The Importance of the Lateral Prefrontal Cortex for Strategic Decision Making in the Prisoner's Dilemma

Alexander Soutschek^{1,2} · Marian Sauter^{1,3} · Torsten Schubert¹

Published online: 4 August 2015
© Psychonomic Society, Inc. 2015

Abstract Previous functional imaging studies investigating the neural basis of strategic decision making in the prisoner's dilemma reported a correlation between cooperative behavior and dorsolateral prefrontal cortex (DLPFC) activity; however, the precise function of the DLPFC in establishing cooperation remains unclear so far. The present study investigated the causal role of the DLPFC in an iterative prisoner's dilemma game with transcranial magnetic stimulation (TMS). We discovered that disrupting the DLPFC with TMS decreased cooperation rates in comparison to control conditions, with this effect being most pronounced when the partner had defected previously. Thus, the current results suggest that the DLPFC contributes to strategic decision making in the prisoner's dilemma game.

Keywords Prisoner's dilemma · Game theory · Strategic decision making · Dorsolateral prefrontal cortex · Cooperation

Strategic decision making represents an important precondition for successful human behavior. Game theory allows analyzing strategic decision making in situations in which the outcomes depend on the choices of two actors. In game theory, a player's strategy defines which option the player chooses,

depending on the second player's decision (Camerer, 2003). A paradigm often used to analyze strategic decision making is the prisoner's dilemma game (PDG). In the PDG, player A is confronted with the decision to either cooperate with (C) or defect from (D) a second player, B. This results in four different outcome types: both players cooperate (CC), Player A cooperates and Player B defects (CD), Player A defects and Player B cooperates (DC), and both players defect (DD). The payoffs are arranged such that Player A can maximize her outcome if she defects when Player B cooperates (unreciprocated defection; DC) compared to mutual cooperation (CC). Importantly, also in the case that Player B defects, Player A's payoff is higher if she defects, too (mutual defection; DD), compared to if she cooperates (unreciprocated cooperation; CD). Consequently, if the PDG is played as a one-shot game, then a rational agent should always defect, regardless of her opponent's decision. If the opponent cooperates, then a player is better off defecting compared to cooperating because unreciprocated defection (DC) leads to higher outcomes than mutual cooperation (CC). The same applies to a defecting opponent because Player A's outcome is higher in case of mutual defection (DD) compared with unreciprocated cooperation (CD). Consequently, DD is the Nash-equilibrium in the PDG (i.e., the only outcome from which each player could only do worse by changing her strategy while the other keeps her strategy). However, the dilemma in the PDG arises from the fact that, while defection is the dominant strategy in the PDG, both players would increase their payoffs in the case of mutual cooperation compared to mutual defection (in concepts of game theory, CC is pareto-superior to DD). This dilemma can be solved when playing the PDG in an iterative instead of a one-shot version because in an iterative PDG a player can adjust her decisions flexibly in response to her opponents' behavior. The best strategy in an iterative PDG versus an unknown opponent is to play tit-for-tat (i.e., to begin

✉ Alexander Soutschek
alexander.soutschek@econ.uzh.ch

¹ Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

² Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, Blumlisalpstr. 10, 8006 Zurich, Switzerland

³ Department of Psychology, Ludwig-Maximilians-Universität, Munich, Germany

with cooperation and then to choose the behavior the opponent had shown in the previous round; Axelrod & Hamilton, 1981). Thus, if the opponent cooperates, then a player should start cooperating too in order to maximize the payoff because mutual cooperation is the best long-term solution of the dilemma for both players.

Functional imaging studies investigating the neural correlates of decision making in the PDG reported a correlation between cooperative decisions and activity in, among other regions, the dorsolateral prefrontal cortex (DLPFC; Emonds, Declerck, Boone, Vandervliet, & Parizel, 2012; J. K. Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004b). Because the DLPFC has been related to strategic decision making (Knoch, Pascual-Leone, Meyer, Treyer, & Fehr, 2006; Kuo, Sjöström, Chen, Wang, & Huang, 2009; Steinbeis, Bernhardt, & Singer, 2012), Emonds et al. (2012) speculated that the DLPFC might be involved in strategically resolving conflicts between self-interest and collective interests in the PDG (see also J. K. Rilling et al., 2004b). However, the precise function of the DLPFC in the PDG remains unclear in this study because this interpretation is based only on the finding of increased DLPFC activity in a PDG compared to a stag hunt game. The stag hunt game is a cooperation game with a similar payoff structure as the PDG, but with no incentive to free-ride because payoffs are lower for DC than for CC outcomes, such that the stag hunt game requires less complex strategic decision-making processes than the PDG. In addition, due to the correlative nature of the fMRI signal, it is not even clear whether the observed DLPFC activity is causally related to decision making in the PDG at all or whether it only represents a passive by-product of the higher processing demands in the PDG than the stag hunt game.

The goal of the present study was to test the causal role of the DLPFC for strategic decision making in the PDG with offline transcranial magnetic stimulation (TMS). Participants played the PDG against a computer partner that used a tit-for-tat strategy. This allowed us to test how participants strategically adjust their choices to the partner's behavior in an experimentally controlled way: if the opponent in a PDG plays a tit-for-tat strategy, then the optimal strategy is not to immediately reciprocate a defective choice of the partner but to continue with cooperation (a so-called generous tit-for-tat strategy; Milinski & Wedekind, 1998). This is because reciprocating defection increases the probability of mutual defection outcomes, while cooperating with a tit-for-tat playing partner is likely to lead to mutual cooperation. From a cognitive perspective, using a generous tit-for-tat strategy may require one to suppress the impulse to reciprocate defection in order to punish a defecting opponent. Such impulse control processes are thought to be related to prefrontal cortex functioning (Kerns et al., 2004; Miller & Cohen, 2001), in line with the functional imaging results of enhanced DLPFC activity during the PDG (Emonds et al., 2012; J. K. Rilling et al., 2004b).

Thus, we expect that disrupting DLPFC functioning with TMS inhibits participants' ability to suppress the impulse to reciprocate a defecting opponent, impairing the use of generous tit-for-tat strategies in the current paradigm. This, in turn, should generally decrease cooperation rates following DLPFC TMS compared with control conditions.

We note that playing the PDG against a human partner may involve partly different processes compared with playing it against a computer partner: for example, it has been argued that mutual cooperation with other humans is experienced as rewarding, as is indicated by activity in the striatum during mutual cooperation (J. Rilling et al., 2002). However, social variants of the PDG also require adjusting one's own decisions to the partner's choices (and, thus, strategic decision making in a game-theory sense of the word). When examining the use of strategies in the PDG, the computer version of the PDG had the advantage that it was possible to define an optimal strategy that leads to stable mutual cooperation. Therefore, the use of a computer instead of a human opponent allowed us to examine how disrupting DLPFC functioning affects strategic decision making in the PDG in an experimentally controlled way.

Method

Participants Fifty-six right-handed volunteers who were recruited at the Humboldt-Universität participated in the study and were randomly assigned to one of four experimental conditions (for details, see below). Fifteen participants were tested in the left PFC TMS condition (9 female; $M_{\text{age}} = 27.1$, $SD_{\text{age}} = 6.2$), 13 in the right PFC TMS condition (8 female; $M_{\text{age}} = 25.9$, $SD_{\text{age}} = 2.9$), 14 in the control TMS condition (10 female; $M_{\text{age}} = 25.2$, $SD_{\text{age}} = 4.5$), and 14 in the no TMS condition (8 female; $M_{\text{age}} = 23.6$, $SD_{\text{age}} = 2.3$). All participants were naïve to the purpose of the study, reported to suffer of no psychiatric disorders, gave informed consent, and were paid 10 euros per hour plus a performance-dependent bonus (see below).

Experimental design and procedure Participants played an iterated prisoner's dilemma game in which they had to choose between cooperating and defecting with a virtual opponent. As a cover story, we told participants that the computer mirrored the behavior of real human opponents in the PDG (however, we did not tell them which algorithm was used). In fact, the computer played a tit-for-tat strategy and cooperated or defected with a probability of 80% if the player had cooperated or defected, respectively, in the previous trial (J. Rilling et al., 2002; J. K. Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004a). This probabilistic algorithm was used in order to induce free-riding behavior of the computer after mutual cooperation. The payoff participants obtained after each trial depended on their own and the computer's

decision: if a player cooperated, she received 2 cents if the computer also cooperated (mutual cooperation; CC) and lost 1 cent if the computer defected (unreciprocated cooperation; CD). If a player defected, then she obtained 3 cents or 0 cents, depending on whether the computer cooperated (unreciprocated defection; DC) or defected (mutual defection; DD), respectively. Thus, a player obtained a higher payoff if she defected, regardless of the interaction partner's choice; at the same time, payoffs were higher for both players in the case of mutual cooperation compared to mutual defection. In order to increase participants' engagement in the task, they received the cumulated outcomes in addition to their basic payment.

At the start of each PDG trial, participants were asked whether they would like to cooperate or defect (1,500 ms). During this interval, participants should indicate their decision by pressing the right or left control key with their right or left index finger on a QWERTZ keyboard. Half of the participants indicated cooperation by pressing the right control key and defection by pressing the left control key; for the other half of participants, response mapping was reversed. After an interval of 1,000 ms, participants received a visual feedback on their own and their opponent's outcome in the current trial (1,500 ms). The next trial started after an intertrial interval of 1,000 ms (see Fig. 1).

Participants performed four blocks in which they played the PDG in 12 consecutive trials. We instructed subjects that in every block they would play the prisoner's dilemma against a new virtual opponent.

TMS The experiment included four between-subject TMS conditions (left PFC vs. right PFC vs. control site vs. no TMS). Because Emonds et al. (2012) reported bilateral PFC activity while participants played the PDG, we decided to

stimulate both the left and the right PFC in the current study. In addition to the PFC, we stimulated participants at the vertex as an active control site in order to control for unspecific TMS effects. In the no TMS condition, participants played the PDG without having been stimulated before. TMS protocol was approved by the local ethics committee.

Before participants played the PDG in the left PFC, right PFC, or control site condition, we applied 480 impulses of rTMS (frequency: 1 Hz, 110% resting motor threshold) using a Focal Bipulse 8-coil with an internal diameter of 5 cm (eXimia TMS; Nexstim Ltd, Helsinki, Finland); TMS parameters were within the recommended guidelines (Rossi, Hallett, Rossini, Pascual-Leone, & Safety of the T. M. S. Consensus Group, 2009). Because the stimulation duration was 8 minutes, the PDG took 4 minutes after the stimulation to guarantee the suppression of excitability of the stimulated brain region during task performance (Robertson, Theoret, & Pascual-Leone, 2003). Participants were stimulated either at the right PFC, the left PFC, or at the vertex as control site, with the coil being held tangentially to the participant's head. The right and left PFC were targeted at the electrode positions F4 and F3, respectively; this procedure has been shown to provide a reliable stimulation of the DLPFC (Baumgartner, Knoch, Hotz, Eisenegger, & Fehr, 2011; Herwig, Satrapi, & Schonfeldt-Lecuona, 2003). Electrode position Pz served as control site (Soutschek, Taylor, Muller, & Schubert, 2013). We confirmed coil positioning in some participants (seven in the right PFC, five in the left PFC, and five in the vertex condition) by using optical tracking via eXimia NBS to determine Montreal Neurological Institute (MNI) coordinates of the stimulation sites based on the individual participants' structural magnetic resonance images. Mean MNI coordinates confirmed the coil to be placed over the right DLPFC ($x = 44$,

Fig. 1 (A) Example trial of the prisoner's dilemma game (PDG): At the beginning of a block we presented the initials of a fictive interaction partner. In each single trial of a block participants were first asked whether they would like to cooperate or to defect. Then, participants received a feedback regarding their own and their interaction partner's outcome. Afterward, the next trial started, in which participants again had

to decide whether or not to cooperate with the same interaction partner. (B) Payoff matrix for the PDG. The subject's choices (player) are listed aside rows, the computer's choices are listed atop columns. Amounts in bold indicate the payoff for the player, amounts in parentheses indicate the payoff for the computer

$y = 25$, $z = 48$), left DLPFC ($x = -43$, $y = 22$, $z = 49$), and paracentral lobule/vertex ($x = 1$, $y = -30$, $z = 78$) in the corresponding experimental conditions (see Fig. 2).

TMS may be experienced as more aversive or irritating at anterior (e.g., the PFC) compared to posterior brain regions like the vertex (Robertson et al., 2003). Although this thought to impact task performance in online rather than in offline experimental designs, we controlled for this potential confound by asking the participants for their perceived aversiveness of the stimulation (on a 5-point Likert scale) after the experiment.

Statistical analysis We analyzed cooperation rates (i.e., number of cooperation decisions / number of cooperation and defection decisions) in the PDG. For tests of significance, we calculated ANOVAs and planned comparisons with a significance threshold of 5%.

Results

Aversiveness ratings We compared participants' reported aversiveness of TMS by a univariate ANOVA including the factor TMS site (right PFC vs. left PFC vs. control). We found that the perceived aversiveness did not differ between the TMS conditions, $F(2, 39) < 1$, $p > .48$. Thus, this analysis provided no evidence that participants perceived TMS over the lateral PFC as more aversive than over the vertex.

Cooperation rates To test the hypothesis that the DLPFC is causally involved in establishing cooperation, we conducted a between-subjects ANOVA including the factor TMS (left PFC vs. right PFC vs. control site vs. no TMS) on cooperation rates (see Fig. 3).

The significant main effect of TMS indicated that cooperation rates differed between the TMS conditions, $F(3, 52) = 4.52$, $p < .01$, $\eta_p^2 = 0.207$: in line with our hypothesis, cooperation rates were lower in the left PFC (52%) and right PFC (58%) relative to the control TMS (70%) condition, both $t_s > 2.33$, $p_s < .05$, as well as relative to the no TMS (71%) condition, both $t_s > 2.23$, $p_s < .05$. No significant difference occurred between the left PFC and the right PFC conditions, $t < 1$, $p > .39$, or between the no TMS and the control TMS conditions, $t < 1$, $p > .92$. Thus, disrupting the left as well as

Fig. 2 Mean stimulation sites (indicated by a black circle) for (A) left lateral prefrontal cortex, (B) right lateral prefrontal cortex, and (C) vertex

Fig. 3 Cooperation rates (in %), plotted as a function of TMS condition (left PFC TMS vs. right PFC TMS vs. control TMS vs. no TMS). Error bars indicate the standard error of the mean

the right PFC resulted in decreased cooperation rates compared to the control TMS and the no TMS conditions.

Given that disrupting PFC activity resulted in lower cooperation rates, we asked whether the TMS effects on cooperation depended on the outcome (i.e., CC, CD, DC, or DD) in the preceding trial. In other words, did TMS modulate participants' response to the computer partner's choices? To test that, we analyzed cooperation rates after each outcome type with a 4×4 mixed-measures ANOVA including the factors TMS and Previous outcome type (post-CC vs. post-CD vs. post-DC vs. post-DD). Three participants had always cooperated and never defected, so that it was not possible to compute cooperation rates after DC and DD trials for them; therefore, the data of these subjects were excluded from the following analysis. We found a significant main effect of Previous outcome type, $F(3, 147) = 21.38$, $p < .001$, $\eta_p^2 = 0.304$: Participants cooperated more frequently following CC (75%) relative to following CD (49%), DC (39%), and DD (48%) outcomes, all $p_s < .001$, whereas no significant differences occurred between post-CD, post-DC, and post-DD trials, $p_s > .05$. A similar result pattern has been observed in previous studies using the PDG (J. Rilling et al., 2002; J. K. Rilling et al., 2012; J. K. Rilling et al., 2004b): since mutual cooperation is the best long-term outcome for both players, subjects have an incentive to continue with cooperation following CC outcomes as long as the partner cooperates, too. We replicated the finding of a significant main effect of TMS, $F(3, 49) = 4.82$, $p < .01$, $\eta_p^2 = 0.228$, and, importantly, also the TMS $\times$ Previous outcome type interaction yielded a significant result, $F(9, 147) = 1.96$, $p < .05$, $\eta_p^2 = 0.107$: cooperation rates were decreased after CD outcomes both in the left PFC (30%) and the right PFC (39%) conditions compared to the control TMS (67%) condition, $t_s > 2.96$, $p_s < .01$, as well as compared to the no TMS (61%) condition, $t_s > 2.16$, $p_s < .05$. However, cooperation rates after CD outcomes did not differ significantly between the left and the right PFC condition, $t <$

1, nor between the control TMS and the no TMS condition, $t < 1$. Moreover, we found no significant differences between the TMS conditions after CC, DC, or DD outcomes, $t_s < 1.97$, $p_s > .05$ (see Fig. 4 and Table 1). Thus, after disrupting PFC functioning, participants defected more often following CD outcomes (unreciprocated cooperation) compared to control site TMS and compared to the no TMS condition.

Given the significant between-subjects effects of TMS on subjects' choices, one may wonder whether this effect is caused by different cooperation-rate levels of the computer in the TMS groups. Theoretically, it is possible that by chance the computer partner defected more often in the PFC TMS groups than in the control groups, leading to the observed higher defection rates in the PFC TMS groups (given that subjects played tit-for-tat). To rule out this potential confound, we tested whether the cooperation rates of the computer partner differed between the TMS conditions. For that purpose, we computed a 4 (TMS) $\times$ 4 (Previous outcome type) mixed-measures ANOVA on the computer's cooperation rates. This ANOVA yielded a main effect of Previous outcome type, $F(3, 147) = 271.88$, $p < .001$, $\eta_p^2 = 0.847$, indicating that the computer cooperated more often following CC (80%) and CD (83%) trials compared with DC (23%) and DD (20%) trials. This is because the tit-for-tat playing computer algorithm cooperated with a probability of 80% when the subject had cooperated before and with a probability of only 20% when the subject had defected in the preceding trial. Importantly, neither the effect of TMS, $p > .98$, nor the TMS $\times$ Previous outcome type interaction, $p > .52$, yielded a significant trend. This shows that the computer's cooperation rates did not differ between TMS groups. Thus, the TMS effects on subjects' choices cannot be explained by differences in cooperation rates of the computer partner between the TMS groups.

Fig. 4 Cooperation rates (in %), plotted as a function of TMS condition (left PFC TMS vs. right PFC TMS vs. control TMS vs. no TMS) and outcome type of the previous trial (post-CC vs. post-CD vs. post-DC vs. post-DD trials). Error bars indicate the standard error of the mean. *Note.* Asterisks illustrate all significant differences between TMS conditions ($p < .05$). CC = mutual cooperation, CD = unreciprocated cooperation, DC = unreciprocated defection, DD = mutual defection

Table 1 Cooperation rates (in %) as a function of TMS (left PFC TMS vs. right PFC TMS vs. control TMS vs. no TMS) and outcome type of the previous trial (post-CC vs. post-CD vs. post-DC vs. post-DD). Standard errors of mean are in brackets.

	Left PFC TMS	Right PFC TMS	Control TMS	No TMS
post-CC	64 (10)	85 (4)	78 (5)	76 (7)
post-CD	30 (5)	39 (8)	67 (5)	61 (7)
post-DC	34 (7)	41 (8)	33 (8)	49 (5)
post-DD	44 (7)	43 (8)	41 (7)	60 (8)

Discussion

The present study investigated the role of the lateral PFC in strategic decision making in the PDG. We found that TMS of the left and right DLPFC resulted in decreased cooperation rates compared to a control site and a no TMS condition. This suggests that the PFC has a causal role in establishing successful cooperation. Importantly, our results indicate that the TMS effects on subjects' willingness to cooperate depends on the computer's behavior because PFC TMS decreased cooperation rates mainly after previous CD trials (i.e., when the player had cooperated and the interaction partner had defected). In line with previous studies, subjects showed higher cooperation rates following CC compared with CD, DC, and DD outcomes (J. Rilling et al., 2002; J. K. Rilling et al., 2004a). This is because mutual cooperation is the best long-term solution to the PDG, such that rational subjects should cooperate as long as the partner reciprocates cooperation. However, also continuing with cooperation even when the partner free-rides in the preceding round represents often a more successful strategy than playing strictly tit-for-tat and is referred to as "generous tit-for-tat strategy" (Milinski & Wedekind, 1998). Cooperation following CD trials is likely to lead to mutual cooperation in the next trial under the condition that the partner plays a tit-for-tat strategy (which was the case in the present study) because the partner will start cooperating, too. In contrast, switching to a defective strategy after CD trials increases the probability of mutual defection because a tit-for-tat playing interaction partner will then, in turn, start defecting, too. Thus, our data suggest that PFC TMS impaired subjects' ability to use a generous tit-for-tat strategy because disrupting DLPFC activity decreased cooperation rates following CD outcomes relative to control conditions. From a cognitive perspective, this finding can be explained by the well-supported role of the DLPFC in impulse control (Kerns et al., 2004; Miller & Cohen, 2001). A precondition of playing a generous tit-for-tat strategy may be the ability to suppress the impulse to reciprocate defection following CD outcomes. In the current study, DLPFC TMS may have inhibited subjects' impulse control, resulting in the observed lower cooperation rates after CD outcomes. This

interpretation is in line with previous results ascribing the DLPFC a role in inhibiting costly punishments of unfair behavior in the dictator game (Brune et al., 2012). In addition, our results extend previous findings on the role of the DLPFC in strategic decision making (Baumgartner et al., 2011; Knoch et al., 2006; Ruff, Ugazio, & Fehr, 2013) by providing evidence for a DLPFC involvement in iterative games in which choices depend on the partner's preceding behavior.

Two potential counterarguments against our interpretation need to be addressed: First, TMS at the prefrontal cortex is often perceived as more aversive than TMS at posterior brain regions. However, this is supposed to affect task performance mainly in online and not – as in the current study – in offline TMS designs (Robertson et al., 2003). Furthermore, our questionnaire data provided no evidence for the assumption that participants perceived TMS at the PFC as more painful than at the vertex. Finally, we found no significant differences between the effects of control TMS and no TMS on behavior, suggesting that TMS per se did not affect performance in the PDG. Therefore, it is reasonable to assume that the observed effects following PFC TMS were caused by disrupting PFC activity with TMS, and not by TMS side-effects.

Second, one may criticize that the version of the PDG we used did not measure behavior in a real “social situation” because participants played the PDG against a computer. However, strategic decision making is required in both computer-based and human variants of the PDG. It is important to note that DLPFC activity was found during computer-based as well as social PDGs (Emonds et al., 2012; J. Rilling et al., 2002), suggesting the DLPFC to have a crucial role in strategic decision making also when the PDG is played against human partners.

The hypothesized role of the DLPFC in the PDG is consistent with a previous fMRI study relating the DLPFC to strategic decision making in situations in which a player's outcome depends also on the partner's choices (Kuo et al., 2009). In the study of Kuo et al., participants played interaction games with a clear formal game-theoretic solution. The authors found increased bilateral DLPFC activity during this task relative to a control condition, consistent with the data of the current study in which TMS of both left and right DLPFC impaired decision making in the PDG. However, although the present findings provide evidence for a causal role of the DLPFC for strategic decision making in the PDG, it remains an open question how exactly the DLPFC contributes to implementing strategic behavior. It has been shown that the use of strategies in the PDG is impaired by concurrent working memory demands (Milinski & Wedekind, 1998). Since the DLPFC is thought to relate to working memory processes (Braver et al., 1997), DLPFC TMS may have impaired strategic decision making in the current study by disrupting the working memory functioning. Alternatively, it is even possible that the left and right DLPFC have dissociable functions in

the PDG, despite the similar behavioral effects of left and right DLPFC TMS. Previous studies on preference-based decision making suggest a hemispheric specialization, with the right DLPFC controlling norm-guided choices (Knoch et al., 2006; Ruff et al., 2013) and the left DLPFC controlling conflicts between immediate and delayed rewards (Figner et al., 2010). Both these functions may be important for decision making in the PDG and lead to similar behavioral effects: the right DLPFC may facilitate cooperative behavior because defecting violates social norms, whereas the left DLPFC fosters cooperation because it increases long-term outcome. Further studies are needed to clarify how exactly the DLPFC contributes to strategic decision making. Furthermore, we would like to note that, since both the left and right PFC was found to be activated in the PDG (Emonds et al., 2012), the contralateral PFC might, in theory, have compensated the inhibition of the stimulated hemisphere. However, the fact that unilateral PFC stimulation significantly affected cooperation rates suggests that the contralateral PFC could not (or at least not fully) compensate TMS effects.

The finding that the DLPFC is relevant for strategic decision making in interaction games may deepen our understanding of how social cooperation among nonrelatives is possible at all. Evolutionary biologists assume that the evolution of reciprocal cooperative behavior in a species presupposes the existence of two cognitive abilities (Axelrod & Hamilton, 1981; Trivers, 1971): first, individuals must be able to discriminate between cooperative conspecifics and free-riders. Second, a mechanism is required that weights long-term over immediate short-term benefits (Frank, 1988). While the first cognitive ability may be related to the temporoparietal junction, which was found to be active in mental-state attribution tasks (Frith & Frith, 2006; Saxe & Kanwisher, 2003; Saxe & Powell, 2006), the current study suggests that the DLPFC is involved in the strategic weighting of long-term over short-term outcomes. In order to successfully establish mutual cooperation, humans must risk short-term losses and cooperate even when the opponent had defected in the previous round. From the perspective of cognitive psychology, the ability to suppress the impulse to defect is part of the broader category of cognitive control processes. Consequently, the current study contributes to the existing literature showing that cognitive control processes are involved in decision making in game-theoretic paradigms (De Neys, Novitskiy, Geeraerts, Ramautar, & Wagemans, 2011; Knoch et al., 2006; Soutschek & Schubert, 2014). The discovery that the PFC, which evolved relatively late in evolutionary history, appears to be a precondition for mutual altruism is thus consistent with the idea that cognitive complexity among different species corresponds to different levels of cooperation (Brosnan, Salwiczek, & Bshary, 2010). Thus, the development of specific brain structures appears to be an important precondition for mutual altruism among nonrelatives.

Altogether, the current study sheds light on the neural basis of cooperative behavior. Establishing stable cooperative behavior presupposes the ability to strategically adopt one's own decisions to the opponent's behavior. Disrupting DLPFC activity decreased cooperation rates, indicating that the DLPFC is causally involved in strategic decision making in the PDG in order to establish mutual cooperation.

Competing interests The authors declare no competing financial interests.

References

- Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390–1396.
- Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C., & Fehr, E. (2011). Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nature Neuroscience*, 14(11), 1468–1474. doi:10.1038/nn.2933
- Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., & Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage*, 5(1), 49–62. doi:10.1006/nimg.1996.0247
- Brosnan, S. F., Salwiczek, L., & Bshary, R. (2010). The interplay of cognition and cooperation. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 365(1553), 2699–2710. doi:10.1098/rstb.2010.0154
- Brune, M., Scheele, D., Heinisch, C., Tas, C., Wischniewski, J., & Gunturkun, O. (2012). Empathy moderates the effect of repetitive transcranial magnetic stimulation of the right dorsolateral prefrontal cortex on costly punishment. *PLoS ONE*, 7(9), e44747. doi:10.1371/journal.pone.0044747
- Camerer, C. F. (2003). Behavioral game theory: Plausible formal models that predict accurately. *Behavioral and Brain Sciences*, 26(2), 157–158.
- De Neys, W., Novitskiy, N., Geeraerts, L., Ramautar, J., & Wagemans, J. (2011). Cognitive control and individual differences in economic ultimatum decision-making. *PLoS ONE*, 6(11), e27107. doi:10.1371/journal.pone.0027107
- Emonds, G., Declercq, C. H., Boone, C., Vandervliet, E. J. M., & Parizel, P. M. (2012). The cognitive demands on cooperation in social dilemmas: An fMRI study. *Social Neuroscience*, 7(5), 494–509.
- Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, 13(5), 538–539. doi:10.1038/nn.2516
- Frank, R. H. (1988). *Passions with reason: The strategic role of the emotions*. New York: Norton.
- Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4), 531–534. doi:10.1016/j.neuron.2006.05.001
- Hervig, U., Satrapi, P., & Schonfeldt-Lecuona, C. (2003). Using the international 10–20 EEG system for positioning of transcranial magnetic stimulation. *Brain Topography*, 16(2), 95–99.
- Kerns, J. G., Cohen, J. D., MacDonald, A. W., III, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, 303(5660), 1023–1026. doi:10.1126/science.1089910303/5660/1023
- Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, 314(5800), 829–832. doi:10.1126/science.1129156
- Kuo, W. J., Sjöström, T., Chen, Y. P., Wang, Y. H., & Huang, C. Y. (2009). Intuition and deliberation: Two systems for strategizing in the brain. *Science*, 324(5926), 519–522. doi:10.1126/science.1165598
- Milinski, M., & Wedekind, C. (1998). Working memory constrains human cooperation in the prisoner's dilemma. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23), 13755–13758.
- Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167–202. doi:10.1146/annurev.neuro.24.1.167
- Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., & Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, 35(2), 395–405.
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004a). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, 22(4), 1694–1703. doi:10.1016/j.neuroimage.2004.04.015
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004b). Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport*, 15(16), 2539–2543.
- Rilling, J. K., DeMarco, A. C., Hackett, P. D., Thompson, R., Ditzel, B., Patel, R., & Pagnoni, G. (2012). Effects of intranasal oxytocin and vasopressin on cooperative behavior and associated brain activity in men. *Psychoneuroendocrinology*, 37(4), 447–461. doi:10.1016/j.psyneuen.2011.07.013
- Robertson, E. M., Theoret, H., & Pascual-Leone, A. (2003). Studies in cognition: The problems solved and created by transcranial magnetic stimulation. *Journal of Cognitive Neuroscience*, 15(7), 948–960. doi:10.1162/089892903770007344
- Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A., & Safety of the T. M. S. Consensus Group. (2009). Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical Neurophysiology*, 120(12), 2008–2039. doi:10.1016/j.clinph.2009.08.016
- Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, 342(6157), 482–484. doi:10.1126/science.1241399
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in “theory of mind”. *NeuroImage*, 19(4), 1835–1842.
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699. doi:10.1111/j.1467-9280.2006.01768.x
- Soutschek, A., & Schubert, T. (2014). Dynamic adjustments of cognitive control during economic decision making. *Acta Psychologica (Amst)*, 152C, 42–46. doi:10.1016/j.actpsy.2014.07.011
- Soutschek, A., Taylor, P. C., Muller, H. J., & Schubert, T. (2013). Dissociable Networks Control Conflict during Perception and Response Selection: A transcranial magnetic stimulation study. *Journal of Neuroscience*, 33(13), 5647–5654. doi:10.1523/JNEUROSCI.4768-12.2013
- Steinbeis, N., Bernhardt, B. C., & Singer, T. (2012). Impulse control and underlying functions of the left DLPFC mediate age-related and age-independent individual differences in strategic social behavior. *Neuron*, 73(5), 1040–1051. doi:10.1016/j.neuron.2011.12.027
- Trivers, R. L. (1971). Evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1), 35–57. doi:10.1086/406755