Toward a Brain-Based Bio-Marker of Guilt

Hongbo Yu[1], Leonie Koban[2,3], Molly J. Crockett[4], Xiaolin Zhou[5,6,7,8,9] and Tor D. Wager[2,10]

[1]Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, California, United States of America. [2]Institute of Cognitive Science, University of Colorado, Boulder, CO, USA. [3]Department of Psychology and Neuroscience, University of Colorado, Boulder, CO, USA. [4]Department of Psychology, Yale University, New Haven, Connecticut, United States of America. [5]School of Psychological and Cognitive Sciences, Peking University, Beijing, China. [6]Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China. [7]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China. [8]Institute of Psychological and Brain Sciences, Zhejiang Normal University, Zhejiang, China. [9]Key Laboratory of Applied Brain and Cognitive Sciences, School of Business and Management, Shanghai International Studies University, Shanghai, China. [10]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire, United States of America.

Neuroscience Insights
Volume 15: 1–3

DOI: 10.1177/2633105520957638
SAGE

ABSTRACT: Guilt is a quintessential emotion in interpersonal interactions and moral cognition. Detecting the presence and measuring the intensity of guilt-related neurocognitive processes is crucial to understanding the mechanisms of social and moral phenomena. Existing neuroscience research on guilt has been focused on the neural correlates of guilt states induced by various types of stimuli. While valuable in their own right, these studies have not provided a sensitive and specific bio-marker of guilt suitable for use as an indicator of guilt-related neurocognitive processes in novel experimental settings. In a recent study, we identified a distributed Guilt-Related Brain Signature (GRBS) based on 2 independent functional MRI datasets. We demonstrated the sensitivity of GRBS in detecting a critical cognitive antecedent of guilt, namely one's responsibility in causing harm to another person, across participant populations from 2 distinct cultures (ie, Chinese and Swiss). We also showed that the sensitivity of GRBS did not generalize to other types of negative affective states (eg, physical and vicarious pain). In this commentary, we discuss the relevance of guilt in the broader scope of social and moral phenomena, and discuss how guilt-related biomarkers can be useful in understanding their psychological and neurocognitive mechanisms underlying these phenomena.

KEYWORDS: Guilt, biomarker, function MRI, social cognition, morality

RECEIVED: August 14, 2020. **ACCEPTED:** August 20, 2020.

TYPE: Commentary

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Hongbo Yu, Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106, USA. Email: hongbo.yu@psych.ucsb.edu; Twitter: @AbrahamYule

COMMENT ON: Yu H, Koban L, Chang LJ, Wagner U, et al. A generalizable multivariate brain pattern for interpersonal guilt. *Cereb Cortex*. 2020;30:3558-3572. doi:10.1093/cercor/bhz326. PubMed PMID: 32083647; PubMed Central PMCID: PMC7232998.

Guilt as a multifaced concept

Guilt, like many other social emotions, is a multifaced psychological construct and is often used equivocally in everyday life. Hurting an innocent person is a paradigmatic scenario in which people feel and express guilt.[1,2] However, even in this case, we may not be dealing with one single *kind* of guilt—it is an open question whether an initial intention to harm influences the quality and magnitude of the guilt an agent later experiences.[3] When we shift our focus to non-social use of the term "guilt," we will see even more diversity and complexity.[4] For example, guilt appeal has been used as an advertising strategy for healthy diets. Some snack brands, instead of using label such as "reduced fat" or "reduced calories" for high-fat, high-calorie food products, directly label them as "reduced guilt" in order to ease customers' worries about the healthfulness of those products.[5] We feel and express guilt when we fail to live up to our personal goals that are not directly related to other individuals or moral norms, such as keeping a healthy diet, working hard for an exam, and physical exercise. Indeed, people report experiencing guilt in their everyday life over almost all the domains of moral violations proposed in the Moral Foundations Theory,[6,7] including harm, unfairness, disloyalty, subversion, degradation, dishonesty, and lack of self-restraint. In fact, violation of self-restraint elicits stronger guilty feelings (on a 5-point Likert scale) than violation of fairness (mean difference = 1.40, SE = 0.21, z-ratio = 6.69, $P < .001$) and violation of honesty principles (mean difference = 0.87, SE = 0.17, z-ratio = 5.22, $P < .001$) (Figure 1), according to a large-scale experience sampling survey.[7]

What strategies should we take to investigate the neurocognitive basis of guilt in the face of its conceptual complexity? An analogy with pain, another multifaceted concept, may be useful to illustrate the approach we are introducing here. In social neuroscience, it has been debated whether physical pain, bodily sensation induced by external nociceptive stimuli, and social "pain" – psychological anguish elicited by social isolation, rejection or empathy – share the same neurocognitive basis. Studies examining the blood-oxygen-level-dependent (BOLD) signals that correlates with each phenomenon have consistently shown overlapping brain areas elicited by physical pain and social "pain."[8] However, the relatively low spatial resolution of BOLD signal hinders the inference from overlapping activations to

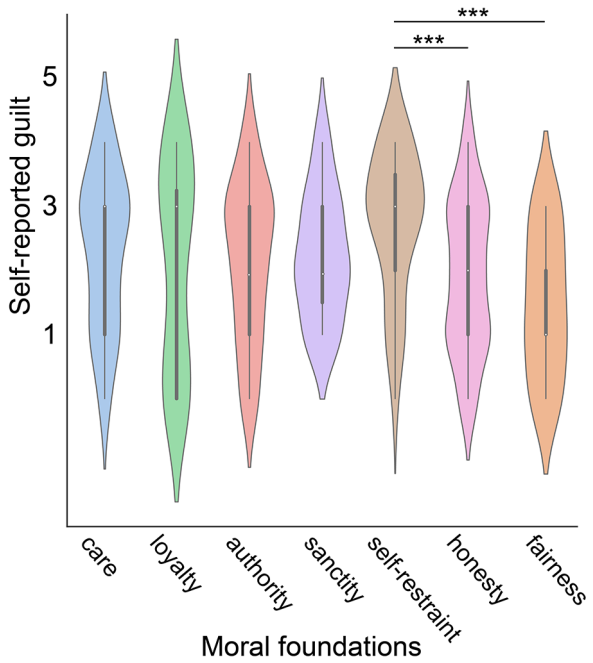

Figure 1. Self-reported guilt following committing an immoral act in everyday life as a function of moral foundations.
***$P < .001$.

overlapping neural representations or psychological constructs.[9] An alternative approach is to develop a multivariate brain-based signature (or bio-marker) of each construct. The idea here is that if the bio-marker of physical pain does not respond to social "pain" and vice versa, then these 2 constructs do not share the same neural representation.[10]

Developing a guilt-related brain signature

Inspired by this approach, we recently identified a multivariate brain-based signature of guilt based on a paradigmatic case of guilt—causing harm to an innocent person.[11] We trained and validated the signature on 2 fMRI datasets. In the training dataset (N = 24, Chinese population), participants and an anonymous co-player performed a perceptual task, where failure would cause pain to the co-player.[12] We induced guilt by manipulating the responsibility of the participants in causing the pain. Specifically, if a participant performed poorly and the co-player performed well, then the performance failure, and the resulting co-player's pain, was caused by the participant. In comparison, if both the participants and the co-player performed poorly, then both of them were responsible for the co-player's pain. Behaviorally, both self-reported guilt and reparation were positively correlated with participants' responsibility. We trained a multivariate Support-Vector-Machine (SVM) classifier to dissociate the sole-responsible and the both-responsible conditions. This classifier, or guilt-related brain signature (GRBS), was not only able to discriminate the 2 conditions on which it was trained, it was also able to discriminate the sole-responsible condition with other closely matched control conditions in the training dataset. Moreover, the predictive power of GRBS was generalizable to an independent test dataset (N = 19; Swiss population) that adopted a similar interpersonal action-monitoring task.[13] We further demonstrated that GRBS did not discriminate different levels of painful thermal stimulation or different degree of vicarious pain, nor did it differentiate recalled guilt from recalled sadness or shame induced by person-specific episodes. Together, these results demonstrate that GRBS satisfies the 3 criteria proposed for bio-markers: sensitivity, specificity, and generalizability.[14] Specifically, GRBS: (1) detects the presence of the "cognitive antecedents" of guilt in social interactions, here operationalized as responsibility; (2) does not discriminate other types of negative experiences, such as physical pain and emotion memory; and (3) detects the presence of the cognitive antecedent on which the signature is trained are present in studies and samples other than the training sample.

Using GRBS as an indicator of guilt-related neurocognitive processes in social-moral decision-making

Guilt-related neurocognitive processes are involved in many social-moral decision-making contexts. However, agents in those contexts are not always aware of or have biases in reporting guilt and guilt-related processes. In these situations, GRBS has the potential to provide an implicit, brain-based measure of guilt-related neurocognitive processes that are not easily forged by the agents. Returning to the self-restraint failure example, one theoretically important question is when people claim that they feel guilty for eating too much or for not working hard enough, are the neurocognitive processes underlying this affective phenomenon the same one as when they feel guilty for hurting their partner? If GRBS could discriminate self-restraint failure from self-restraint success, then we would be more confident that we are talking about the same *kind* of emotion in interpersonal and intrapersonal scenarios.[4]

Guilt is also relevant to moral evaluations of actions and character. When evaluating the moral status of an action or the moral character of an agent, the agent's inner states, such as attention and emotion accompanying their action, usually play an integral role.[15] Take hypocrisy as an example. A commonly held conceptualization characterizes a hypocrite as someone whose behaviors fall short of their expressed attitudes regarding some moral standards, namely "saying one thing, doing another."[16] Note, however, that this conceptualization speaks only to observable behaviors, irrespective of the mental states of the agent who behaves this way. Some philosophers, however, make the distinction between deceptive and *akratic* hypocrite.[17]

Deceptive hypocrites "appears moral while, if possible, avoiding the cost of actually being moral."[18] These hypocrites do not genuinely care about the moral standards that they publicly preach or cite to blame others, and therefore deserve the moral objections that laypeople assign to hypocrites.[16] *Akratic* hypocrites, on the other hand, do genuinely care about the moral standards that they preach, but occasionally give in to temptations at the time of decision-making, perhaps due to weakness-of-the-will. A hallmark of

akratic hypocrites, therefore, is their feelings of conflict and guilt when they realize that what they do violates the moral standards they genuinely believe to be relevant and valuable.[17] Judging and treating deceptive and *akratic* hypocrites differently according to their mental states (ie, moral conflict, guilt) seems fairer and leaves room for moral education and self-improvement.[19] Behavioral measures alone are difficult, if not impossible, to distinguish these 2 types of hypocrites, because self-reported conflicted feelings and guilt can be easily faked. Applying the GRBS to neural response patterns associated with moral decision-making may offer a way to gauge the guilt-related neurocognitive processes involved and therefore provides a way to characterize the extent of deceptive versus *akratic* hypocrisy.[20]

Understanding the diversity and complexity of guilt via the brain-based signature approach

There are some limitations to GRBS that are worth noting. First, GRBS was trained on the datasets where the experimental designs emphasized the detection of cognitive antecedents of guilt (ie, responsibility) rather than sustained feelings of guilt. Therefore, GRBS performed at chance level in predicting post-task self-reported guilt.[11] To develop a brain-based signature more sensitive to the experiential component of guilt, future studies should adopt experimental tasks that allow the participants to interact with or confronted by the victims whom they harm, in reality or virtual reality.[21]

Another way to extend the research on biomarkers of guilt is to develop brain-based signatures that are sensitive to other modes of guilt that do not directly involve agency or responsibility. For example, survivors of natural disasters or human atrocities often report intense guilty feelings for other victims who do not survive or suffer more seriously.[22] Some individuals with severe depression express feeling guilty for their mere existence in the world.[23] Descendants and fellow citizens of former prosecutors (eg, war criminals, human right abusers, etc.) are deeply concerned about the crimes that their ancestors or ingroup members, but not themselves, are responsible.[24] When someone carries out harm under coercion, are guilt-related neurocognitive processes suppressed due to their diminished sense of agency?[25] Ascertaining the resemblance between GRBS and the neural representations of these various modes of guilt experiences, and developing brain-based signatures for those other modes of guilt, will advance our understanding of structure and taxonomy of this complex affective phenomenon.

Authors' Note

Leonie Koban is also affiliated with INSEAD, Fontainebleau, France and Paris Brain Institute, Paris, France.

Author Contributions

HY wrote the first draft of the paper. LK, MJC, XZ and TDW edited the manuscript.

ORCID iD

Hongbo Yu https://orcid.org/0000-0002-3384-7772

REFERENCES

1. Baumeister RF, Stillwell AM, Heatherton TF. Guilt: an interpersonal approach. *Psychol Bull*. 1994;115:243-267. doi:10.1037/0033-2909.115.2.243.
2. Tangney JP, Dearing RL. *Shame and Guilt*. New York: The Guilford Press; 2003.
3. McGraw KM. Guilt following transgression: an attribution of responsibility approach. *J Pers Soc Psychol*. 1987;53:247-256. doi:10.1037/0022-3514.53.2.247.
4. McGee D, Giner-Sorolla R. How guilt serves social functions from within [published online ahead of print 2019]. *Moral Psychol Guilt*.
5. Hur J, Jang SS. Anticipated guilt and pleasure in a healthy food consumption context. *Int J Hosp Manag*. 2015;48:113-123.
6. Graham J, Haidt J, Koleva S, et al. Moral foundations theory: the pragmatic validity of moral pluralism. In: Devine P, Plant A, ed. *Advances in Experimental Social Psychology*. Vol. 47. Cambridge: Academic Press; 2013:55-130.
7. Hofmann W, Wisneski DC, Brandt MJ, Skitka LJ. Morality in everyday life. *Science*. 2014;345:1340-1343.
8. Eisenberger NI. The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nat Rev Neurosci*. 2012;13:421-434.
9. Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*. 2006;10:424-430.
10. Woo C-W, Koban L, Kross E, et al. Separate neural representations for physical pain and social rejection. *Nat Commun*. 2014;5:5380.
11. Yu H, Koban L, Chang LJ, et al. A generalizable multivariate brain pattern for interpersonal guilt. *Cereb Cortex*. 2020;30:3558-3572.
12. Yu H, Hu J, Hu L, Zhou X. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc Cogn Affect Neurosci*. 2014;9:1150-1158. doi:10.1093/scan/nst090.
13. Koban L, Corradi-Dell'Acqua C, Vuilleumier P. Integration of error agency and representation of others' pain in the anterior insula. *J Cogn Neurosci*. 2013;25:258-272.
14. Woo C-W, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. 2017;20:365.
15. Stohr KE. Moral cacophony: when continence is a virtue. *J Ethics*. 2003;7:339-363.
16. Laurent SM, Clark BAM. What makes hypocrisy? Folk definitions, attitude/behavior combinations, attitude strength, and private/public distinctions. *Basic Appl Soc Psych*. 2019;41:104-121.
17. Bartel C. Hypocrisy as either deception or akrasia. *Philos Forum*. 2019;50:269-281.
18. Batson CD, Thompson ER. Why don't moral people act morally? Motivational considerations. *Curr Dir Psychol Sci*. 2001;10:54-57. doi:10.1111/1467-8721.00114.
19. Bommarito N. *Inner Virtue*. New York: Oxford University Press; 2017.
20. Yu H, Contreras-Huerta LS, Prosser AMB, Apps MAJ, Hofmann W, Sinnott-Armstrong W, Crockett MJ. The conflicted conscience of hypocrites (under review).
21. Yu H, Duan Y, Zhou X. Guilt in the eyes: eye movement and physiological evidence for guilt-induced social avoidance. *J Exp Soc Psychol*. 2017;71:128-137. doi:10.1016/j.jesp.2017.03.007.
22. O'Connor LE, Berry JW, Weiss J, Schweitzer D, Sevier M. Survivor guilt, submissive behaviour and evolutionary theory: the down-side of winning in social comparison. *Br J Med Psychol*. 2000;73:519-530.
23. Ratcliffe M. *Experiences of Depression: A Study in Phenomenology*. Oxford: Oxford University Press; 2014.
24. Branscombe NR, Slugoski B, Kappen DM. *Collective Guilt: What It Is and What It Is Not*. Cambridge: Cambridge University Press; 2004.
25. Caspar EA, Ioumpa K, Keysers C, Gazzola V. Obeying orders reduces vicarious brain activation towards victims' pain. *NeuroImage*. 2020;222:117251.