


Neural and Cognitive Signatures of Guilt Predict Hypocritical Blame



Hongbo Yu^{1,2}, Luis Sebastian Contreras-Huerta^{3,4,5},
Annayah M. B. Prosser^{1,3,6}, Matthew A. J. Apps^{3,5},
Wilhelm Hofmann⁷, Walter Sinnott-Armstrong^{8,9,10,11},
and Molly J. Crockett^{1,12}

¹Department of Psychology, Yale University; ²Department of Psychological and Brain Sciences, University of California Santa Barbara; ³Department of Experimental Psychology, University of Oxford; ⁴Wellcome Centre for Integrative Neuroimaging, University of Oxford; ⁵Centre for Human Brain Health, School of Psychology, University of Birmingham; ⁶Department of Psychology, University of Bath; ⁷Department of Psychology, Ruhr University Bochum; ⁸Center for Cognitive Neuroscience, Duke University; ⁹Department of Philosophy, Duke University; ¹⁰Kenan Institute for Ethics, Duke University; ¹¹Duke Institute for Brain Sciences, Duke University; and ¹²Department of Psychology, Princeton University

Psychological Science
1–19

© The Author(s) 2022

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09567976221122765

www.psychologicalscience.org/PS



Abstract

A common form of moral hypocrisy occurs when people blame others for moral violations that they themselves commit. It is assumed that hypocritical blamers act in this manner to falsely signal that they hold moral standards that they do not really accept. We tested this assumption by investigating the neurocognitive processes of hypocritical blamers during moral decision-making. Participants (62 adult UK residents; 27 males) underwent functional MRI scanning while deciding whether to profit by inflicting pain on others and then judged the blameworthiness of others' identical decisions. Observers (188 adult U.S. residents; 125 males) judged participants who blamed others for making the same harmful choice to be hypocritical, immoral, and untrustworthy. However, analyzing hypocritical blamers' behaviors and neural responses shows that hypocritical blame was positively correlated with conflicted feelings, neural responses to moral standards, and guilt-related neural responses. These findings demonstrate that hypocritical blamers may hold the moral standards that they apply to others.

Keywords

moral hypocrisy, blame, conflicted feeling, guilt, lateral prefrontal cortex, open materials

Received 8/13/21; Revision accepted 8/2/22

Moral hypocrisy is commonly reviled. Philosophers and psychologists have identified multiple forms of hypocrisy (Alicke et al., 2013; Effron et al., 2018; Graham et al., 2015; Monin & Merritt, 2012). One form of hypocrisy involves double standards or discrepancies between judgments of oneself and others, such as claiming that certain actions are forbidden for others but permissible for oneself (Graham et al., 2015; Valdesolo & DeSteno, 2007). Another involves discrepancies between moral judgments and behaviors, such as “saying one thing and doing another” (Dover, 2019; Howe & Monin, 2017; Laurent & Clark, 2019). A paradigmatic example of the latter is *hypocritical blame*, where someone blames others for transgressions that they themselves previously

committed (Tognazzini & Coates, 2018). Consider politicians who have extramarital affairs while condemning adultery in other people or environmental activists who jet to exotic destinations but vociferously shame others for flying.

Philosophers argue that hypocritical blame is morally wrong because hypocrites do not really care about the

Corresponding Authors:

Hongbo Yu, Department of Psychological and Brain Sciences,
University of California Santa Barbara
Email: hongbo.yu@psych.ucsb.edu

Molly J. Crockett, Department of Psychology, Princeton University
Email: molly.crockett@yale.edu

moral standards that they express and apply to other people and therefore do not feel any conflict or guilt when their own behaviors fall short of these moral standards (Kittay, 1982; Szabados & Soifer, 1999). On this view, hypocrites condemn other people's moral failures as a trick to convince observers (or even themselves) that they do care about moral standards, whereas their transgressive behavior implies a lack of commitment to those same standards (Fritz & Miller, 2018; Todd, 2019; Wallace, 2010).

Similarly, psychological research demonstrates that laypeople judge hypocrisy to be morally wrong, partly because they see hypocrites as falsely signaling a commitment to moral standards that they do not actually possess (Effron et al., 2018; Jordan et al., 2017; Laurent & Clark, 2019). Vignette studies show that a mere discrepancy between moral judgment and behavior is sufficient to produce perceptions of hypocrisy. Importantly, perceptions of hypocrisy generate other inferences about moral character: Hypocrites are perceived as dislikable, immoral, and untrustworthy (Barden et al., 2005; Jordan et al., 2017; O'Connor et al., 2020).

Although much is known about the psychology of perceiving hypocrisy in other people, less is known about the psychology underlying hypocritical blame itself. Central to many philosophical and psychological accounts as well as common folk intuitions of hypocrisy is the assumption that hypocrites do not think of themselves as blameworthy for violating the standards that they blame other people for violating, and so they do not feel any conflict or guilt when they act hypocritically (Tognazzini & Coates, 2018). How accurate are such assumptions? That is, are discrepancies between moral judgments of others and one's own moral behavior necessarily attributable to a lack of caring about one's own moral standards? No previous studies have empirically examined this critical assumption. Here, we investigated the possibility that at least some people who engage in hypocritical blame do feel that the moral standards they apply to others are also binding for themselves. Some hypocritical blame might arise from weakness of will if people succumb to temptation when they choose to perform actions that they later condemn in others, even though they really do believe that both their own and others' acts are morally wrong.

One challenge in studying moral hypocrisy is the ample disagreement over what counts as hypocrisy—even philosophers lack consensus over a definition (Fritz & Miller, 2018; Kittay, 1982; Szabados & Soifer, 2004; Wallace, 2010). To sidestep this problem, we measured folk intuitions about hypocrisy. We operationalized hypocritical blame as blaming other people for making the same decisions that one has made previously oneself,

Statement of Relevance

Hypocrites blame other people for moral violations they themselves have committed. Common perceptions of hypocrites assume they are disingenuous and insincere. However, the mental states and neurocognitive processes underlying hypocritical blamers' behaviors are not well understood. We showed that people who hypocritically blamed others reported stronger feelings of moral conflict during moral decision-making, had stronger neural responses to moral standards in lateral prefrontal cortex, and exhibited more guilt-related neurocognitive processes associated with harming others. These findings suggest that some hypocritical blamers do care about the moral standards they use to condemn other people but sometimes fail to live up to those standards themselves, contrary to the common philosophical and folk perception.

and we then verified that such a discrepancy between blame judgments and behavior meets participants' own definition of hypocrisy. By grounding our study of moral hypocrisy in folk intuitions, we did not need to commit to a single normative account of hypocrisy, of which there are many. Instead, we simply investigated the neurocognitive processes of people who appear hypocritical to most observers.

A second challenge in determining whether people who engage in hypocritical blame actually care about moral standards is that such concerns are easy to fake (Batson et al., 1997, 1999, 2002; Batson & Thompson, 2001; FeldmanHall et al., 2012). Therefore, behavioral and self-report measures alone are not adequate for probing the underlying moral beliefs and concerns. To address this challenge, we triangulated self-reports, behavior, and brain activity to obtain convergent evidence. Our method can reveal relations between hypocritical blame and neural representations of moral standards and guilt for violating those moral standards. These neural representations should be more difficult—if not impossible—to fake.

Past work has demonstrated the engagement of the dorsolateral prefrontal regions in representing social and moral norms (e.g., fairness), and more specifically, activity patterns in these regions suggest that they may represent moral standards distinctively from material values (Buckholtz & Marois, 2012; Carlson & Crockett, 2018; Crockett et al., 2017; Zoh et al., 2022). We therefore predicted that this region would be responsible for the

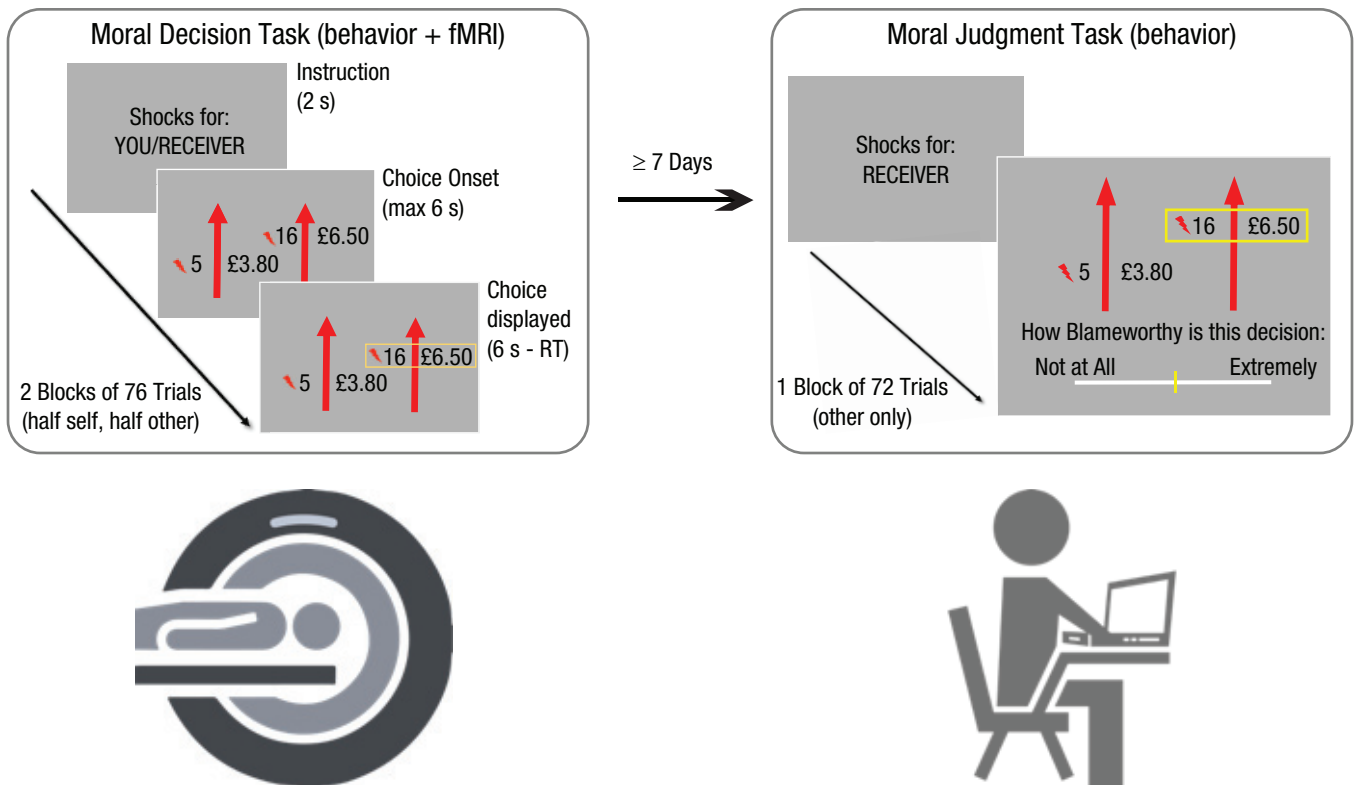


Fig. 1. Procedure of the moral decision task and the moral judgment task. In the moral decision task (left), participants chose between a harmful option and a helpful option. The harmful option contained more money and more electric shocks. On half of the trials, the shocks were for the decider (*self* condition), whereas on the other half, the shocks were for the receiver (*other* condition). Participants were informed that one of their choices would be randomly selected and implemented at the end of the scanning session. We created the trial set so that the difference in shocks and difference in money were orthogonal across the trials. At least 1 week after the scanning session (right), we presented a series of decisions to the participants and asked them to judge how blameworthy it would be for someone to choose the harmful options on those trials (*other* condition only). Although this was not made explicit to the participants, the set of trials that they judged were the same as those that they had faced during the functional MRI (fMRI) session. In both the decision-making task and the moral judgment task, participants' decisions and judgments were private and unobserved.

representation of moral rules and standards. When people violate moral standards that they care about, they experience feelings of conflict and guilt (Baumeister et al., 1994; Green et al., 2012; Lythe et al., 2015; Tangney et al., 2007; Yu et al., 2014; Zahn et al., 2009). Here, we tested the hypothesis that some blamers do find the moral standards that they apply to other people also binding for themselves. If this hypothesis is correct, then these blamers would show strong neural responses to moral standards and guilt when they make the same decisions that they later blame other people for making (Bartel, 2019; Mele, 1989; O'Connor et al., 2020).

Open Practices

Materials, behavioral data, and analysis codes for these studies have been made publicly accessible on OSF (<https://osf.io/ardcu>). The design and analysis plans for the studies were not preregistered.

Method

Overview of research

We tested our predictions in a paradigm that allowed us to quantify hypocritical blame in terms of a discrepancy between moral behavior and moral blame judgments of the same behavior in other people. In Study 1 (behavioral; $N = 188$), we described this paradigm to a separate group of participants and verified that our operationalization of hypocritical blame was indeed perceived as hypocritical by our participants. In Study 2, participants ($N = 62$) completed a moral decision task in the functional MRI (fMRI) scanner (Crockett et al., 2017), where they decided whether to seek profit by inflicting pain on either themselves or an anonymous receiver (Fig. 1, left). At least 1 week after scanning, participants returned to the lab and completed a moral blame task, where we presented them with the same set of trials that they had previously seen in the moral

decision task (Fig. 1, right). On each trial, we highlighted the more harmful option (inflicting pain on the receiver for profit) and asked them to judge how blameworthy it would be for other people to choose that option. We calculated, for each participant, a hypocritical-blame index that quantified discrepancies between their current blame judgments and the choices they made a week earlier.

Participants

For Study 1, 188 adult U.S. participants, who were recruited from Amazon Mechanical Turk, completed the task (125 males; mean age = 36.3 years). This study was approved by the Yale University Human Subjects Committee (Approval No. HSC 2000022385). For Study 2, 80 healthy volunteers between the ages of 18 and 38 years were recruited from the University of Oxford and the local Oxford, UK, community. This recruitment was designed to achieve a target sample size, after accounting for expected participant dropout (based on Crockett et al., 2017), of 64 participants, which is sufficient for detecting brain-behavior correlations (r s) of .3 with 80% power. Data collection was terminated once we reached the predetermined sample size ($N = 80$).

The study was conducted at the Wellcome Centre for Integrative Neuroimaging (WIN) and the University of Oxford Department of Experimental Psychology and was approved by the University of Oxford ethics committee (Approval No. R50262/RE001). All participants gave written informed consent and were compensated for their time. Individuals who had a history of neurological or neuropsychiatric disorders, used psychoactive medication or drugs, were pregnant, or had studied psychology for more than 2 years were excluded from participation; those who had previously participated in studies involving social interaction or electric shocks were also excluded because of concerns that prior experience with similar experimental settings would influence psychological and brain processes in the current task. Eight participants had excessive head motion in the scanner (3-mm translation or 3° rotation within one scanning session). Two participants expressed doubts regarding whether the receiver participant would actually receive the electric shocks. One participant's neuroimaging data were not registered because of technical issue of the scanner. Four participants failed to show up to the moral judgment session. Three participants produced judgment data that were uninterpretable (i.e., delivering more shocks for less money was judged less blameworthy). These participants were excluded from further analysis, leaving 62 participants for the final fMRI data analysis (27 males; mean age = 22.7 years).

Moral decision task and fMRI testing session

Prior to attending the fMRI testing session, participants first completed a battery of online personality questionnaires (data to be reported separately). At least 1 week later (range = 7–74 days, $Mdn = 13$ days), they attended an MRI scanning session at WIN. After giving informed consent, participants went through a pain-thresholding procedure (for details, see Crockett et al., 2014). The purpose of this procedure was twofold: (a) to familiarize participants with experience of the shocks, which they would later take into account in their decision-making, and (b) to determine the physical intensity of the shocks so that their subjective intensity was matched across participants. After the pain-thresholding procedure, participants were informed that they would be randomly assigned to roles of either decider or receiver using a procedure that has been described in detail elsewhere (Crockett et al., 2014, 2017). In reality, participants were always assigned to the role of decider, and the role of receiver was played by a confederate.

Following role assignment, participants received instructions for the moral decision task (full instructions for this task can be found in the Supplemental Material available online in the “Instructions for the Decision Task” section), answered comprehension questions, and practiced outside the scanner for six trials. Participants were informed that one of their choices would be randomly selected and implemented at the end of the scanning session. They were informed that their choices would be anonymous and that they would not meet or interact with the receiver. This was done to minimize concerns about reputation or reciprocity in their decision-making. They then completed the moral decision task in the fMRI scanner. In this task, participants made a series of binary choices in which one option contained more electric shocks and amounts of money (i.e., harmful option), and the other option contained fewer shocks and less money (i.e., helpful option). The money was always for the participant (i.e., decider), but the shocks were allocated to the receiver in half of the trials (i.e., *other* condition) and to the participant in the other half of the trials (i.e., *self* condition).

The procedure for generating the choice options is described in detail elsewhere (Crockett et al., 2017). Specifically, we created a set of 72 trials according to the criteria reported by (Crockett et al., 2017). Four catch trials, in which the more harmful option contained a smaller amount of money, were inserted, resulting in 76 trials in the set. Half of the trials were randomly selected to present the more harmful option on the right-hand side of the screen; in the other half of the trials, the more harmful option was presented

on the left-hand side of the screen. Then the 76 trials were duplicated to constitute the *self* and the *other* conditions. Trials were then distributed into two scanning runs of 76 trials each; each run contained equal numbers of *self* and *other* trials. Four different trial sets were created in this way, which were randomly assigned to the participants. Our trial optimization procedure (Crockett et al., 2017) ensured that the amount of profit that would result from choosing the harmful option was uncorrelated with the amount of shocks ($|r| < .07$, $ps > .525$).

After the moral decision task, participants exited the scanner, and one trial from the task was randomly selected and implemented outside the scanner. Participants then completed a series of self-report measures about their experiences during the decision-making task, including a measure of conflicted feeling they experienced about their choices (1 = *not at all*, 7 = *very much*). They also answered debriefing questions that assessed their beliefs of the experimental setup. In the debriefing session, we did not include questions that explicitly asked whether the participants doubted about the veracity of the paradigm, as these questions may cue feelings of doubt in the participants. Instead, we included indirect questions about the clarity of our instructions regarding the presence of the receiver, the delivery of electric shocks to the receiver, and the confidentiality of the participants' decisions. Participants responded to these questions on 7-point Likert scales (1 = *yes fully*, 7 = *no not at all*). The majority of participants chose 1, *yes fully*; very few chose 2 and only two chose 3. The average scores of these questions were below 1.1. None of the participants' ratings on any of these questions were equal to or higher than the midpoint of the scale, indicating that our instructions were clear.

In addition to these questions, participants also provided open-ended comments about the study, where they could express their feelings, experience, and questions regarding any part of the study. We read these comments and flagged expressions of doubt regarding the presence of the receiver or the delivery of electric shocks to the receiver. Two participants explicitly mentioned "doubt" and being "skeptical" about the experimental setup (e.g., whether the receiver existed). These participants were therefore excluded. Two other participants did not explicitly mention doubt or skepticism, but their responses to the open-ended questions suggested that they had considered whether the receiver was present. In our analysis, we included these two participants to maximize the size and diversity of our sample. However, excluding them did not change the patterns of results (see Table S2 in the Supplemental Material).

Moral judgment task and behavioral testing session

At least 1 week after the fMRI session, participants were invited to participate in a behavioral experiment session at their earliest convenience. During the behavioral experiment session, participants completed a battery of behavioral and psychophysiological tasks, the first of which was a moral judgment task. In this task, participants were presented with a subset of the choice sets that they faced in the scanning session, namely all of the trials in the *other* condition in which the money was for the decider and the shocks were for the receiver. We did not explicitly measure whether participants recognized these trials. However, if they indeed remembered their choices in most of the trials, we would expect to see much less hypocritical blame than what we actually observed, because people have a tendency to appear consistent (Gawronski, 2012), and exhibiting hypocrisy is something people hate (Jordan et al., 2017). In both the decision-making task and the moral judgment task, participants' decisions and judgments were private and unobserved.

On each trial of the moral judgment task, the harmful option (i.e., more money for the decider and more shocks for the receiver) was highlighted. Participants were asked to judge the blameworthiness of each harmful choice on that specific trial on a visual analog scale ranging from *not at all blameworthy* to *extremely blameworthy*. Full instructions for this task can be found in the Supplemental Material in the "Instructions for the Judgment Task" section.

MRI acquisition and preprocessing

We performed fMRI scanning on a 3-T Siemens Prisma scanner at WIN at The University of Oxford. Functional images were obtained with multiband T2*-weighted echo-planar imaging (EPI) sequence. The EPI images were acquired in an ascending manner at an oblique angle ($\sim 30^\circ$) to the anterior commissure–posterior commissure (AC-PC) plane to minimize signal dropout in the orbitofrontal areas. The following acquisition parameters were used: 72 slices in interleaved ascending order, matrix size = 108×108 , voxel size = $2 \times 2 \times 2 \text{ mm}^3$ with 1-mm gap, echo time (TE) = 30 ms, repetition time (TR) = 1,570 ms, flip angle = 70° , field of view (FOV) = $216 \times 216 \text{ mm}^2$. The structural image was taken using a magnetization-prepared rapid gradient echo (MPRAGE) sequence with 192 slices (TR = 1,900 ms, TE = 3.97 ms, field of view = $192 \times 192 \text{ mm}^2$, voxel size = $1 \times 1 \times 1 \text{ mm}^3$ resolution). We also acquired a field map (short TE = 4.92 ms, long TE = 7.38 ms, TR = 482.0 ms,

resolution = $2 \times 2 \times 2 \text{ mm}^3$, FOV = $219 \times 219 \text{ mm}^2$) to correct distortions in the functional images.

MRI data were preprocessed and analyzed using SPM software (Version 12; www.fil.ion.ucl.ac.uk/spm). Functional images were realigned and unwarped with reference to the field map and coregistered to the participant's own structural image. The structural images underwent routine preprocessing steps, including segmentation, bias correction, and spatial normalization to the Montreal Neurological Institute (MNI) template. Finally, images were spatially smoothed with an SPM default Gaussian kernel (8-mm full-width at half-maximum).

General Linear Model (GLM) 1: model of trial-wise anticipated blameworthiness judgments

We constructed the first GLM to obtain, for each participant, a representation map of trial-wise judgments of how blameworthy it would be to choose the more harmful option relative to the less harmful option (or participant-specific moral standards). In this model, blood-oxygen-level-dependent (BOLD) signals were regressed on four critical first-level regressors containing the onsets of *self* trials where the left (a) or right (b) option was selected and *other* trials where the left (c) or right (d) option was selected. Duration of these four regressors was set to the participant's reaction time on that trial (i.e., the interval between the presentation of options and the button press). For this analysis, what we looked for at the individual level was the neural signal at the time of decision-making that scaled with the relative blameworthiness of the more harmful option relative to the less harmful option, as judged by the participants themselves. To this end, for the two regressors corresponding to the trials in the *other* condition, we included a parametric modulator that contained each participant's judgment on that trial of the relative blameworthiness of the more versus less harmful option (collected in the behavioral session at least a week following scanning). In GLM 1, this relative blameworthiness rating was treated as an attribute of the trial, independently of what the participant actually chose on that trial. The parametric analysis, therefore, captured the neural correlates of a crucial element of moral decision-making: assessing the relative blameworthiness of choosing the more profitable but more harmful option, relative to the less harmful and profitable option. We expected this evaluative process to be present during all choices, regardless of what the participant ultimately chose. We additionally included regressors of no interest corresponding to onsets of button presses, cue for transitions between conditions,

and missing trials, as well as six nuisance regressors to control for head motion.

For the second-level (or group-level) analysis, we included participants' degree of hypocritical blame as a parametric regressor while controlling for their moral preferences (κ_{other}). We constructed the GLM this way deliberately because the second-level modulator, hypocritical blame, was dependent on participants' behaviors. The brain activities captured by the first-level GLM therefore should not depend on behavior. This approach is widely adopted in fMRI studies of value-based choice in moral (Crockett et al., 2017), social (Ruff & Fehr, 2014), and economic (Rangel et al., 2008) domains. We defined participant-specific representation of blame as the positive effect of the blameworthiness-judgment parametric modulator, collapsing across help and harm decisions. For whole-brain analyses, we adopted family-wise error (FWE) corrected $p_{\text{FWE}} < .05$ at the cluster level and uncorrected $p < .001$ at peak voxel level. The cluster with the largest cluster-level p_{FWE} value (.040) that still passed this threshold had 113 contiguous voxels.

GLM 2: model of decision parameters

For each participant, we built a GLM to obtain representation maps of the objective amounts of money difference (Δmoney) and shock difference (Δshock) between the harmful option and the helpful option in the *self* and *other* conditions. In this model, we included the same four critical regressors as in GLM 1. Each of these regressors was further associated with four parametric modulators: the amount of money and shocks for the harmful option and the helpful option. Critically, custom scripts ensured that these two parametric modulators competed for variance during the estimation, rather than being serially orthogonalized as is standard in SPM. We included the same set of regressors of no interest as in GLM 1. To obtain representation maps of Δmoney and Δshock , we defined first-level contrast of $\text{money}_{\text{harmful option}} > \text{money}_{\text{helpful option}}$ and $\text{shock}_{\text{harmful option}} > \text{shock}_{\text{helpful option}}$ for both the *self* and *other* regressors, collapsing across help and harm decisions. Individual representation maps to Δmoney and Δshock were used, together with the guilt signature (see below), to calculate guilt-related pattern expression associated with each representation map.

Multivariate analysis with guilt-related brain signature

We adopted a multivariate decoding approach to probe guilt-related neurocognitive processes associated with

choice attributes Amoney and Ashock. Specifically, we utilized a previously validated brain-based signature of guilt (Yu et al., 2020). On the basis of two independent neuroimaging data sets that used interpersonal interactions to evoke guilt, Yu et al. (2020) identified a guilt-related brain signature (GRBS) that discriminated conditions associated with different levels of interpersonal guilt. Specifically, in the training data set, participants were either completely or partially responsible for an anonymous stranger's pain (Yu et al., 2014). Participants' self-reported guilt feelings were positively associated with their responsibility. The GRBS was trained to discriminate the completely responsible from the partially responsible conditions and was able to do so accurately (accuracy = 88%) in a cross-validated manner (i.e., leave-one-subject-out cross-validation). Moreover, the predictive power (or *sensitivity*) of GRBS can be generalized to a neuroimaging study that adopted a similar interpersonal harm task, in which the participants were from a different cultural background relative to those in the training data set (Koban et al., 2013). However, a useful brain-based signature of a psychological construct (e.g., guilt in social interactions) should not be sensitive to other negative experiences; otherwise, it would be difficult to infer what neurocognitive processes the signature captures (Wager et al., 2013; Woo et al., 2017). To demonstrate that this signature does not simply pick up any negatively valenced experiences but is sensitive only to guilt experience elicited in social interactions (i.e., *specificity*), in Yu et al. (2020) we tested its predictive power when applied to a few other fMRI data sets in which the participants had negatively valenced experience but were not engaged in live social interactions. These include physical pain, vicarious pain, and recall of past negatively valenced experiences. In Yu et al. (2020), we found that the GRBS cannot distinguish different levels of physical pain or vicarious pain, but it can distinguish different levels of guilt in a social interactive task. This suggests that the GRBS is specific to guilt. In addition, we found evidence that the GRBS is specific to guilt experienced during a real social interaction: The GRBS could not distinguish the brain activity pattern when participants recalled a past experience of guilt from the brain activity pattern when the participants recalled a past experience of sadness or shame.

Results

Blaming others for behaving similarly to oneself appears hypocritical to observers

In Study 1, participants ($N = 188$) read a vignette describing a protagonist who took part in an experiment in

which they decided to deliver 10 electric shocks to another person in exchange for \$1. After the protagonist made his decision, he learned of another person who made the same decision. In the *discrepant* condition ($n = 95$), the protagonist judged this person's decision to be extremely morally blameworthy, whereas in the *consistent* condition ($n = 93$), the protagonist judged this decision to be not at all morally blameworthy. Participants were then asked to indicate to what extent they thought the protagonist was hypocritical, moral, and trustworthy on 7-point Likert scales (1 = *not at all*, 7 = *extremely*; full text of vignettes can be found in the Supplemental Material in the "Vignettes for Folk Intuitions of Discrepant Blamers" section).

As expected, participants in the discrepant condition judged the protagonist to be more hypocritical ($M = 6.1$, $SD = 1.5$) than participants in the consistent condition ($M = 3.0$, $SD = 2.3$), $t(181) = 11.4$, $p < .001$, Cohen's $d = 1.64$ (Fig. 2a). In addition, participants judged the protagonist to be less moral ($M = 2.4$, $SD = 1.7$; Fig. 2b) and less trustworthy ($M = 2.2$, $SD = 1.6$; Fig. 2c) in the discrepant condition relative to the consistent condition—moral: $M = 2.8$, $SD = 1.8$, $t(181) = -2.22$, $p = .028$, Cohen's $d = 0.24$; trustworthy: $M = 2.9$, $SD = 1.7$, $t(181) = -3.39$, $p < .001$, Cohen's $d = 0.42$. These findings demonstrate that people ascribe hypocrisy on the basis of a mere behavioral description of a discrepancy between blame and behavior, even in the absence of information about the discrepant blamer's mental state. In addition, we found that people infer that discrepant blamers are less moral and trustworthy than consistent blamers, controlling for the amount of harm inflicted.

Quantifying hypocritical blame in the laboratory

Having confirmed in Study 1 that observers perceive hypocrisy when someone blames another person for making the same decision they made themselves, in Study 2, we sought to examine the neural correlates of moral decision-making in so-called hypocritical blamers. Because we assessed participants' actual decisions and moral judgments of other people's decisions in the same set of trials (see the Method section for details), we were able to quantify each participant's degree of hypocritical blame by comparing (a) their own likelihood of making a harmful decision on each trial of the decision task with (b) the blameworthiness they assigned to another person for a harmful decision on each trial of the moral judgment task. Specifically, for (a), we computed each participant's likelihood of harming another person on a given trial with a well-established computational model for this type of moral decision-making (Crockett et al., 2014, 2015, 2017). In

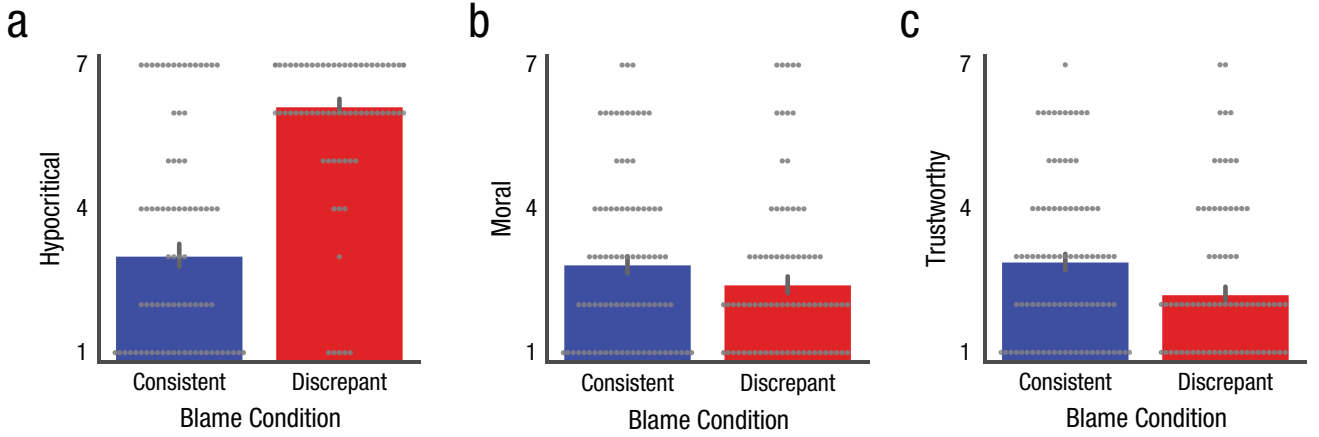


Fig. 2. Folk intuitions about hypocritical blame. An independent group of participants ($N=188$) judged a protagonist who blamed another person for making the same decision that they made themselves to be more hypocritical (a), less moral (b), and less trustworthy (c) than a protagonist who blamed consistently. Red and blue bars indicate means, and dots indicate individual data. Error bars indicate *SEM*.

this model, the probability of choosing to inflict pain for profit is a softmax transformation of the subjective value of the harmful relative to the helpful option, ΔV . For participant j and trial i ,

$$\Delta V_{ij} = (1 - \kappa_j) \Delta \text{money}_{ij} - \kappa_j \Delta \text{shock}_{ij}$$

$$p(\text{harm})_{ij} = \left(\frac{1}{1 + e^{-\beta_j \times \Delta V_{ij}}} \right) (1 - 2\varepsilon_j) + \varepsilon_j.$$

Here, Δmoney and Δshock denote the additional money and shocks associated with the harmful relative to the helpful option. The harm-aversion parameter, κ , indicates the relative weight participants place on shocks over money in the valuation process. κ takes on different values in the *self* and *other* conditions (κ_{self} and κ_{other}), thereby reflecting different decision preferences when harming oneself versus harming another person. We used a softmax function to convert ΔV into the probability of choosing the harmful option over the helpful option. Here, β is a participant-specific inverse temperature parameter that characterizes the steepness of the softmax curve. We also included a lapse-rate parameter ε that captures task-irrelevant noise such as inattention (cf. Crockett et al., 2014). Because the task and computational framework has been established in a number of studies across multiple labs, we did not plan to compare multiple models in a data-driven manner. However, when estimating the model without the lapse rate, we found that it had a higher Bayesian information criterion (2,963) than the one with the lapse rate (2,833), indicating that including the lapse rate improved the model.

We replicated key findings from past studies using this task (Crockett et al., 2014, 2017; Volz et al., 2017): Specifically, on the aggregate level, participants were more averse to harming other people than harming themselves for money (see Fig. S1 in the Supplemental Material). For our analysis of hypocritical blame, we focused here solely on participants' decisions to inflict pain on other people for profit. We observed wide variation in harm aversion: Some participants refused to inflict a single additional shock on a stranger for a profit of £19, whereas others were willing to inflict 20 additional shocks on a stranger in exchange for 10 pence. Because κ_{other} was not normally distributed (Kolmogorov-Smirnov test = 0.11, $p = .05$), in the following we use Spearman tests for the correlational analyses that involved κ_{other} .

After computing each participant's likelihood of choosing the harmful option for each trial in the decision task, we computed a hypocritical-blame score for each participant by combining the data from the moral judgment task with the data from the decision-making task. Specifically, we computed the hypocritical-blame score by summing across trials the amount of blame assigned on each trial of the judgment task, weighted by the participants' own likelihoods of choosing selfishly on the same trial in the decision task:

$$\text{hypocritical blame}_j = \sum_i \text{blame}_{ij} \times p(\text{harm})_{ij}.$$

Here, blame_{ij} denotes participant j 's blameworthiness judgment on trial i of the judgment task, and $p_{ij}(\text{harm})$ is the likelihood that participant j would choose the harmful option on the same trial in the moral decision-making

task computed on the basis of the value function and softmax function. Weighting the blame judgments by the choice likelihoods captured the logic that it is more hypocritical for someone to blame another person for an action that they themselves have previously taken confidently than tentatively or accidentally (Alicke et al., 2013; Dover, 2019; Laurent & Clark, 2019; Wallace, 2010). According to our definition of hypocritical blame, it is critical to keep blameworthy judgments on the same scale across participants. Therefore, we did not normalize blameworthiness judgments within participants before calculating hypocritical blame because doing so would take away our ability to compare participants' blame tendency on the same scale.

As Figure 3 illustrates, participants who assign a high level of blame on trials where they themselves are likely to choose the harmful option will have a high hypocritical-blame score based on our model (Fig. 3b and 3d). In contrast, participants who almost never assign blame on the trials where they themselves are

likely to choose the harmful option will have a low hypocritical-blame score based on our model (Fig. 3a and 3c). Given this operationalization of hypocritical blame, 97% of participants displayed at least some level of hypocritical blame. However, we observed a wide range of individual variation in the degree of hypocritical blame that could be described by a normal distribution ($M = 12.9$, $SD = 7.5$; Shapiro-Wilk normality test, $p = .248$; Fig. 3e). Given the way we defined hypocritical blame, it is not surprising that hypocritical blame was negatively correlated with κ_{other} (Spearman's $\rho = -.57$, $t = -5.42$, $p < .001$; for this and the following Spearman correlations, the sample size is 62; Fig. 3f). Therefore, we controlled for κ_{other} in all subsequent analyses involving hypocritical-blame scores (robustness checks can be found in "Robustness Test for Statistically Controlling for κ_{other} " and Table S3 in the Supplemental Material). Our definition of hypocritical blame takes advantage of the fact that people do not blame identically and captures the individual

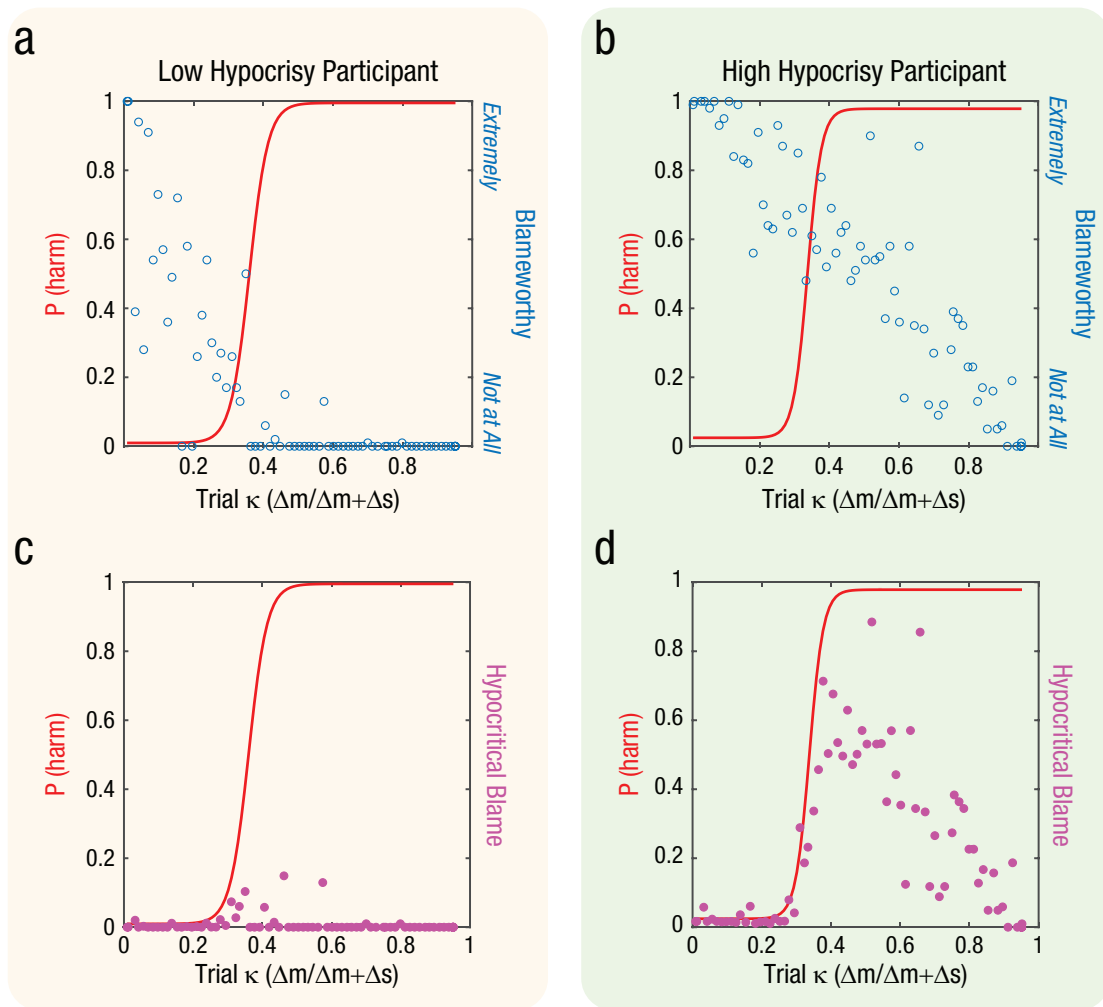


Fig. 3. (continued on next page)

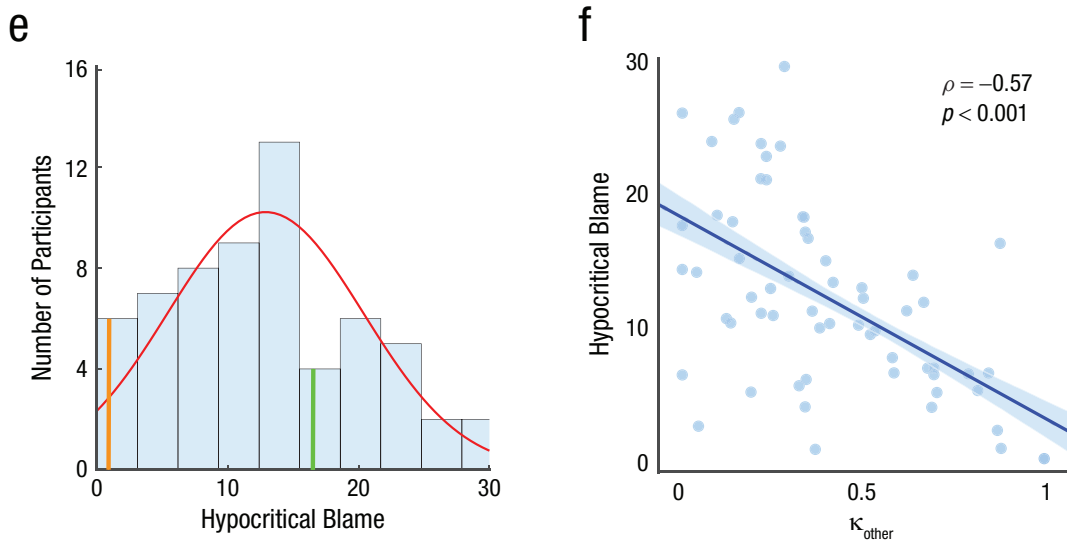


Fig. 3. Definition and distribution of hypocritical blame. The top and middle rows show results for two representative participants, one whose behavior is described in (a) and (c) and the other in (b) and (d), illustrating how hypocritical blame is defined in our formula. These two participants have comparable harm aversion in the *other* condition ($\kappa_{\text{other}} = 0.37$ and $\kappa_{\text{other}} = 0.39$, respectively). The x -axis indicates the κ value [$\Delta\text{money}/(\Delta\text{money} + \Delta\text{shock})$] of each trial. A higher trial κ value means that choosing the more harmful option will confer a large amount of monetary gain for the decider by increasing the receiver's harm by a small degree. The y -axes on the left indicate the model-derived probability of choosing the harmful option. The red curves are the best-fitting softmax curves. In (a) and (b), the blue dots indicate participants' blameworthiness judgment ratings. In (c) and (d), the magenta dots represent the product of blameworthiness judgment and the model-derived, participant-specific probability of choosing the more harmful option on the same trial (weighted blame). In our definition, the participant's hypocritical blame is the sum of weighted blame across all the trials. As can be seen, the participant described in (b) exhibited more hypocritical blame by our definition than the participant described in (a) because they indicated substantial degree of blameworthiness judgment on the trials in which they were very likely to choose the harmful option themselves. (e) Distribution and interindividual variability of hypocritical blame. The brown and the green vertical lines indicate the hypocritical-blame score of the low (hypocritical blame = 0.7) and high (hypocritical blame = 16.8) hypocritical blame participants, respectively. (f) Hypocritical-blame score was negatively correlated with κ_{other} (Spearman's $\rho = -.57$, $p < .001$). Error band represents *SEM*.

differences in the discrepancy between one's moral behaviors and their propensity to assign blame to others. Indeed, when we controlled for participants' average blame, the negative correlation between κ_{other} and hypocritical blame became stronger (Spearman's $\rho = -.81$, $t = -10.78$, $p < .001$). Note that hypocritical blame was not correlated with participants' overall blameworthiness judgments (Spearman's $\rho = .09$, $p = .476$).

Hypocritical blame is associated with conflicted moral decision-making

We next turned to data concerning the mental states of hypocritical blamers when they blamed other people for what they themselves had done. To test the possibility that at least some people who engage in hypocritical blame do feel that the moral standards they apply to others are also binding for themselves, we first examined the prediction that those who exhibit more hypocritical blame, relatively to those who behave similarly

but do not find it blameworthy, feel more conflicted about their own decision-making when failing to live up to their own moral standards.

To obtain a measure of conflicted feelings, after participants completed the moral decision-making task in the scanner, we asked them to indicate their subjective perception of "facing a moral dilemma" during the task on a 7-point Likert scale. Conflicted feelings varied substantially across participants, covering the entire range of the scale ($M = 3.7$, $SD = 1.6$). Controlling for participants' moral behavior (as indexed by κ_{other}), we observed a positive correlation between conflicted feelings and hypocritical blame, Spearman's $\rho = .39$, $t = 3.22$, $p = .002$ (Fig. S2 in the Supplemental Material). This means that, consistent with our prediction, results showed that hypocritical blamers reported feeling more rather than less conflict during their moral decisions, controlling for the decisions they actually made.

Next, we tested a hypothesis that subjective feelings of conflict result from failures to live up to one's own

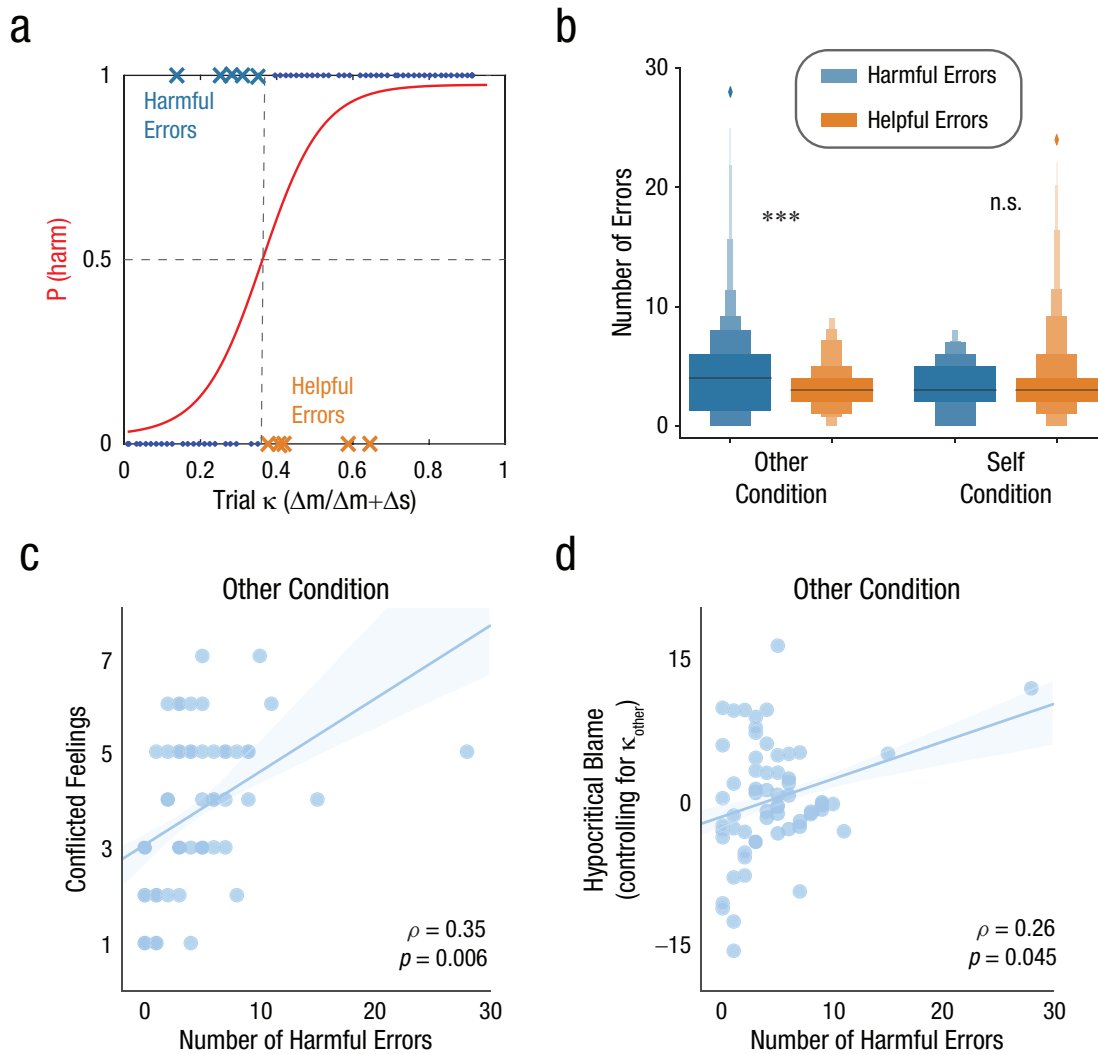


Fig. 4. Hypocritical blame is associated with harmful errors during moral decision-making and with subsequent conflicted feelings. (a) The choices made by a representative participant in the *other* condition. The x -axis indicates trial κ , and the y -axis indicates the probability of choosing the harmful option. Blue dots indicate “correct” choices, in which the chosen option has higher subjective value for the decider than the unchosen option. Errors occur when the decider chooses the option that has the lower subjective value than the other option. The green and yellow checks indicate harmful errors and helpful errors, respectively. (b) Numbers of harmful and helpful errors in the *self* and *other* condition. (c, d) The number of harmful errors in the *other* condition is positively correlated with conflicted feelings (c) and hypocritical blame (d) after analyses control for κ_{other} and the number of erroneous help choices. *** $p < .001$. Error bands in (c) and (d) represent *SEM*.

moral standards. Our decision model specified that participants would sometimes make choice errors, in which they select the option that is less subjectively valuable to them (see “Analysis of the Number of Harmful and Helpful Errors” in the Supplemental Material). Choice errors are particularly likely to occur around a participant’s indifference point, where participants face the highest levels of decision conflict (Fig. 4a). Past work distinguishes between “harmful errors” (erroneously choosing the harmful option) and “helpful errors” (erroneously choosing the helpful option) and suggests that

participants who value helping over harming should be more likely to make harmful errors (Hutcherson et al., 2015). Accordingly, here we found that harmful errors were more common than helpful errors during moral decision-making (*other* condition: $\beta = -0.33$, $SE = 0.09$, 95% confidence interval [CI] = $[-0.51, -0.14]$, $z = -3.52$, $p < .001$) but not during nonmoral decision-making (*self* condition: $\beta = 0.09$, $SE = 0.09$, 95% CI = $[-0.10, 0.27]$, $z = 0.91$, $p = .36$; choice-by-condition interaction: $\beta = 0.41$, $SE = 0.13$, 95% CI = $[0.15, 0.67]$, $z = 3.12$, $p = .002$; Fig. 4b).

Harmful errors during moral decision-making represent cases in which participants harm other people even though helping is more aligned with their overall moral preferences. One potential explanation of these errors is weakness of will—that is, these participants succumbed to temptation to gain money even though they judged their harmful acts to be morally wrong. If hypocritical blame arises from weakness of will, harmful errors should be positively correlated with both feelings of moral conflict and with hypocritical blame. We found support for both predictions: The number of harmful errors in the *other* condition was positively correlated with both conflicted feelings (Fig. 4c) and hypocritical blame (Fig. 4d) after we controlled for participants' moral preferences (κ_{other}) and the number of helpful errors—Spearman's $\rho = .35$, $t = 2.85$, $p = .006$ for conflicted feelings; Spearman's $\rho = .26$, $t = 2.05$, $p = .045$ for hypocritical blame. Conflicted feelings were not correlated with helpful errors in the *other* condition—Spearman's $\rho = .03$, $t = 0.25$, $p = .808$ —nor with harmful or helpful errors in the *self* condition—Spearman's $\rho = .02$, $t = 0.18$, $p = .855$ for harmful errors; Spearman's $\rho = .04$, $t = 0.28$, $p = .779$ for helpful errors—suggesting that posttask subjective feelings of conflict arose from failures to live up to one's own moral standards specifically rather than from choice errors more generally.

Hypocritical blame is positively correlated with neural representations of moral standards

We next investigated how brain activity during moral decision-making related to individual differences in hypocritical blame. If participants' inferences in Study 1 are accurate—that is, if hypocritical blamers are indeed less moral and trustworthy—then hypocritical blame should be associated with weak or absent neural representations of moral standards. In contrast, if hypocritical blame arises from failing to live up to one's own moral standards, neural representations of moral standards should be observable in hypocritical blamers. To test these predictions, we modeled participants' neural activity at decision onset, focusing on the *other* condition in which participants traded off money for themselves against pain to others, because the equivalent trials in the *self* condition would not be considered morally blameworthy. In GLM 1 (see the Method section), decision onsets were modulated by a first-level parametric regressor indicating each participant's own judgments of blameworthiness for selecting the harmful option on each trial, which was obtained in the moral judgment task a week after scanning (Fig. 5a).

Participants with higher moral standards should, on average, judge harmful choices to be more blameworthy than participants with lower standards (Crockett et al., 2017). This was indeed what we found—participants' overall blameworthiness judgments were positively associated with their harm aversion in the *other* condition, $\beta = 35.69$, $SE = 5.71$, $b = 0.29$, 95% CI for $b = [0.20, 0.38]$, $t(60) = 6.25$, $p < .001$. However, because individuals vary in how they assign blame judgments as a function of pain inflicted and profits gained (Siegel et al., 2017), a more precise index of moral standards would consider each participant's trial-specific blame judgments rather than their overall propensity to blame. Because each participant provided a blame judgment for every trial they faced in the scanner, we therefore treated these judgments as participant-specific, trial-wise indicators of moral standards.

To examine the relationship between brain responses to moral standards and individual differences in hypocritical blame, we conducted a group-level analysis that included participants' degree of hypocritical blame as a second-level parametric regressor while controlling for their moral preferences (κ_{other}). A positive effect of this group-level parametric regressor would identify voxels in which neural responses to moral standards (i.e., blameworthiness judgments of harm choices) scale positively with participants' hypocritical blame. In a whole-brain analysis ($p_{\text{FWE}} < .05$, whole-brain corrected at the cluster level after voxel-wise thresholding at $p < .001$), the parametric contrast revealed a positive relationship between hypocritical blame and BOLD responses to moral standards in the anterior cingulate cortex (ACC), bilateral lateral prefrontal cortex (LPFC), and left inferior temporal cortex (Fig. 5b; Table S1 in the Supplemental Material). The reverse contrast did not identify any significant cluster, indicating that there were no brain regions showing a significant negative relationship between hypocritical blame and BOLD responses to moral standards.

The positive relationship between hypocritical blame and BOLD responses to moral standards (i.e., blameworthiness judgments of harm choices) in ACC, LPFC, and left inferior temporal cortex could not be explained by decision difficulty (i.e., the reverse of subjective value difference between the chosen and the unchosen options): Responses in these areas did not scale with trial-wise decision difficulty, and the strength of responses to decision difficulty did not predict individual differences in hypocritical blame (see "GLM 3" and Fig. S3 in the Supplemental Material). Thus, hypocritical blame was positively associated with neural responses to moral standards during moral decision-making.

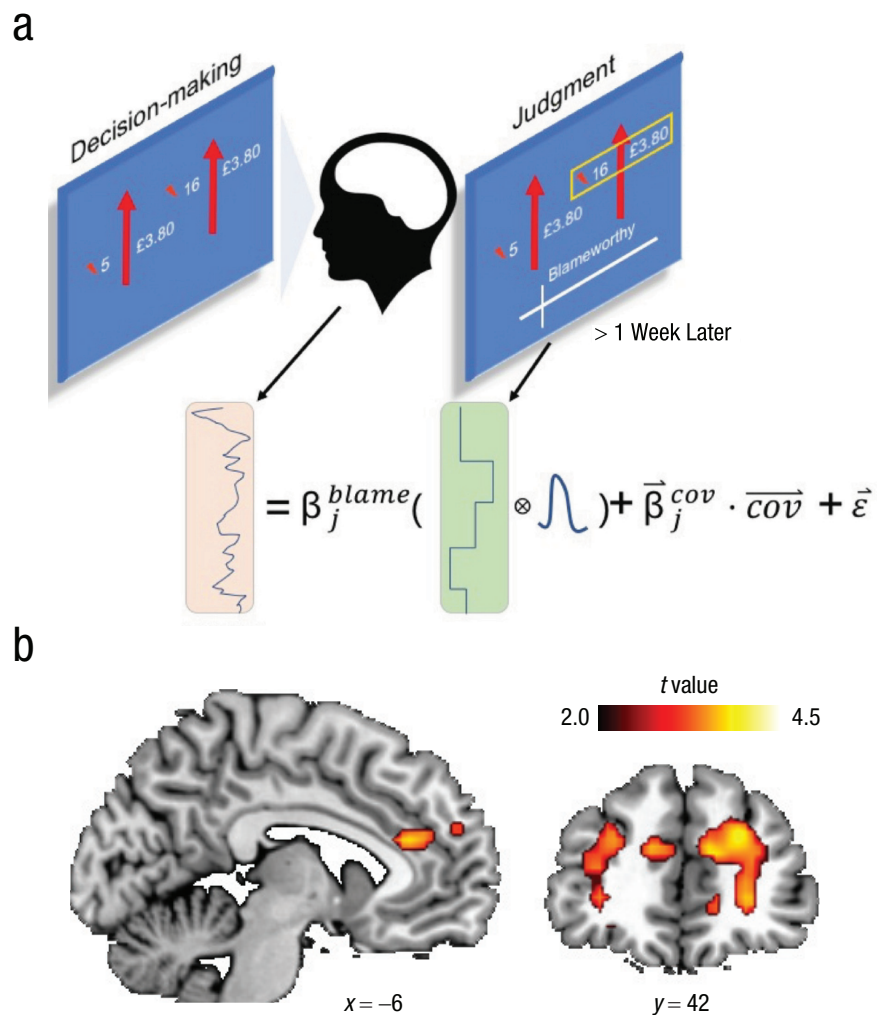


Fig. 5. Hypocritical blame is associated with stronger neural responses to moral standards. (a) Analysis strategy. In a general linear model, we regressed blood-oxygen-level-dependent (BOLD) responses at decision trial onset against the blameworthiness judgment made on the corresponding trial in the judgment task (at least 1 week later). (b) At decision onset, responses to trial-wise blameworthiness in anterior cingulate cortex (ACC) and bilateral lateral prefrontal cortex (LPFC) positively correlated with hypocritical blame (family-wise error corrected $p < .05$, whole-brain corrected at the cluster level after voxel-wise thresholding at $p < .001$). HRF = hemodynamic response function.

Hypocritical blame is associated with a neural signature of guilt during moral decision-making

Folk intuitions about hypocritical blamers suggest that hypocritical blamers are less moral than people who behave equally badly but do not find their behavior blameworthy. This implies that hypocritical blamers will experience weak (or no) feelings of guilt when they consider harming other people for profit. In contrast, our hypothesis that some hypocritical blame is due to weakness of will predicts a positive relationship between guilt and hypocritical blame. To avoid inducing demand effects or social desirability concerns, we

did not ask participants to directly report feelings of guilt during the moral decision-making task. Instead, we leveraged a multivariate brain-based classifier trained on two independent data sets to identify brain states that positively predict guilt evoked by interpersonal interactions (GRBS; Yu et al., 2020). GRBS corresponds to a distributed brain network that exhibits only weak spatial similarity with other social and affective brain signatures, suggesting that guilt is associated with a specific pattern of neural activity (Yu et al., 2020). Signatures of different psychological constructs have different distributions of prediction weights across the brain. The primary goal of developing a brain signature is to predict neurocognitive processes and

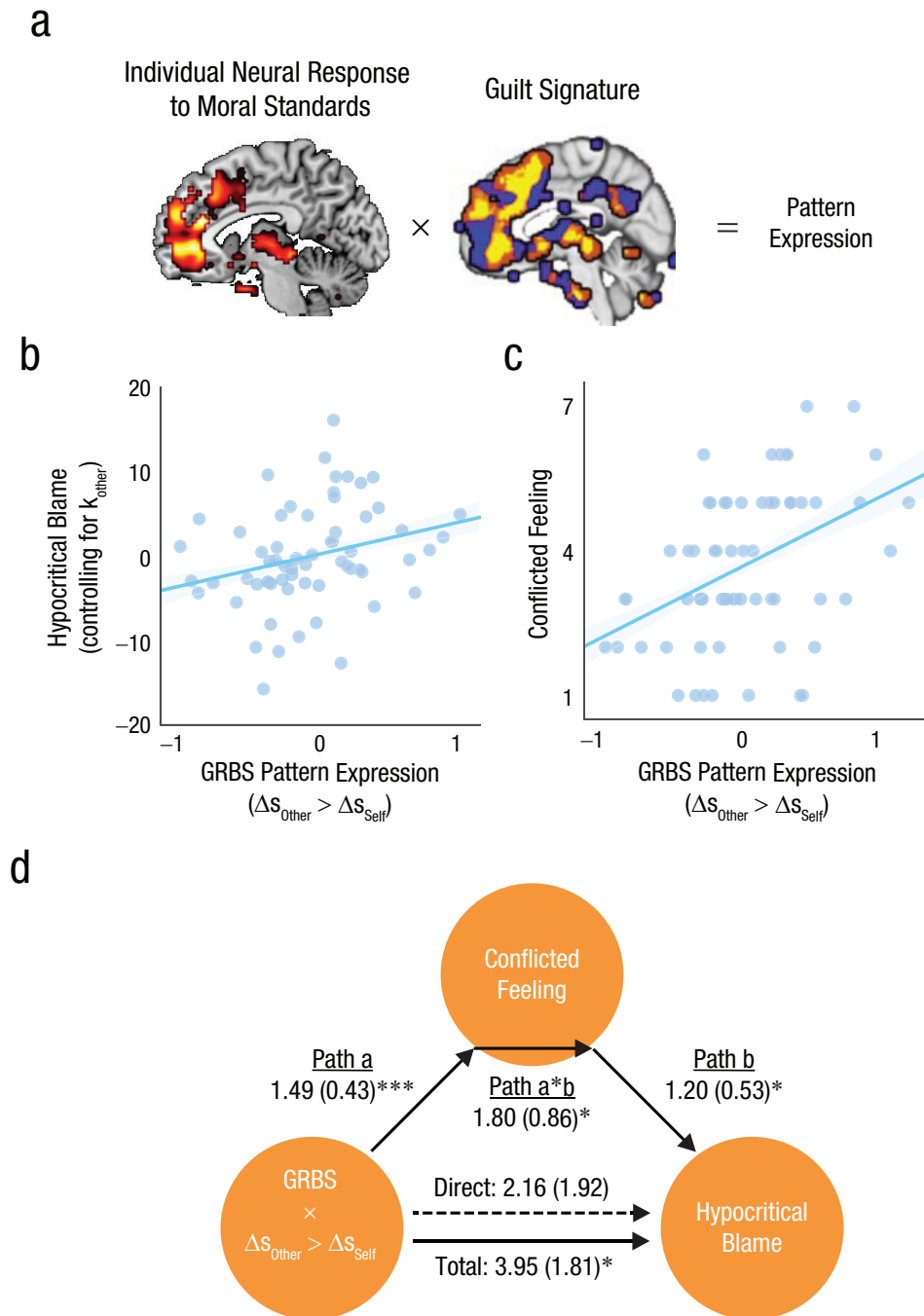


Fig. 6. Hypocritical blame is associated with neural signature of guilt at the time of decision. (a) Schematic illustration of the definition of pattern expression. (b, c) Guilt-related brain signature (GRBS) pattern expression associated with the brain responses to shocks for the receiver (relative to shocks for themselves) was positively correlated both with hypocritical blame (b) and conflicted feeling (c) Error bands represent *SEM*. (d) A mediation analysis provided further evidence that conflicted feeling mediates the relationship between guilt-related processes associated with receiver's harm and hypocritical blame Coefficients are unstandardized. * $p < .05$, *** $p < .001$.

behaviors of out-of-sample observations based on brain activation patterns, not to localize structures whose univariate activation strength is sensitive to experimental manipulations (Wager et al., 2013; Woo et al., 2017).

Therefore, the signature is a map of prediction weights that is distributed across the whole brain (Fig. 6a). The GRBS offers a sensitive and specific brain-based indicator of guilt-related neurocognitive processes. In the

context of our moral decision-making task, guilt-related neurocognitive processes should be positively associated with the amount of additional pain inflicted (Δ shock), and/or negatively correlated with the amount of additional profit gained (Δ money), because these factors are positively and negatively correlated with blameworthiness judgments, respectively (Crockett et al., 2017; Siegel et al., 2017), and guilt is reliably evoked by anticipated blame (Baumeister et al., 1994; Ferguson et al., 1997; Parkinson & Illingworth, 2009).

A separate GLM was constructed to obtain brain response maps associated with Δ shock and Δ money in the *self* and the *other* condition (GLM 2; see the Method section). The dot product of GRBS and participants' brain maps of Δ shock and Δ money is a scalar value, pattern expression, which reflects the spatial resemblance between neural responses to Δ shock and Δ money and the brain-based signature of guilt (Fig. 6a). After calculating these values, we asked whether guilt pattern expressions associated with shocks and money in the *other* condition, relative to those in the *self* condition (i.e., Δ shock_{other} > Δ shock_{self} for shocks and Δ money_{other} > Δ money_{self} for money), were positively or negatively correlated with hypocritical blame, controlling for moral preference (κ_{other}). Contrary to folk intuitions and consistent with our hypothesis, guilt pattern expressions evoked by shocks—Fig. 6b; GRBSA_{shock}, Spearman's $\rho = .34$, $t(62) = 2.74$, $p = .008$ —were positively correlated with hypocritical blame. We did not observe a similar relationship for guilt pattern expressions evoked by money—GRBSA_{money}, Spearman's $\rho = .09$, $t = 0.69$, $p = .492$. Together, these findings reflect a positive relationship between guilt pattern expressions evoked by harm to other people and hypocritical blame.

Because self-reported feelings of conflict about the moral decision task were positively associated with hypocritical blame, we next considered whether those conflicted feelings could be explained by guilt-related neurocognitive processes at the time of decision. Consistent with this prediction, our results showed a positive correlation between self-reported conflicted feelings and GRBS pattern expression related to shocks—Fig. 6c; for GRBSA_{shock}, Spearman's $\rho = .42$, $p < .001$; for GRBSA_{money}, Spearman's $\rho = -.06$, $p = .655$. Next, we tested whether self-reported feelings of conflict mediated the positive relationship between GRBSA_{shock} and hypocritical blame in a mediation analysis (Imai et al., 2010) in which GRBSA_{shock} was entered as the independent variable, participants' conflicted feelings was the mediator, and hypocritical blame was the dependent variable. As in the regression analysis, κ_{other} was included as a covariate. We found a significant mediation effect of conflicted feelings (Fig. 6d; average mediation effect = 1.80, 95% CI = [0.36, 3.77], $p = .02$; average direct effect = 2.16,

95% CI = [-0.97, 5.61], $p = .186$), indicating that guilt-related neurocognitive processes evoked by harm to other people may lead to participants' conflicted feelings after the task, which in turn positively predict hypocritical blame.

A note on achieved power

We did not preregister our studies. Given the observed data, the sample sizes were generally good for the analysis we carried out. For Study 1, the achieved powers (assuming $\alpha = .05$) for the comparison of hypocrisy and trustworthiness across groups were 1.00 and .81, respectively. The effect size of the comparison of morality was smaller ($d = 0.24$), so the achieved power for this analysis was low (.37). We acknowledge that this result should be interpreted with caution. For Study 2, our critical analysis was the correlation between hypocritical blame and guilt pattern expressions evoked by shocks. Given the effect size of the correlation, our final sample ($N = 62$) achieved a power of .8 in detecting the correlation.

Discussion

In this study, we developed a laboratory paradigm to precisely quantify hypocritical blame, in which people blame others for committing the same transgressions they committed themselves (Todd, 2019). At the core of this operationalization of hypocrisy is a discrepancy between participants' moral judgments and their behaviors in a moral decision-making task. Therefore, we measured participants' choices in an incentivized moral decision-making task that they believed had real impact on their own monetary payoff and painful electric shocks delivered to a receiver. We then compared those choices with moral judgments they made a week later of other people in the same choice context. By comparing participants' judgments with their own behaviors, we were able to quantify the degree to which they judge other people more harshly for making the same choices they themselves made previously (i.e., hypocritical blame).

We found that hypocritical blame was positively associated with (a) feelings of moral conflict arising from failures to live up to one's own moral standards and (b) neural representations of moral standards and guilt during moral decision-making. This suggests that at least some instances of hypocritical blame are not attributable to a lack of moral standards for oneself but instead may manifest from failures to consistently live up to one's own moral standards.

Our results also shed new light on the neural basis of moral standards. Past work has implicated lateral and

medial prefrontal regions in representing moral values (Crockett et al., 2017; FeldmanHall et al., 2015; Qu et al., 2019; van Baar et al., 2019; for a review, see Carlson & Crockett, 2018). Here, we extend these findings by showing that activity in these regions is also sensitive to the spectrum of moral attitudes behind the same external behavioral pattern. Holding constant the behavioral preference for avoiding harm to others (κ_{other}), we found that LPFC and ACC responses to moral standards predicted how far one's moral judgments ultimately deviated from their behavior. This relationship is in line with our finding that individuals who exhibited more hypocritical blame also reported more conflicted feelings during decision-making, and it fits well into a broader context of LPFC and ACC functions in conflict monitoring in both nonmoral domains (Botvinick et al., 2001; Etkin et al., 2006) and moral domains (Buckholz et al., 2015; Carlson & Crockett, 2018; Van Bavel et al., 2015).

To further characterize the neurocognitive processes underlying hypocritical blame, we adopted a multivariate approach to the neuroimaging data, applying an independently trained brain signature of interpersonal guilt to the representation of moral standards. This brain-based signature enabled us to draw conclusions about the underlying neurocognitive processes from patterns of activity in the whole brain, thereby avoiding inferring cognitive processes based solely on the location of brain activations (Poldrack, 2011; Wager et al., 2013). Some theorists have postulated that apparently hypocritical blamers are not really hypocritical if they experience guilty feelings for violating the moral standards, because hypocrites do not sincerely endorse the moral value in question but express it only for strategic reasons (Bartel, 2019; Bell, 2013; Wallace, 2010). Our multivariate analysis provided empirical evidence supporting this theoretical conjecture by showing that guilt-related processes evoked by consideration of harm to other people tracks individual differences in hypocritical blame. Our mediation analysis further uncovered the guilt-related processes associated with the encoding of harm as a potential source of conflicted feelings reported by more hypocritical blamers.

Taken together, our findings may have implications for lay and philosophical debates about who can legitimately blame others. According to some accounts, the legitimacy of blame depends not just on facts about the target of blame but also on facts about the blamer (Todd, 2019; Tognazzini & Coates, 2018). For example, some researchers argue that hypocritical blame is illegitimate because "blame carries with it a kind of practical commitment to critical self-scrutiny" (Wallace, 2010, p. 326). This argument implies that all hypocritical blamers lack legitimacy to blame other people if they do not blame themselves as much as they blame others.

However, our findings from Study 2 suggest that at least some instances of apparently hypocritical blame may not in fact be disqualifying for standing to blame, the normative requirements blamers have to meet so that their blame is appropriate and fitting (Todd, 2019), because they are associated with signs of self-scrutiny (e.g., feeling guilt and regret for one's past wrongdoing) that are required for standing to blame (Bell, 2013; Wallace, 2010).

Several limitations of this work are worth noting. First, our findings do not directly speak to the precise neurocognitive mechanisms that give rise to moral judgments that conflict with past moral behaviors. Although our data provide evidence that at least some instances of hypocritical blame cannot be explained by an absence of moral standards, we cannot definitively conclude that such cases are attributable to weakness of will (Bartel, 2019; Batson & Thompson, 2001; Mele, 1989), though our finding that hypocritical blame is positively associated with harmful errors does provide some preliminary evidence for this claim. Future studies might investigate this question with causal interventions designed to induce mental fatigue (Inzlicht et al., 2014; Schmidt et al., 2012). Second, our work cannot shed light on the neural mechanisms that underlie the construction of hypocritical blame judgments. While the current investigation focused on the neurocognitive processes underlying moral decision-making and how they are related to subsequent hypocritical blame, future studies could scan participants while they are making blame judgments that are either consistent or inconsistent with their past behavior. Third, we can draw conclusions about hypocritical blame only in the context of physical harm, which is just one of many types of moral transgressions (Graham et al., 2013; Schein & Gray, 2017). However, in philosophy (McKinnon, 1991; Moberg, 1987; Shklar, 1984) and psychology (Ji et al., 2006; Yousaf & Gobet, 2013), hypocritical blame has also been widely associated with violation of religious beliefs, sexual norms, and moral concerns unrelated to physical harm (e.g., purity, loyalty, self-discipline). Future work could adopt our computational operationalization and multivariate neuroimaging approach to investigate whether similar or distinct neurocognitive processes characterize hypocritical blame across a variety of moral domains, thereby ascertaining the conceptual and mechanistic complexity involved.

To conclude, we developed a model of hypocritical blame that allowed us to quantify a tendency for people to blame others for the same behaviors that they themselves committed and test several common assumptions about hypocritical blamers. In marked contrast to the intuitions of observers, our findings showed that hypocritical blamers' self-reports and neural activity during moral decision-making is consistent with the

presence rather than absence of moral standards. Participants with higher levels of hypocritical blame reported more intense conflicted feelings and showed heightened guilt-related neural responses when considering harming other people. Thus, contrary to the common assumption that hypocritical blamers do not really accept the moral standards that they apply to other people, our data suggest that many people who engage in hypocritical blame do hold and care about moral standards and apply these moral standards to themselves as well as to others.

Transparency

Action Editor: Leah Somerville

Editor: Jamin Halberstadt

Author Contributions

H. Yu: conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualization, writing – original draft, writing – review and editing.

L. S. Contreras-Huerta: investigation, methodology, project administration, writing – review and editing.

A. M. B. Prosser: investigation, methodology, project administration, writing – review and editing.

M. A. J. Apps: conceptualization, writing – review and editing.

W. Hofmann: conceptualization, writing – review and editing.

W. Sinnott-Armstrong: conceptualization, writing – review and editing.

M. J. Crockett: conceptualization, funding acquisition, project administration, resources, supervision, writing – review and editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by grants from the John Templeton Foundation (Beacons Project and No. 61495), the Academy of Medical Sciences (SBF001\1008), the Oxford University Press John Fell Fund, and the Wellcome Trust Institutional Strategic Support Fund (204826/Z/16/Z) awarded to M. J. Crockett. H. Yu was supported by The Royal Society Newton International Fellowship (NF160700) and a Theresa Seessel Endowed Fellowship, Yale University. M. A. J. Apps was supported by a Biotechnology and Biological Sciences Research Council David Phillips Fellowship (BB/R010668/1).

Open Practices

This article has received the badge for Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

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

Luis Sebastian Contreras-Huerta <https://orcid.org/0000-0001-7747-590X>

Annayah Prosser <https://orcid.org/0000-0003-2381-9556>

Matthew A. J. Apps <https://orcid.org/0000-0001-5793-2202>

Walter Sinnott-Armstrong <https://orcid.org/0000-0003-2579-9966>

Molly J. Crockett <https://orcid.org/0000-0001-8800-410X>

Acknowledgments

We thank Linda Skitka, Jenifer Siegel, and members of the Crockett Lab for suggestions on study design, data analysis, and comments on an earlier version of the manuscript. We also thank Michel-Pierre Coll, Anne-Marie Nussberger, Mary Montgomery, Talia Longthorne, Eloise Copland, Heather Koh, Cassandra Popham, and Wenchuan Wu for assistance in data collection.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/09567976221122765>

References

- Alicke, M. D., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, *26*(5), 673–701. <https://doi.org/10.1080/09515089.2012.677397>
- Barden, J., Rucker, D. D., & Petty, R. E. (2005). “Saying one thing and doing another”: Examining the impact of event order on hypocrisy judgments of others. *Personality and Social Psychology Bulletin*, *31*(11), 1463–1474.
- Bartel, C. (2019). Hypocrisy as either deception or akrasia. *The Philosophical Forum*, *50*(2), 269–281.
- Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, *72*(6), 1335–1348. <https://doi.org/10.1037/0022-3514.72.6.1335>
- Batson, C. D., & Thompson, E. R. (2001). Why don't moral people act morally? Motivational considerations. *Current Directions in Psychological Science*, *10*(2), 54–57. <https://doi.org/10.1111/1467-8721.00114>
- Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, *83*(2), 330–339. <https://doi.org/10.1037/0022-3514.83.2.330>
- Batson, C. D., Thompson, E. R., Seufferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–537.
- Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267. <https://doi.org/10.1037/0033-2909.115.2.243>
- Bell, M. (2013). The standing to blame: A critique. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 263–281). Oxford Academic.
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

- Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655–661.
- Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O., & Marois, R. (2015). From blame to punishment: Disrupting prefrontal cortex activity reveals norm enforcement mechanisms. *Neuron*, *87*(6), 1369–1380.
- Carlson, R. W., & Crockett, M. J. (2018). The lateral prefrontal cortex and moral goal pursuit. *Current Opinion in Psychology*, *24*, 77–82.
- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences, USA*, *111*(48), 17320–17325. <https://doi.org/10.1073/pnas.1408988111>
- Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature Neuroscience*, *20*(6), 879–885. <https://doi.org/10.1038/nn.4557>
- Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., Grosse-Rueskamp, J. M., Dayan, P., & Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology*, *25*(14), 1852–1859. <https://doi.org/10.1016/j.cub.2015.05.021>
- Dover, D. (2019). The walk and the talk. *Philosophical Review*, *128*(4), 387–422.
- Effron, D. A., O'Connor, K., Leroy, H., & Lucas, B. J. (2018). From inconsistency to hypocrisy: When does “saying one thing but doing another” invite condemnation? *Research in Organizational Behavior*, *38*, 61–75.
- Etkin, A., Egner, T., Peraza, D. M., Kandel, E. R., & Hirsch, J. (2006). Resolving emotional conflict: A role for the rostral anterior cingulate cortex in modulating activity in the amygdala. *Neuron*, *51*(6), 871–882.
- FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *Neuroimage*, *105*, 347–356.
- FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. <https://doi.org/10.1016/j.cognition.2012.02.001>
- Ferguson, T. J., Oothof, T., & Stegge, H. (1997). Temporal dynamics of guilt: Changes in the role of interpersonal and intrapsychic factors. *European Journal of Social Psychology*, *27*(6), 659–673. [https://doi.org/10.1002/\(SICI\)1099-0992\(199711/12\)27:6<659::AID-EJSP837>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1099-0992(199711/12)27:6<659::AID-EJSP837>3.0.CO;2-K)
- Fritz, K. G., & Miller, D. (2018). Hypocrisy and the standing to blame. *Pacific Philosophical Quarterly*, *99*(1), 118–139.
- Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition*, *30*(6), 652–668.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.
- Graham, J., Meindl, P., Koleva, S., Iyer, R., & Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, *9*(3), 158–170. <https://doi.org/10.1111/spc3.12158>
- Green, S., Ralph, M. A. L., Moll, J., Deakin, J. F. W., & Zahn, R. (2012). Guilt-selective functional disconnection of anterior temporal and subgenual cortices in major depressive disorder. *Archives of General Psychiatry*, *69*(10), 1014–1021.
- Howe, L. C., & Monin, B. (2017). Healthier than thou? “Practicing what you preach” backfires by increasing anticipated devaluation. *Journal of Personality and Social Psychology*, *112*(5), 718–753.
- Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451–462.
- Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2010). Causal mediation analysis using R. In H. Vinod (Ed.), *Advances in social science research using R* (pp. 129–154). Springer.
- Inzlicht, M., Schmeichel, B. J., & Macrae, C. N. (2014). Why self-control seems (but may not be) limited. *Trends in Cognitive Sciences*, *18*(3), 127–133.
- Ji, C.-H. C., Pendergraft, L., & Perry, M. (2006). Religiosity, altruism, and altruistic hypocrisy: Evidence from protestant adolescents. *Review of Religious Research*, *48*(2), 156–178.
- Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368.
- Kittay, E. F. (1982). On hypocrisy. *Metaphilosophy*, *13*(3–4), 277–289. <https://doi.org/10.1111/j.1467-9973.1982.tb00685.x>
- Koban, L., Corradi-Dell'Acqua, C., & Vuilleumier, P. (2013). Integration of error agency and representation of others' pain in the anterior insula. *Journal of Cognitive Neuroscience*, *25*(2), 258–272.
- Laurent, S. M., & Clark, B. A. M. (2019). What makes hypocrisy? Folk definitions, attitude/behavior combinations, attitude strength, and private/public distinctions. *Basic and Applied Social Psychology*, *41*(2), 104–121.
- Lythe, K. E., Moll, J., Gethin, J. A., Workman, C. I., Green, S., Ralph, M. A. L., Deakin, J. F. W., & Zahn, R. (2015). Self-blame-selective hyperconnectivity between anterior temporal and subgenual cortices and prediction of recurrent depressive episodes. *JAMA Psychiatry*, *72*(11), 1119–1126.
- McKinnon, C. (1991). Hypocrisy, with a note on integrity. *American Philosophical Quarterly*, *28*(4), 321–330.
- Mele, A. R. (1989). Akritic feelings. *Philosophy and Phenomenological Research*, *50*(2), 277–288.
- Moberg, D. O. (1987). Holy masquerade: Hypocrisy in religion. *Review of Religious Research*, *29*(1), 3–24.
- Monin, B., & Merritt, A. (2012). Moral hypocrisy, moral inconsistency, and the struggle for moral integrity. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 167–184). American

- Psychological Association. <https://doi.org/10.1037/13091-009>
- O'Connor, K., Effron, D. A., & Lucas, B. J. (2020). Moral cleansing as hypocrisy: When private acts of charity make you feel better than you deserve. *Journal of Personality and Social Psychology*, *119*(3), 540–559. <https://doi.org/10.1037/pspa0000195>
- Parkinson, B., & Illingworth, S. (2009). Guilt in response to blame from others. *Cognition and Emotion*, *23*(8), 1589–1614.
- Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697.
- Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J.-C. (2019). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLOS Biology*, *17*(6), Article e3000283. <https://doi.org/10.1371/journal.pbio.3000283>
- Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545–556. <https://doi.org/10.1038/nrn2357>
- Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562.
- Schein, C., & Gray, K. (2017). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. <https://doi.org/10.1177/1088868317698288>
- Schmidt, L., Lebreton, M., Cléry-Melin, M.-L., Daunizeau, J., & Pessiglione, M. (2012). Neural mechanisms underlying motivation of mental versus physical effort. *PLOS Biology*, *10*(2), Article e1001266. <https://doi.org/10.1371/journal.pbio.1001266>
- Shklar, J. N. (1984). *Ordinary vices*. Harvard University Press.
- Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. <https://doi.org/10.1016/j.cognition.2017.05.004>
- Szabados, B., & Soifer, E. (1999). Hypocrisy, change of mind, and weakness of will: How to do moral philosophy with examples. *Metaphilosophy*, *30*(1–2), 60–78. <https://doi.org/10.1111/1467-9973.00112>
- Szabados, B., & Soifer, E. (2004). *Hypocrisy: Ethical investigations*. Broadview Press.
- Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology*, *58*, 345–372.
- Todd, P. (2019). A unified account of the moral standing to blame. *Noûs*, *53*(2), 347–374.
- Tognazzini, N., & Coates, D. J. (2018). Blame. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy*. <https://plato.stanford.edu/archives/fall2018/entries/blame/>
- Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689–690.
- van Baar, J. M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*, Article 1483. <https://doi.org/10.1038/s41467-019-09161-6>
- Van Bavel, J. J., FeldmanHall, O., & Mende-Siedlecki, P. (2015). The neuroscience of moral cognition: From dual processes to dynamic systems. *Current Opinion in Psychology*, *6*, 167–172.
- Volz, L. J., Welborn, B. L., Gobel, M. S., Gazzaniga, M. S., & Grafton, S. T. (2017). Harm to self outweighs benefit to others in moral decision making. *Proceedings of the National Academy of Sciences, USA*, *114*(30), 7963–7968.
- Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388–1397.
- Wallace, R. J. (2010). Hypocrisy, moral address, and the equal standing of persons. *Philosophy & Public Affairs*, *38*(4), 307–341.
- Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377.
- Yousaf, O., & Gobet, F. (2013). The emotional and attitudinal consequences of religious hypocrisy: Experimental evidence using a cognitive dissonance paradigm. *The Journal of Social Psychology*, *153*(6), 667–686.
- Yu, H., Hu, J., Hu, L., & Zhou, X. (2014). The voice of conscience: Neural bases of interpersonal guilt and compensation. *Social Cognitive and Affective Neuroscience*, *9*(8), 1150–1158. <https://doi.org/10.1093/scan/nst090>
- Yu, H., Koban, L., Chang, L. J., Wagner, U., Krishnan, A., Vuilleumier, P., Zhou, X., & Wager, T. D. (2020). A generalizable multivariate brain pattern for interpersonal guilt. *Cerebral Cortex*, *30*(6), 3558–3572.
- Zahn, R., Moll, J., Paiva, M., Garrido, G., Krueger, F., Huey, E. D., & Grafman, J. (2009). The neural basis of human social values: Evidence from functional MRI. *Cerebral Cortex*, *19*(2), 276–283. <https://doi.org/10.1093/cercor/bhn080>
- Zoh, Y., Chang, S. W., & Crockett, M. J. (2022). The prefrontal cortex and (uniquely) human cooperation: A comparative perspective. *Neuropsychopharmacology*, *47*(1), 119–133.