NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

as a manuscript

Oksana Zinchenko

NEUROBIOLOGICAL MECHANISMS OF SOCIAL PUNISHMENT AS A COOPERATION PROMOTER

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Psychology HSE

Academic supervisor:
Candidate of Biological Sciences,
Vasily Klucharev

Moscow 2019

Table of contents

PUBLICATIONS AND APPROBATION OF RESEARCH

The studies that comprise this thesis are the following:

First-tier publications[1]

1) Zinchenko O., Arsalidou M. Brain responses to social norms: Meta-analyses of fMRI studies // Hum. Brain Mapp. 2018. T. 39. № 2. C. 955-970.
2) Zinchenko, O., Klucharev, V. Commentary: The Emerging Neuroscience of Third-Party Punishment // Frontiers in Human Neuroscience. 2017. T.11. C. 512.

Second-tier publications[2]

3) Zinchenko, O., Belianin, A., Klucharev, V. Neurobiological Mechanisms of Fairness-Related Social Norm Enforcement: a Review of Interdisciplinary Studies // Zhurnal Vysshei Nervnoi Deyatelnosti Imeni I.P. Pavlova. 2018. T. 68. № 1. C. 16-27.
4) Zinchenko, O., Belianin, A., Klucharev, V. The role of the temporoparietal and prefrontal cortices in third-party punishment: a tDCS study // Psychology. Journal of the Higher School of Economics. 2019. (in print).

Reports at the conferences

1) IEEE International Symposium «Video and Audio Signal Processing in the Context of Neurotechnologies», June 30 – July 2, 2016 (Saint Petersburg, Russian Federation). Report: *Neurobiological mechanisms of social punishment*.

[1] First-tier publications include papers indexed in the Web of Science (Q1 or Q2) or Scopus (Q1 or Q2) databases, as well as peer-reviewed collections of conferences that appear in CORE rankings (ranks A and A*).

[2] Second-tier publications are papers published in journals included on HSE's list of high quality journals or indexed in the Web of Science (Q3 or Q4) or Scopus (Q3 or Q4) databases, as well as peer-reviewed collections of conferences appearing in CORE rankings (rank B).

2) Annual Meeting of the Society for Neuroeconomics, August 28 – 30, 2016 (Berlin, Germany). Report: *The role of the temporo-parietal junction and dorsolateral prefrontal cortex in third-party punishment of norm violations*.

3) Cognition, Computation, Communication and Perception (CCCP) Conference 3: “Theoretical and Neurobiological Bases of Higher Cognitive Functions”, September 2016 (Moscow, Russian Federation). Report: *Investigation of the interaction between temporo-parietal junction and lateral prefrontal cortex in third-party punishment of norm violations*.

4) Annual Conference of Society for Neuroeconomics, October 6 – 8, 2017 (Toronto, Canada). Report: *Fronto-parietal coupling of brain rhythms during third-party punishment*.

5) 44th Annual Meeting of the Society for Philosophy and Psychology, July 11th – July 14th, 2018 (Ann Arbor, Michigan, MI). Report: *Brain mapping of social norms: fMRI meta-analyses*.

This work has been carried out in Centre for Cognition and Decision Making, Institute for Cognitive Neuroscience, National Research University Higher School of Economics, Moscow, Russian Federation.

RESEARCH TOPIC

Cooperation supports the social order in society. The enforcement of cooperation in groups is one of the most fundamental issues in behavioral economics and psychology, and many studies shed light on various factors that are important in sustaining and enforcing cooperation in society, such as social norms. Social norms are not self-enforcing by nature but require specific mechanisms to sustain them, such as social punishment carried out by a third party. Although many studies in neuroscience, psychology, and behavioral economics have investigated the general role of social norms and social punishment and their brain representations, no concordant map of brain activations has been presented for brain responses to social norms. Neither have the neural dynamics underlying third-party punishment as an enforcer of cooperation and social norms been investigated in detail. The aim of the present study was to perform a meta-analysis to provide a neurocognitive map of concordant brain responses to social norms and their violations and to investigate the causal role of brain regions in third-party punishment as a mechanism to enforce social norms. The meta-analysis showed concordant activation in the anterior cingulate and the right medial frontal gyrus (right dorsolateral prefrontal cortex or rDLPFC) in relation to social norm representation and concordant activation in the right insula and claustrum in relation to norm violation. A fundamental neural model for third-party punishment (Krueger, Hoffman, 2016) suggests that the rDLPFC and right temporoparietal junction (rTPJ) are key regions the activation of which underlie third-party punishment of norm violations. We used transcranial direct current stimulation (tDCS) to independently or jointly disrupt rTPJ and rDLPFC activity during the third-party dictator game. We found that anodal tDCS of the rDLPFC did not modulate the third-party punishment. However, we found a significant effect of anodal tDCS of the rTPJ, which decreased the third-party punishment for moderately unfair splitting of the resources, while joint stimulation of the rTPJ (by anodal tDCS) and rDLPFC (by cathodal tDCS) produced only a marginal effect. Our results support the profound role of the rTPJ in the initiation of social punishment.

CONTENTS

Introduction

Social norms and the mechanisms of their enforcement: behavioral findings

Human society greatly depends on social norms, which work as a mechanism supporting cooperation. Social norms can be defined as implicit or explicit rules that are formed to govern interactions within groups and that are considered appropriate within a society. Some examples of social norms include common courtesy and culturally appropriate manners (Sherif and Sherif, 1953). Importantly, in human societies cooperation is mainly based on social norms (Fehr and Fischbacher, 2004).

Different kinds of social norms regulate individual behavior, one of which is the *norm of fairness* (Elster, 1989). The norm of fairness in democratic societies is usually considered a norm of equality (Elster, 1989). A common approach to investigate social norms is to use interactive economic games, such as the ultimatum game introduced by Güth and colleagues (1982; see Gabay et al., 2014 for a review), the Prisoner's dilemma (Dickinson et al, 2015), and the dictator game (Tammi, 2013). Such games allow different distributions of financial transfers between players. For instance, in the dictator game there are two players, one of whom (the *dictator*) is given the opportunity to distribute monetary units (MUs) between herself and another player (the *recipient*) (Tammi, 2013). Behavioral studies have robustly demonstrated that many people who play economic games prefer fair distributions to unequal ones (Guth et al., 1982; Kahneman et al., 1986; Forsythe et al., 1994; Engel, 2010).

However, people do not always conform to social norms and sometimes tend to violate them to maximize their own interests. Such violations usually meet with increasing social pressure to conform to the norms. Psychological studies suggest that the violation of social norms could result in the exclusion of the norm violator from the group or in other less harsh forms of social disapproval (Schachter, 1951; Sherif, Sherif, 1953). It follows that the behavior conflicting with social norms can have dramatic consequences. Social disapproval and social exclusion enforce norm compliance; in fact, even the possibility of such sanctions could increase norm compliance (Ruff et al., 2013).

Such behavior—a tendency to spend one's own resources to punish norm violations (e.g., unfair distributions of MUs that violate the *norm of fairness*)—is called *social punishment* (Fehr, Fischbacher, 2004; Ruff et al., 2013). Social punishment can be demonstrated experimentally based on material costs only, for example, when people spend some MUs from their own budget to punish a norm violator. It can also be expressed as social disapproval (Carpenter, Seki 2011; Masclet et al. 2003; Guala, 2012), which is more common in social life (e.g., reprimands, social exclusion, etc.). Behavioral economics studies suggest that social punishment is usually meted out by individuals who are directly affected by the norm violations of others (i.e., *second parties*). Yet, individuals who are not directly affected by the norm violations of others (*third parties*) are also willing to punish norm violators at their own expense (Fehr, Fischbacher, 2004). It has been shown that norm violation behavior (such as unfair behavior in the case of the *norm of fairness*) leads to negative emotions, such as anger (Batson et al., 2007; Pedersen, 2012), guilt (Wagner et al, 2011), and embarrassment (Melchers et al., 2015), that could drive individuals to punish their opponent at the expense of monetary reward or to consider the opponent guilty. Overall, social punishment as the "propensity of cooperative individuals to spend some of their resources penalizing norm violators" (Zinchenko, Klucharev, 2017) is the main mechanism supporting social norms in large social groups.

Neurofunctional model of social norms and norm violations

Because social norms are so important in maintaining social order, further investigation is crucial to understand the roots of human behavior in different social contexts. Montague and Lohrehz (2007) propose a neurofunctional model of social norms based on a review of studies exploring neural correlates of adherence to shared social norms. They suggest that the brain can flexibly adjust behavior according to existing social norms, similar to other forms of adaptive behavior. To successfully interact with others in any social group, the following steps are necessary: 1) to have a representation of the norm, 2) to have a mechanism detecting violations of this norm, and 3) to have the chance to look at the current situation from a third-party perspective to be able to maintain norm compliance (Montague, Lohrenz, 2007; Xiang et al., 2013). We adopted this model to perform the first meta-analysis of neuroimaging studies of social norms (Zinchenko, Arsalidou, 2018).

Third-party punishment as a mechanism of norm enforcement: a comparison with second-party punishment and the model of neural activation

In addition to investigating social norms in general, it is particularly critical to study the mechanisms of enforcement, implementation, and compliance, including social punishment. *Third-party punishment* is a special form of social punishment that is unique to human culture (Riedl et al., 2012) and that has not been observed in other primates, including chimpanzees. While the majority of neuroimaging studies investigate the neural basis of second-party punishment, there are not many studies about the neural mechanisms of third-party punishment. Importantly, third-party punishment is crucial for establishing cooperation in larger social groups. Therefore, studies of third-party punishment are of practical importance and are relevant in the modern urbanized world.

Neuroimaging and brain stimulation studies provide some insights on the neural mechanisms of third-party punishment. It has been shown that second- and third-party punishment have different neural mechanisms (Strobel et al., 2011) and that only some regions, such as the ventral striatum, share a common activation for both types of punishment (Stallen et al., 2018). For instance, the lateral prefrontal cortex (LPFC)—and its subpart the DLPFC—is casually involved in both types of social punishment but in slightly different ways. The right LPFC (rLPFC) is involved in both voluntary and sanction-induced norm compliance in the case of second-party punishment (Ruff et al, 2013). In the case of third-party punishment, rDLPFC activity correlates with the evaluation of the responsibility for committing norm violations (Buckholtz et al., 2008). In particular, the emotional evaluation of the personal responsibility that results in third-party punishment correlates with activity of the amygdala, the medial prefrontal cortex, and the posterior part of the cingulate cortex (Buckholtz et al., 2008).

Neuroimaging studies suggest that several distinct brain networks are consistently recruited during third-party punishment (Krueger, Hoffman, 2016). According to Krueger and Hoffman's model (2016), these brain networks include the *central-executive*, *mentalizing,* and *salience* networks. The mentalizing network is responsible for the ability to imagine thoughts and possible actions of others and mainly relies on individual experience, while the activity of the central-executive network is required for our cognitive control, working memory, task-switching, planning, etc. Hypothetically, in accordance with the predictions of Krueger and Hoffman's model, third-party punishment decisions start with the activation of the salience network (insula, amygdala, and dorsal anterior cingulate), which allows the *detection of norm violations* and consequently generates an aversive response. Next, the default mode network (TPJ, dorsomedial prefrontal cortex or dMPFC) integrates the perceived harm and inference of intentions into an *assessment of blame*. Finally, the central executive network (DLPFC) converts the blame signal into a specific *punishment decision*.

Neural mechanisms of third-party punishment: neuroimaging and brain stimulation studies

Most previous studies focus on the brain correlates of third-party punishment and practically ignore the interactions between the large-scale brain networks. A recent brain stimulation study shows that transcranial magnetic stimulation (rTMS) of the rDLPFC increased third-party punishment, while psychometric methods have provided evidence of a correlation between an individual empathy index and the intensity of third-party punishment (Brune et al., 2012). These results may suggest that the DLPFC integrates all signals from the previous steps of the decision-making process, including the emotional emphatic responses.

It follows that suppression of the DLPFC should lead to increased third-party punishment only if the activity of the DLPFC underlies the final evaluation of the costs of the punishment decision. If so, suppression of the DLPFC should decrease the perceived costs of social punishment and therefore increase third-party punishment. The previous TMS study did not disentangle material and moral costs (Brune at al., 2012);

third parties punished the norm violator and helped the victim at the same time. Therefore, the role of the DLPFC in third-party punishment remains largely unclear.

Considering other main brain regions from the model (Krueger, Hoffman, 2016), the previous brain stimulation studies provided a more coherent interpretation of the role of the rTPJ in third-party punishment. It has been shown that rTMS of the rTPJ decreases third-party punishment of outgroup members (Baumgartner et al., 2014). This supports Krueger and Hoffman's model (2016) of third-party punishment and indicates the vital role of the rTPJ in the processing of emotional information during social punishment. This interpretation is in line with extensive meta-analyses that demonstrated the involvement of the rTPJ in mentalizing and empathy (Van Overwalle, 2009; Garrigan, Adlam, Langdon, 2016).

A seminal functional magnetic resonance imaging (fMRI) study of third-party punishment has demonstrated a functional interaction between the rDLPFC and the rTPJ (Buckholtz et al., 2008). This study suggests that the activation of the rTPJ before a punishment decision is followed by simultaneous deactivation of the rDLPFC and results in the follow-up activation of the rDLFPC when the final decision is made. Taking into account these findings (Buckholtz et al., 2008), we speculate that the chronometry of the third-party punishment decision is as follows. The information about the harm (a degree of norm violation) and the intentions (intentional versus unintentional norm violations) are processed in the salience network (anterior cingulate, anterior insula) and the mentalizing network (rTPJ). Subsequently, the resulting information is transferred to the DLPFC to calculate the final decision, considering the context of the situation and the self-maximization (if the punishment decision is costly).

Recent neuroimaging studies focus not only on the functional role of the exact brain region but also on the interaction between different brain regions (e.g., Treadway et al., 2014; Bellucci et al., 2017). Similarly, Feng and colleagues (2018) analyze resting-state fMRI data using graph theory and support Krueger and Hoffman's model of the key brain nodes involved in third-party punishment. Another fMRI study investigates task-related brain activity and supports the main role of the mentalizing (TPJ and dMPFC) and central-executive (LPFC) systems in third-party punishment (Bellucci et al., 2017). Importantly, this study demonstrates that the dMPFC receives the incoming signals only from the TPJ, while the activity of the dMPFC and its functional co-activation with the dLPFC correlate with the degree of third-party punishment (Bellucci et al., 2017). According to these findings, the TPJ is considered to be an integrative node, receiving the information from other sub-regions.

The primary role of the mentalizing and central-executive networks in third-party punishment is supported by traumatic brain injury studies. Glass and colleagues (2016) show that damage to these cortical regions decreased the intensity of third-party punishment and altruistic compassion. However, to date the functional connectivity before or during social punishment has not been investigated using electrophysiological methods with high time resolution. To our knowledge, the electroencephalogram studies reported only the inter-brain connectivity between the receiver's and the punisher's brain activity during third-party punishment using a hyperscanning approach (Astolfi et al., 2015; Ciaramidaro et al., 2018).

In summary, we reviewed the key neuroimaging studies of social norms and social norm enforcement, focusing particularly on social punishment and third-party punishment. We identified the following gaps in the research on social norms and social punishment, which we addressed in a series of studies: 1) no meta-analyses have been performed to identify the key brain regions concordantly activated in relation to representations of social norms and their violations; 2) previous studies robustly demonstrated the role of the mentalizing and central-executive networks in third-party punishment, but brain stimulation has not been used to demonstrate a causal relationship between the aforementioned networks and third-party punishment or to investigate interaction between the mentalizing and central-executive networks.

Research goals

1) To perform a meta-analysis of neuroimaging studies of fMRI modality to identify the key regions related to information processing in social norms (the representation of social norms and norm violations).

2) To perform a brain stimulation study to investigate the functional interactions of the rDLPFC and the rTPJ during third-party punishment decisions.

3) To identify the functional roles of the rDLPFC and the rTPJ in third-party punishment decisions.

Overview

In the following overview of the research project, we briefly describe the individual studies. In the first study (Study 1, Attachment A of the thesis), the meta-analysis of neuroimaging studies was performed to identify concordant activation for social norm representations and norm violations. In the second study (Study 2, Attachment B of the thesis), a systematic review of studies on social punishment was performed to identify key behavioral and neural features specific to mechanisms of fairness-related social norm enforcement, focusing on third-party punishment. In the third study (Study 2, Attachment C of the thesis), the hypotheses relating to tDCS were formulated, and a tDCS study was conducted to investigate the neural dynamics underlying third-party punishment decisions. Following, we discuss the contributions of these studies (theoretical, methodological, and empirical novelty).

Theoretical novelty

The results of the extensive systematic review and meta-analysis contributed to the field of theoretical models of social punishment and social norms, which allows extending the model of neural correlates of social norms and their enforcement.

Methodological novelty

For the first time, a meta-analytic methodology was used to verify current models of neural mechanisms underlying social norms and their enforcement based on a concordant map of brain activations across a range of fMRI studies.

Empirical novelty

For the first time, a tDCS study on third-party punishment has been conducted using anodal stimulation of the rTPJ and rDLPFC and using the novel experimental protocol of joint stimulation. The findings suggest that the anodal tDCS of the rTPJ led to decreased punishment for moderately unfair offers, while the joint stimulation led to a marginal increase in punishment. For the first time, a study has focused on the neural dynamics underlying third-party punishment, which has been tested using brain stimulation methods.

Methodology and study design

Part I (Meta-analysis). To identify the brain activations concordant with the representation of social norms and of their violation, we performed an activation likelihood estimation (ALE) analysis of 36 fMRI studies (Zinchenko, Arsalidou, 2018). Due to the different spatial and temporal resolution of fMRI, positron emission tomography (PET), and magnetoencephalography (MEG), we excluded PET and MEG studies from the search criteria. Therefore, we analyzed only fMRI studies to achieve homogeneity of the imaging data.

We performed a literature search among fMRI studies and looked for the behavioral tasks that reflected the definition of *social norm representation* as a commonly expected appropriate behavior in a certain situation (Cialdini, Goldstein, 2004; Montague, Lohrehz, 2007). Therefore, the tasks we looked for in the literature search included references to both moral and social norms and accepting rules or normative principles (“good” versus “bad” or “neutral”; “moral” versus “semantic”, etc.). Then we defined *norm violations* as behavioral deviations from shared social norms (i.e., inappropriate behavior), for instance, perceived unfairness (Buckholtz, Marois, 2012; Chang, Sanfey, 2008). Many of the selected studies focus on brain responses to norm violations and the influence of norm violations on decision making. For instance, some studies focus on negative moral emotions, such as guilt (Wagner et al, 2011) or embarrassment, as a consequence of norm violations (Takahashi et al. 2004; Takahashi et al. 2008). We created a map of concordant brain activations that reflect the general responses to global normative judgements, or social norms, and their violations.

Importantly, the observed concordant brain activations can be driven not only by the motivation to follow the social norm or to enforce it but also by the ambiguity in the trust game (Li, Turmunkh & Wakker, 2019) or by the reciprocity motive in the ultimatum game (Hoffman, McCabe, and Smith, 1996). To control for these factors, our next literature search focused on general neural correlates of social norm representation and norm violations, regardless of the specific tasks. To control for the effect of different motivations, from the set of eligible articles we included different tasks related to social norm representation and norm violations (a description of the tasks and contrasts can be found in Table 1, Zinchenko and Arsalidou, 2018). The foci (brain coordinates) extracted from these tasks were included into statistical analysis. Such a meta-analysis allows quantitative verification of which areas are concordantly active across all such tasks and shows neural correlates of involvement in global normative judgements.

Specifically, we used ALE, which is typically applied to perform whole-brain, random-effects voxel-wise imaging analyses (Eickhoff et al., 2009; 2012; 2017). It uses brain coordinates combined from different neuroimaging studies (fMRI; PET; MEG-MRI) to create a probabilistic map of activation that is thresholded and compared to random spatial distribution. Such an algorithm provides statistical maps of brain activations that are involved in a cognitive function. ALE is based on the three-dimensional coordinates derived from the statistical contrasts of fMRI studies. Overall, the meta-analysis examined the concordance of brain regions associated with the cognitive processing of social norms.

In total, the articles we analyzed in the meta-analysis report on 993 participants. We analyzed the experiments from 18 articles for the representation of social norms and the experiments from 29 articles for norm violations, which resulted in 47 experiments equaling 387 foci.

Part II (Neurocognitive model of third-party punishment). Next, we performed an extensive systematic review of neuroimaging and brain stimulation studies about enforcing the norm of fairness. We included 63 articles in the analysis to identify commonalities and differences in neural correlates of the mechanisms involved in enforcing the norm of fairness for second- and third-party punishment, focusing on the specific brain activation underlying third-party punishment. We concluded that two main networks—the mentalizing and central-executive networks—play a major role in third-party punishment decisions. Our analysis showed that the brain mechanisms involved in decisions to punish non-cooperative individuals in order to restore the *norm of fairness* require further investigation using combined neuroimaging and brain stimulation methods to differentiate the role of the DLPFC and the TPJ and to clarify their functional interaction.

Part III (Brain stimulation study). The results of our systematic review (*Part II*) suggest that the rDLPFC and rTPJ are the key integrative cortical regions that process information regarding the necessity of third-party punishment and mentalizing during the punishment decision, respectively. The interaction of these regions during third-party punishment has not been investigated. Therefore, it is important to clarify how the key nodes of the mentalizing and executive networks interact with each other during social punishment. To that end, we performed a tDCS study to probe this interaction and stimulate these regions simultaneously in a reciprocal (antagonistic) manner.

tDCS, also known as “micropolarisation” (Rusinov, 1977; Shelyakin, 2006), is a non-invasive brain stimulation technique that allows the modulation of the activity in specific regions of the cortex (Nitsche, Paulus, 2000, Nitsche et al., 2003; Paulus, 2011). tDCS is based on the application of weak electrical currents of 0.4–3 mA intensity to the scalp from the anode to the cathode. Anodal tDCS typically depolarizes (excites) neurons, and cathodal tDCS typically hyperpolarizes (inhibits) neurons. We applied tDCS with 1.5 mA intensity using an “offline” protocol to investigate the aftereffects of tDCS on third-party punishment. Participants were exposed to 15 minutes of tDCS, after which they completed the third-party punishment dictator game. Participants (*sanctioners*) were able to punish dictators using a budget of 20 MUs. The budget was renewed for each round, and all points not invested in the punishment were converted into a monetary payoff and paid to the participant after the experiment.

Only healthy, right-handed subjects with at least one year of university education (undergraduate students) were invited to participate in the study. Participants who reported having brain trauma in last three years, neurological and psychiatric conditions, or any metallic particles (tooth implants, etc.) were excluded from the study.

In total, we tested 23 right-handed participants (mean age=21.5 years, range=18–27 years, 7 males) for Study 1 to test the independent stimulation (anodal tDCS of the rTPJ; anodal tDCS of the rDLPFC; sham) and 21 right-handed subjects (mean age=22.79 years, range=18–27 years, 10 males) for Study 2 (joint anodal–cathodal tDCS of the rTPJ and the DLPFC). Each subject participated in only one of the two studies. The map of the electric current distribution and the electrode locations can be seen in Figure 1. Participants came to the laboratory three times for three different stimulation sessions; the representation of the sessions can be seen in Figure 2. Five subjects were excluded based on the exclusion criteria, did not punish at least once or demonstrated only antisocial punishment in fair trials. We analyzed the data from 20 and 19 subjects for the first and second studies, respectively.

KEY RESULTS AND CONCLUSION

Key results

Part I (Meta-analysis). We investigated the data from 993 participants for the meta-analysis; 52% of the participants were right-handed females, aged 23.89±6.28 years. The literature search allowed us to analyze the general brain responses to social norms (i.e., social expectations) and sub-analyze brain responses for the categories of “social norm representation” and “norm violations”. We also conducted contrast analyses between sub-categories to define the brain activations specific to each category. The results of the literature search and the timeline of the studies are presented in Figure 3.

On the general map of the brain responses to social norms we detected five clusters. The largest cluster was found in the right insula (BA 13), followed by the left medial frontal gyrus (Brodmann Area, BA 32) that extended to the cingulate gyrus (BA 32), and the right superior and middle frontal gyri (BA 9 and BA 10). Other regions included the left insula and the claustrum. The map is presented in Figure 4.

The concordant map of brain responses to social norm representations included the left anterior cingulate and the right medial frontal gyrus (BA 10). The map is presented in Figure 5.

The meta-analysis of the “norm violation” category revealed five suprathreshold clusters for norm violations, with the one with the highest likelihood of being detected in the right insula (BA 13), followed by the right cingulate gyrus (BA 32), the left insula (BA 13) and the claustrum, and the right middle and superior frontal gyri (BA 9 and 10). The map is presented in Figure 6.

Interestingly, a conjunction analysis did not reveal any common clusters between the social norm representation and norm violation categories. This suggests the possible independent brain representations behind these sub-categories, in accordance with the Montague and Lohrenz (2007) model. Compared to norm violation, social norm representation showed greater concordance in the anterior cingulate gyri (BA 32) and right medial frontal gyrus (BA 10), whereas compared to social norm representation, norm violation showed greater concordance in the right insula and claustrum and in the more dorsal parts of the cingulate gyrus (BA 24, 32). In summary, the findings suggest that the rDLPFC plays a key role in social norm representations and the detection of norm violations.

Part II (Neurocognitive model of third-party punishment). In accordance with our research goals, we performed a systematic review of behavioral, neuroimaging, and brain stimulation studies to identify the main open research questions in the third-party punishment research. The results that were briefly described in the Introduction section of this thesis were published in Zinchenko, Belyanin, and Klucharev (2018). Based on the previous fMRI study (Buckholtz et al., 2008), we speculated that an enhancement of TPJ activity with the simultaneous suppression of DLPFC activity should lead to increased third-party punishment due to the possible enhancement of the antagonistic TPJ–DLPFC interaction. Therefore, we suggested that a simultaneous application of tDCS to the TPJ and DLPFC should enhance such antagonistic interaction between these two regions and increase third-party punishment. Such a behavioral effect of tDCS could reflect changes in the functional connectivity between the TPJ and the DLPFC. Therefore, a combined non-invasive brain stimulation–neuroimaging study is needed to uncover the neural dynamics underlying third-party punishment.

Part III (Brain stimulation study). Based on the results of our review paper, we formulated the new research hypotheses, which have been published in Zinchenko and Klucharev (2017). Therefore, we conducted a tDCS experiment in which we tested the classic stimulation protocols with anodal tDCS stimulation of the rDLPFC and the rTPJ separately and the novel simultaneous stimulation protocol of the enhancement of TPJ activity with the simultaneous suppression of DLPFC activity. However, we observed only a trend relating to the effect of the joint stimulation tDCS protocol ($p=0.055$). When the rTPJ was activated and the rDLPFC was simultaneously deactivated, we observed a trend of increased third-party punishment. We suggest that tDCS is not the ideal method to study interactions of the rDLPFC and rTPJ. In the future, online transcranial alternating current stimulation could be used to study the synchronization and desynchronization of these brain regions. Nevertheless, we observed the effect of the anodal stimulation of the rTPJ, which led to decreased punishment for moderately unfair splitting of the resources (p=0.006). A recent study involving anodal tDCS of the rTPJ shows that subjects were assigned less blame for accidental harm during a moral judgment task (Sellaro et al., 2015), while a meta-analysis suggests that the rTPJ showed significant activation when one makes one's own moral decisions (Garrigan, Adlam, Langdon, 2016). Overall, rTPJ activity can reflect an analysis of the consequences of the third-party's own decision and of how harmful it would be for others. Therefore, anodal stimulation of the rTPJ area could exaggerate the latter process and consequently lead to diminished punishment.

One of the important findings of our tDCS study is that anodal tDCS had an effect on moderately unfair splitting of the resources (30:10) only: when third-party punishment of unfair splits created a Pareto optimal distribution of MUs (10:10:10) and it was impossible to improve the income of one player without worsening the incomes of the other players, while the punishment in other conditions led to advantageous and disadvantageous inequity. Pareto optimality is a state of allocation of resources where it is impossible to improve the income of one player without worsening the incomes of the other players. Therefore, in our study social punishment for other splits (0:40, 15:25, 20:20, 25:15, 35:5, and 40:0) would lead to advantageous and disadvantageous inequity. Following this, we suggest that anodal tDCS led to decreased moral costs, which resulted in decreased punishment.

Provisions for the defense

1) According to our meta-analysis of fMRI studies, social norm representation is robustly associated with activity of the anterior cingulate and right DLPFC, while norm violation is associated with the activation of the right insula and claustrum.

2) The Krueger and Hoffman model (2016), along with the results of our extensive systematic review and our meta-analysis, suggests the key role of the DLPFC and the TPJ in monitoring social norms and their enforcement. However, according to our tDCS study, anodal tDCS of the rDLPFC does not lead to changes in third-party punishment.

3) According to the tDCS study, anodal tDCS of the rTPJ decreases third-party punishment for moderately unfair splitting of the resources. We suggest that during the dictator game rTPJ activity underlies the initiation of the decision to punish, while activation of the rDLPFC becomes important in the latest stages of decision making.

Conclusion

We conducted the first meta-analysis of neuroimaging studies on social norms and their violations. The results suggest that social norm representation is linked to the activation of the anterior cingulate gyri and the rDLPFC and that norm violations are coded by the activation of the right insula and claustrum. Based on this, we proposed a neurocognitive model of social norms for healthy adults suggesting that the temporoparietal-medial-prefrontal circuit controls the emotional responses to norm violations and regulates the subsequent punishment of norm violators. The results of the brain stimulation study suggest that anodal tDCS of the rTPJ decreases the third-party punishment for moderately unfair splitting of the resources, while joint stimulation of the rTPJ (by anodal tDCS) and rDLPFC (by cathodal tDCS) produces only a marginal effect. This study demonstrates that anodal tDCS of the rTPJ decreases third-party punishment for moderately unfair behavior when the participants have an opportunity to restore equality in their social groups. Overall, the study findings support the critical role of the temporoparietal-medial-prefrontal circuit in third-party punishment. These findings can be used in future studies on social norms and the mechanisms of their enforcement in healthy subjects.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Professor Vasily Klucharev, for giving me opportunity to conduct and complete the PhD study at the Centre for Cognition and Decision Making.

My sincere thanks goes to Dr. Marie Arsalidou, who is my coauthor on “Brain responses to social norms: Meta-analyses of fMRI studies”, for the valuable contribution.

I would like to thank Dr. Alexis Belyanin, Dr. Matteo Feurra and Dr. Anna Shestakova for their valuable comments and help while designing and conducting these studies.

Many thanks to Dr. Beatriz Martin-Luengo and Dr. Ksenia Panidi for reviewing this thesis.

I acknowledge HSE University Basic Research Program and Russian Academic Excellence Project '5-100' for the financial support through the research unit.

REFERENCES

1. Astolfi, L., Toppi, J., Casper, C., Freitag, C., Mattia, D., Babiloni, F., Ciaramidaro, A., Siniatchkin, M. Investigating the neural basis of empathy by EEG hyperscanning during a Third Party Punishment // Conf Proc IEEE Eng Med Biol Soc. 2015. 5384-5387.

2. Baumgartner, T., Götte, L., Gügler, R., Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement // Human Brain Mapping. T. 33. № 6. C. 1452-1469.

3. Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R., Knoch, D. Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction // Soc Cogn Affect Neurosci. 2014. T. 9. № 5. C. 653-660.

4. Bellucci, G., Chernyak, S. V., Goodyear, K., Eickhoff, S. B., Krueger, F. Neural signatures of trust in reciprocity: A coordinate-based meta-analysis // Human brain mapping. 2016. T. 38. № 3. C 1233–1248.

5. Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K., Grafman, J., Krueger, F. Effective connectivity of brain regions underlying third-party punishment: Functional MRI and Granger causality evidence // Soc Neurosci. 2017. T. 12. № 2. C. 124-134.

6. Bendor J, Swistak P. The Evolution of Norms // American Journal of Sociology. 2001. T. 106. № 6. C. 1493–1545.

7. Brüne, M., Scheele, D., Heinisch, C., Tas, C., Wischniewski, J., Güntürkün, O. Empathy moderates the effect of repetitive transcranial magnetic stimulation of the right dorsolateral prefrontal cortex on costly punishment // PloS ONE. 2012. T. 7. № 9. e44747.

8. Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., Marois, R. The neural correlates of third-party punishment // Neuron. 2008. T. 60. № 5. C. 930-940.

9. Buckholtz, J.W., Marois, R. The roots of modern justice: cognitive and neural foundations of social norms and their enforcement // Nat Neurosci. 2012. T.15, № 5. C. 655-661.

10. Chang, L. J., Sanfey, A. G. Great expectations: neural computations underlying the use of social norms in decision-making // Social cognitive and affective neuroscience. 2011. T. 8. №3. C. 277–284.

11. Cialdini, R.B., Goldstein, N.J. Social influence: compliance and conformity // Annu Rev Psychol. 2004. T. 55. C. 591-621.

12. Ciaramidaro, A., Toppi, J., Casper, C., Freitag, C. M., Siniatchkin, M., Astolfi, L. Multiple-Brain Connectivity During Third Party Punishment: an EEG Hyperscanning Study // Scientific Reports. 2018. T. 8. №1. C. 6822.

13. Dickinson, D. L., Masclet, D., Villeval, M. C. Norm enforcement in social dilemmas: An experiment with police commissioners // Journal of Public Economics. 2015. T. 126. C. 74– 85.

14. Eickhoff, S., Laird, A., Grefkes, C., Wang, L., Zilles, K., Fox, P. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty // Human Brain Mapping. 2009. T. 30. № 9. C. 2907– 2926.

15. Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., Fox, P. T. Activation likelihood estimation revisited // NeuroImage. 2012. T. 59. № 3. C. 2349– 2361.

16. Eickhoff, S. B., Laird, A. R., Fox, P. M., Lancaster, J. L., Fox, P. T. Implementation errors in the GingerALE Software: Description and recommendations // Human Brain Mapping. 2017. T. 38. № 1. C. 7– 11.

17. Elster, J. Social Norms and Economic Theory // The Journal of Economic Perspectives. 1989. T. 3. № 4. C. 89–117.

18. Fehr, E., Fischbacher, U. Third-party punishment and social norms // Evol. Hum. Behav. 2004. T. 25. № 2. C. 63–87.

19. Feng, C., Luo, Y.J., Krueger, F. Neural signatures of fairness-related normative decision making in the ultimatum game: a coordinate-based meta-analysis // Hum Brain Mapp. 2015. T.36. № 2. C. 591-602.

20. Gabay, A. S., Radua, J., Kempton, M. J., Mehta, M. A. The ultimatum game and the brain: A meta-analysis of neuroimaging studies // Neuroscience and Biobehavioral Reviews. 2014. T. 47. C. 549– 558.

21. Garrigan, B., Adlam, A.L., Langdon, P.E. The neural correlates of moral decision-making: A systematic review and meta-analysis of moral evaluations and response decision judgements // Brain Cogn. 2016. T. 108. C. 88-97.

22. Guth, W., Schmittberger, R. Schwarze, B. An experimental analysis of ultimatum bargaining // Journal of Economic Behavior and Organization. 1982. T. 3. № 4. C. 367-388.

23. Hoffman, E., McCabe, K. Smith, V.L. Social Distance and Other-Regarding Behaviour in Dictator Games // American Economic Review. 1996. T. 86. C. 653-660.

24. Krueger, F. Hoffman, M. The Emerging Neuroscience of Third-Party Punishment // Trends in Neurosciences. 2016. T. 39. № 8. C. 499-501.

25. Li, C., Turmunkh, U., Wakker , P.P. Trust as a decision under ambiguity // Exp Econ. 2019. T. 22. C. 51.

26. Melchers, M., Markett, S., Montag, C., Trautner, P., Weber, B., Lachmann, B., Reuter, M. Reality TV and vicarious embarrassment: An fMRI study // Neuroimage. 2015. T. 109. C. 109– 117.

27. Montague, P. R., Lohrenz, T. To detect and correct: Norm violations and their enforcement // Neuron. 2007. T. 56. № 1. C. 14– 18.

28. Nitsche, M. A., Paulus, W. Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation // The Journal of physiology. 2000. T. 527. C. 633-639.

29. Nitsche, M.A., Paulus, W. Sustained excitability elevations induced by transcranial DC motor cortex stimulation in humans // Neurology. 2001. T. 57. № 10. C. 1899-1901.

30. Nitsche, M.A., Nitsche, M.S., Klein, C.C., Tergau, F., Rothwell, J.C., Paulus, W. Level of action of cathodal DC polarisation induced inhibition of the human motor cortex // Clin Neurophysiol. 2003. T. 114. № 4. C. 600-604.

31. Paulus, W. Transcranial electrical stimulation (tES - tDCS; tRNS, tACS) methods // Neuropsychol Rehabil. 2011. T. 21. № 5. C. 602-617.

32. Ruff, C.C., Ugazio, G., Fehr, E. Changing social norm compliance with noninvasive brain stimulation // Science. 2013. T.342. № 6157. C. 482-484.

33. Pedersen, E. J. The roles of empathy and anger in the regulation of third-party punishment // Open Access Theses. 2012. 377.

34. Riedl, K., Jensen, K., Call, J., Tomasello, M. No third-party punishment in chimpanzees // Proceedings of the National Academy of Sciences of the United States of America. 2012. T. 109. № 37. C. 14824-14829.

35. Rusinov, B.C. The functional significance of the electrical processes of the brain. M. 1977. C.363-373.

36. Sellaro, R., Güroglu, B., Nitsche, M.A., van den Wildenberg, W.P., Massaro, V., Durieux, J., Hommel, B., Colzato, L.S. Increasing the role of belief information in moral judgments by stimulating the right temporoparietal junction // Neuropsychologia. 2015. T. 77. C. 400-408.

37. Schachter, S. Deviation, rejection, and communication // Journal of Abnormal Psychology. 1951. T. 46, № 2. C. 190– 207.

38. Shelyakin, A. M. Micropolarization of the brain / A. M. Shelyakin, G. N. Ponomarenko; by ed. O. V. Bogdanova. SPb. : IIC Baltika, 2006. - 223 c.

39. Sherif, M., Sherif, C. W. Groups in harmony and tension. An integration of studies on intergroup relations // New York: Harper and Brothers. 1953.

40. Stallen, M., Rossi, F., Heijne, A., Smidts, A., De Dreu, C. K.W., Sanfey, A.G. Neurobiological Mechanisms of Responding to Injustice // J. Neurosci. 2018. T. 38. № 12. C. 2944-2954.

41. Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., Kirsch, P. Beyond revenge: Neural and genetic bases of altruistic punishment // Neuroimage. 2011. T. 54. № 1. C. 671–680.

42. Tammi, T. Dictator game giving and norms of redistribution: Does giving in the dictator game parallel with the supporting of income redistribution in the field? // The Journal of Socio-Economics. 2013. T. 43. C. 44–48.

43. Treadway, M.T., Buckholtz, J.W., Martin, J.W., Jan, K., Asplund, C.L., Ginther, M.R., Jones, O.D., Marois, R. Corticolimbic gating of emotion-driven punishment // Nat Neurosci. 2014. T. 17. № 9. C. 1270–1275.

44. Van Overwalle, F. Social cognition and the brain: a meta-analysis // Hum Brain Mapp. 2009. T. 30. № 3. C. 829-858.

45. Wagner, U., N'diaye, K., Ethofer, T., Vuilleumier, P. Guilt-specific processing in the prefrontal cortex // Cerebral Cortex. 2011. T. 21. № 11. C. 2461–2470.

46. Xiang, T., Lohrenz, T., Montague, P. R. Computational substrates of norms and their violations during social exchange // Journal of Neuroscience. 2013. T. 33. №3. C. 1099– 1108.

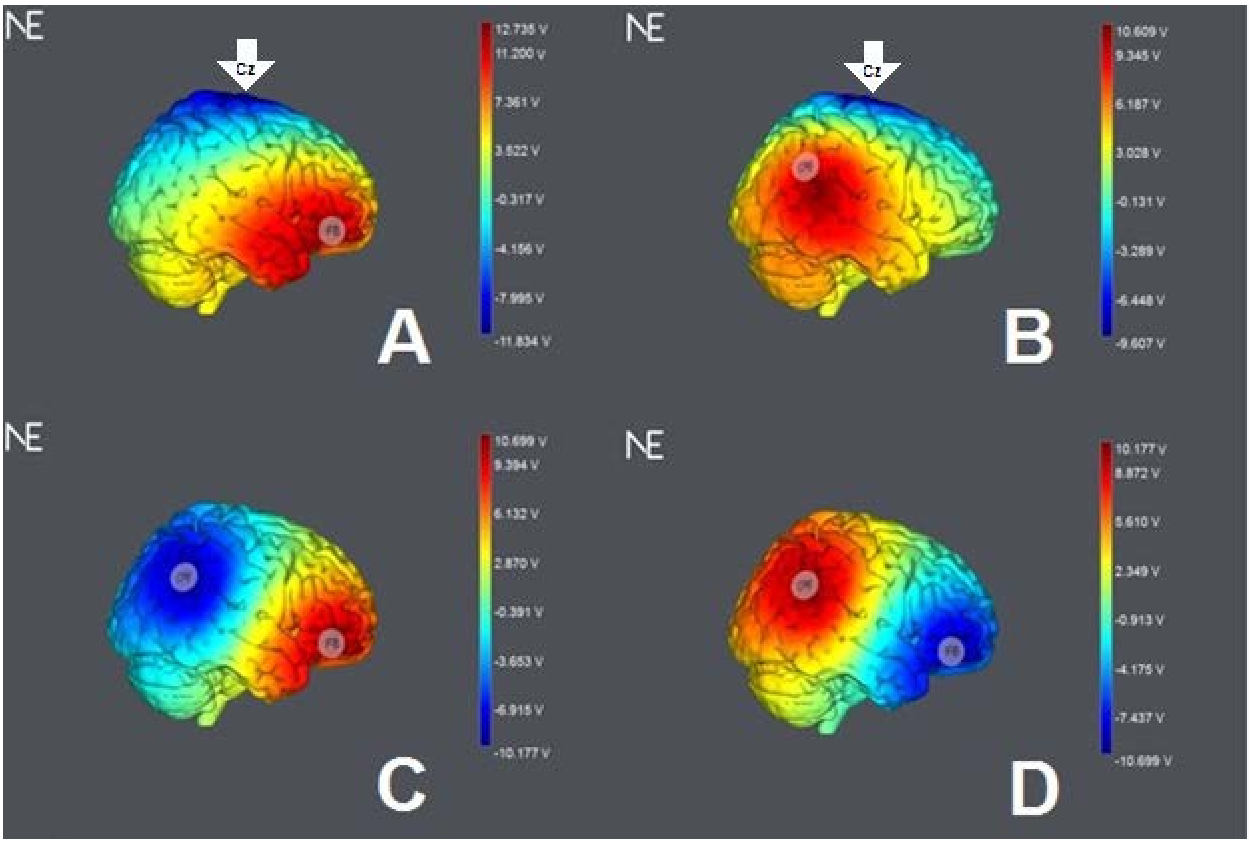
47. Zinchenko, O., Klucharev, V. Commentary: The Emerging Neuroscience of Third-Party Punishment // Frontiers in Human Neuroscience. 2017. T.11. C. 512.

48. Zinchenko O., Arsalidou M. Brain responses to social norms: Meta-analyses of fMRI studies // Hum. Brain Mapp. 2018. T. 39. № 2. C. 955-970.

49. Zinchenko, O., Belianin, A., Klucharev, V. Neurobiological Mechanisms of Fairness-Related Social Norm Enforcement: a Review of Interdisciplinary Studies // Zhurnal Vysshei Nervnoi Deyatelnosti Imeni I.P. Pavlova. 2018. T. 68. № 1. C. 16-27.

ATTACHMENTS

Figure 1. Configuration of the electrodes in the tDCS study overlaid over electric current distributions (NIC Neuroelectrics StarStim 8, Magnitude, V).



A: anodal tDCS of the rDLPFC (F8), reference electrode Cz (Condition 1.1 in paper);
B: anodal tDCS of the rTPJ (CP6), reference electrode Cz (Condition 1.2 in paper);
C: anodal tDCS of the rDLPFC (F8) and cathodal tDCS to the rTPJ (CP6) (Condition 2.1 in paper);
D: cathodal tDCS of the rDLPFC (F8) and anodal tDCS to the rTPJ (CP6) (Condition 2.2 in paper).

Figure 2. Representation of the sessions of the tDCS study (active stimulation and sham).

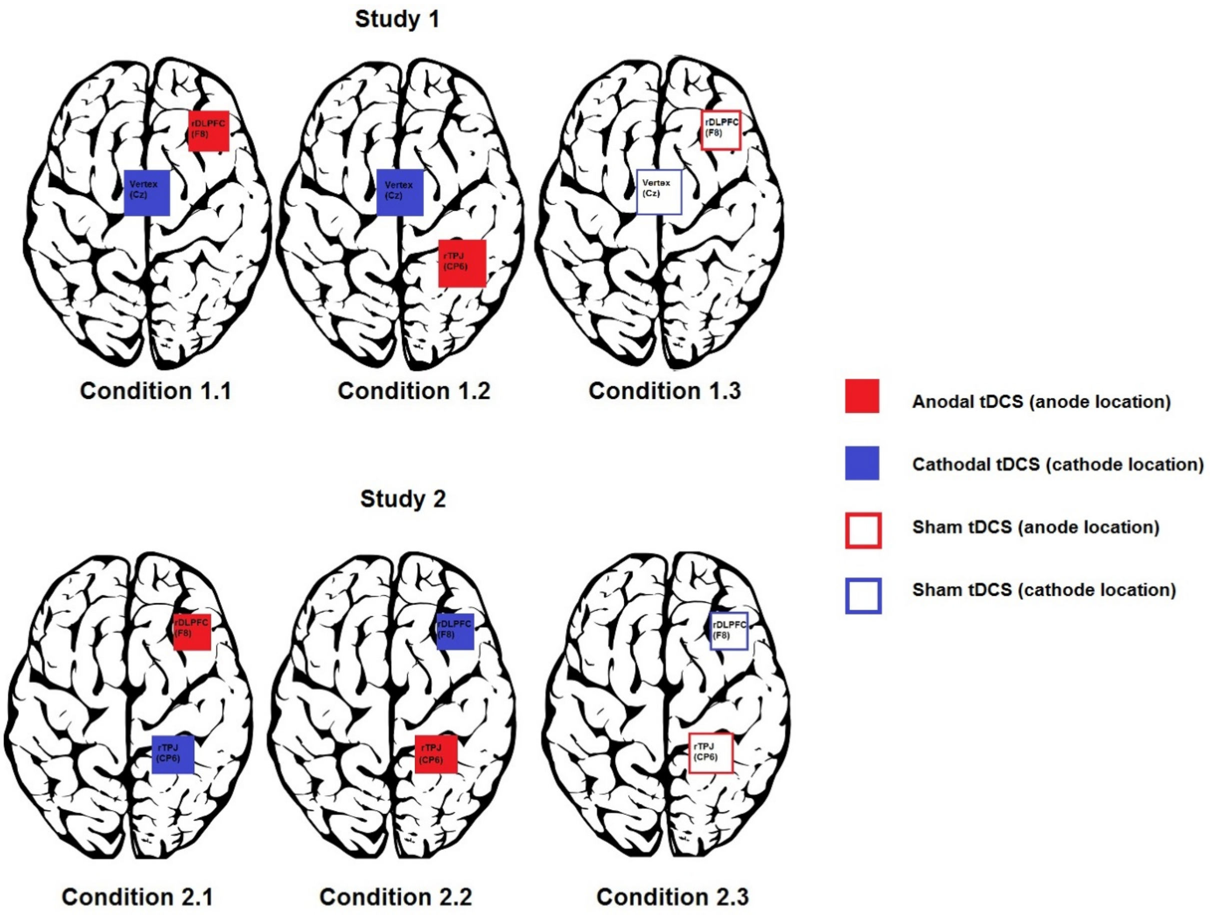
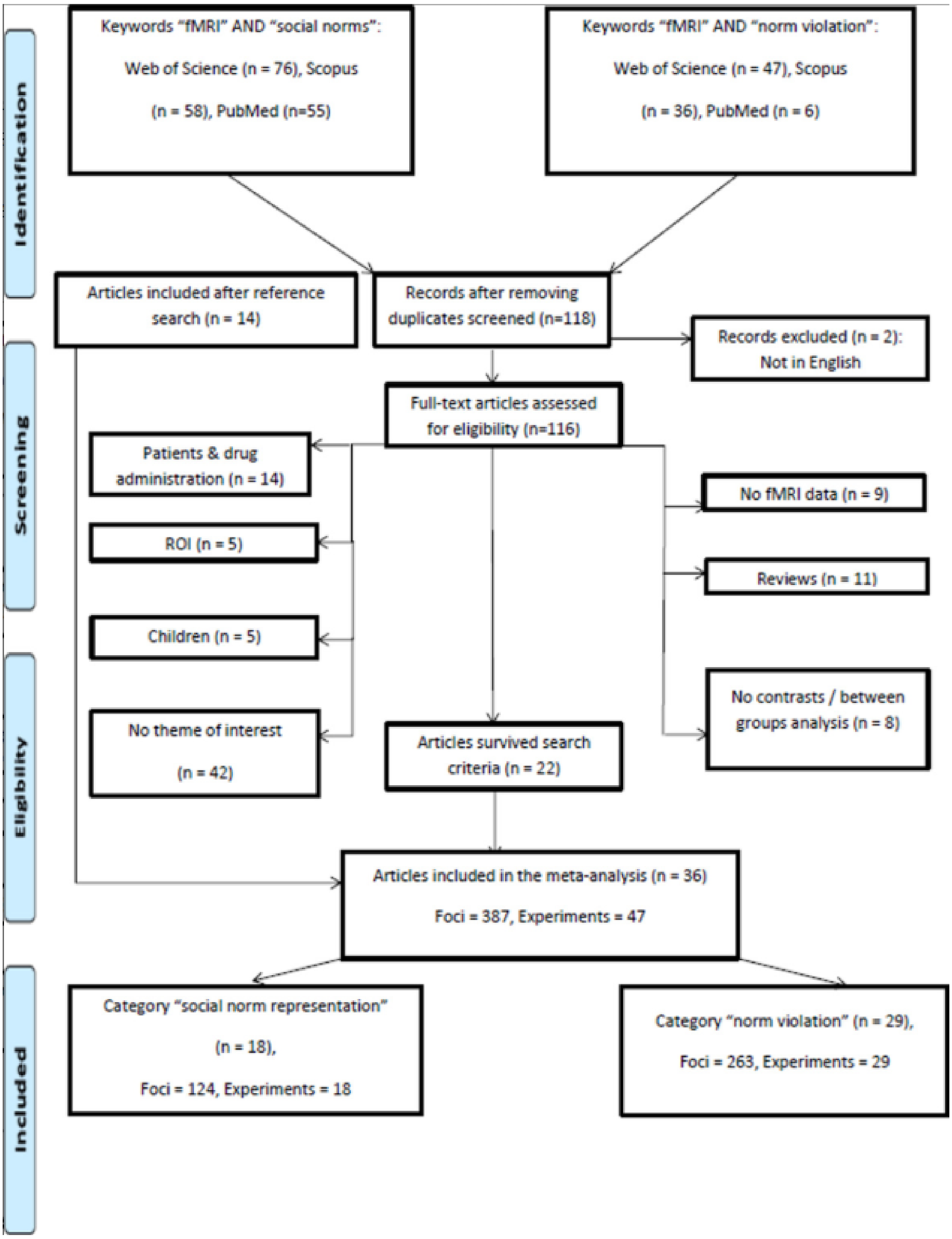


Figure 3. Literature search and step-by-step description of studies' selection for the meta-analysis.


Identification
Screening
Eligibility
Included
Keywords "fMRI" AND "social norms": Web of Science (n = 76), Scopus (n = 58), PubMed (n=55)
Keywords "fMRI" AND "norm violation": Web of Science (n = 47), Scopus (n = 36), PubMed (n = 6)
Articles included after reference search (n = 14)
Records after removing duplicates screened (n=118)
Records excluded (n = 2): Not in English
Full-text articles assessed for eligibility (n=116)
Patients & drug administration (n = 14)
ROI (n = 5)
Children (n = 5)
No theme of interest (n = 42)
No fMRI data (n = 9)
Reviews (n = 11)
No contrasts / between groups analysis (n = 8)
Articles survived search criteria (n = 22)
Articles included in the meta-analysis (n = 36) Foci = 387, Experiments = 47
Category "social norm representation" (n = 18), Foci = 124, Experiments = 18
Category "norm violation" (n = 29), Foci = 263, Experiments = 29

Note: this figure is published in *Zinchenko O. O., Arsalidou M. Brain responses to social norms: Meta-analyses of fMRI studies // Human Brain Mapping. 2018. Vol. 39. No. 2. P. 955-970*

Figure 4. The concordant map of the activations to general "social norms" category of tasks.

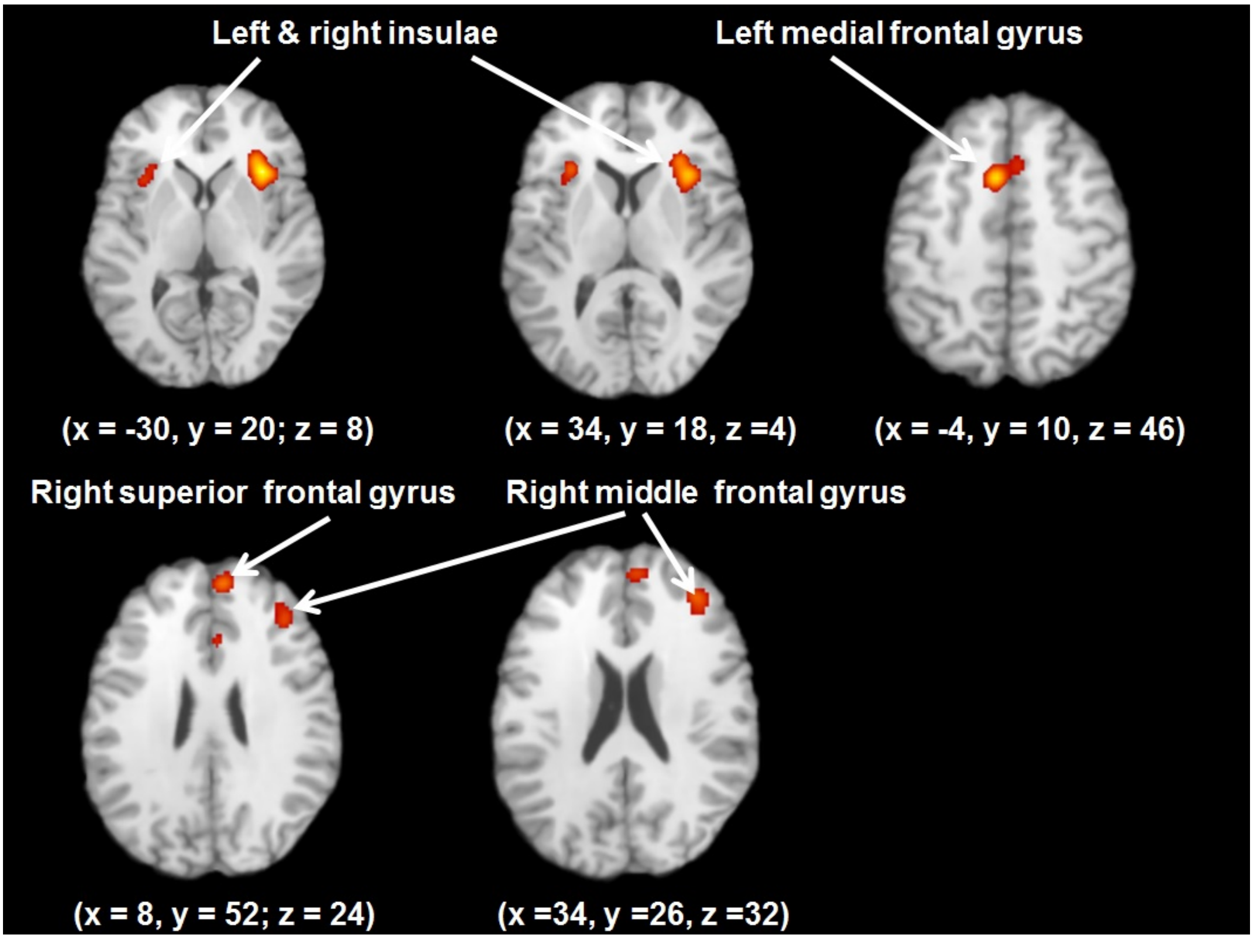


Figure 5. The concordant map of the activations to “social norm representation” category of tasks.

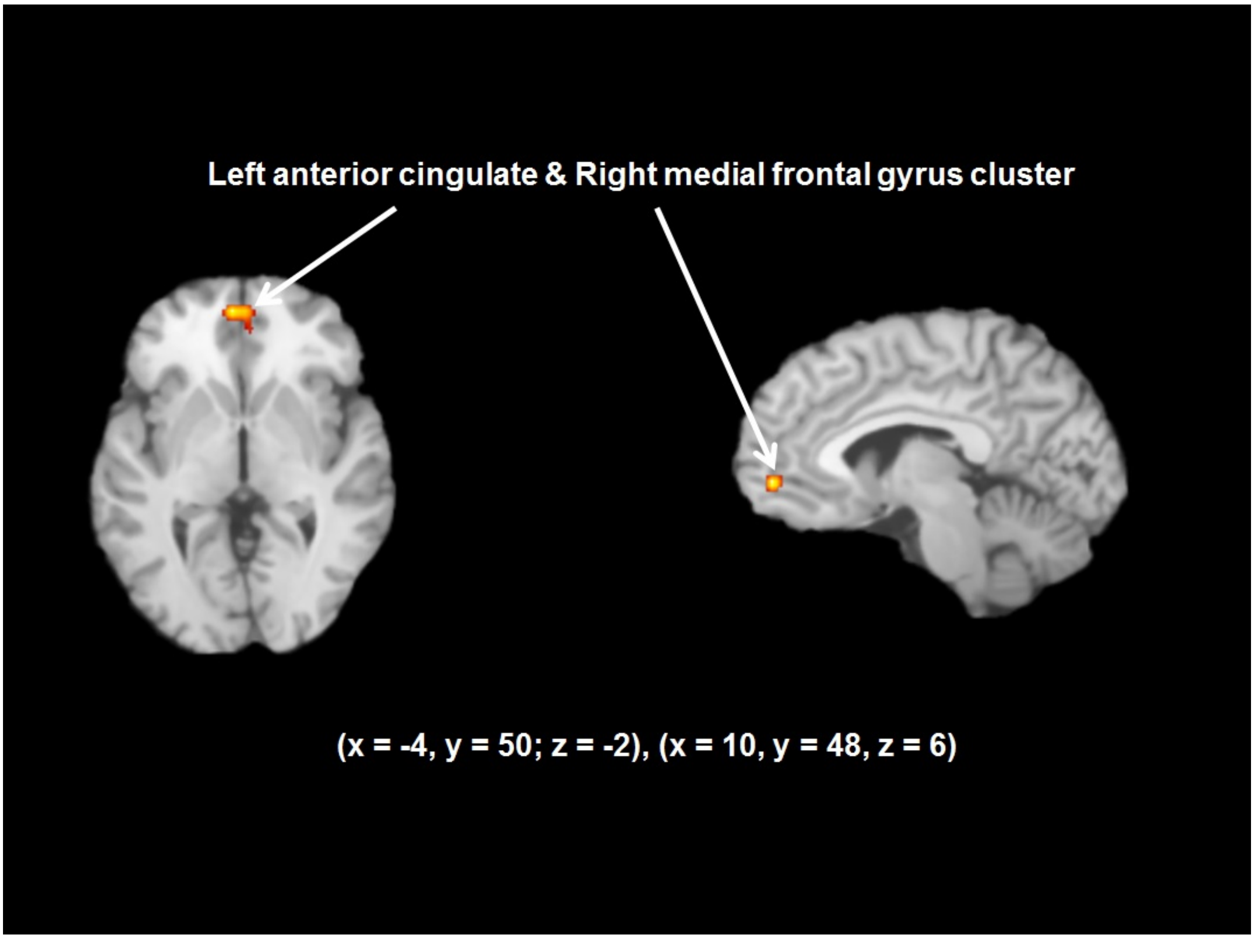

Left anterior cingulate & Right medial frontal gyrus cluster
(x = -4, y = 50; z = -2), (x = 10, y = 48, z = 6)

Figure 6. The concordant map of the activations to “norm violation” category of tasks.