Title: Manipulation of Pro-Sociality and Rule Following with Non-invasive Brain Stimulation

Running Title: Pro-Sociality and Rule Following

Authors: Jörg Gross[1,2,*,†], Franziska Emmerling[3,4,†], Alexander Vostroknutov[5], Alexander T. Sack[3]

Affiliations

[1]Institute of Psychology, Leiden University, The Netherlands.

[2]Center for Experimental Economics and Political Decision Making, University of Amsterdam, The Netherlands.

[3]Department of Cognitive Neuroscience, Maastricht University, The Netherlands.

[4]Department of Experimental Psychology, University of Oxford, United Kingdom.

[5]Center for Mind & Brain Sciences, University of Trento, Italy.

† These authors contributed equally to this work

[*] Corresponding author: Jörg Gross, mail@joerg-gross.net

Abstract

Decisions are often governed by rules on adequate social behaviour. Recent research suggests that the right lateral prefrontal cortex (rLPFC) is involved in the implementation of internal fairness rules (norms), by controlling the impulse to act selfishly. A drawback of these studies is that the assumed norms have to be deduced from behaviour and that norm-following and pro-sociality are indistinguishable. Here, we directly confronted participants with a rule that demanded to make advantageous or disadvantageous monetary allocations for themselves or another person. To disentangle its functional role in rule following and pro-sociality, we divergently manipulated the rLPFC by applying cathodal or anodal transcranial direct current stimulation (tDCS). Cathodal tDCS increased participants' rule-following, even of rules that demanded to lose money or hurt another person financially. In contrast, anodal tDCS led participants to specifically violate more often those rules that were at odds with what participants chose freely. Brain stimulation over the rLPFC thus did not simply increase or decrease selfishness. Instead, by disentangling rule following and pro-sociality, our results point to a broader role of the rLPFC in flexibly adapting social behaviour by integrating the costs and benefits of social rules in order to align decisions with internal goals.

Keywords: decision making, prosocial behaviour, obedience, lateral prefrontal cortex

Introduction

Rules play a vital role in human societies. Adhering to speeding limits, not littering, or customs like shaking hands help to organize and regulate everyday life. Rules often demand to restrict goal-directed behaviour. For example, waiting in front of a red traffic light or standing in a queue in the supermarket interferes with the internal goal to proceed towards one's destination, or to not waste more time than strictly necessary.

Likewise, in the social domain, norms about fairness, morality, or pro-sociality often demand to restrict selfishness. The right lateral prefrontal cortex (LPFC) has been causally linked to the implementation of pro-social norms[1-7]. For example, brain stimulation, both with transcranial magnetic stimulation (TMS)[3,8] and transcranial direct current stimulation (tDCS)[9] over the right LPFC led to higher acceptance rates of unfair offers in the Ultimatum Game. In this game, participants have to make the decision to accept or reject an offer from another participant about splitting a sum of money. In case of rejection, both participants earn nothing. Applying cathodal stimulation, believed to decrease excitability of neurons in the targeted brain region[10,11], increased the propensity to accept highly unequal and thus unfair offers. One interpretation of these findings, that has been put forward, is that participants under cathodal TMS and tDCS were less able to resist the economic temptation to accept low offers, since 'something is still better than nothing'[1-7,9,12-14]. Resonating with this interpretation, participants under cathodal TMS also made faster decisions[14], which was interpreted as a sign for a quick uncontrolled selfish impulse guiding decision-making. At the same time, anodal tDCS over the right LPFC, believed to increase excitability of neurons in the targeted brain region[10,11], led to more social

norm compliance[13]. From this perspective, the right LPFC exhibits executive control over the impulse to act selfishly and allows to align behaviour with norms and rules. A different functional role of the right LPFC in normative social decision making has been recently put forward by Buckholtz[15]. Rather than simply implementing 'impulse control', Buckholtz argues that the LPFC performs a value based cost-benefit analysis by weighing and integrating the cost and benefits of actions, rules, personal goals, past experience, and situational factors and frames. In line with this interpretation, the LPFC has been broadly associated with adaptive behaviour that enables humans to flexibly react to external stimuli in order to implement internal goals, rather than just follow fixed stimulus-response patterns[12,16-21] and integrates thought and action in the pursuit of these goals[18-25]. Further, while brain stimulation over the right LPFC shifted decisions towards more selfishness or more pro-sociality depending on the stimulation, it did not affect the underlying fairness perception across multiple studies[9,12-14]. This suggests that brain stimulation over the right LPFC led to a misalignment of thought and action. Further, Greene et al.[26] have shown that lying exhibits LPFC activity, while honesty did not and FeldmanHall et al.[27] found a positive correlation between LPFC activity and the extent of selfishness. Both findings are at odds with the selfish impulse control hypothesis of LPFC recruitment. Also, difficult moral dilemmas, that require to find a compromise between norms and welfare maximization, have been associated with greater LPFC activity[28], pointing to a value-based integrative function of the LPFC.

Here, we aim to experimentally disentangle these conflicting hypotheses on the role of the LPFC in social behaviour, using transcranial direct current stimulation (tDCS). In the experiment, participants repeatedly choose to maximize their payoff, the payoff of another person, or to act selfishly or pro-socially. In one part of the experiment, we

confront participants with a rule that demands which option to choose. The rule is sometimes aligned with what participants would have chosen without a rule (i.e. the rule coincides with their internal goals), is consequence neutral, or demands to choose an action that does not coincide with intrinsic behaviour. If the right LPFC is critically involved in suppressing selfish impulses, decreasing neural excitability of this brain area with cathodal stimulation should lead to more selfishness and rule violations when the rule demands to restrict selfish payoff maximization, while anodal stimulation should lead to more rule-following, even when the rule demands to restrict selfish payoff maximization or hurt another person. Whereas if the right LPFC plays a broader role in aligning behaviour with internal goals, we should see the opposite pattern: More rule-following under cathodal brain stimulation and less rule-following under anodal brain stimulation, in particular when the rule is at odds with internal goals.

Methods

Subjects. Participants were recruited from the subject pool of the Behavioral and Experimental Economics Lab (BEElab) at Maastricht University and were invited via e-mail. Experiments were conducted with the informed consent of 103 healthy adult subjects (mean age = 21.4 +/− 3.0, 56 female) who were free to withdraw from participation at any time. The study was approved by the local ethical committee of the Faculty of Psychology and Neuroscience, Maastricht University and all methods were performed in accordance with the relevant guidelines and regulations.

Experimental Procedure. Upon arriving at the lab, participants were seated in individual cubicles in front of a computer screen. Four to six participants completed the experiment simultaneously to ensure that they trusted their decisions to impact

another real human individual. In the experiment, participants had to decide repeatedly whether to drag a ball with the mouse to either the left or right side of the computer screen into a blue or orange box (Fig. 1).

Across three blocks, the decisions had real financial consequences either for the participant (‘me’-block), for another real but unknown person (‘other person’-block), or both, the participant and another person (‘me vs. other person’-block). For example, in a given trial of the ‘me’- or ‘other person’-block, dragging the ball to the left side would yield 10 cents, while dragging the ball to the right side would yield 0 cents (Fig. 1a) for the participant or the other person, respectively.

In each trial of the ‘me vs. other person’-block participants had two options to distribute a sum of money between themselves and the other person (so called mini dictator game, Fig. 1b). For example, dragging the ball to the left side would yield 10 cents for the participant but 0 cents for the other person, while dragging the ball to the right side would yield 5 cents to the participant and 5 cents for the other person.

In one part of the experiment, participants could freely choose to opt for the action they preferred (‘free’ part). In the other part, a simple and arbitrary rule was given to the participants (‘rule’ part). The rule was to always drag the ball either to the left or right side of the screen (counterbalanced across participants), regardless of the consequence (“The rule is to put each ball in the blue (orange) area”). Violating the rule would, thus, sometimes have real negative consequences (financial costs for oneself or the other person), no consequence (following the rule would yield the same outcome as violating the rule), or real positive consequences (financial benefits for oneself or the other person).

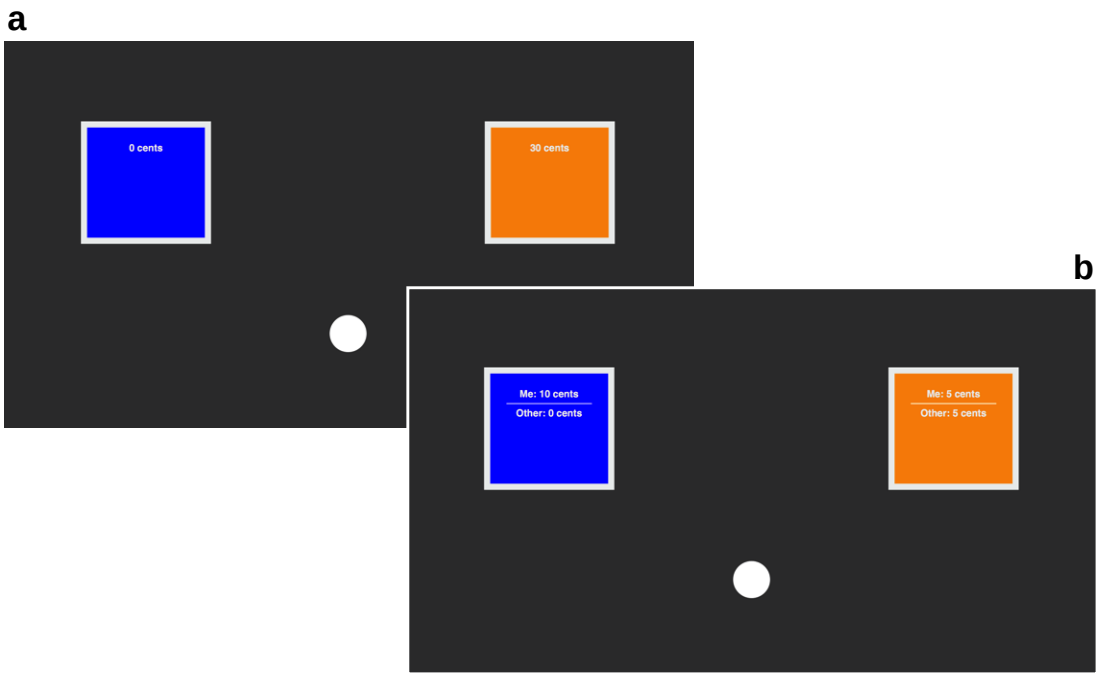

Fig. 1. Experimental setup. Participants repeatedly had to drag a ball in either the blue or orange area. In the 'me'-block and 'other person'-block, the decision had real financial consequences for the participant or another person, respectively, that changed across trials (see **a** for an example-trial). In the 'me vs. other person'-block, participants had to decide between allocating a sum of money between themselves and another person. The sum, as well as the allocation-choice changed across rounds (see **b** for an example-trial). In the 'free'-part, participants freely decided to drag the ball in either area, whereas in the 'rule'-part a simple rule was given to the participant. Half of the participants were instructed to always place the ball in the blue area, whereas the other half was instructed to always place it in the orange area.

After finishing the main task, participants made a series of fairness judgements in which a hypothetical person A distributed a sum of money between herself and another hypothetical person B. Participants rated each allocation on a fairness scale from -3 (completely unfair) to 3 (completely fair). After answering demographics questions, participants were finished. Participation took around 40 minutes. At the end of the experiment, the sum of money was paid to both the participant and the other person, according to the decisions the participant made in the experiment.

Decision consequences. Dragging the ball to either the blue or orange area could lead to the following consequences in euro-cents: [-30, -10, -5, 0, 5, 10, 30] in the 'me'-block and 'other person'-block. Table 1 shows all trial combinations each participant was confronted with in each block.

In each trial of the 'me vs. other person'-block participants had two options to distribute a sum of money between themselves and the other person (Fig. 1b). The sum of money could take the following values in euro-cents: [-30, -20, -10, 10, 20, 30], and could be distributed in the following way:

-30: [-30,0], [0,-30],

-20: [-20,0], [-10,-10], [0,-20],

-10: [-10,0], [-5,-5], [0,-10],

10: [10,0], [5,5], [0,10],

20: [20,0], [10,10], [0,20],

30: [30,0], [0,30]

Participants had to make choices for all possible combinations of these allocations (in total 28 trials for 'free'-part and 'rule'-part, respectively). Participants, thus, repeatedly faced the option to either distribute a sum of money more selfishly (taking a bigger share of the money), or more pro-socially (giving more or splitting the amount equally). When faced with a rule, the rule demanded to take the selfish option in half of the trials and the pro-social option in the other half of the trials.

Table 1.
'Me'-block and 'other person'-block trials. Combinations of outcomes for which participants had to make decisions for themselves ('me'-block) or another person ('other person'-block).

		yellow						
		-30	**-10**	**-5**	**0**	**5**	**10**	**30**
	-30				1			
	-10		1	1	1	1	1	
	-5		1	1	1	1	1	
blue	**0**	1	1	1	1	1	1	1
	5		1	1	1	1	1	
	10		1	1	1	1	1	
	30				1			

tDCS manipulation. To test the involvement of the right LPFC on rule adherence, we used a double-blind placebo-controlled tDCS design. Participants (n = 103) were randomly assigned to three tDCS conditions. TDCS is a non-invasive brain modulation technique using micro-currents believed to manipulate the resting membrane potential of neurons in the targeted brain region[29-31]. In a placebo/sham condition (n = 36) the skin sensations accompanying real stimulation can be mimicked, while no real stimulation is administered. Therefore, participants cannot differentiate the sham condition from real modulation. The right LPFC was manipulated with either cathodal (n = 32) or anodal tDCS (n = 35) over F4 as determined by the international 10/20-EEG system (Fig. 2). TDCS was applied by 5x7cm standard electrodes, at an intensity of 2mA, and with 30s ramping phases. Stimulation was applied during task execution for 30 minutes. No manipulation was induced in the sham condition. There was no significant difference in in the

distribution of sex (chi square test, $\chi^2(3) = 0.37$, p = 0.83) or age (one-way ANOVA, F(2) = 15.7, p = .0.43) across tDCS condition.

Fig. 2. tDCS setup. Anode or cathode placed over F4 (international 10/20 EEG system), reference over contralateral mastoid.

Results

First, we calculated the sum of money each subject accumulated in the 'me'-block and 'other person'-block in the 'free'-part and entered the data into a censored regression model fitted in R. Unsurprisingly, during sham, when participants were not restricted by a rule, they overwhelmingly chose options that would yield the most money for themselves or the other person. This was not significantly altered by the two active stimulation conditions, showing that participants were still able to identify and willing to choose the option that is most beneficial for themselves or the other person during cathodal and anodal tDCS (Fig. 3, Table 2 & 3).

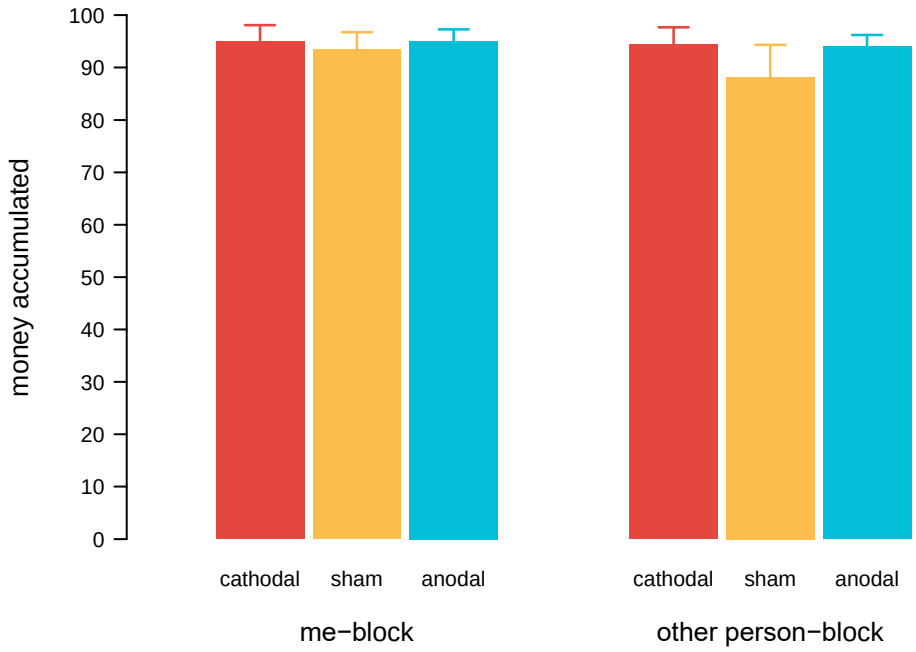

Fig. 3. Free decisions. Money accumulated for oneself (‘me’-block) or another person (‘other person’-block). 100% is the maximum that can be earned by always choosing the option that would yield more. Error bars show the standard errors of the mean.

Table 2.
‘Me’-trials (free decisions).
Censored regression predicting the earnings for oneself in the ‘free’-part, depending on the tDCS condition.

	Estimate	Std. error	t-value	p-value
Intercept (cathodal tDCS)	242.56	28.18	8.61	< 0.01
sham tDCS	-14.04	28.96	-0.49	0.63
anodal tDCS	-23.62	28.66	-0.82	0.41

Table 3.
‘Other person’-trials (free decisions).
Censored regression predicting the earnings for the other person in the ‘free’-part, depending on the tDCS condition.

	Estimate	Std. error	t-value	p-value
Intercept (cathodal tDCS)	238.81	28.70	8.32	< 0.01
sham tDCS	-16.90	31.13	-0.54	0.59
anodal tDCS	-6.71	31.45	-0.21	0.83

When a rule was in place, 36 participants followed the rule unconditionally independent of the tDCS condition (chi square test, $\chi^2(2) = 1.46$, p = .48). We, hence, focused on those subjects, who violated the rule depending on the decision consequences (for a model incorporating also unconditional rule followers, see Supplementary Material). Conditional rule followers followed the rule nearly without exception when it yielded positive consequences (Fig. 4). When adhering to the rule would have negative consequences and, thus, did not coincide with what participants would choose freely, rule adherence dropped from 98% to 44%, averaging across all brain stimulation conditions (Wilcoxon Signed Rank test, W = 6198, p < .001). However, participants under cathodal tDCS still followed the rule 52% of the time, while it was only followed 32% of the time under anodal tDCS (Fig. 4, Mann–Whitney U test, U = 630, p = .02). Decisions under cathodal and sham tDCS did not differ significantly (Mann–Whitney U test, U = 1021, p = .55).

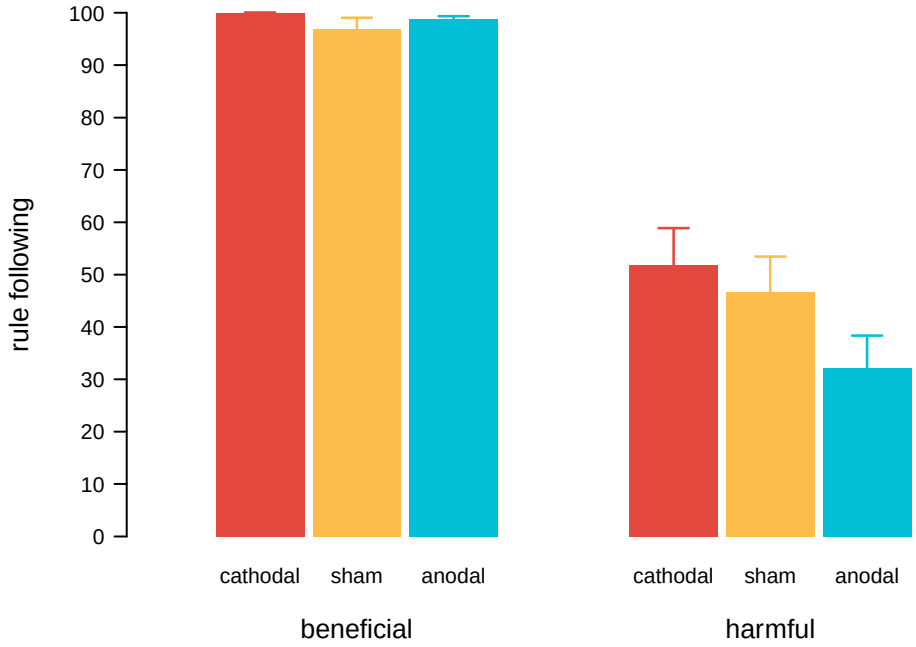

Fig. 4. Rule following. Average rule following in percentage when following the rule was either beneficial, i.e. demanding to take the option that would yield more money to oneself ('me'-block) or another person ('other person'-block), or when the rule was harmful, i.e. demanding to take the option that would hurt oneself or the other person financially. Error bars show the standard errors of the mean.

This pattern was consistent across 'me'-trials (Fig. 5a), and 'other person'-trials (Fig. 5b). We aggregated the number of times participants followed the rule for the 'me'-block and the 'other person'-block for each type of consequence: negative (meaning that the rule demanded to take the 'harmful' option), neutral (meaning that following the rule yielded the same outcome as violating the rule), and positive (meaning that the rule was to take the option that would benefit oneself or the other person financially).

Thus, we had three values for each participant in each block, measuring the average obedience to the rule (see Supplementary Information for additional analyses). To account for the dependencies within subjects, we fitted two (Bayesian) random intercept regression models using JAGS/R to the 'me'-trial and 'other person'-trial data, respectively. Non-informative Gaussian priors (m=0, sd=100) were used for

each predictor and non-informative uniform priors (range 0 to 100) for the error terms. We used three parallel chains. For every estimated coefficient, the potential scale reduction factor (Gelman and Rubin Diagnostic) was below 1.05, indicating good mixing of the three chains and, thus, high convergence. Regression tables reported below show estimated coefficients together with the 95% confidence interval (CI). Note that, since non-informative priors were used, a 95% CI that only contains negative or positive values can be interpreted as significant at a $p = .05$ two-sided threshold from a frequentist perspective. Fitting the models using restricted maximum likelihood (REML) as implemented in the lme4 package in R revealed similar estimates and the same statistical inferences.

While participants overwhelmingly followed the rule when it was beneficial to them or the other person across all three tDCS conditions, participants under cathodal compared to anodal tDCS followed more 'harmful' rules, both when the participant herself had to bear the consequences (random-effects regression, reduction in rule following for negative consequences, 95% CI: [-0.31,-0.01], Table 4) or when another person was affected (random-effects regression, reduction in rule following for negative consequences, 95% CI: [-0.40,-0.08], Table 5), with sham being in the middle, but not significantly different from the stimulation conditions. Note that the observed increase in rule-following under cathodal tDCS, when following the rule led to negative consequences to oneself, is at odds with the idea that cathodal brain stimulation leads to more payoff-maximization decisions due to uncontrolled selfish impulses.

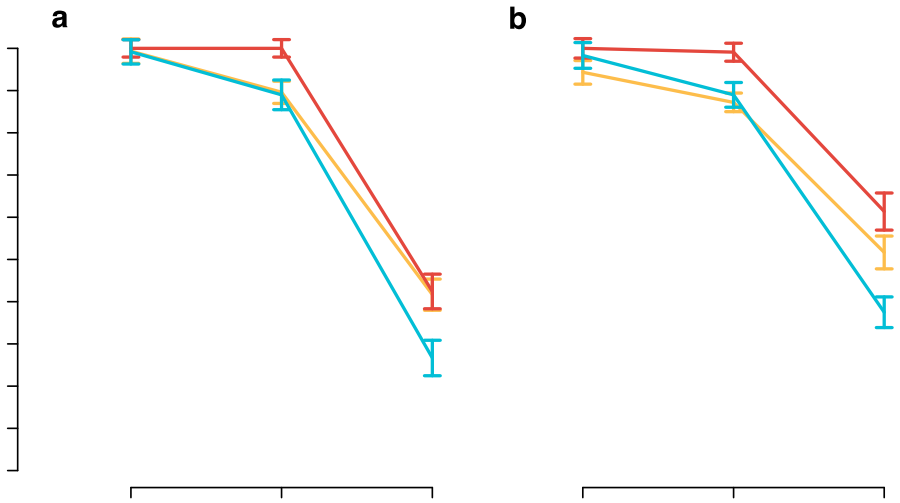

Fig. 5. Rule following across blocks depending on the consequence. Average rule following across tDCS conditions (red = cathodal tDCS, yellow = sham, blue = anodal tDCS) in percentage when the consequence of the rule was either negative (following the rule led to a worse outcome), neutral (following or violating the rule led to the same outcome), or positive (following the rule led to a better outcome), separately for decisions that either affected the participant herself (**a**) or another person (**b**). Error bars show the standard errors of the mean.

Table 4.
'Me'-trials (confronted with a rule).
Random intercept regression predicting the propensity to follow rules in 'me'-trials, depending on the tDCS condition.

	Estimate	SD	95% CI
Intercept (cathodal tDCS – negative)	0.42	0.053	[0.32, 0.523]
sham tDCS	-0.01	0.073	[-0.15, 0.13]
anodal tDCS	-0.16	0.077	[-0.31, -0.01]
neutral consequence	0.58	0.071	[0.44, 0.71]
positive consequence	0.58	0.071	[0.43, 0.71]
sham tDCS x neutral	-0.10	0.097	[-0.29, 0.09]
anodal tDCS x neutral	0.05	0.103	[-0.15, 0.25]
sham tDCS x positive	0.01	0.097	[-0.19, 0.19]
anodal tDCS x positive	0.15	0.103	[-0.06, 0.35]
random intercept variance	0.08	0.034	[0.00, 0.13]

Table 5.
'Other person'-trials (confronted with a rule).
Random intercept regression predicting the propensity to follow rules in 'other person'-trials, depending on the tDCS condition.

	Estimate	SD	95% CI
Intercept (cathodal tDCS – negative)	0.61	0.057	[0.51, 0.73]
sham tDCS	-0.10	0.078	[-0.25, 0.06]
anodal tDCS	-0.24	0.083	[-0.40, -0.08]
neutral consequence	0.38	0.069	[0.24, 0.51]
positive consequence	0.39	0.070	[0.25, 0.52]
sham tDCS x neutral	-0.02	0.094	[-0.21, 0.16]
anodal tDCS x neutral	0.14	0.100	[-0.06, 0.33]
sham tDCS x positive	0.04	0.095	[-0.15, 0.22]
anodal tDCS x positive	0.22	0.100	[0.03, 0.42]
random intercept variance	0.13	0.027	[0.08, 0.19]

To more directly test the effect of rules on selfishness, we looked at selfishness across tDCS conditions when making decisions on distributions between oneself and another person ('me vs. other person'-block). When choosing freely, participants accumulated significantly more money for themselves under cathodal tDCS of the right LPFC compared to anodal tDCS of this brain area (Fig. 6a, Mann-Whitney U test, U = 404, p = .04). Thus, we replicated the previously observed effect that cathodal tDCS over the right LPFC leads to more selfish decisions[3,6,9,12-14].

To analyse how being faced with a rule changed selfishness, we looked at the earnings that participants accumulated at the expense of the other person when a rule was in place and compared it to the to the accumulated earnings when participants were free to decide. Note that in the rule-part, the rule demanded to choose the selfish option in half of the trials and the pro-social option in the other half of the trials.

Selfishness was significantly reduced under cathodal compared to anodal tDCS (Fig. 6b, Mann–Whitney U test, U = 307, p = .03). The rule, that dictated more pro-social decisions as compared to what participants chose freely in the free part (Fig. 6a), led participants to give 33% more to the other person on average (and thus took 33% less for themselves), during cathodal stimulation of the right LPFC. Thus, the confrontation with a rather pro-social rule was able to attenuate the increased selfishness of participants under cathodal tDCS, while participants stayed more consistent with their free choices under anodal tDCS.

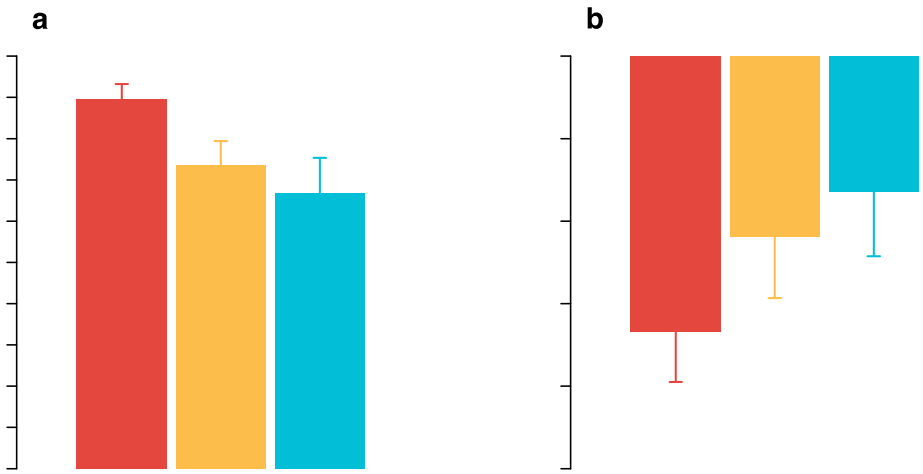

Fig. 6. Selfishness and change in selfishness due to following the rule. (**a**) Average amount of money accumulated for oneself at the expense of another person across tDCS conditions in the free-part (0% corresponds to an equal and fair split of the money, 100% means maximal selfishness). (**b**) Change in selfishness, as measured by the difference in accumulated earnings between the free-part and the rule-part, when faced with a rule that demanded to take the pro-social option in half of the trials. Error bars show the standard errors of the mean.

After the main task, participants made fairness judgements for several hypothetical money allocations between a person A and a person B. Neither cathodal, nor anodal

tDCS altered the fairness perception of participants (Fig. 7 & Table 6). In line with earlier findings[12-14,32], this suggests that brain stimulation led participants to make different decisions without changing the underlying evaluation process.

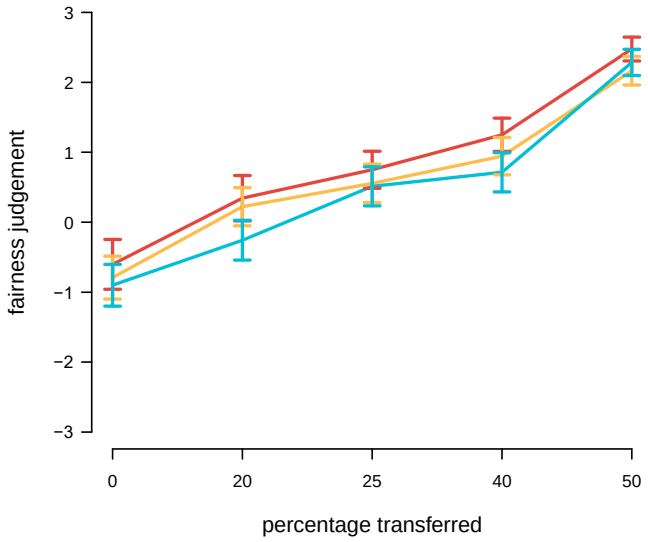

Fig. 7. Fairness judgements. Average fairness judgements on a scale from -3 to 3 depending on the amount transferred by a hypothetical person A to a hypothetical person B for each tDCS treatment (red = cathodal tDCS, yellow = sham, blue = anodal tDCS).

Table 6.
Fairness evaluations.
Random intercept regression predicting fairness ratings depending on different money allocations, and tDCS condition.

	Estimate	SD	95% CI
Intercept (cathodal tDCS)	-0.73	0.19	[-1.09, -0.36]
sham tDCS	-0.20	0.26	[-0.72, 0.30]
anodal tDCS	-0.39	0.26	[-0.90, 0.11]
percentage transferred	6.04	0.34	[5.39, 6.72]
sham tDCS x percentage transferred	-0.20	0.47	[-1.10, 0.72]
anodal tDCS x percentage transferred	0.15	0.47	[-0.80, 1.05]
random intercept variance	0.88	0.08	[0.74, 1.04]

Discussion

Rules often take the form of external restrictions on the pursuit of own goals, and sometimes demand to take actions that are against one's own will[33]. We often follow rules, nevertheless[34]. Here we provided evidence for a causal involvement of the right LPFC in rule following with social consequences. When freely deciding, participants made decisions that would yield the most benefits for them or others and manipulating the right LPFC did not change that. When an arbitrarily imposed rule coincided with this internal goal, people followed the rule overwhelmingly, irrespective of brain stimulation. However, when the rule was to hurt oneself or another person financially, anodal tDCS over this brain area led to more rule breaking, compared to more rule following under cathodal tDCS.

At the same time, under cathodal tDCS, participants made rather selfish choices in allocating a sum of money between them and another person. Being confronted with a rule that demanded to split the money more pro-socially, selfishness was, however, significantly reduced. Thus, although cathodal tDCS led to more damage towards oneself or another person due to high rule-following of a costly rule, a rather 'pro-social' rule in the 'me vs. other person'-block was able to make participants choose the socially desirable option more often. Under anodal tDCS on the other hand, participants stayed more consistent with their free choices when a rule was in place. Following the hypothesis in the literature, that cathodal stimulation to the right LPFC leads to more selfish payoff maximization[3,9,12-14], we should have seen more rule-breaking under cathodal tDCS and increased selfishness, regardless of the rule. Instead our results suggest that the right LPFC is critically involved in shifting behaviour from a more rule-based decision mode that is less sensitive to consequences (i.e. hurting another person or being pro-social towards another person)

to decisions that are more focused on outcomes and consequences in light of internal goals. This result is in line with the idea that the LPFC is important for a cost-benefit integration of external restrictions and own goals, rather than controlling selfish impulses[15].

By disentangling the role of rules in pro-social decisions with regards to the LPFC, and demonstrating that tDCS can systematically modify the alignment between internal goals and external restrictions, these results may be able to reconcile seemingly contradictory observations and opposing views in the literature regarding the automaticity of pro-social behaviour. While selfishness has been seen as the impulse that needs to be controlled by executive control instances[8,9], on the flipside, some scholars argued that pro-social behaviour is impulsive and rational-economic reasoning towards payoff maximization is controlled by secondary control processes[35-37]. Further, some studies have observed lower pro-social behaviour after cathodal brain stimulation[6,8,9,12], while others have observed higher pro-social behaviour[2,5,13].

Our results also resonate with a recent brain stimulation study showing that anodal tDCS over the right LPFC increases honesty when honesty is in conflict with material gain [38]. Based on our results and interpretation, this finding may be explained by internal goals (honesty) that are in conflict with the economic temptation to cheat. This conflict is resolved in favour of internal goals (honesty), when anodal tDCS is applied. As in previous studies, the internal goals or intrinsic behaviour have to be deduced post-hoc and we hence can only speculate about it, while in our design we can directly compare behaviour under no rule and rule when the rule is either aligned or in conflict with voluntary behaviour. Maréchal et al.[38] further found no difference of tDCS in participants that cheated to the extreme. This resonates with our finding

that unconditional rule following is not affected by tDCS, and suggests that individual differences exists in the extent to which a situation is perceived as a conflict between motives that needs trading off (and therefore the recruitment of the right LPFC) or not. While we find differences between the two active brain stimulation protocols, the difference to the sham condition were smaller and not significant. Future studies may be needed to investigate these comparisons further.

Both, phylogenetically and ontogenetically, the LPFC is one of the latest developing brain regions[39-42] and its major role has been implied in adaptive behaviour that enables humans, as compared to other vertebrates, to flexibly react to external stimuli in order to implement goals, rather than just follow fixed stimulus-response patterns[12,16-21]. Following rules regardless of its consequence can be seen as just reacting to an external stimulus, whereas weighing the costs and benefits of a rule, and deciding to follow it depending on its consequences, is arguably a much more adaptive behaviour. We demonstrate that the right LPFC is involved in trading off internal goals with external restrictions, helping us to violate rules when they just demand to hurt someone without any other benefits. These results may further our understanding of the neurobiological basis of normative decision making in the social domain. Instead of a simple dichotomy of subcortical brain areas that drive selfishness, and the LPFC controlling these 'selfish impulses', our results provide a more nuanced explanation of the function of the LPFC in human social behaviour, in line with the broader cognitive literature, suggesting that the right LPFC plays a crucial part in flexibly reacting to the social environment, by trading off internal goals with external restrictions.

References

1. Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G. & Fehr, E. The Neural Signature of Social Norm Compliance. *Neuron* **56,** 185–196 (2007).
2. Balconi, M. & Canavesio, Y. High-frequency rTMS on DLPFC increases prosocial attitude in case of decision to support people. *Social Neuroscience* **9,** 82–93 (2013).
3. Baumgartner, T., Knoch, D., Hotz, P., Eisenegger, C. & Fehr, E. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nat Neurosci* **14,** 1468–1474 (2011).
4. Christov Moore, L. & Iacoboni, M. Self-other resonance, its control and prosocial inclinations: Brain–behavior relationships. *Human Brain Mapping* **37,** 1544–1558 (2016).
5. Christov Moore, L., Sugiyama, T., Grigaityte, K. & Iacoboni, M. Increasing generosity by disrupting prefrontal cortex. *Social Neuroscience* (2016). doi:10.1080/17470919.2016.1154105
6. Wout, M. V., Kahn, R. S., Sanfey, A. G. & Aleman, A. Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *Neuroreport* **16,** 1849–1852 (2005).
7. Yamagishi, T. *et al.* Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 5582–5587 (2016).
8. Knoch, D. *et al.* Disruption of Right Prefrontal Cortex by Low-Frequency Repetitive Transcranial Magnetic Stimulation Induces Risk-Taking Behavior. *J Neurosci* **26,** 6469–6472 (2006).
9. Knoch, D. *et al.* Studying the Neurobiology of Social Interaction with Transcranial Direct Current Stimulation—The Example of Punishing Unfairness. *Cereb. Cortex* **18,** 1987–1990 (2008).
10. Cohen Kadosh, R. Modulating and enhancing cognition using brain stimulation: Science and fiction. *Journal of Cognitive Psychology* **27,** 141–163 (2015).
11. Fertonani, A. & Miniussi, C. Transcranial Electrical Stimulation: What We Know and Do Not Know About Mechanisms. *Neuroscientist* 1073858416631966 (2016). doi:10.1177/1073858416631966
12. Strang, S. *et al.* Be nice if you have to – the neurobiological roots of strategic fairness. *Soc Cogn Affect Neurosci* **10,** 790–796 (2015).
13. Ruff, C. C., Ugazio, G. & Fehr, E. Changing Social Norm Compliance with Noninvasive Brain Stimulation. *Science* **342,** 482–484 (2013).
14. Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V. & Fehr, E. Diminishing Reciprocal Fairness by Disrupting the Right Prefrontal Cortex. *Science* **314,** 829–832 (2006).
15. Buckholtz, J. W. Social norms, self-control, and the value of antisocial behavior. *Current Opinion in Behavioral Sciences* 122–129 (2015). doi:10.1016/j.cobeha.2015.03.004
16. Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 7338–7343 (2005).
17. Kadota, H. *et al.* The role of the dorsolateral prefrontal cortex in the inhibition of stereotyped responses. *Exp Brain Res* **203,** 593–600 (2010).
18. Kesner, R. P. & Churchwell, J. C. An analysis of rat prefrontal cortex in

mediating executive function. *Neurobiology of Learning and Memory* **96,** 417–431 (2011).
19. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review of neuroscience* 167–202 (2001). doi:10.1146/annurev.neuro.24.1.167
20. Braver, T. S., Paxton, J. L., Locke, H. S. & Barch, D. M. Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 7351–7356 (2009).
21. Stokes, M. G. *et al.* Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* **78,** 364–375 (2013).
22. Petrides, M. Lateral prefrontal cortex: architectonic and functional organization. *Proc R Soc B* **360,** 781–795 (2005).
23. Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A. & Braver, T. S. Global Connectivity of Prefrontal Cortex Predicts Cognitive Control and Intelligence. *J Neurosci* **32,** 8988–8999 (2012).
24. Domenech, P. & Koechlin, E. Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences* **1,** 101–106 (2015).
25. Koechlin, E., Ody, C. & Kouneiher, F. The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science* **302,** 1181–1185 (2003).
26. Greene, J. D. & Paxton, J. M. Patterns of neural activity associated with honest and dishonest moral decisions. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 12506–12511 (2009).
27. FeldmanHall, O. *et al.* Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Soc Cogn Affect Neurosci* **7,** 743–751 (2012).
28. Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M. & Cohen, J. D. The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron* **44,** 389–400 (2004).
29. Kuo, M.-F., Polanía, R. & Nitsche, M. in *Transcranial Direct Current Stimulation in Neuropsychiatric Disorders* 29–46 (Springer International Publishing, 2016). doi:10.1007/978-3-319-33967-2_3
30. Nitsche, M. A. *et al.* Transcranial direct current stimulation: State of the art 2008. *Brain Stimulation* **1,** 206–223 (2008).
31. Nitsche, M. A. *et al.* Shaping the effects of transcranial direct current stimulation of the human motor cortex. *Journal of Neurophysiology* **97,** 3109–3117 (2007).
32. Dambacher, F. *et al.* Reducing proactive aggression through non-invasive brain stimulation. *Soc Cogn Affect Neurosci* **10,** 1303–1309 (2015).
33. Milgram, S. Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology* **67,** 371–378 (1963).
34. Arendt, H. *Eichmann in Jerusalem*. (Penguin, 2006).
35. Bear, A. & Rand, D. G. Intuition, deliberation, and the evolution of cooperation. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 936–941 (2016).
36. Rand, D. G. & Epstein, Z. G. Risking Your Life without a Second Thought: Intuitive Decision-Making and Extreme Altruism. *PLoS ONE* **9,** e109687 (2014).
37. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489,** 427–430 (2012).
38. Maréchal, M. A., Cohn, A., Ugazio, G. & Ruff, C. C. Increasing honesty in humans with noninvasive brain stimulation. *Proc. Natl. Acad. Sci. U.S.A.* **6,** 201614912 (2017).

39. Shaw, P., Kabani, N. J. & Lerch, J. P. Neurodevelopmental trajectories of the human cerebral cortex. *J Neurosci* **28,** 3586–3594 (2008).
40. Stout, D. The Evolution of Cognitive Control. *Topics in Cognitive Science* **2,** 614–630 (2010).
41. Fuster, J. M. The Prefrontal Cortex — An Update: Time Is of the Essence. *Neuron* **30,** 319–333 (2001).
42. Kolb, B. *et al.* Experience and the developing prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 17186–17193 (2012).

Supplementary Information

Additional Analysis

We specifically wanted to test the role of the right LPFC in rule following when the rule did not coincide with what participants would choose in the 'free'-part (i.e. rules that demanded to financially hurt oneself or the other person), while showing that behaviour is unchanged when internal goals and the rule coincide (i.e. rules that are beneficial or neutral).
However, due to aggregating the data across consequences we lost possibly valuable variability related to the degree of how beneficial or harmful following the rule really was (see experimental setup & design).
We therefore also fitted two more complex models to the unaggregated data, using the binary trial-by-trial response variable (0 = not following the rule, 1 = following the rule). As predictor, we used the continuous rule consequence variable, that varied between -30 (following the rule would lead to a loss of 30 cents) and +30 (following the rule would lead to earning 30 cents more than violation the rule). In this regression, we included the observations of all participants and dummy-coded unconditional rule following (participants who followed the rule across all trials without being influenced by its consequence at all).
To account for the dependencies within subjects, we fitted two (Bayesian) random intercept binomial regression models using JAGS/R to the 'me'-trial and 'other person'-trial data, respectively. Non-informative Gaussian priors (m=0, sd=100) were used for each predictor and non-informative uniform priors (range 0 to 100) for the error terms. We used three parallel chains. For every estimated coefficient, the potential scale reduction factor (Gelman and Rubin Diagnostic) was below 1.05, indicating good mixing of the three chains and thus high convergence. Regression tables reported below show estimated coefficients (log-odds) together with the 95% confidence interval (CI, also called highest density interval in the Bayesian framework). Note that, since non-informative priors were used, a 95% CI that only contains negative or positive values can be interpreted as significant at a $p = .05$ two-sided threshold from a frequentist perspective. Fitting the models using restricted maximum likelihood (REML) as implemented in the lme4 package in R revealed similar estimates and resulted in the same statistical inferences.
Table S1 and Figure S1a show the fitted model for 'me'-trials. As can be seen, the probability to follow the rule increased the more beneficial the rule was, up to 100% for rules that would yield beneficial outcomes to the participant (consequence ≥ 0) in all three tDCS conditions. However, participants under cathodal and sham tDCS, compared to anodal tDCS, had a higher likelihood to follow harmful rules and, therefore, had a steeper increase in rule obedience towards more beneficial consequences.
Table S2 and Figure S1b shows the fitted model for 'other person'-trials. Again, the probability to follow the rule increased the more beneficial the rule was, up to 100% for rules that would yield beneficial outcomes to the other person (consequence ≥ 0) in all three tDCS conditions. However, participants under cathodal, compared to anodal, tDCS had again a higher likelihood to follow harmful rules.

Table S1.
‘Me’-trials (‘confronted with a rule).
Random intercept binomial regression predicting the likelihood to follow rules in ‘me’-trials, depending on the tDCS condition.

	Estimate	SD	95% CI
Intercept (cathodal tDCS)	3.36	0.72	[1.94, 4.78]
sham tDCS	-0.21	0.97	[-2.13, 1.68]
anodal tDCS	-0.66	1.02	[-2.65, 1.34]
lost by following	0.25	0.02	[0.20, 0.29]
full adherence	31.02	16.78	[11.96, 68.89]
sham tDCS x lost by following	-0.01	0.03	[-0.07, 0.05]
anodal tDCS x lost by following	0.07	0.03	[0.01, 0.14]
random intercept variance	3.06	0.38	[2.37, 3.85]

Table S2.
‘Other person’-trials (confronted with a rule).
Random intercept binomial regression predicting the likelihood to follow rules in ‘other person’-trials, depending on the tDCS condition.

	Estimate	SD	95% CI
Intercept (cathodal tDCS)	5.23	0.94	[3.46, 7.15]
sham tDCS	-2.28	1.19	[-4.63, 0.04]
anodal tDCS	-2.59	1.23	[-5.14, -0.26]
lost by following	0.23	0.03	[0.18, 0.28]
full adherence	29.82	17.05	[9.82, 67.62]
sham tDCS x lost by following	-0.07	0.03	[-0.12, -0.01]
anodal tDCS x lost by following	-0.01	0.03	[-0.07, 0.05]
random intercept variance	3.46	0.47	[2.66, 4.48]

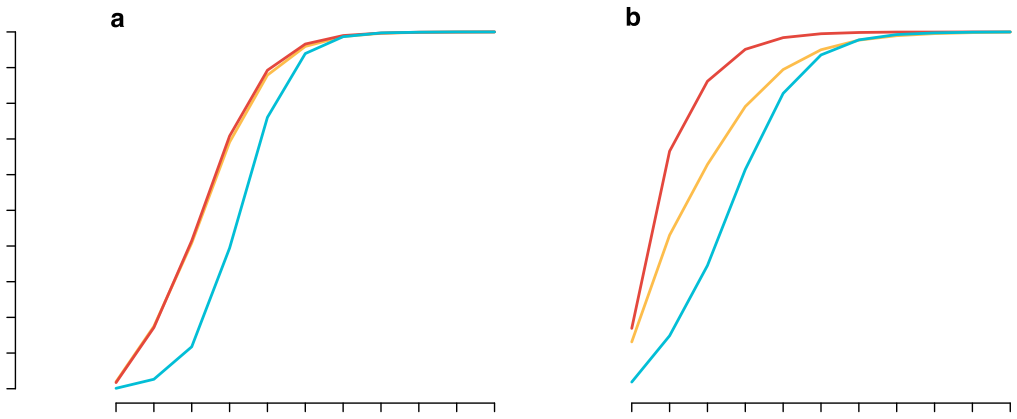

Figure S1. Predicted probability to follow the rule as a function of the consequence of rule-following (red = cathodal tDCS, yellow = sham, blue = anodal tDCS), when the consequences of the rules would affect the participant (**a**) or another person (**b**).