frontiers in Human Neuroscience

GENERAL COMMENTARY
published: 24 October 2017
doi: 10.3389/fnhum.2017.00512

# Commentary: The Emerging Neuroscience of Third-Party Punishment

*Oksana Zinchenko*[1]* *and Vasily Klucharev*[1,2]

[1] Centre for Cognition and Decision Making, National Research University Higher School of Economics, Moscow, Russia,
[2] Department of Psychology, National Research University Higher School of Economics, Moscow, Russia

**Keywords: third-party punishment, default mode network, central-executive network, transcranial direct current stimulation, temporoparietal junction, dorsolateral prefrontal cortex, functional connectivity, social norms**

**A commentary on**

**The Emerging Neuroscience of Third-Party Punishment**
*by Krueger, F., and Hoffman, M. (2016). Trends Neurosci. 39, 499–501. doi: 10.1016/j.tins.2016.06.004*

More than a decade of neuroimaging research has established that several distinct brain networks are consistently recruited during *social punishment*, i.e., the propensity of cooperative individuals to spend some of their resources penalizing norm violators. Studies in behavioral economics have shown that social punishment can explain why genetically unrelated individuals are often able to maintain high levels of socially beneficial cooperation (Fehr and Gächter, 2002; de Quervain et al., 2004; Gureck et al., 2006). In particular, social norms can be reinforced by parties that are directly affected by norm violators ("*second parties*" punishment—SPP) and parties that are financially unaffected ("*third parties*" —TPP) (Fehr and Fischbacher, 2004). Importantly, norm violations often do not hurt other people directly. Thus, third-party sanctions are particularly effective at reinforcing group norms that regulate human behavior (Bendor and Swistak, 2001; Fehr and Fischbacher, 2004).

Pioneering behavioral studies have showed that strong emotions trigger the willingness to punish norm violators (Hirshleifer, 1987; Frank, 1988; Fehr and Gächter, 2002); in particular, TPP is motivated by both empathy toward the victim and anger toward the norm violator (Batson et al., 2007; Pedersen, 2012). Recently, neuroimaging studies have demonstrated a critical role of *executive* (the dorsolateral prefrontal cortex, DLPFC) and *mentalizing* (the temporoparietal junction, TPJ) brain regions in TPP (Baumgartner et al., 2012; Bellucci et al., 2016). Thus, neuroscience studies could help to further develop psychological theories of TPP by clarifying the specific neurocognitive mechanisms triggering punishment decisions in various social contexts.

Recently, Krueger and Hoffman (2016) reviewed and summarized the roles of three brain networks that are activated during TPP: the salience network (SN), the default mode network (DMN), and the central executive network (CEN). First, they suggested that the SN (the insula, amygdala, and dorsal anterior cingulate) detects and generates an aversive experience that initiates TPP. Second, the authors argued that the DMN (the medial prefrontal cortex, posterior cingulate cortex, and TPJ) integrates the perceived harm and inference of intentions into an assessment of blame. Finally, they proposed that the CEN (the dorsolateral prefrontal cortex and posterior parietal cortex) converts the blame signal into a specific punishment decision.

Interestingly, these three networks partially overlap with those underlying the detection of norm violations in other social contexts. There is a growing cognitive neuroscience literature on a neural mechanism that detects when individual behavior or beliefs differ from those of others (for reviews,

OPEN ACCESS

**Edited by:**
*Xiaolin Zhou,*
*Peking University, China*

**Reviewed by:**
*Hongbo Yu,*
*University of Oxford, United Kingdom*
*Matthew Ginther,*
*Court of Federal Claims, United States*

***Correspondence:**
*Oksana Zinchenko*
*ozinchenko@hse.ru*

**Received:** *10 July 2017*
**Accepted:** *09 October 2017*
**Published:** *24 October 2017*

**Citation:**
*Zinchenko O and Klucharev V (2017) Commentary: The Emerging Neuroscience of Third-Party Punishment. Front. Hum. Neurosci. 11:512. doi: 10.3389/fnhum.2017.00512*

see Izuma, 2013; Klucharev and Shestakova, 2015). A number of neuroimaging studies have demonstrated that the activity of the SN, DMN, and CEN encodes perceived deviations from group norms (Klucharev et al., 2009; Berns et al., 2010; Campbell-Meiklejohn et al., 2010; Izuma and Adolphs, 2013). In particular, the insula, dorsal anterior cingulate, medial prefrontal cortex, posterior cingulate cortex, and DLPFC have all been implicated in norm monitoring. Interestingly, many of these studies also reported norm-monitoring activity in the ventral striatum (Klucharev et al., 2009; Crockett et al., 2013; Xiang et al., 2013), which is a key region implicated in reward valuation. Despite the fact that the ventral striatum was not mentioned by Krueger and Hoffman (2016), recent studies have also implicated this region in TPP (Strobel et al., 2011; Hu et al., 2015), which further indicates that these two lines of research (detection of norm violations and TPP) share common neural mechanisms and should be further integrated.

However, amygdala activity was reported only in SPP and TPP studies (Buckholtz et al., 2008; Yu et al., 2015; Ginther et al., 2016). This can be explained by the financial losses and harms associated with this paradigm. TPJ activity also seems to be specific to the context of TPP (Baumgartner et al., 2012, 2014). A recent quantitative review suggested that the TPJ consists of functionally and spatially distinct neuroanatomical sub-regions specializing in different cognitive processes (Schurz et al., 2017). It has been hypothesized that the TPJ supports the processing of social contexts that require the representation of (a) the social context (stimuli) and (b) the context provided by attention, memory, and language (Carter and Huettel, 2013). These convergent processes constitute a theory of mind. This ability to make inferences about other people's mental states, which is associated with the TPJ, is critical to the ability to blame them for violations of complex context-dependent social norms. Thus, to uncover the neural mechanisms of TPP, it is essential to clarify the neurocomputational mechanism that allows the TPJ (as a part of the DMN) to link norm-violation detection (SN) to specific punishments (CEN).

Interestingly, TPJ activity during TPP is paralleled by an initial deactivation of the DLPFC (Buckholtz et al., 2008). This indicates functionally opposed neural activity in these two regions. The DLPFC demonstrates a biphasic neural activity—following initial deactivation, it increases activity—when subjects make the final decision to punish "based on assessed responsibility and blameworthiness" (Buckholtz et al., 2008, p. 935). Thus, it is important to explain the "antagonistic" relationship between the DMN (TPJ) and CEN (DLPFC). Many recent studies have evaluated functional and effective connectivity during SPP (Yu et al., 2015) and TPP (Treadway et al., 2014; Bellucci et al., 2016). They demonstrated that the lateral regions of the prefrontal cortex receive an input from the TPJ during SPP (Yu et al., 2015), while the dorsomedial prefrontal cortex plays the role of a hub, coordinating DLPFC and TPJ activity during the decision stage of TPP (Bellucci et al., 2016). Neuroimaging studies have demonstrated that the temporoparietal-medial-prefrontal circuit suppresses the amygdala during evaluations of unintentional harm (Treadway et al., 2014; Yu et al., 2015) in both SPP and TPP or boosts amygdala activity and strengthens its connectivity with the lateral prefrontal regions (during TPP) when a harm is intentional (Treadway et al., 2014). This suggests that the temporoparietal-medial-prefrontal circuit gates the emotional responses to norm violations and regulates subsequent reactive punishment.

These recent findings raise intriguing and testable questions for future research, e.g., in the use non-invasive brain stimulation to further verify fMRI findings. There is evidence suggesting that transcranial current stimulation could effectively modulate within- and between-network interactions. For example, transcranial alternating current stimulation induced oscillatory desynchronization between the medial frontal and parietal cortices and, therefore, affected value-based decisions but not closely matched perceptual decisions (Polanía et al., 2015). Simultaneous anodal transcranial direct current stimulation of the DLPFC, together with cathodal stimulation of the supraorbital region, led to changes in the default mode network and frontal-parietal networks (Keeser et al., 2011) and increased synchrony within the focused attention network (Peña-Gómez et al., 2012). According to Buckholtz et al. (2008), the CEN exerts an inhibitory influence over the DMN in order to program decisions about an appropriate punishment. Thus, a person could use a simultaneous application of transcranial direct or alternating current stimulation to the TPJ and DLPFC in order to modulate an antagonistic CEN/DMN interaction during TPP. We speculate that an enhancement of TPJ activity, along with the simultaneous suppression of DLPFC activity, should enhance an antagonistic CEN/DMN interaction and lead to increased TPP. The aforementioned behavioral effect should be associated with changes in the functional connectivity between the TPJ and DLPFC. A combined non-invasive brain stimulation-neuroimaging approach could further uncover the complex intrinsic network dynamics in the brain, which underlies TPP.

## AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct and intellectual contribution to the work, and approved it for publication.

## FUNDING

The study has been funded by the Russian Academic Excellence Project "5-100."

## REFERENCES

Batson, C. D., Kennedy, C. L., Nord, L. A., Stocks, E. L., Fleming, D. Y. A., Marzette, C. M., et al. (2007). Anger at unfairness: is it moral outrage? *Eur. J. Soc. Psychol.* 37, 1272–1285. doi: 10.1002/ejsp.434

Baumgartner, T., Götte, L., Gügler, R., and Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Hum. Brain Mapp.* 33, 1452–1469. doi: 10.1002/hbm.21298

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R. R., and Knoch, D. (2014). Diminishing parochialism in intergroup conflict by disrupting the

right temporo-parietal junction. *Soc. Cogn. Affect. Neurosci.* 9, 653–660. doi: 10.1093/scan/nst023

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K., et al. (2016). Effective connectivity of brain regions underlying third–party punishment: functional MRI and Granger causality evidence. *Soc. Neurosci.* 10, 1–11. doi: 10.1080/17470919.2016.1153518

Bendor, J., and Swistak, P. (2001). The evolution of norms. *Am. J. Sociol.* 106, 1493–1545. doi: 10.1086/321298

Berns, G. S., Capra, C. M., Moore, S., and Noussair, C. (2010). Neural mechanisms of the influence of popularity on adolescent ratings of music. *Neuroimage* 49, 2687. doi: 10.1016/j.neuroimage.2009.10.070

Buckholtz, J., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., et al. (2008). The neural correlates of third-party punishment. *Neuron* 60, 930–940. doi: 10.1016/j.neuron.2008.10.016

Campbell-Meiklejohn, D. K., Bach, D. R., Roepstorff, A., Dolan, R. J., and Frith, C. D. (2010). How the opinion of others affects our valuation of objects. *Curr. Biol.* 20, 1165–1170. doi: 10.1016/j.cub.2010.04.055

Carter, R. M., and Huettel, S. A. (2013). A nexus model of the temporal-parietal junction. *Trends Cogn. Sci. (Regul. Ed).* 17, 328–336. doi: 10.1016/j.tics.2013.05.007

Crockett, M., Apergis-Schoute, A., Herrmann, B., Lieberman, M., Müller, U., Robbins, T., et al. (2013). Serotonin modulates striatal responses to fairness and retaliation in humans. *J. Neurosci.* 33, 3505–3513. doi: 10.1523/JNEUROSCI.2761-12.2013

de Quervain, D. J. F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004). The neural basis of altruistic punishment. *Science* 305, 1254–1258. doi: 10.1126/science.1100735

Fehr, E., and Fischbacher, U. (2004). Third-party punishment and social norms. *Evol. Hum. Behav*. 25, 63–87. doi: 10.1016/S1090-5138(04)00005-4

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Frank, R. (1988). *Passions within Reason: The Strategic Role of the Emotions*. New York, NY: Norton.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., et al. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *J. Neurosci.* 36, 9420–9434. doi: 10.1523/JNEUROSCI.4499-15.2016

Gureck, O., Irlenbusch, B., and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science* 312, 108–111 doi: 10.1126/science.1123633

Hirshleifer, J. (1987). "On the emotions as guarantors of threats and promises," in *The Latest on the Best*, ed J. Dupre (Cambridge, MA: MIT Press), 307–326.

Hu, Y., Strang, S., and Weber, B. (2015). Helping or punishing strangers: neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Front. Behav. Neurosci.* 9:24. doi: 10.3389/fnbeh.2015.00024

Izuma, K. (2013). The neural basis of social influence and attitude change. *Curr. Opin. Neurobiol.* 23, 456–462. doi: 10.1016/j.conb.2013.03.009

Izuma, K., and Adolphs, R. (2013). Social manipulation of preference in the human brain. *Neuron* 78, 563–573. doi: 10.1016/j.neuron.2013.03.023

Keeser, D., Meindl, T., Bor, J., Palm, U., Pogarell, O., Mulert, C., et al. (2011). Prefrontal transcranial direct current stimulation changes connectivity of resting-state networks during fMRI. *J. Neurosci.* 31, 15284–15293. doi: 10.1523/JNEUROSCI.0542-11.2011

Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., and Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron* 61, 140–151. doi: 10.1016/j.neuron.2008.11.027

Klucharev, V., and Shestakova, A. (2015). "Social influence and persuasion and message propagation," in *Brain Mapping: An Encyclopedic Reference*, ed A. W. Toga (San Diego, CA: Academic Press), 251–258. doi: 10.1016/B978-0-12-397025-1.00189-5

Krueger, F., and Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends Neurosci.* 39, 499–501. doi: 10.1016/j.tins.2016.06.004

Peña-Gómez, C., Sala-Lonch, R., Junqué, C., Clemente, I. C., Vidal, D., Bargalló, N., et al. (2012). Modulation of large-scale brain networks by transcranial direct current stimulation evidenced by resting-state functional MRI. *Brain Stimul.* 5, 252–263. doi: 10.1016/j.brs.2011.08.006

Pedersen, E. J. (2012). *The Roles of Empathy and Anger in the Regulation of Third-Party Punishment*. Open Access Theses. 377. Available online at: http://scholarlyrepository.miami.edu/oa_theses/377

Polanía, R., Moisa, M., Opitz, A., Grueschow, M., and Ruff, C. C. (2015). The precision of value-based choices depends causally on fronto-parietal phase coupling. *Nat. Commun.* 6:8090. doi: 10.1038/ncomms9090

Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., and Sallet, J. (2017). Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: a review using probabilistic atlases from different imaging modalities. *Hum. Brain Mapp*. 38, 4788–4805. doi: 10.1002/hbm.23675

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., et al. (2011). Beyond revenge: neural and genetic bases of altruistic punishment. *Neuroimage* 54, 671–680. doi: 10.1016/j.neuroimage.2010.07.051

Treadway, M. T., Buckholtz, J. W., Martin, J. W., Jan, K., Asplund, C. L., Ginther, M. R., et al. (2014). Corticolimbic gating of emotion-driven punishment. *Nat. Neurosci.* 17, 1270–1275. doi: 10.1038/nn.3781

Xiang, T., Lohrenz, T., and Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *J. Neurosci.* 33, 1099a−1108a. doi: 10.1523/JNEUROSCI.1642-12.2013

Yu, H., Li, J., and Zhou, X. (2015). Neural substrates of intention–consequence integration and its impact on reactive punishment in interpersonal transgression. *J. Neurosci.* 35, 4917–4925. doi: 10.1523/JNEUROSCI.3536-14.2015

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.