(1,17)<1$ and $F(1,17)=1.86$, $p=0.19$, respectively; Losses: $F(1,17)<1$ and $F(1,17)<1$).

**Functional brain data—**First, we replicated the results of Experiment I (Mentalization × Motivation ANOVA). The main effect of Mentalization for the human opponent data revealed that all regions that showed a significant main effect of Opponent's Response in experiment I also showed this effect in experiment II ($q_{FDR}<0.05$); these were bilateral TP, TPJ and MPFC, plus right VLPFC and FG (Fig. 4a).

Second, using the Mentalization by Opponent Type ANOVA, we determined which regions were activated more during human-opponent than computer-opponent runs. Regions showing significant activation of Human Show>No-Show *and* a main effect of Opponent Type were bilateral TPJ, TP, MPFC, and right FG (Fig. 4b). Activation in these regions during Human-Opponent runs was greater than for Computer-Opponent runs (see Fig. 4c; Table 2). Importantly, VLPFC showed no effect of Opponent Type (Fig. 4b–c and Table 2). However, this region showed a significant main effect of Mentalization similar to the other four brain regions, and planned contrasts of both human- and computer-opponent runs, activation during 'Show' were significantly greater than during 'No-Show' events (see Table 2).

## Discussion

Using an a-priori defined interval during a competitive two-person fMRI game, we delineated brain areas that were selective to implicit 'on-line' mentalization, to motivation or to both processes. The 'on-line' mentalization-related areas were defined as areas showing the combined effect of Opponent's Response/Mentalization in experiment I ('Show' vs. 'No-Show') and Opponent Type in experiment II (Human vs. Computer) regardless of the Motivation effect. This 'on-line' implicit mentalization network includes bilateral dorsal MPFC, TPJ, TP, and right FG and closely corresponds to the network previously reported to underlie explicit 'off-line' mentalization. However, only two regions out of this network were independent of the motivation factor effect, the MPFC and the TP, suggesting their unique role in mentalization.

## The mentalization-selective network

Our finding that dorsal MPFC is selective to mentalization vs. motivation processing during interpersonal competitive interactions agrees with previous 'on-line' and 'off-line' mentalization studies (Frith and Frith 2003; Amodio and Frith 2006). It has been suggested that the role of this region in social cognition is to integrate information regarding values of actions and outcomes (Amodio and Frith 2006). However, our result challenges the view of the MPFC as also involved in processing reward value (Delgado et al. 2000; Bush et al. 2002; Pochon et al. 2002; Knutson et al. 2003). Importantly, none of these studies included ongoing interaction with another human player. Our study suggests that during an interpersonal competitive game entailing both motivation (i.e. outcome value of gains and losses) and mentalization (i.e. value of action in relation to the other), the MPFC is more involved in the latter. In this context it is important to mention a recent hypothesis that the dorso-ventral MPFC is part of a neuronal network subserving processes related to self-reflection and projection (Buckner and Caroll 2007; Schmitz and Johnson 2007). Interestingly, Buckner and Caroll include mentalization as part of the self-appraisal neural network while Schmitz and Johnson discuss reward processing in this context. By using the Domino paradigm we were able to distinguish between these two self-relevant processes during a social competitive situation, again demonstrating the involvement of the dorsal MPFC in mentalization more than in reward-related motivational processes.

Our finding that the TP's activation is selective to mentalization vs. motivation is also consistent with other reports (for review see Frith and Frith 2003). These authors suggest that the TP's

role in mentalization relates to retrieving relevant semantic information using adjustable behavioral scripts that are applied within the mentalization framework.

As mentioned above, the mentalization-related areas were identified as areas that showed main effects of Mentalization, or Opponent's Response, and Opponent Type regardless of a Motivation main effect. Regarding the Mentalization main effect, all the areas identified had greater activation during the 'Show' than the 'No-Show' events. We suggest that while both conditions entail mentalization, the 'Show' events require more information processing since the player uses new information about the opponent to update his/her representation of the opponent's state of mind (i.e. both strategy and potential next moves). This is due to the fact that the opponent obtains new information about the player during these events (e.g. if the player bluffed or played fairly). From the perspective of the player, the opponent might use this information to change his/her strategy. Thus, the player has to take more information into account when updating his/her representation of the opponent, requiring greater levels of mentalization (Fig. 1c). In relation to the Opponent Type main effect, as we expected the mentalization areas showed greater activation when players played against what they believed was a human agent than a random, non-strategic, non-human computerized one, consistent with our understanding of the mentalization processes as crucial to social, inter-personal interactions (Adolphs 2003). In contrast, the VLPFC activation was sensitive to the Mentalization/Opponent's Response main effect (as with the mentalization areas) as well as to the Motivation effect, but was not human-selective and thus not integral to the human mentalization network. Its activation pattern is consistent with an assumed lateral PFC role in 'top-down' cognitive control to support goal- and rule-directed behavior flexibly (Miller and Cohen 2001;Bunge 2004). This regulation might be required in processing values of outcomes while playing against either human or computer.

Notably, the few neuroimaging studies to date that directly investigated implicit 'on-line' mentalization using cooperative (McCabe et al. 2001; Rilling et al. 2004) and competitive (Gallagher et al. 2002) games, demonstrated activation consistent with the 'off-line' mentalization circuit in MPFC (McCabe et al. 2001; Gallagher et al. 2002; Rilling et al. 2004) and posterior STS (Rilling et al. 2004 only), but not in TP. The different results are likely due to prior analysis of an interval not primarily concerned with mentalization (McCabe et al. 2001), using PET with its lower temporal resolution (Gallagher et al. 2002), and single-shot games that limited players' ability to assess an opponents' state of mind and predict their behavior based on prior observations (Rilling et al. 2004). Other neuroimaging studies of social interactions focused on aspects differing from mentalization as reflected during cooperative games (e.g. Rilling et al. 2002; Sanfey et al. 2003; Delgado et al. 2005; King-Casas et al. 2005).

## The motivation-selective network

Brain areas that showed motivation-related activation during the Response to Outcome interval showed greater neural response to gains than losses. Theses areas included, among other areas, the bilateral nucleus accumbens and OFC, which are known to be integral to reward processing (e.g. Schultz 2001; O'Doherty 2004). In the context of the Domino game, one can view gains as rewards, since the player is disposing of an extra domino chip (overt gains) and avoiding losses (relative gains), which brings him/her closer to potentially winning the game (that entails monetary reward). The fact that the activations in reward-related brain areas were specific to motivation processes (across mentalization levels), emphasizes the validity of our ANOVA analysis (Mentalization by Motivation) in the Response to Outcome interval.

## The mentalization and motivation networks overlap

Two of the regions that are usually indicated in explicit ‘off-line’ mentalization and are sensitive to implicit ‘on-line’ mentalization processes in our study are also sensitive to motivation (i.e. TPJ and FG). TPJ is consistently activated by ‘off-line’ mentalization tasks, and its role in mentalization is believed to be related to prediction of others' actions (Frith and Frith 2003). On the contrary, our finding of right FG activation during mentalization agrees with only a few prior mentalization studies. These used social attribution tasks where subjects explicitly attributed mental states to animated shapes (Castelli et al. 2000; Schultz et al. 2003) and reported activation overlapping with the fusiform face area (Kanwisher et al. 1997). Although a role for face-selective FG in social behavior is likely (Haxby et al. 2000; Rotshtein et al. 2005) our and other mentalization studies demonstrate its activation even when no face stimuli were presented, suggesting its general involvement in social behavior-related semantic memory processes, of which facial perception represents only one aspect (Schultz et al. 2003; Schultz 2005).

Given the TPJ and FG activation in both mentalization and motivation, we hypothesize that these regions help prepare for next step in the game by integrating information related to mentalization and motivation aspects of interpersonal behavior. These regions' integrative role was previously demonstrated for sensory information. In recent fMRI studies where participants identified objects presented across different modalities, Beauchamp et al. (2004) showed that posterior STS/MTG and ventral temporal regions integrated different information types concerning complex objects, both within and across sensory modalities. Such integration also occurred for audiovisual presentation of speech in STS, with implications for social behavior (Wright et al. 2003). Our results suggest, for the first time, a more general role of these temporal regions in integrating non-sensory, higher-order social information.

Alternative interpretation of our results in the TPJ and FG involves these regions' role in perceiving social and motivational stimuli. While FG is known to be involved in face perception (Haxby et al. 2000; Rotshtein et al. 2005) STS may function in perceiving and encoding stimulus importance (although this was documented only in relation to auditory stimuli in animals, e.g. Rutkowski and Weinberger (2005)). We cannot provide conclusive evidence to favor integrative vs. perceptual theories.

## Study limitations

There were several limitations to our study. First, we measured mentalization-related activation using differences in activation between the ‘Show’ and ‘No-Show’ events. However, since the overt losses and gains given in the ‘Show’ events are more salient than the relative outcomes of the ‘No-Show’, it is possible that the former evoked more emotion-motivation laden responses than the latter. The more salient events might entail various processes such as emotion regulation (Lieberman 2007; Ochsner and Gross 2005) and self-reflection (Lane et al. 1997; for review Lieberman (2007)) and even increased arousal. Noticeably, the post-scan debriefing indicated that the potential difference in affective load between outcome events is not supported by a difference in the players' awareness levels of different emotions attached to outcomes. Ideally, future studies should include control ‘No-Show’ events that match the ‘Show’ events on their emotional loads. Second, as implied earlier, it is most likely that players were thinking about their opponent (i.e. mentalizing) throughout the game and not only during the Response to Outcome interval. We assume mentalization is utilized the most during this interval since players receive new information about their opponent and use this information to update their mental representation of him/her. However, we did not explicitly ask players either during or after playing the game when they were thinking about their opponent the most. Third, counter to our expectation, players did not distinguish between a human and a computer-

opponent with regard to trying to predict their next move based on opponent's last move, at least based on the players' responses to the Likert scale statements. The latter might be insufficiently sensitive to capture subtle differences. It is possible that players refer to both human and computer as equivalent opponents and did not believe they were playing against a human or a computer-opponent. Ongoing research will elucidate this. Finally, future research into motivational processes of social behavior will benefit from objective autonomic measurement (such as heart rate or galvanic skin response) of players' responses to gains and losses.

## Summary

Overall, using a unique fMRI paradigm characterized by inter-personal competitive interactions, we demonstrated that a selective brain network, previously shown to subserve explicit ‘off-line’ mentalization, also underlies implicit ‘on-line’ mentalization. Within this network, only the MPFC and TP were mentalization-selective when compared to motivational processes of gains and losses, highlighting the involvement of these regions in attributing mental states to other humans during interpersonal, competitive interactions. Conversely, the TPJ and FG activations were related to mentalization and motivation suggesting a more general role in social behavior. We emphasize that the Domino paradigm involves complex social interaction that simulates real-life situations and therefore encompasses several cognitive and affective processes, of which mentalization and motivation are two major components, respectively. Clearly, other processes are also involved in the game, including working memory, anticipation of outcome (Kahn et al. 2002) and decision-making. Since implicit ‘on-line’ mentalization is key in this paradigm and its network is not well established, this is the primary focus of the current report.

## Acknowledgments

The authors would like to thank Drs. Kristen McKiernan Miller, Michael Stevens and Brian Knutson for their helpful comments on earlier versions of the manuscript. This work was partially supported by a Hartford Hospital grant (PI: M. Assaf).

## References

Adolphs R. Cognitive neuroscience of human social behaviour. Nature Reviews. Neuroscience 2003;4:165–178. doi:10.1038/nrn1056.

Amodio D, Frith C. Meeting of minds: the medial frontal cortex and social cognition. Nature Reviews. Neuroscience 2006;7:268–277. doi:10.1038/nrn1884.

Beauchamp MS, Lee KE, Argall BD, Martin A. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 2004;41:809–823. doi:10.1016/S0896-6273(04)00070-4. [PubMed: 15003179]

Buckner RL, Carroll DC. Self-projection and the brain. Trends in Cognitive Sciences 2007;11:49–57. doi:10.1016/j.tics.2006.11.004. [PubMed: 17188554]

Bunge SA. How we use rules to select actions: a review of evidence from cognitive neuroscience. Cognitive, Affective & Behavioral Neuroscience 2004;4:564–579.

Bush G, Vogt BA, Holmes J, Dale AM, Greve D, Jenike MA, et al. Dorsal anterior cingulate cortex: a role in reward-based decision making. Proceedings of the National Academy of Sciences of the United States of America 2002;99:523–528. doi:10.1073/pnas.012470999. [PubMed: 11756669]

Castelli F, Happe F, Frith U, Frith C. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. NeuroImage 2000;12:314–325. doi:10.1006/nimg.2000.0612. [PubMed: 10944414]

Dale AM, Buckner RL. Selective averaging of rapidly presented individual trials using fMRI. Human Brain Mapping 1997;5:329–340. doi:10.1002/(SICI)1097-0193(1997)5:5<329::AIDHBM1>3.0.CO;2-5.

Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience 2005;8:1611–1618. doi:10.1038/nn1575.

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA. Tracking the hemodynamic responses to reward and punishment in the striatum. Journal of Neurophysiology 2000;84:3072–3077. [PubMed: 11110834]

Dolan RJ. Emotion, cognition, and behavior. Science 2002;298:1191–1194. doi:10.1126/science.1076358. [PubMed: 12424363]

First, MB.; Spitzer, RL.; Gibbon, M.; Williams, JBW. Structured Clinical Interview for DSM-IV-TR axis I disorders, research version, patient edition. (SCID-I/P). Biometrics Research, New York State Psychiatric Institute; New York: 2002.

Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. Human Brain Mapping 1995;3:165–189. doi:10.1002/hbm.460030303.

Frith U, Frith CD. Development and neurophysiology of mentalizing. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 2003;358:459–473. doi:10.1098/rstb.2002.1218.

Gallagher HL, Jack AI, Roepstorff A, Frith CD. Imaging the intentional stance in a competitive game. NeuroImage 2002;16:814–821. doi:10.1006/nimg.2002.1117. [PubMed: 12169265]

Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 2002;15:870–878. doi:10.1006/nimg.2001.1037. [PubMed: 11906227]

Haxby JV, Hoffman EA, Gobbini MI. The distributed human neural system for face perception. Trends in Cognitive Sciences 2000;4:223–233. doi:10.1016/S1364-6613(00)01482-0. [PubMed: 10827445]

Kahn I, Yeshurun Y, Rotshtein P, Fried I, Ben-Bashat D, Hendler T. The role of the amygdala in signaling prospective outcome of choice. Neuron 2002;33:983–994. doi:10.1016/S0896-6273(02)00626-8. [PubMed: 11906703]

Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience 1997;17:4302–4311. [PubMed: 9151747]

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. Getting to know you: reputation and trust in a two-person economic exchange. Science 2005;308:78–83. doi:10.1126/science.1108062. [PubMed: 15802598]

Knutson B, Fong GW, Bennett SM, Adams CM, Hommer D. A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. NeuroImage 2003;18:263–272. doi:10.1016/S1053-8119(02)00057-5. [PubMed: 12595181]

Lane RD, Fink GR, Chau PML, Dolan RJ. Neural activation during selective attention to subjective emotional responses. Neuroreport 1997;8:3969–3972. doi:10.1097/00001756-199712220-00024. [PubMed: 9462476]

Lieberman MD. Social cognitive neuroscience: a review of core processes. Annual Review of Psychology 2007;58:259–289. doi:10.1146/annurev.psych.58.110405.085654.

McCabe K, Houser D, Ryan L, Smith V, Trouard T. A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Sciences of the United States of America 2001;98:11832–11835. doi:10.1073/pnas.211415698. [PubMed: 11562505]

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 2001;24:167–202. doi:10.1146/annurev.neuro.24.1.167.

Ochsner KN, Gross JJ. The cognitive control of emotion. Trends in Cognitive Sciences 2005;9:242–249. doi:10.1016/j.tics.2005.03.010. [PubMed: 15866151]

O'Doherty JP. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. Current Opinion in Neurobiology 2004;14:769–776. doi:10.1016/j.conb.2004.10.016. [PubMed: 15582382]

Phan KL, Wager T, Taylor SF, Liberzon I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage 2002;16:331–348. doi:10.1006/nimg.2002.1087. [PubMed: 12030820]

Pochon JB, Levy R, Fossati P, Lehericy S, Poline JB, Pillon B, et al. The neural system that bridges reward and cognition in humans: an fMRI study. Proceedings of the National Academy of Sciences

of the United States of America 2002;99:5669–5674. doi:10.1073/pnas.082111099. [PubMed: 11960021]

Ressler N. Rewards and punishments, goal-directed behavior and consciousness. Neuroscience and Biobehavioral Reviews 2004;28:27–39. doi:10.1016/j.neubiorev.2003.10.003. [PubMed: 15036931]

Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C. A neural basis for social cooperation. Neuron 2002;35:395–405. doi:10.1016/S0896-6273(02)00755-9. [PubMed: 12160756]

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. The neural correlates of theory of mind within interpersonal interactions. NeuroImage 2004;22:1694–1703. doi:10.1016/j.neuroimage.2004.04.015. [PubMed: 15275925]

Rotshtein P, Henson RN, Treves A, Driver J, Dolan RJ. Morphing Marilyn into Maggie dissociates physical and identity face representations in the brain. Nature Neuroscience 2005;8:107–113. doi:10.1038/nn1370.

Rutkowski RG, Weinberger NM. Encoding of learned importance of sound by magnitude of representational area in primary auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 2005;102:13664–13669. doi:10.1073/pnas.0506838102. [PubMed: 16174754]

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. The neural basis of economic decision-making in the ultimatum game. Science 2003;300:1755–1758. doi:10.1126/science.1082976. [PubMed: 12805551]

Schultz RT. Developmental deficits in social perception in autism: the role of the amygdala and fusiform face area. International Journal of Developmental Neuroscience 2005;23:125–141. doi:10.1016/j.ijdevneu.2004.12.012. [PubMed: 15749240]

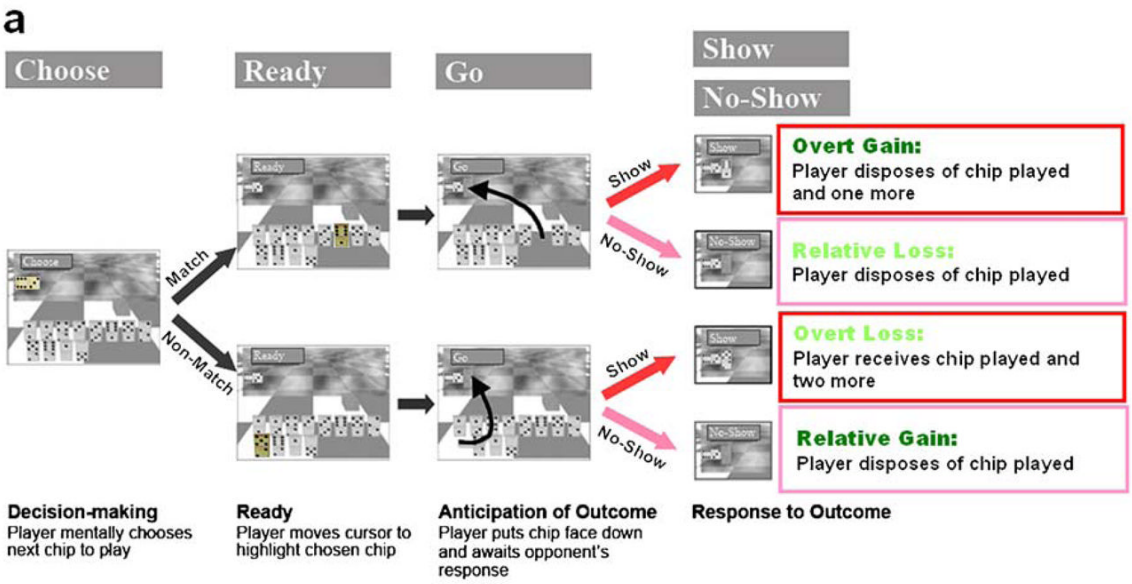
Schultz RT, Grelotti DJ, Klin A, Kleinman J, Van der Gaag C, Marois R, et al. The role of the fusiform face area in social cognition: implications for the pathobiology of autism. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 2003;358:415–427. doi:10.1098/rstb.2002.1208.

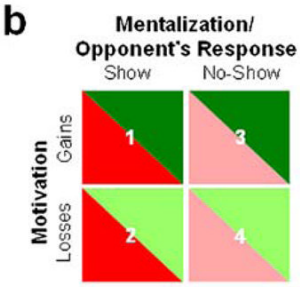
Schmitz TW, Johnson SC. Relevance to self: a brief review and framework of neural systems underlying appraisal. Neuroscience and Biobehavioral Reviews 2007;31:585–596. doi:10.1016/j.neubiorev.2006.12.003. [PubMed: 17418416]

Schultz W. Reward signaling by dopamine neurons. The Neuroscientist 2001;7:293–302. [PubMed: 11488395]

Talairach, J.; Tournoux, P. A co-planar stereotaxic atlas of a human brain. Thieme; New York: 1988.

Wright TM, Pelphrey KA, Allison T, McKeown MJ, McCarthy G. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cerebral Cortex (New York, N.Y.) 2003;13:1034–1043. doi:10.1093/cercor/13.10.1034.




c

| Motivation: Player's gains or losses | Mentalization: Player processing of Opponent's… |
|---|---|
| 1. Overt gain | Last response and his/her knowledge about player's choice |
| 2. Overt loss | Last response and his/her knowledge about player's choice |
| 3. Relative gain | Last response |
| 4. Relative loss | Last response |

**Fig. 1.**

Domino game paradigm. Panel **a** depicts the domino game sequence and corresponding consequences. In the beginning of each game round the player (participant scanned) must decide what chip he/she will play next (decision-making interval) and move the cursor to this chip when instructed (ready interval). The chip can either match the opponent's chip (i.e. have one of the numbers match those on the opponent's chip, 6:3 in this example; upper panel, 6:1) or not (lower panel, 5:2). After placing the selected chip face down next to the opponent's chip, he/she awaits the opponent's response (anticipation of outcome interval). The opponent can either challenge the player's choice (‘show’; *red arrows*) or not (‘no-show’; *pink arrows*). Based on the player's choice and the opponent's response there are four possible consequences for each round (response to outcome interval): show match (*overt gain*); no-show match (*relative loss*, as the player could have been rewarded if challenged); show non-match (*overt loss*) and no-show non-match (*relative gain*, as the player could have been punished if challenged). Note: *colors of boxes/wording* correspond to panel b color schemes. Panel **b** outlines the ANOVA design for the response to outcome interval analyses. *Red tones* correspond to mentalization/opponent's response effect and *green* to motivation. Panel **c** describes the suggested mentalization and motivation mechanism of each event. The opponent's chip and samples of matching and non-matching chips are highlighted for demonstration purposes only. In the actual scan the game board and all chips are colored. Also all chips are the same size and color

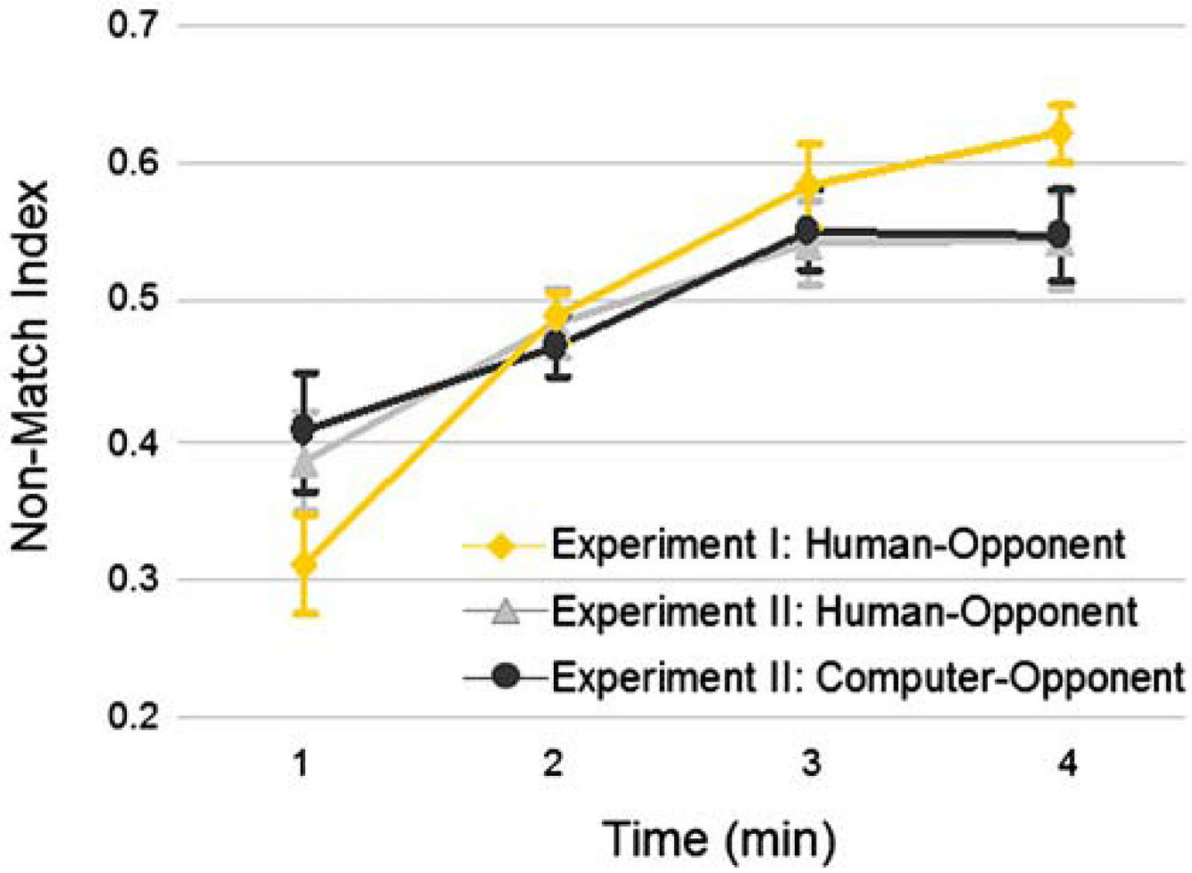



**Fig. 2.**

Players' choices as a function of time during experiments I and II. Non-match index (number of non-match choices divided by the number of non-match and match choices) for games played against human in experiment I (*orange*), and against human (*light gray*) and computer (*dark gray*) opponent in experiment II are plotted for each minute of the game (averaged for all games for all subjects). No significant differences were found between the two groups (experiment I and II) when playing against human and between human and computer opponents in experiment II

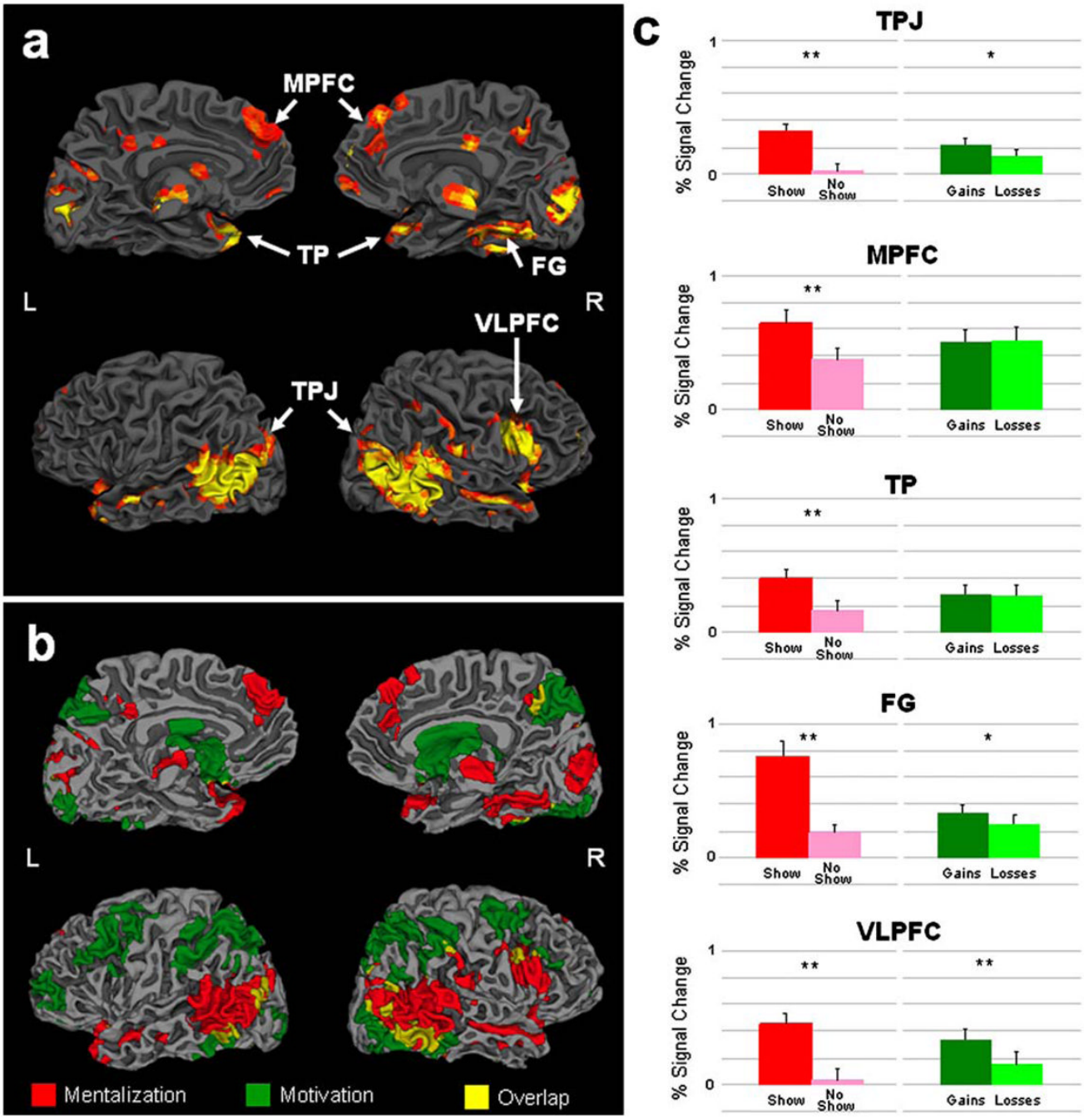



**Fig. 3.**

Main effects of mentalization and motivation (experiment I). Panel a depicts the activation map of random effects ANOVA showing brain regions with a significant main effect of mentalization/opponent's response ($n$=19, q(FDR)<0.005). These regions included the temporoparietal junction (TPJ), temporal pole (TP), medial prefrontal cortex (MPFC), ventrolateral prefrontal cortex (VLPFC) and fusiform gyrus (FG). Other brain regions presented in this map are midbrain, cuneus, postcentral gyrus and posterior cingulate cortex. Panel **b** shows the brain networks activated during mentalization (*red*), motivation (*green*) and overlapping regions (*yellow*). As shown in the graphs (panel **c**), all the mentalization regions had significantly more activation during 'show' (*red bars*) than 'no-show' (*pink bars*) events

regardless of losses and gains. Some of the brain regions (TPJ, FG and VLPFC) also demonstrated significant main effect of motivation, such that activations related to gains (*dark green*) were higher than activation during losses (*light green*). (Note: since bilateral regions, such as TPJ and TP, showed the same patterns of activations, graphs presented here are averaged percent signal change of right and left activations.) $^{**}p<0.001$; $^{*}p<0.05$; L=left; R=right hemisphere

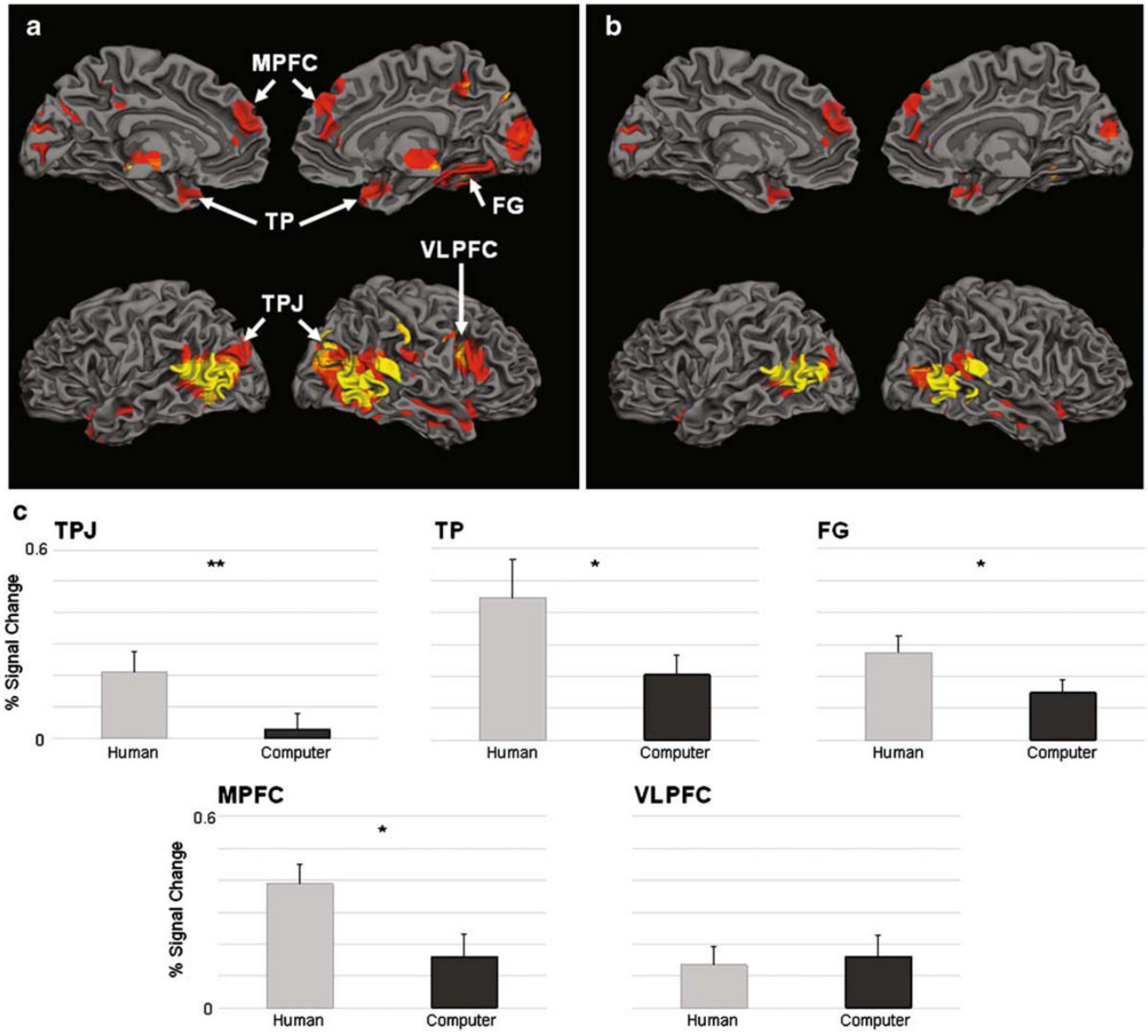



**Fig. 4.**

Mentalization network (experiment II). Activation map of random effects ANOVA showing brain regions with significant main effect of mentalization during the human-opponent runs ($n$=18, q(FDR)<0.05). Notably, these brain regions are almost identical to the brain regions activated by the same analysis in study I (see Fig. 3a.) Panel **b** depicts brain regions showing a significant main effect of mentalization during the human opponent runs ($n$=18, qFDR<0.05) masked with regions showing a significant main effect of opponent type ($p$<0.05) such that activations during human runs are greater than during computer runs. Note that right VLPFC does not appear in this map. Panel **c** shows the response to outcome interval activations during the human-opponent games (*light gray bars*) and the computer-opponent games (*dark gray bars*). All regions but the VLPFC showed significant main effect of opponent type (i.e. human vs. computer) such that signal was greater during the human-opponent compared to the computer-opponent games. (Note: since bilateral regions, such as TPJ and TP, showed the same patterns of activations, graphs presented here are averaged percent signal change of right

and left activations.) TPJ, temporoparietal junction; TP, temporal pole, FG, fusiform gyrus; VLPFC, ventrolateral prefrontal cortex; MPFC, medial prefrontal cortex; **$p$<0.005; *$p$<0.05; $k$>50

## Table 1

Experiment I: brain regions activated during the response to outcome interval

| Anatomic location of maximum activation | Talairach coordinates | | | Z Score |
|---|---|---|---|---|
| | x | y | z | |
| Main effect of mentalization | | | | |
| *$q_{(FDR)}$<0.005; Show>no-show* | | | | |
| L TPJ (MTG/STS) | −42 | −72 | 15 | 6.78 |
| R TPJ (MTG/STS) | 45 | −75 | 18 | 6.15 |
| L TP | −42 | −7 | −14 | 5.47 |
| R TP | 48 | 0 | −18 | 5.01 |
| R FG | 30 | −53 | −7 | 5.25 |
| R VLPFC | 48 | 19 | 24 | 6.06 |
| MPFC | 6 | 48 | 31 | 4.27 |
| L midbrain | −6 | −26 | −4 | 5.04 |
| R midbrain | 6 | −26 | −1 | 5.73 |
| Cuneus | 0 | −87 | 10 | 6.69 |
| R postcentral Gyrus (BA 1) | 65 | −16 | 26 | 4.39 |
| Main effect of motivation | | | | |
| *$q_{(FDR)}$<0.005; Gains>losses* | | | | |
| L NAcc | −15 | 8 | −8 | 6.18 |
| R NAcc | 12 | 6 | −8 | 6.29 |
| L OFC | −18 | 40 | −15 | 3.74 |
| R OFC | 24 | 51 | −13 | 4.04 |
| L MFG (BA 6) | −36 | 5 | 49 | 4.23 |
| R MFG (BA 6) | 39 | 11 | 46 | 4.39 |
| L postcentral gyrus (BA 2/40) | −56 | −35 | 49 | 4.52 |
| R postcentral gyrus (BA 2) | 53 | −27 | 48 | 4.31 |
| L precuneus | −15 | −64 | 61 | 4.42 |
| R precuneus | 12 | −59 | 50 | 4.11 |
| L cerebellum | −36 | −65 | −25 | 4.23 |
| R cerebellum | 30 | −68 | 22 | 4.77 |

*L* left, *R* right, *TPJ* temporoparietal junction, *TP* temporal pole, *FG* fusiform gyrus, *VLPFC* ventrolateral prefrontal cortex, *MPFC* medial prefrontal cortex, *BA* Brodmann region, *NAcc* nucleus accumbens, *OFC* orbitofrontal cortex, *MFG* middle frontal gyrus, *FDR* false discovery rate

**Table 2**

Experiment II: A 2×2 ANOVA (opponent type by mentalization) of the response to outcome interval BOLD signal in ROIs identified in experiment I

| 1. Anatomic location | 2. Talairach coordinates | | | 3. Mentalization S>NS | | 4. *Opponent type H>C* | | 5. *Human S>NS* | | 6. Computer S>NS | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | x | y | z | *F(1,17)* | *p* | *F(1,17)* | *p* | *t(17)* | *p* | *t(17)* | *P* |
| R TPJ[a] | 53 | −61 | 6 | 58.4 | .0001 | *7.7* | *.01* | *7.3* | *.0001* | 6.0 | .0001 |
| L TPJ | −45 | −60 | 14 | 35.3 | .0001 | *8.0* | *.01* | *5.8* | *.0001* | 4.9 | .0001 |
| | −62 | −55 | 5 | 18.7 | .0001 | *10.4* | *.005* | *3.2* | *.005* | 4.2 | .001 |
| R TP[a] | 27 | 8 | −21 | 15.0 | .001 | *6.6* | *.02* | *3.8* | *.002* | 2.3 | .03 |
| L TP | −29 | 11 | −21 | 19.6 | .0001 | *5.2* | *.03* | *3.8* | *.001* | 3.1 | .006 |
| | −21 | 5 | −20 | 7.1 | .01 | *9.1* | *.008* | *2.4* | *0.03* | 1.8 | n.s |
| R FG[a] | 36 | −49 | 16 | 14.0 | .002 | *6.3* | *.02* | *5.7* | *.0001* | 2.2 | .04 |
| MPFC | −9 | 47 | 14 | 11.1 | .004 | *5.7* | *.03* | *3.3* | *.004* | 2.3 | .04 |
| | −6 | 47 | 6 | 15.4 | .001 | *7.6* | *.01* | *2.4* | *.03* | 2.7 | .01 |
| R VLPFC[a] | 48 | 7 | 27 | 23.9 | .0001 | *0.2* | *n.s* | *5.0* | *.0001* | 3.6 | .002 |

Main effects are reported in columns 3–4 and planned contrasts in columns 5–6. Coordinates reported are the points of maximal activation of the contrast human show>no-show (*column 5, in italics*) and/or main effect of opponent type (human>computer, *column 4, in italics*) within the same clusters of activation. Note that the VLPFC is the only region that did not have a significant main effect of opponent type, indicating no differences in activation between human and computer runs

*L* left, *R* right, *TPJ* temporoparietal junction, *TP* temporal pole, *FG* fusiform gyrus, *VLPFC* ventrolateral prefrontal cortex, *MPFC* medial prefrontal cortex, *ACC* anterior cingulate cortex, *S* 'show' opponent's response, *NS* 'no-show' opponent's response, *H* human-opponent, *C* computer-opponent

[a]Maximal activation in both contrasts overlap