frontiers in Neuroscience

GENERAL COMMENTARY
published: 29 June 2017
doi: 10.3389/fnins.2017.00374

Commentary: Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment

*Anne-Marie Nußberger[1], Mary Montgomery[1], Yingyi Luo[2,3] and Hongbo Yu[1,4]**

[1] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom, [2] Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China, [3] Faculty of Science and Engineering, Waseda University, Tokyo, Japan, [4] Center for Brain and Cognitive Sciences, Peking University, Beijing, China

Keywords: third-party punishment, fMRI neuroimaging, intention, consequences of actions, multivariate pattern analysis

A commentary on

Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment
by Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., et al. (2016). J. Neurosci. 36, 9420–9434. doi: 10.1523/JNEUROSCI.4499-15.2016

OPEN ACCESS

Edited by:
Bernd Weber,
University of Bonn, Germany

Reviewed by:
Frank Krueger,
George Mason University, United States

****Correspondence:***
Hongbo Yu
hongbo.yu@psy.ox.ac.uk

Specialty section:
This article was submitted to Decision Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *27 February 2017*
Accepted: *15 June 2017*
Published: *29 June 2017*

Citation:
Nußberger A-M, Montgomery M, Luo Y and Yu H (2017) Commentary: Parsing the Behavioral and Brain Mechanisms of Third-Party Punishment. Front. Neurosci. 11:374. doi: 10.3389/fnins.2017.00374

Third-party punishment (TPP) is an important safeguard of social cooperation (Fehr and Gächter, 2002; Buckholtz and Marois, 2012). Examples range from a teacher sending their students out of classroom because they misbehaved, to a judge sending a person criminal to prison because they committed murder. Investigating the psychological and neural mechanisms of TPP decisions has significant practical implications, such as advising when and how third parties (e.g., judges) might be susceptible to affective and cognitive constraints inherent to human nature (Krueger and Hoffman, 2016).

Two types of information have been identified as crucial for TPP decisions, namely the *mental state* of the suspect and the consequential *harm* caused to the victim (Cushman, 2008; Schein and Gray, in press). Previous research has suggested that brain areas associated with mentalizing and affective processing are recruited during TPP, but determining the specific contribution of each of these areas has been a challenge. This is partly due to limitations of using scenario-based paradigms in fMRI (Buckholtz et al., 2008; Treadway et al., 2014). Furthermore, given the relatively low temporal resolution of fMRI (Serences, 2004), it has proven difficult to dissociate signals related to the respective processing of mental state and harm from signals related to their integration and translation into a specific punishment, as previous studies presented information related to these different aspects all at a time (an alternative solution to the problem of low temporal resolution is using electroencephalogram, see Yoder and Decety, 2014; Hesse et al., 2016). Moreover, prior designs have usually included only two levels of mental states (intentional vs. unintentional) and damage (harm vs. no harm), whereas both factors could vary greatly in real-world settings. Where different magnitudes of harm were actually included (e.g., Buckholtz et al., 2008; Krueger et al., 2014), formal analyses of the neural correlates for the varying levels of harm have been missing.

In a recent publication, Ginther et al. (2016) address these methodological challenges by introducing three key innovations to TPP research. First, their design effectively orthogonalizes three crucial components of TPP judgments,—namely harm and mental state processing, their integration, and the ultimate punishment decision—by presenting them at separate stages of a trial (Ginther et al., 2016, Figure 1). This is in contrast to previous scenario-based studies on TPP or moral judgment where these stages were not separated (Greene et al., 2004; Buckholtz et al., 2008; Young and Saxe, 2008). Additionally, the authors balanced the presentation order within

each participant to dissociate each component's processing from their integration, which inevitably arises with sequential presentations of mental state and harm information.

Second, a parametric manipulation of mental state and harm allowed the authors to characterize response-profiles of brain areas sensitive to the respective components more precisely. For instance, activations in orbitofrontal cortex (OFC) and dorsolateral prefrontal cortex (DLPFC) were higher for harm than for mental state evaluations. However, activations in OFC were best accounted for by a quadratic relationship with level of harm suggesting that OFC activity reflects decision-difficulty (being greater for intermediate levels of harm). In contrast, activations in DLPFC were best accounted for by a linear response to level of harm (Ginther et al., 2016, Tables 3,4) which may reflect culpability as a different decision-aspect. Ginther et al.'s findings also indicate a possible need to reinterpret previous neuroimaging findings that have been solely based on main effect contrasts (e.g., Harm > No-Harm, Intentional > Unintentional, see Buckholtz et al., 2008; Young and Saxe, 2008; Treadway et al., 2014; Yu et al., 2015), as such contrast may miss out important and interesting response profiles.

Third, multivariate pattern analysis (MVPA) was used to dissociate spatially overlapping neural ensembles that serve different functions. In line with previous research (Buckholtz et al., 2008; Young and Saxe, 2008; Treadway et al., 2014), the authors found mental state and harm evaluations were associated with response in bilateral superior temporal sulcus (STS) and temporal parietal junction (TPJ)—core areas linked to representing others' cognitive and affective states (Van Overwalle, 2009). Crucially, by using MVPA, Ginther et al. could show that activation patterns in both TPJ and STS are differentially associated with mental state and harm evaluations (cf. Ginther et al., 2016, p. 9,428). This is an important advance over previous studies using univariate analyses that may not have the sensitivity to dissociate spatially overlapping but functionally distinct neural ensembles (Woo et al., 2014).

Moreover, MVPA proved useful in clarifying the function of DLPFC in punishment decisions. In line with prior studies that have suggested a consistent relationship between DLPFC and TPP (Knoch et al., 2006; Buckholtz et al., 2015), the authors found increased activation in right DLPFC during the decision stage. However, in contrast to previous findings (Buckholtz et al., 2008), activation strength in DLPFC did not correlate with level of punishment. The results from the MPVA offered a potential explanation for this, as they showed that level of punishment predicted *patterns* of DLPFC neural activity, rather than activation *strength* (Ginther et al., 2016; Figures 6B,C).

Moving forward, the design devised by Ginther et al. would also allow investigating how sequential presentation-orders of mental state and harm information may influence punishment decisions. It has been demonstrated that the way of presenting an action's consequences may influence the judgment of mental states underlying that action, a phenomenon named the "Knobe Effect" (Knobe, 2003). Thus, future studies could test how differences in narrative frames modulate punishment decision-formation. In regard to neuroimaging methodology and analysis, future studies could further broaden our understanding of the neural mechanisms in TPP by focusing more on the functional and effective brain networks involved in moral/legal judgment (e.g., Bellucci et al., 2017), using techniques such as psychophysiological interaction and Granger causality modeling. Another avenue future studies could pursue, is to combine the novel design presented by Ginther et al. with brain lesion or virtual lesion (e.g., transcranial magnetic stimulation) approaches to investigate the causal role of specific brain areas in TPP (e.g., Buckholtz et al., 2015; Glass et al., 2016). The separation of different processing stages could be further enhanced by adopting techniques that allow for higher temporal resolutions (e.g., electroencephalography or magnetoencephalography). Put in context with paradigms adopted in previous neuroimaging studies of TPP, a systematic meta-analysis comparing scenario-based (e.g., Buckholtz et al., 2008; Treadway et al., 2014; Ginther et al., 2016) and interaction-based (e.g., Feng et al., 2016) designs could shed light on the question in how far neurobiological processes are common across paradigms, or paradigm-specific and as such potentially not genuine signatures of TPP.

Overall, the work by Ginther et al. expands the scope of neuroscientific research on TPP by (i) insightfully revising the scenario-based paradigm such that different processing stages can be separated and by (ii) introducing sophisticated data analysis techniques to better characterize encoding profiles of relevant brain areas. The novel findings from this study provide empirical evidence for numerous theoretical accounts of the neural basis of TPP and raise intriguing and testable questions for future research: How do people move from integrated mental state and harm information to a definitive punishment choice? How do the affective components of harm evaluation, as well as our pre-existing social norms, bias the neural encoding of mental states (e.g., Knobe, 2003)? And what kind of scenario narration (in verbal or non-verbal form) could minimize such biases in the legal system? We believe that seeking answers to these questions will not only advance our understanding of the neurobiological basis of TPP, but also of other psychological faculties that both enable, and are cultivated by, the concept of justice (Rawls, 1971).

AUTHOR CONTRIBUTIONS

AN, MM, YL, and HY have written the manuscript, AN and HY have revised the mansucript.

ACKNOWLEDGMENTS

We would like to thank Dr. Molly Crockett, Dr. Patricia Lockwood, and the reviewer for their helpful comments on our manuscript. HY is supported by a Newton International Fellowship from the British Academy (NF160700).

REFERENCES

Bellucci, G., Chernyak, S., Hoffman, M., Deshpande, G., Dal Monte, O., Knutson, K. M., et al. (2017). Effective connectivity of brain regions underlying third-party punishment: functional MRI and Granger causality evidence. *Soc. Neurosci.* 12, 124–134. doi: 10.1080/17470919.2016.1153518

Buckholtz, J. W., and Marois, R. (2012). The roots of modern justice: cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* 15, 655–661. doi: 10.1038/nn.3087

Buckholtz, J. W., 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

Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., et al. (2015). From blame to punishment: disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron* 87, 1369–1380. doi: 10.1016/j.neuron.2015.08.023

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 353–380. doi: 10.1016/j.cognition.2008.03.006

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

Feng, C., Deshpande, G., Liu, C., Gu, R., Luo, Y.-J., and Krueger, F. (2016). Diffusion of responsibility attenuates altruistic punishment: a functional magnetic resonance imaging effective connectivity study. *Hum. Brain Mapp.* 37, 663–677. doi: 10.1002/hbm.23057

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

Glass, L., Moody, L., Grafman, J., and Krueger, F. (2016). Neural signatures of third-party punishment: evidence from penetrating traumatic brain injury. *Soc. Cogn. Affect. Neurosci.* 11, 253–262. doi: 10.1093/scan/nsv105

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., and Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44, 389–400. doi: 10.1016/j.neuron.2004.09.027

Hesse, E., Mikulan, E., Decety, J., Sigman, M., Garcia, M. D. C., Silva, W., et al. (2016). Early detection of intentional harm in the human amygdala. *Brain* 139, 54–61. doi: 10.1093/brain/awv336

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis* 63, 190–194. doi: 10.1093/analys/63.3.190

Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., and Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science* 314, 829–832. doi: 10.1126/science.1129156

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

Krueger, F., Hoffman, M., Walter, H., and Grafman, J. (2014). An fMRI investigation of the effects of belief in free will on third-party punishment. *Soc. Cogn. Affect. Neurosci.* 9, 1143–1149. doi: 10.1093/scan/nst092

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Schein, C., and Gray, K. (in press). The theory of dyadic morality: reinventing moral judgment by redefining harm. *Pers. Social Psychol. Rev.* doi: 10.1177/1088868317698288

Serences, J. T. (2004). A comparison of methods for characterizing the event-related BOLD timeseries in rapid fMRI. *Neuroimage* 21, 1690–1700. doi: 10.1016/j.neuroimage.2003.12.021

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

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp.* 30, 829–858. doi: 10.1002/hbm.20547

Woo, C.-W., Koban, L., Kross, E., Lindquist, M. A., Banich, M. T., Ruzic, L., et al. (2014). Separate neural representations for physical pain and social rejection. *Nat. Commun.* 5, 5380. doi: 10.1038/ncomms6380

Yoder, K. J., and Decety, J. (2014). Spatiotemporal neural dynamics of moral judgment: a high-density ERP study. *Neuropsychologia* 60, 39–45. doi: 10.1016/j.neuropsychologia.2014.05.022

Young, L., and Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage* 40, 1912–1920. doi: 10.1016/j.neuroimage.2008.01.057

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.