Check for updates

Received: 6 July 2020 | Revised: 19 August 2021 | Accepted: 21 August 2021

DOI: 10.1002/hbm.25648

RESEARCH ARTICLE

WILEY

Who initiates punishment, who joins punishment? Disentangling types of third-party punishers by neural traits

Thomas Baumgartner[1] | Jan Hausfeld[1,2] | Miguel dos Santos[1] | Daria Knoch[1]

[1]Department of Social Neuroscience and Social Psychology, Institute of Psychology, University of Bern, Bern, Switzerland

[2]CREED, Amsterdam School of Economics, University of Amsterdam, Amsterdam, Netherlands

Correspondence
Thomas Baumgartner and Daria Knoch, Department of Social Neuroscience and Social Psychology, Institute of Psychology, University of Bern, Fabrikstrasse 8, Bern CH-3012, Switzerland.
Email: thomas.baumgartner@unibe.ch (T. B.) and daria.knoch@unibe.ch (D. K.)

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 100019_166006; Typhaine Foundation

Abstract

The act of punishing unfair behavior by unaffected observers (i.e., third-party punishment) is a crucial factor in the functioning of human societies. In everyday life, we see different types of individuals who punish. While some individuals initiate costly punishment against an unfair person independently of what other observers do (independent punishers), others condition their punishment engagement on the presence of another person who punishes (conditional punishers). Still others do not want to partake in any sort of punishment (nonpunishers). Although these distinct behavioral types have a divergent impact on human society, the sources of heterogeneity are poorly understood. We present novel laboratory evidence on the existence of these three types. We use anatomical brain characteristics in combination with stated motives to characterize these types. Findings revealed that independent punishers have larger gray matter volume in the right temporo-parietal junction compared to conditional punishers and nonpunishers, an area involved in social cognition. Conditional punishers are characterized by larger gray matter volume in the right dorsolateral prefrontal cortex, a brain area known to be involved in behavioral control and strategic reasoning, compared to independent punishers and nonpunishers. Finally, both independent punishers and nonpunishers are characterized by larger gray matter volume in an area involved in the processing of social and monetary rewards, that is, the bilateral caudate. By using a neural trait approach, we were able to differentiate these three types clearly based on their neural signatures, allowing us to shed light on the underlying psychological mechanisms.

KEYWORDS
brain anatomy, conditional punisher, neural trait, punishment types, third-party punishment

1 | INTRODUCTION

Human societies depend on the maintenance of elementary social norms. Individuals' willingness to sanction norm violations at a personal cost enforce and/or even promote many of these social norms

Thomas Baumgartner and Jan Hausfeld contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

(e.g., Balliet, Mulder, & Van Lange, 2011; Fehr & Gächter, 2002; Henrich et al., 2006). When group members have the opportunity to engage in punishment, punishment suffers from the known problem that some individuals are not willing to punish but still reap the benefits from other individuals who do so, that is, second-order free-rider problem (e.g., Fehr & Gächter, 2002; Panchanathan & Boyd, 2004). In contrast to the vast literature on the first-order free-rider problem (i.e., contributions to a public good) and the accompanying typology of contributors (see e.g., conditional cooperation by Fischbacher, Gächter, and Fehr (2001)), the existence of different types of punishers has barely been explored (two exceptions are Kamei (2014), Molleman, Kölle, Starmer, and Gächter (2019), although several other studies relate punishment to other behaviors or preferences, e.g., Brañas-Garza, Espín, Exadaktylos, and Herrmann (2014), Falk, Fehr, and Fischbacher (2005), Herrmann, Thöni, and Gächter (2008), and Kurzban, DeScioli, and O'Brien (2007)).

This neglect on the punisher typology comes as a surprise, as it seems rather natural to assume that there is heterogeneity in whether other people's punishment choices affect the own punishment choice: Some people want to ensure punishment is implemented and punish independently of other people's punishment behavior. Without this type of "independent punishers," no punishment would occur and norms would no longer be enforced. Other people might prefer to condition their own punishment choice on whether another person punishes. This "conditional punisher" type both potentially engages in costly norm-enforcement, but by following others' actions, also risks that punishment might not occur at all. When no one else does anything, it is easier to feel that taking action is not necessary, or even appropriate. This would be in line with the conceptual framework of El Zein, Bahrami, and Hertwig (2019), who propose that the choice to participate in collective decision-making minimizes the material and psychological burden of an individual's responsibility. Most previous studies on second-party punishment (two exceptions, i.e., Kamei, 2014, Molleman et al., 2019) and all studies on third-party punishment have neglected conditional punishers, and instead just allowed for two types (punishers or nonpunishers). Permitting conditional punishment could lead to more subjects sanctioning compared to only binary yes-or-no punishment choices, that is, if people choose "conditional punishment" when available but choose "no punishment" in a binary choice. In contrast, the potentially strategic motives for participating in group-punishment are negligible for people who punish independently of others' choices and people who would never punish, that is, "nonpunishers."

We aim to identify and characterize these different types of third-party punishers (independent punishers, conditional punishers, and nonpunishers) using an objective individual trait measure (structural brain characteristics) and using stated motives for further characterization. Neuroanatomical differences are useful markers for explaining individual variability because structural differences are relatively stable over time in healthy adult subjects, demonstrate a high individual specificity, and have been shown to be useful in predicting individual differences in various traits, skills, and behaviors (e.g., Kanai & Rees, 2011; Nash, Gianotti, & Knoch, 2015; Valizadeh, Liem, Merillat, Hanggi, & Jancke, 2018). Importantly, neural traits are objectively indexed, brain-based measures that are free from personal biases and demand characteristics. Thus, behavioral performance is left unadulterated by the act of completing trait measures and vice versa.

We investigate punishment behavior in a third-party setting. The two experimental studies on a similar typology of punishers, that is, Kamei (2014) and Molleman et al. (2019), use a second-party setting and find that people tend to mimic others' punishment decisions, even though some people are unaffected by others' choices. Further, based on the expressed level of the punishers' anger, Molleman et al. (2019) suggest that "compared to independent punishers, preferences of conditional punishers might perhaps reflect a more deliberative attitude, with behavior relatively less likely to be driven by negative emotions." This suggestion strengthens the idea of the conditional punisher acting strategically, and shows that different motives and emotions affect the punishment decision. In contrast to second-party punishment, third-parties' motives for punishment are less ambiguous: Third-party punishment is often considered an altruistic or prosocial act (e.g., Fehr & Fischbacher, 2004; Kurzban et al., 2007; Mathew & Boyd, 2011), whereas antisocial motives such as retaliation often drive second-party punishment (Carpenter & Matthews, 2012; Zhou, Jiao, & Zhang, 2017). In fact, some researchers view the existence of third-party punishment as the decisive factor for the enforcement of social norms in human societies (Fehr & Fischbacher, 2004; Henrich et al., 2006, 2010; Mathew & Boyd, 2011).

In our study, we developed a new third-party punishment paradigm to identify three distinct behavioral punishment types (see Figure 1). Participants form groups of four who observe a very unfair behavior in an unrelated interaction between two other people, that is, defecting after observing cooperating in a sequential prisoner's dilemma. As a novel feature, the four potential punishers can indicate their punishment choice for the norm violator by choosing between three different options separating the following punishment types: First, people who want to participate in the punishment of unfair behavior, but only if another person punishes as well ("conditional punisher"). By conditioning their behavior on others' punishment decisions, these people strategically choose to share the punishment burden, but also risk that the norm transgressor might get away without punishment if no one takes the lead. Second, this "letting the norm transgressor get away" would be unacceptable for people who are willing to incur costs even if no one else punishes ("independent punishers"). By doing so, these (potentially solitary) altruistic punishers make sure that unfair behavior is sanctioned and encourage a more cooperative environment (O'Gorman, Henrich, & Vugt, 2009). Third, a last group of people does not punish at all ("nonpunishers").

Neural traits associated with certain functions provide inferences of both how and why people differ. The current research thus sought to answer the following: Can we identify the neural signatures underlying different behavioral types in third-party punishment behavior? And, once identified, what do these signatures reveal about the psychological processes driving these three types' punishment behavior?

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

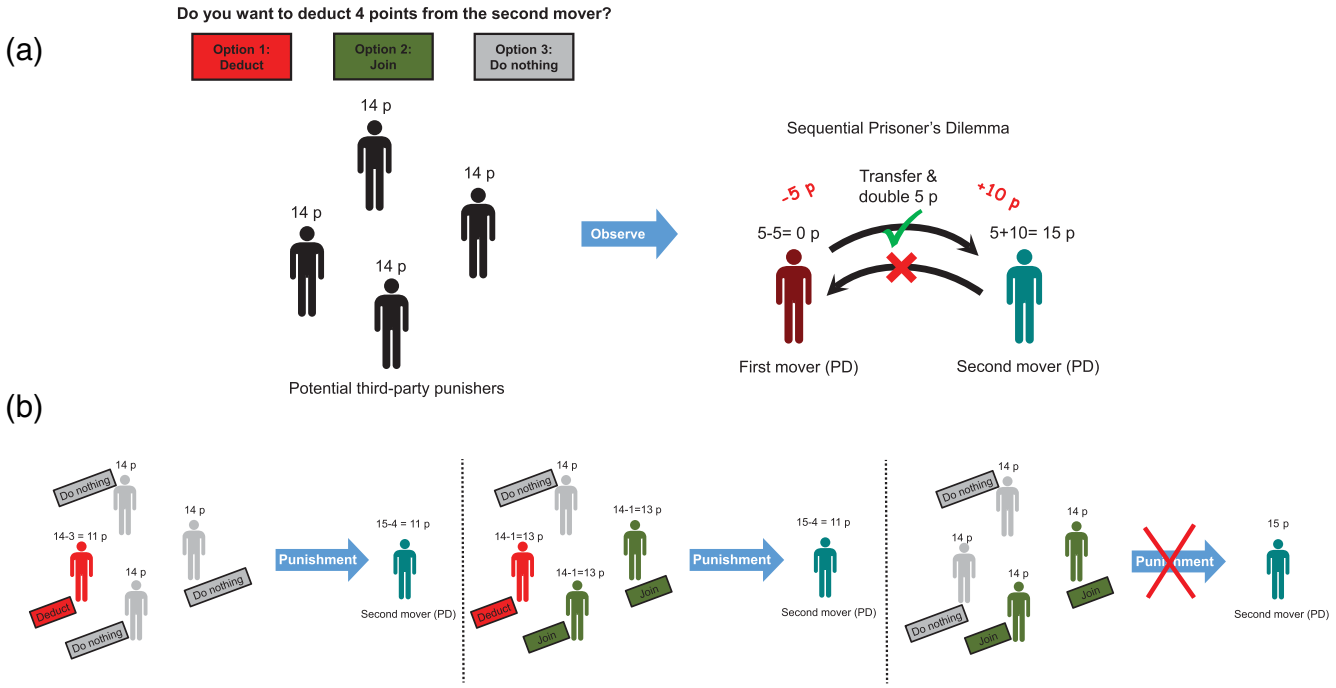

FIGURE 1 Experimental paradigm and possible choice combinations. (a) Four potential third-party punishers are within one group (left side) and are each endowed with 14 points. They observe an interaction between two other players in a sequential prisoner's dilemma, in which a first mover transfers points and a second mover (sequentially) does not transfer points back (right side). The potential punishers can choose between three different options on whether to sanction the defecting second mover in the PD: Option 1: I deduct points (i.e., to punish independently, red button); Option 2: I join the deduction of points (i.e., to punish conditionally, green button) or Option 3: I do nothing (i.e., no punishing, gray button). The third-party punishers were aware that (i) punishment only takes place if at least one of the third-party punishers decides to punish independently (chooses Option 1), (ii) the costs for punishment (3 points in total) were shared equally among all players taking part in the punishment (choose option 1 or 2), and (iii) the defecting second mover loses 4 points if punishment takes place. (b) Three potential choice scenarios and associated consequences for punishment. In the left scenario, one subject decides to "deduct points" (Option 1, red) and no other subject "joins." Thus, the second mover of the PD is punished and the sole punisher has to pay the total costs of 3 points for punishment by him/herself. In the middle scenario, one subject decides to "deduct points" (Option 1, red) and two subjects decide to "join" (Option 2, green). Thus, the second mover is punished, and the total costs of 3 points for punishment are divided among the red and green subjects, that is, all three pay 1 point. In the scenario on the right hand side, two subjects decide to "join" (Option 2, green), but no subject chooses to "deduct points" (Option 1, red). Thus, the second mover does not get punished

Third-party punishment has been explored in previous task-dependent functional neuroimaging and brain stimulation studies. These studies indicated that brain areas associated with social cognition (e.g., temporo-parietal junction, TPJ), emotion processing (e.g., amygdala, insula), reward processing (caudate), and behavioral control and strategic reasoning (dorsolateral regions of the prefrontal cortex, DLPFC) play an important role (for reviews and meta-analyses, see Bellucci, Camilleri, Iyengar, Eickhoff, & Krueger, 2020, Buckholtz & Marois, 2012, Krueger & Hoffman, 2016). For example, studies using task-dependent functional neuroimaging and brain stimulation report that punishment decisions of third-parties rely on mentalizing regions such as the TPJ. It has been suggested that the TPJ plays a critical role in inferring the intentions of the perpetrator and representing the victim's needs and appreciations (Baumgartner, Gotte, Gugler, & Fehr, 2012; Buckholtz et al., 2008; Gerfo et al., 2019; Ginther et al., 2016). Further, previous neuroimaging studies also consistently show that DLPFC is implicated in the decision process. The DLPFC is activated when subjects decided to punish and coded the amount of punishment assigned to the perpetrator (Buckholtz et al., 2008; Ginther et al., 2016; Strobel et al., 2011; Zhong, Chark, Hsu, & Chew, 2016). Moreover, inhibiting the function of the right DLPFC by rTMS reduced third-parties' punishment of wrongful acts (Buckholtz et al., 2015). Finally, it has been suggested that reward-related areas of the caudate (in particular dorsal parts) play a role in motivating and reinforcing the punishment act in third-parties (Baumgartner et al., 2012; Hu, Strang, & Weber, 2015; Strobel et al., 2011), whereas emotion-related areas (e.g., amygdala, anterior insula) are thought to detect the norm violation and to generate an aversive emotional response (Buckholtz & Marois, 2012; Civai, Crescentini, Rustichini, & Rumiati, 2012; Ginther et al., 2016).

Although these mentioned studies help to understand the neural mechanism of third-party punishment, none of these studies differentiated between independent punishers, conditional punishers, and nonpunishers. Moreover, most of the previous studies focused on task-dependent brain processes. Thus, we know little to nothing about the distinct (task-independent) neural traits that help characterizing these third-party punishment types.

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Based on the neuroimaging findings on third-party punishment mentioned above and the discussed possible motives driving third-party's punishment choice, we derive the following tentative hypotheses. Among the regions playing a critical role in third-party punishment, the DLPFC and the TPJ might be particularly interesting candidates driving third-parties' punishment choice. The DLPFC is known to play a critical role in behavioral control and strategic decision-making (e.g., Baumgartner, Dahinden, Gianotti, & Knoch, 2019; Gianotti, Nash, Baumgartner, Dahinden, & Knoch, 2018; Soutschek, Sauter, & Schubert, 2015; Spitzer, Fischbacher, Herrnberger, Gron, & Fehr, 2007; Steinbeis, Bernhardt, & Singer, 2012; Yamagishi et al., 2016). We speculate that conditional punishers have a more strategic nature which might be driven by a larger DLPFC (i.e., larger gray matter volume) compared to the other two types. Further, studies in the field of cooperation and prosocial behavior have associated task-dependent and task-independent brain characteristics of the TPJ with altruistic choices (e.g., Baumgartner et al., 2019; Hare, Camerer, Knoepfle, & Rangel, 2010; Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012; Park et al., 2017). We hence speculate that independent punishers have a more altruistic nature/inclination which might be driven by a larger TPJ (i.e., larger gray matter volume) compared to the two other types. Since the previous studies only allow inferring tentative hypotheses, we run exploratory analyses to investigate the neural traits of independent, conditional, and nonpunishers, focusing on brain areas shown to be critically involved in third-party punishment.

2 | MATERIALS AND METHODS

2.1 | Participants and procedure

We recruited 104 students from the University of Bern for participation in this study in the role of third-party punishers. Two of these participants had to be excluded due to artifacts in the anatomical brain data. Thus, we analyzed a final sample of 102 participants (46 female, 56 male, average 23.34 years old with *SD* 2.89). Students of economics, psychology, and social sciences were excluded from participation to reduce the possibility of prior knowledge of the concept of third-party punishment. We recruited participants for one academic year. The goal was to reach a sample size of about 80–100 subjects since two recent methodological studies on fMRI research concluded that sample sizes close to $N = 80$ (Geuter, Qi, Welsh, Wager, & Lindquist, 2018) or $N = 100$ (Turner, Paul, Miller, & Barbey, 2018) are required to reliably and reproducibly recover brain regions with medium effect sizes. Data were analyzed after the collection was complete. All participants in the role of third-party punishers were right-handed, nonsmokers, and reported no history of psychological disorders or neurological or cardiovascular disease. The study was approved by the local ethics committee and participants signed informed written consent prior to the participation in the study. Subjects received a show-up fee of 40 Swiss francs and the points that they earned during the experiment with a currency conversion rate of 2 points (p) = 1 CHF. The experiment was programmed in z-tree (Fischbacher, 2007) and subjects were recruited via ORSEE (Greiner, 2015). Neuroimaging (structural brain data, see below) and behavioral data collections were conducted in different sessions. Behavioral data collections were conducted in a behavioral laboratory with 24 interconnected computer terminals. Participants were randomly assigned to cubicles where they could make their decisions in isolation from others. Before starting the experiment, control questions ensured participants' understanding of the game.

2.2 | Paradigm

The study used a newly developed version of a third-party punishment paradigm. The third-party consisted of four subjects (potential punishers) who observed an interaction between two other subjects (first mover and second mover) who played a sequential prisoner's dilemma (PD) game. Notably, the first and second mover in the PD were present in the laboratory and were paid according to their decisions and the decisions of the third-parties punishers (see below). The first and second mover in the PD both received an initial endowment of 5 points. Then, the first mover could decide whether to transfer 5 points to the second mover or to keep all the points. The second mover learned about the first mover's decision and then had to decide him/herself whether to transfer back 5 points or to keep all points (see Figure 1a, right side). Notably, all transferred points are multiplied by two, for example, if both players transfer points, both receive 10 points in total. If only one person transfers points, the transferring person would receive 0 points while the person keeping the points would receive 15 points. If neither person transfers points, both keep 5 points.

The four potential punishers learned about the underlying structure of the interaction between the first mover and second mover in the PD. Each potential punisher received an initial endowment of 14 points. Their task was to decide whether they want to deduct 4 points from the second mover, in case the first mover decided to transfer points and the second mover kept the points, that is, the first mover cooperated, while the second mover defected. More specifically, they were asked: "Do you want to deduct 4 points from the second mover?" As a novel feature of the paradigm, each potential punisher could decide between the following three options that allowed separating the three predicted punishment types: *I deduct points* (Option 1, i.e., to punish independently); *I join the deduction of points* (Option 2, i.e., to punish conditionally); *I do nothing* (Option 3, i.e., not to punish, see Figure 1a, left side).

Critically, the four potential punishers were aware of the following consequences of the different options. They knew that 4 points would be deducted from the defecting second mover in the PD if at least one person chose Option 1 (i.e., to punish independently). This punishment was not additive and one subject choosing Option 1 was sufficient for punishment to be implemented, regardless of the number of punishers. Implementing punishment yielded a total cost of

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3 points, but this total cost would be shared equally among all players who participate in the punishment act, that is, subjects who chose either Option 1 (i.e., to punish independently) or Option 2 (i.e., to punish conditionally). Therefore, the per subject cost to deduct 4 points from the second mover was 3 points, 1.5 points, 1 point, and 0.75 points in Case 1, 2, 3, and 4 subjects participated in the punishment act, respectively (see Figure 1b, left and middle). Importantly, if subjects chose Option 2 (i.e., to punish conditionally), but no one chose Option 1 (i.e., to punish independently), punishment was not implemented (Figure 1b, right). The potential punishers' decision to deduct points was implemented only if the first mover had decided to transfer points and second mover had refused to transfer points. The first and second mover in the PD were aware that a group of four potential punishers could deduct 4 points from the second mover. The potential punishers were aware that the two other subjects knew that only the second mover in the PD could be punished.

2.3 | Ratings of unfairness and motives for the punishment choice

After deciding on the punishment of the defecting second mover (when the first mover cooperated), the potential punishers had to indicate the perceived unfairness of the second mover's behavior (on a 5-point Likert-scale, 1 = very unfair, 5 = very fair). Additionally, the potential punishers had to rate their agreement with eight motive statements. The statements were taken partially from Balafoutas, Grechenig, & Nikiforakis (2014) and self-created and aim to elicit potential motives underlying the punishment choice. For example, participants had to answer the following motive statements (for all statements, please see Table S1): (1) My choice was the morally right thing to do. (2) The second mover violated a social norm. (3) Deducting points from the second mover does not help anyone. Participants had to indicate how much they agreed with each motive statement (5-point Likert-scale, 1 = no agreement, 5 = full agreement). Finally, only conditional punishers had to rate their agreement with the following three motive statements thought to play a role for this type (5-point Likert-scale, 1 = no agreement, 5 = full agreement): (1) I wanted to help people deducting points by reducing the costs for them. (2) I did not want to be the only person who deducted points. (3) It was morally correct to support someone who chose to deduct points.

2.4 | Acquisition of anatomical brain data

Anatomical brain data was acquired on a Siemens MAGNETOM Prisma 3.0 Tesla whole-body scanner using a 64-channel head coil. T1-weighted 3D-modified driven equilibrium Fourier transformation (MDEFT) images were acquired from each subject (176 slices, field of view: 256 × 256 × 176, slice thickness: 1 mm, no gap, repetition time: 7.93 ms, echo time: 2.49 ms, flip angle: 16°).

2.5 | Preprocessing of anatomical brain data

Anatomical brain data was preprocessed with the computational anatomy toolbox (CAT 12, version 1450, Dahnke, Yotter, & Gaser, 2013) implemented in the statistical parametric mapping software (SPM 12, version 7487). CAT 12 is documented and freely available online (http://www.neuro.uni-jena.de/cat/) and covers diverse morphometric methods. Here we focused on voxel-based morphometry (VBM)—a well-established whole-brain technique capable of discovering subtle, regionally specific changes in gray matter volume by averaging across subjects. This method is based on high-resolution structural three-dimensional magnetic resonance images, registered in standard space, and is designed to find significant regional gray matter differences throughout the whole brain by applying voxel-wise statistics within the context of Gaussian random fields (Ashburner & Friston, 2000). Preprocessing of the data involved spatial normalization (to a MNI template), segmentation into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), modulation, and spatial smoothing with a Gaussian kernel (full width at half maximum = 8 mm; Ashburner, 2007; Ashburner & Friston, 2000, 2005). In detail, the segmentation approach is based on an Adaptive Maximum A Posterior technique without the need for a priori information of tissue probabilities, uses a Partial Volume Estimation with a simplified mixed model of at most two tissue types (Tohka, Zijdenbos, & Evans, 2004), and applies a classical Markov Random Field approach, which incorporates spatial prior information of adjacent voxels into the segmentation estimation (Rajapakse, Giedd, & Rapoport, 1997). Finally, gray matter segmentations are modulated by scaling with the amount of volume changes due to spatial normalization, so that the total amount of gray matter in the modulated image remains the same, as it would be in the original image.

2.6 | Statistical analyses of psychometric data

Statistical analyses were run with the SPSS (version 25). See Section 3 for details about the statistical tests conducted, including independent *t*-tests as well as univariate ANOVAs with between-subject factor behavioral types (independent punishers, conditional punishers, and nonpunishers). Results were considered significant at the level of $p < .05$ (two-tailed), except for the eight motive statements all participants had to answer. Here, we applied a correction for multiple comparisons according to Bonferroni, resulting in a corrected *p*-value of 0.00625 (0.05/8). Thus, only findings reaching this *p*-value survive Bonferroni correction. We use the abbreviation *SD* for standard deviation. As effect size measure eta^2 is reported, which is a measure of explained variance.

2.7 | Statistical analyses of anatomical brain data

In order to explore whether the behavioral types can be characterized by anatomical brain characteristics, we performed univariate analysis

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of variance (ANOVA) with between-subject factor behavioral types (independent punishers, conditional punishers, and nonpunishers) on the smoothed gray matter volume images in SPM 12. As is required for volumetric brain analyses, age, gender, and individual brain size were included in the design matrix as covariates of no interest to model and thus regress out any effects correlated with these covariates.

Given our a priori hypotheses (as outlined in the introduction), we focused in our analyses of anatomical group differences on brain areas playing a role in third-party punishment, including areas involved in mentalizing (TPJ), emotion processing (amygdala, insula, and anterior cingulate cortex), behavioral control and strategic decision-making (middle frontal gyrus), and reward processing (caudate nucleus). To this end, a bilateral mask comprising the amygdala, insula, anterior cingulate cortex, the middle frontal gyrus, the caudate, and TPJ (see below) was created, defined using the Automated Anatomical Labeling Atlas (Tzourio-Mazoyer et al., 2002), and implemented using the WFU Pickatlas toolbox in SPM 12 (Maldjian, Laurienti, Kraft, & Burdette, 2003). For the anatomical definition of the bilateral TPJ, we included the bilateral angular and superior temporal gyrus (posterior to $y = -40$), because several recent meta-analyses have consistently reported activation peaks from studies on mentalizing in this part of the brain (Carter & Huettel, 2013; van Overwalle, 2009). The same anatomical mask of the TPJ was used for small volume correction in previous studies on social decision-making (e.g., Hutcherson, Bushong, & Rangel, 2015).

In order to control for multiple comparisons, we used nonparametric permutation statistics based on threshold-free cluster enhancement (TFCE, Smith & Nichols, 2009) as implemented in the SPM 12 toolbox TFCE (version 201, http://dbm.neuro.uni-jena.de/tfce/). The idea of the TFCE approach is to combine focal effects with large voxel heights as well as broad effects (clusters). In contrast to common approaches that use cluster-based thresholding no initial (and arbitrary) cluster-forming threshold is necessary. TFCE takes a raw statistic image (e.g., t or f maps) and estimates a voxel-wise metrics (the TFCE values) by combining spatially distributed cluster size and height information. As recommended for volumetric data, we applied a cluster-size weighting of $E = 0.5$ and a height weighting of $H = 2.0$, and the Smith permutation method with 5,000 permutations (Smith & Nichols, 2009). Using voxel-level nonparametric permutation testing (based on the estimated TFCE values), voxel-wise p values are computed, which are family-wise error (FWE) corrected. All analyses in the manuscript report FWE-corrected (at p <.05) effects across the whole brain (whole-brain FWE corrected) or across the small volume mask defined above (SV-FWE corrected). In case of significant (FWE-corrected) effects in the univariate ANOVA (see above), we extracted gray matter volume values from regions demonstrating anatomical differences, regressed out age, gender and individual brain size (the covariates of no interest), and conducted post-hoc pairwise comparisons in SPSS (version 25) in order to explore the direction and specificity of the discovered anatomical differences between the three behavioral types.

2.8 | Discriminant analyses of motive statements and anatomical brain data

Finally, in order to get a more integrative view of all variables (structural brain characteristics and motive statements) characterizing the three types, we employed discriminant analyses. Discriminant analyses allow examining how well each individual subject can be classified into different types, based on motives, structural brain characteristics, or both. For that purpose, discriminant analyses determines the most parsimonious way to separate the types and discards predictors, which add little to the discrimination of the types. We use a stepwise estimation using Wilks' lambda for the inclusion of predictors (motives or structural brain characteristics or both), that is, in each step the predictor minimizing Wilk's lambda enters if it explains sufficient additional variance. The discriminant analyses involve deriving linear combinations of the included predictors, that is, discriminant functions, which yield coefficients that can be used to calculate a score for a respective discriminant function. The magnitude of the standardized coefficients indicate how strongly the predictors affect the score. Based on these estimated discriminant functions, individual discriminant scores for every subject and function can be calculated. Subsequently the centroids for each type can be determined and every subject can be classified based upon the distance to the centroids of each type. We use leave-one-out predictions by repeating the mentioned analyses n times with $n - 1$ subjects (i.e., the subject to be classified is not involved in the estimation of the discriminant functions) and we use equal priors, that is, we assume a 33.3% probability to be of either type. Discriminant analyses were calculated in SPSS (version 25).

3 | RESULTS

3.1 | Frequency of the behavioral types

The behavioral data consists of 102 participants in the role of the potential punisher. With our design, we can identify distinct types of third-party punishment, that is, independent punishers, conditional punishers and nonpunishers. We find that 19.6% participants are independent punishers, 24.5% are conditional punishers, and 55.9% chose not to punish at all (Figure 2a). Note that, we use the subscripts $_{\text{Indep}}$ for Independent Punishers, $_{\text{Cond}}$ for Conditional Punishers, and $_{\text{Non}}$ for Nonpunishers in the analysis section.

3.2 | Characterization of the behavioral types with psychometric ratings

In order to characterize the behavioral types, we explored the relation between the types, the perception of unfairness, and the agreement with different statements about the underlying motives. We use univariate ANOVAs with between-subject factor behavioral types for the analyses and subsequent post-hoc independent t-tests.

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

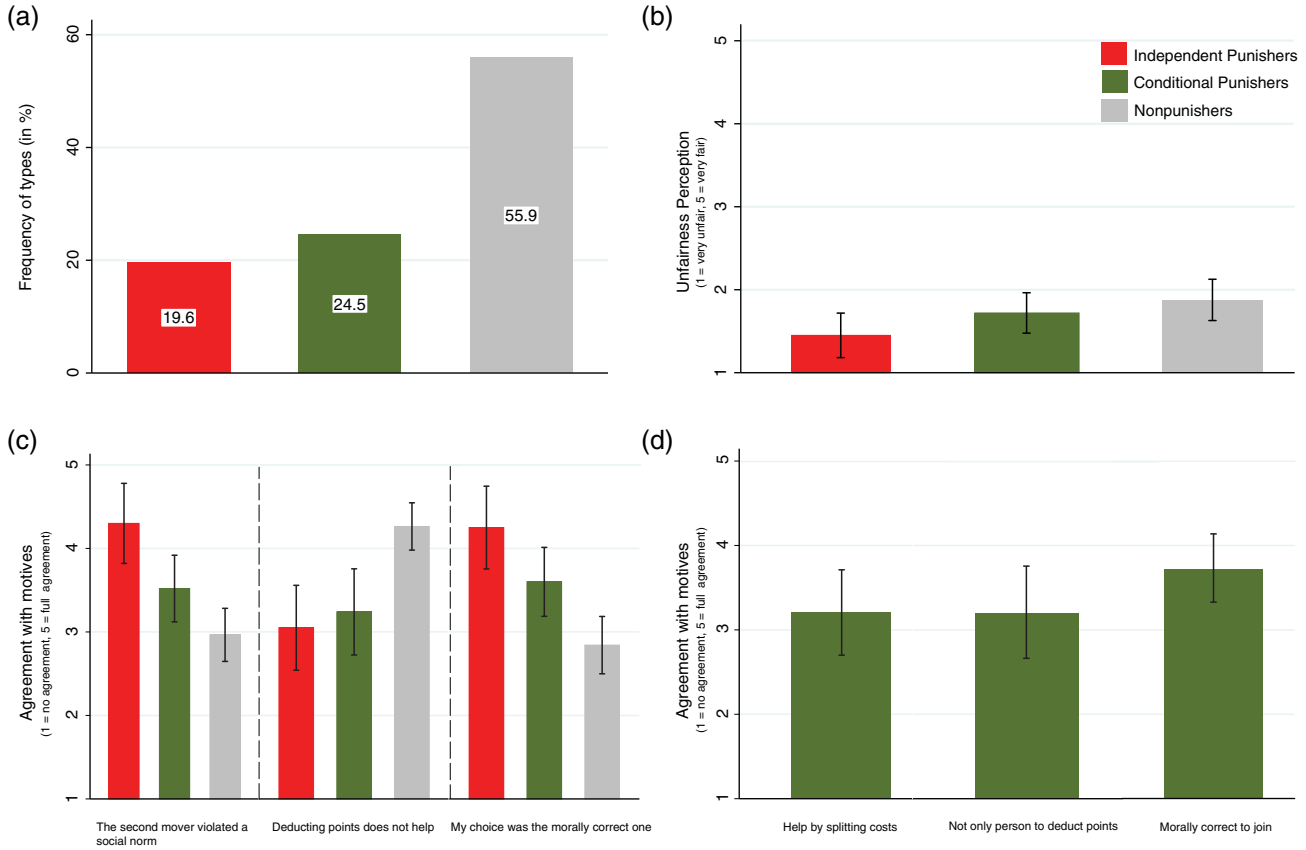

FIGURE 2 Percentage of behavioral types and psychometric ratings of unfairness and motives. The bar graph in (a) illustrates the percentage of the observed behavioral types: independent punishers (19.6%), conditional punishers (24.5%) and nonpunishers (55.9%). The bar graph in (b) illustrates that all three types perceived the defecting behavior of the second mover in the PD as very unfair. The bar graph in (c) illustrates that the three types differed in their agreement with several motive statements, that is, whether the second mover violated a social norm, whether the deduction of points does not help and whether the own choice was the morally correct thing to do (all univariate ANOVAs: p <.001). The bar graph in (d) illustrates that conditional punishers showed a moderate to high agreement on motive statements about collective action and shared responsibility, which are thought to play a role for this particular type. Independent punishers are depicted in red, conditional punishers in green and nonpunishers in gray color (see color legend). Error bars depict 95% confidence intervals

All participants first had to indicate how unfair they perceived the behavior of the defecting second mover in the PD (on a 5-point Likert scale, 1 = very unfair, 5 = very fair). Even though the third-parties behaved differently (and are of different types), the perception of unfairness did not differ between the three types ($F_{(2,99)}$ = 2.05, p = *.134*). In fact, all types considered the behavior of a defecting second mover as very unfair (all means are below 1.9, see Figure 2b). Further, all participants also indicated their agreement on a variety of statements (eight statements, on a 5-point Likert-scale, 1 = no agreement, 5 = full agreement) targeted at specific motives underlying the punishment choice. In order to control for multiple comparisons, we applied a Bonferroni correction (0.05/8 = 0.00625, see Section 2) and report in the manuscript only motives that survived this correction procedure (Table S1 for the statistical analyses to all eight motive statements). We find that the three types differed in their agreement about whether "the second mover in the PD violated a social norm" ($F_{(2,99)}$ = 10.46, p = .0001, eta^2 = 0.17). Independent punishers agreed almost entirely with the motive statement and nonpunishers only moderately, with conditional punishers being in-between (mean$_{Indep}$ = 4.3, mean$_{Cond}$ = 3.52, and mean$_{Non}$ = 2.96). Post-hoc independent t-tests confirmed the difference in pairwise comparisons ($p_{Indep\ vs.\ Cond}$ = .016, $p_{Indep\ vs.\ Non}$ <.0001, $p_{Cond\ vs.\ Non}$ = .047, see Figure 2c). The behavioral types also differed in their opinion on whether "the deduction of points helps no one" ($F_{(2,99)}$ = 11.85, p <.0001, eta^2 = 0.19). Here, nonpunishers agreed to a very large degree and more than both conditional and independent punishers (mean$_{Non}$ = 4.26, mean$_{Cond}$ = 3.24, mean$_{Indep}$ = 3.05; $p_{Cond\ vs.\ Non}$ = .0004, $p_{Indep\ vs.\ Non}$ = .0001), while independent punishers and conditional punishers only agreed moderately with no significant differences between those two types ($p_{Indep\ vs.\ Cond}$ = .610, see Figure 2c). Further, we also asked whether the "own choice was the morally correct thing to do." Independent punishers agreed most (55% of which agreed entirely) followed by conditional punishers and nonpunishers (mean$_{Indep}$ = 4.25, mean$_{Cond}$ = 3.60, mean$_{Non}$ = 2.84; $F_{(2,99)}$ = 10.98, p < 0.0001, eta^2 = 0.18; $p_{Indep\ vs.\ Cond}$ = .050, $p_{Indep\ vs.\ Non}$ = .0001, $p_{Cond\ vs.\ Non}$ = .012, see Figure 2c). Thus, the three behavioral types agreed to different degrees with the motives that we expected to be important based upon the description of the types in the introduction.

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/bdm.2564 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Finally, we explored to what extent the motives for participating in collective actions apply to conditional punishers. Note that, only conditional punishers answered these motive statements because they target specific aspects of conditional behavior. Here, we find that conditional punishers demonstrated a moderate to high agreement on the three motives, providing further evidence for the correct classification of this type. More specifically, conditional punishers agreed moderately with the motives of helping the deducting person ("I wanted to help the people deducting points by lowering the costs for them," mean = 3.2, $SD = 1.22$) and not wanting to punish all by themselves ("I did not want to be the only person to deduct points," mean = 3.2, $SD = 1.32$). Further, conditional punishers agreed most with the motive "it was morally correct to support someone who decided to deduct points" (mean = 3.72, $SD = 0.98$, see Figure 2d).

So far, we characterized the types based on each motive independently. In order to provide a more integrative view, we will now explore whether and to what extent we can classify a participant based upon the motives using a discriminant analysis and leave-one-out predictions (see Figure 4a). Note that only the eight motives all participants answered could be used in this analysis. In a first step, the analysis kept only those motives that explain sufficient variance by minimizing Wilk's Lambda. Three out of 8 motives survive this first step and these motives explain 37.7% of the variance in the three types. The three motives are the same as discussed above (also displayed in Figure 2c) and are used to create two discriminant functions, which separate the types maximally. Figure 4a shows both function scores for each subject and the centroids for each type. The first and second function combined are significant ($\chi^2 = 46.4$, $p < .001$) while the second function itself is not ($\chi^2 = 1.96$, $p = .375$). The first function yields negative (standardized) coefficients for the motive statements concerning the moral correctness of the own choice (−.53) and the perceived norm violation of the second mover (−0.52), and yields a positive (standardized) coefficient for agreeing that deducting points does not help (.58). As can be inferred from Figure 4a, the first function is especially good at discriminating the nonpunishers and independent punishers. We classify each participant using leave-one-out predictions. In total, 60.8% of subjects are classified correctly, which is considerably better than the 33.3% chance level. More specifically, 65% of independent punishers, 28% of conditional punishers and 73.7% of nonpunishers are classified correctly (see Figure 4a). This suggests that the motives are particularly good at classifying nonpunishers and independent punishers.

3.3 | Characterization of the behavioral types with structural brain characteristics

To explore whether the three types can be characterized by distinct neural signatures, we performed quantitative morphometric analyses of T1-weighted anatomical images using VBM implemented in the computational anatomy toolbox (CAT 12). VBM is a whole brain technique capable of discovering subtle, regionally specific changes in gray matter volume (see material and methods for details).

We applied univariate analysis of variance (ANOVA) in SPM 12 with gray matter volume as dependent variable and behavioral types (independent punishers, conditional punishers, and nonpunishers) as between-subject factor. We controlled in the analysis for individual brain size, gender, and age, as is required in volumetric analyses. In order to control for multiple comparisons, we used nonparametric permutation statistics based on threshold-free cluster enhancement (see Section 2 for details). Findings revealed that three brain areas showed volumetric differences (at $p < .05$, SV-FWE corrected) between the three behavioral types (Figure 3a–d), including the right TPJ (TPJ: $x = 47$, $y = -66$, $z = 35$, peak F-value: 17.71, peak TFCE-value: 21099), the right DLPFC (DLPFC: $x = 45$, $y = 11$, $z = 44$, peak F-value = 14.31, peak TFCE-value: 10966), and the bilateral dorsal caudate (left caudate: $x = -14$, $y = -6$, $z = 23$, peak F-value: 19.48, peak TFCE-value: 29133; right caudate: $x = 20$, $y = -3$, $z = 23$, peak F-value: 14.84, peak TFCE-value: 11233). In order to explore the direction and specificity of the observed anatomical differences, we extracted the gray matter volume values from these significant regions (for the extraction thresholded at $p < .001$, uncorrected), regressed out age, gender and individual brain size (the covariates of no interest), and conducted post-hoc pairwise comparison in SPSS. Below we report the findings of parametric statistics. See Table S2 for nonparametric statistics. Notably, all findings hold, irrespective of whether we conduct parametric or nonparametric tests.

Pairwise comparisons between the three behavioral types revealed a distinctive pattern in the right TPJ, right DLPFC and bilateral dorsal caudate. As shown in Figure 3a, independent punishers demonstrated larger gray matter volume in the right TPJ (in the area of the angular gyrus) than both conditional punishers ($F_{(1,43)} = 12.48$, $p = .001$, $\text{eta}^2 = 0.225$) and nonpunishers ($F_{(1,75)} = 11.66$, $p = .001$, $\text{eta}^2 = 0.135$). Conditional punishers and nonpunishers did not differ significantly with respect to gray matter volume in the right TPJ ($F_{(1,80)} = 0.531$, $p = .468$, $\text{eta}^2 = 0.007$).

As shown in Figure 3b, conditional punishers demonstrated larger gray matter volume in the right DLPFC than independent punishers ($F_{(1,43)} = 13.31$, $p = .001$, $\text{eta}^2 = 0.236$) and nonpunishers ($F_{(1,80)} = 7.62$, $p = .007$, $\text{eta}^2 = 0.087$). Independent punishers and nonpunishers did not differ significantly with respect to gray matter volume in the right DLPFC ($F_{(1,75)} = 1.42$, $p = .237$, $\text{eta}^2 = 0.019$).

Finally, as shown in Figure 3c/d, conditional punishers demonstrated lower gray matter volume in the bilateral dorsal caudate than both independent punishers (left caudate: $F_{(1,43)} = 16.57$, $p < .001$, $\text{eta}^2 = 0.278$; right caudate: $F_{(1,43)} = 8.72$, $p = .005$, $\text{eta}^2 = 0.169$) and nonpunishers (left caudate: $F_{(1,80)} = 7.86$, $p = .006$, $\text{eta}^2 = 0.090$; right caudate: $F_{(1,80)} = 6.973$, $p = .010$, $\text{eta}^2 = 0.080$). Independent punishers and nonpunishers did not differ significantly with respect to gray matter volume in the bilateral caudate (left caudate: $F_{(1,75)} = 2.52$, $p = .116$, $\text{eta}^2 = 0.033$; right caudate: $F_{(1,75)} = 1.29$, $p = .259$, $\text{eta}^2 = 0.017$).

As for the motive statements, we conducted a discriminant analysis with the extracted gray matter volume values of the four discovered brain areas (adjusted for brain size, age, and gender) in order to

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

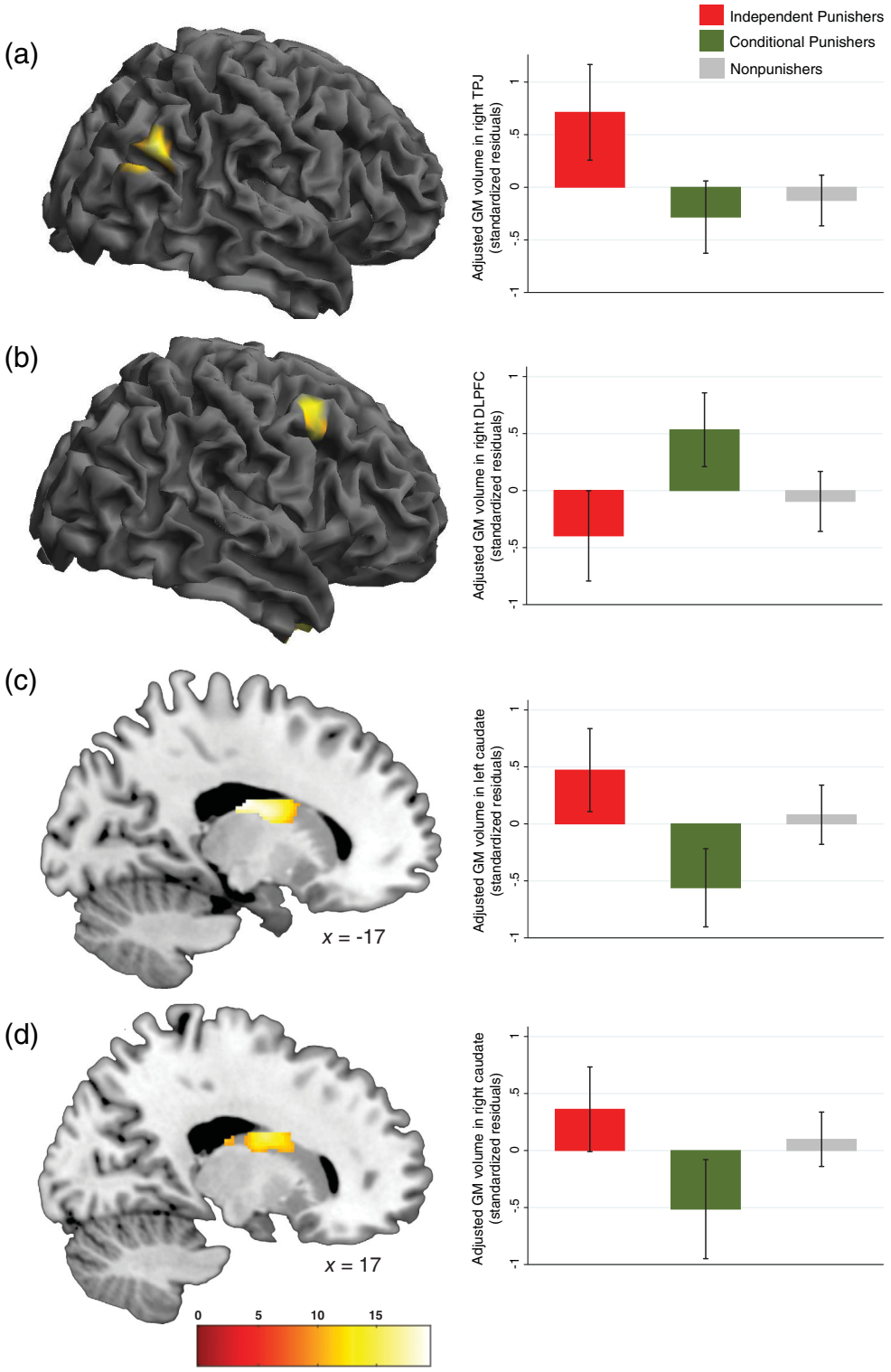

FIGURE 3 Structural brain characteristics in the right TPJ, right DLPFC and bilateral caudate demonstrate significant differences between the three behavioral types. Depicted in (a) are the structural differences in the right TPJ (SV-FWE corrected at p <.05), which were qualified by larger gray matter volume in independent punishers compared to the other two behavioral types. Depicted in (b) are the structural differences in the right DLPFC (SV-FWE corrected at p <.05), which were qualified by larger gray matter volume in conditional punishers compared to the two other behavioral types. Depicted in (c/d) are the structural differences in the left and right dorsal caudate (SV-FWE corrected at p <.05), which are qualified by larger gray matter volume in both independent punishers and nonpunishers compared to conditional punishers. Note that for display purposes, all SV-FWE corrected findings (based on the univariate ANOVA) are depicted at an uncorrected p-value (p <.001) using F-maps. Bar graphs illustrate gray matter volume values based on the depicted regions (encompassing all voxels at a p-value of <.001, as displayed), broken down for the three behavioral types. These values are adjusted for the covariates of no interests (individual brain size, age, and gender) and z-standardized. Independent punishers are depicted in red, conditional punishers in green and nonpunishers in gray color (see color legend). Error bars depict 95% confidence intervals

provide a more integrative view and to examine how good each individual participant can be classified. When we employ the discriminant analysis, the right TPJ, right DLPFC, and the left caudate add sufficient explanatory power (based on Wilks' lambda), resulting in the exclusion of the right caudate. The three included areas explain 30.6% of the variance in the three behavioral types and are used to create two discriminant functions, which separate the types maximally. Figure 4b shows both function scores for each subject and the centroids for each type. The first function and the second function combined are significant ($\chi^2 = 35.73$, p <.001) while the second function itself is not ($\chi^2 = 4.311$, $p = .116$). The first function yields positive (standardized) coefficients for the right TPJ (.58) and the left caudate (.59) and a negative (standardized) coefficient for the right DLPFC (−.56). As can be inferred from Figure 4b, the first discriminant

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

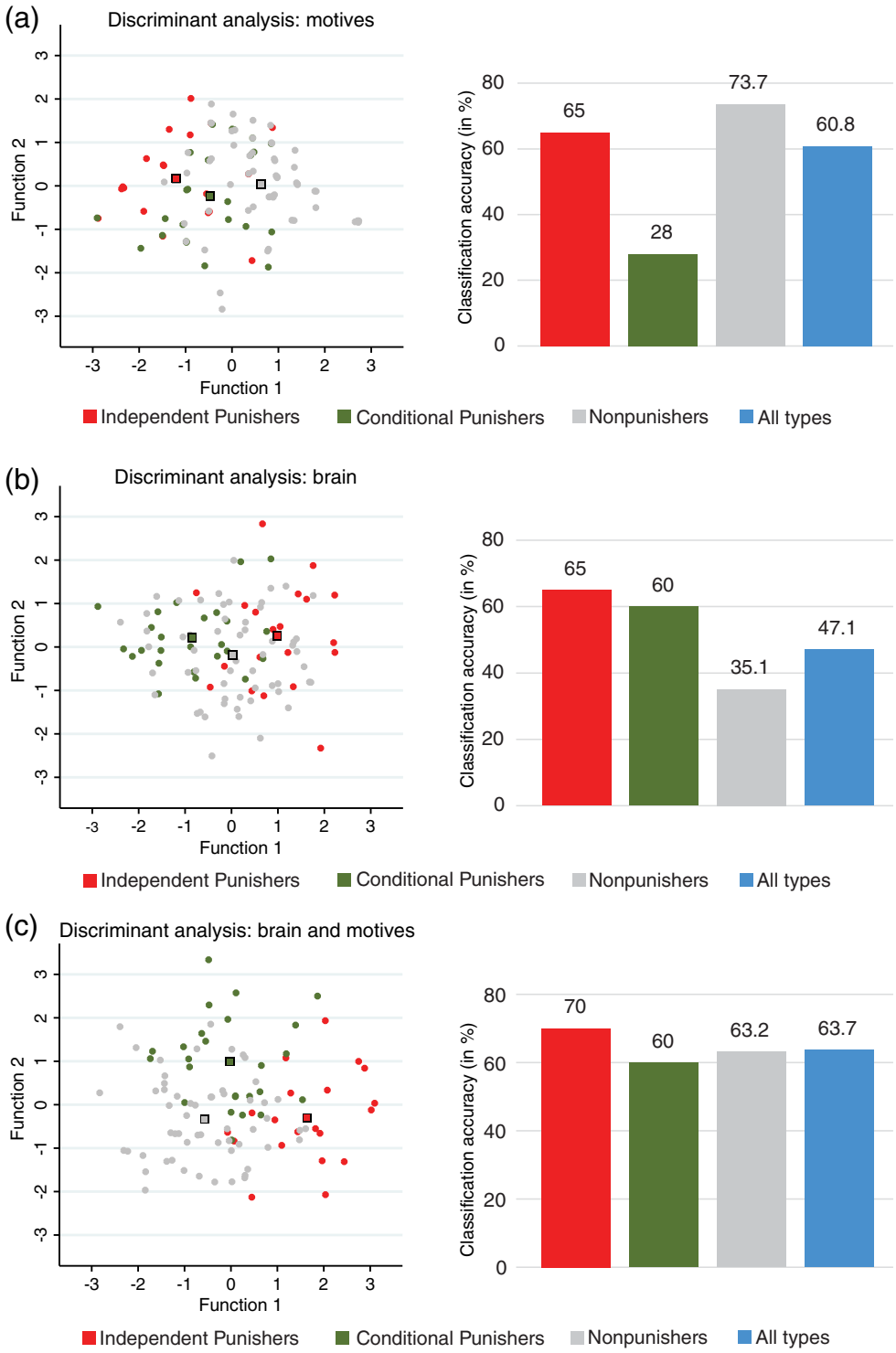

FIGURE 4 Discriminant analyses based on motives and structural brain characteristics. Depicted in (a) are the results of the discriminant analysis calculated with the motive statements. Depicted in (b) are the results of the discriminant analysis calculated with the structural brain characteristics depicted in Figure 3, using the extracted gray matter volume values of the right TPJ, right DLPFC, and left caudate (adjusted for brain size, age, and gender). Depicted in (c) are the results of the discriminant analysis calculated with both the motive statements and structural brain characteristics. The scatterplots illustrate the estimated discriminant functions and the discriminant scores of each individual subject (slightly jittered), together with the centroids of each behavioral type (squares). The bar graphs depict the classification accuracy (in %) for each behavioral type and for all types together. Note that, each participant was classified using leave-one-out predictions. Independent punishers are depicted in red, conditional punishers in green, and nonpunishers in gray, and all types in blue (see color legend)

function is especially good at discriminating conditional punishers and independent punishers. The leave-one-out predictions yield a total of 47.1% of subjects being classified correctly, which is better than the 33.3% chance level. More specifically, 65% of independent punishers, 60% of conditional punishers and 35.1% of nonpunishers are correctly classified (see Figure 4b). In contrast to the motives, structural brain characteristics seem to be better at classifying conditional punishers but less good at classifying nonpunishers.

Finally, to complete the picture, we employed a discriminate analysis with all eight motive statements and four brain areas (using the extracted and adjusted gray matter volume measures). Notably, all the motives and brain areas included in the other two discriminant analyses also yield sufficient discriminating power to be included in the combined motive and brain analysis (using Wilks' lambda for the inclusion of predictors). These six predictors explain 56.5% of the variance and are used to create two discriminant functions. Figure 4c shows

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

both function scores for each subject and the centroids for each type. The first and second function combined are significant ($\chi^2 = 80.438$, p <.001) and the second function itself is also significant ($\chi^2 = 27.629$, p <.001). The first function is mainly affected by the negative (standardized) coefficient for the motive of "deduction does help no one" (−.417) and positive (standardized) coefficients for the right TPJ (.462), the perceived norm violation of the second mover (.444), and the moral correctness of the own choice (.567). As can be inferred from Figure 4c, the first function distinguishes independent punishers and nonpunishers well. The second function is mainly affected by the positive (standardized) coefficient for the right DLPFC (.524) and negative (standardized) coefficient for the left caudate (−.582). As can be inferred from Figure 4c, the second function discriminates between conditional punishers and the other two behavioral types. The leave-one-out predictions show that at least 60% of every type are correctly classified, with a total accuracy of 63.7% (which is better than the 33.3% chance level). Here, 70% of independent punishers, 60% of conditional punishers and 63.2% of nonpunishers are correctly classified (see Figure 4c). Thus, it seems that the combination of both structural brain characteristics and motives is particularly well suited to classify the three behavioral types.

4 | DISCUSSION

Third-party punishment has proven to be a valuable mechanism to enforce norms, increase cooperation rates and to deter free-riders (Fehr & Fischbacher, 2004). The literature on the typology of third-party punishment behavior has so far differentiated only between punishers and nonpunishers and neglected that there are different types of punishers, that is, punishers can be further subdivided into initiators and joiners. To fill this gap, we designed a novel third-party punishment paradigm that allows identifying these distinct punishers. We found that slightly more than half of the subjects (55%) never punished, while the other half of the subjects were divided into independent (20%) and conditional (25%) punishers. By using a neural trait approach, we were able to characterize these three types by their neural signature and shed light on possible underlying psychological mechanisms and motives. Our results showed that independent punishers were characterized by larger gray matter volume in the right TPJ compared to both other types. Further, we found that both independent punishers and nonpunishers were characterized by larger gray matter volume in bilateral caudate than conditional punishers. Finally, only conditional punishers were characterized by larger gray matter volume in the right DLPFC. These structural differences were paralleled by differences in subjective motives underlying the punishment choices. To get an integrative view, we used discriminant analyses to classify the types in leave-one-out predictions. We found that anatomical brain characteristics are able to classify roughly half of the subjects into the correct types, while a combination of both brain characteristics and motives performs best and classifies more than 60% of subjects correctly.

The TPJ, in particular in the area of the right angular gyrus, has been shown to play a crucial role in social cognition, such as mentalizing, perspective-taking, or self-other distinction (Carter & Huettel, 2013; Frith & Frith, 2012; Lamm, Bukowski, & Silani, 2016; Steinbeis, 2016). Further, third-party punishment studies have consistently shown that mentalizing processes during the punishment decision are associated with this part of the TPJ (e.g., Baumgartner, Schiller, Rieskamp, Gianotti, & Knoch, 2013; Buckholtz et al., 2008; Gerfo et al., 2019; Ginther et al., 2016). Our findings showing that independent punishers have a larger TPJ than the other two types complements previous fMRI research and provides first evidence that larger gray matter volume in the right TPJ increases third-parties' propensity to punish norm transgressors, possibly due to an increased capacity for social cognition that helps independent punishers to mentalize with the victim and to represent the victim's needs and appreciation. On a broader perspective, this structural finding in the right TPJ also complements studies in related fields on altruistic and cooperative choices (Baumgartner et al., 2019; Gianotti, Dahinden, Baumgartner, & Knoch, 2019; Morishima et al., 2012). These studies provide evidence that task-independent brain characteristics (structure and baseline activation) in the right TPJ are associated with altruistic choices in the dictator game and cooperative choices in the public goods game. It seems that similar neural traits in the right TPJ are associated with the inclination of third-parties to behave altruistically and to initiate punishment of wrongdoers even if no one else punishes.

In addition to larger gray matter volume in the right TPJ, independent punishers were further characterized by larger gray matter volume in the bilateral caudate compared to conditional punishers. Interestingly, nonpunishers had similar structural brain characteristics in the bilateral caudate as independent punishers and also showed significant differences compared to conditional punishers. This might seem puzzling at first sight: Why do these two types, who demonstrate diametrically opposed punishment choices, have similar structural characteristics in the bilateral caudate? The caudate is an important part of the reward system (Rademacher, Schulte-Rüther, Hanewald, & Lammertz, 2015). Processing rewards plays a major role in goal-directed behavior and motivation. The caudate is implicated in the processing of rewards that accrue as a result of goal-directed behavior or decisions (e.g., Brassen, Gamer, Peters, Gluth, & Buchel, 2012; Delgado, 2007; O'Doherty et al., 2004). Importantly, the caudate encodes both nonsocial (monetary) and social rewards in the same area (Gu et al., 2019; Izuma, Saito, & Sadato, 2008; Wake & Izuma, 2017). For example, the caudate has been shown to encode social rewards in diverse situations, such as mutually cooperating with other individuals (Park et al., 2017; Rilling et al., 2002), getting a fair offer (Tabibnia, Satpute, & Lieberman, 2008), punishing unfairness in the role of a second-party (De Quervain, Fischbacher, Treyer, & Schellhammer, 2004) and third-party (Baumgartner et al., 2012; Hu et al., 2015; Strobel et al., 2011), and giving charitable donations (Harbaugh, Mayr, & Burghart, 2007; Moll et al., 2006). Similarly, numerous studies using diverse nonsocial paradigms (e.g., lotteries, gambling tasks, slot machine tasks, etc.) have shown a role of the

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

caudate in encoding nonsocial monetary rewards (e.g., Arsalidou, Vijayarajah, & Sharaev, 2020; Bjork, Smith, Chen, & Hommer, 2010; Brassen et al., 2012; Hardin, Pine, & Ernst, 2009; Hosking et al., 2017). Collectively, the reviewed literature lead us to speculate that the increased volume in the bilateral caudate in independent punishers as well as nonpunishers (compared to conditional punishers) indicate an increased inclination to seek rewards (social or nonsocial) as a driving force in decision-making, possibly due to an increased capacity of the reward system to influence goal-directed behavior. But, why are independent punishers motivated by social rewards, whereas nonpunishers are more motivated by nonsocial (monetary rewards)? Here a recent neuroimaging study by Park and colleagues is enlightening (Park et al., 2017). It showed that generous choices are associated with enhanced functional connectivity between the TPJ and reward-related areas in the caudate. Thus, it is conceivable that the observed volumetric differences in the TPJ between independent punishers and nonpunishers (as discussed above) might explain why independent punishers are motivated by social rewards, whereas nonpunishers are motivated by nonsocial (monetary) rewards. Summing up, decision-making in these two behavioral types seems to be driven by strong preferences for rewards (social or nonsocial)—an interpretation that is further strengthened by the self-reported punishment motives: Whereas nonpunishers agreed highly with the notion that punishment does not help and only reduces monetary payment, independent punishers vastly agreed that the second mover has violated a social norm and that punishing him/her is the morally correct thing to do.

So far, we were able to characterize independent punishers and nonpunishers with structural brain characteristics of the TPJ and bilateral caudate and argued that both types might be inclined to seek rewards as the driving force in decision-making. In contrast, these structural brain patterns were not characteristic of conditional punishers, and therefore decision-making in this type might not (or to a lesser extent) be driven by social or monetary rewards. Thus, the question arises: What is the driver of those who prefer to condition their own punishment choice on whether another person punishes? Why do they prefer to punish collectively and thereby risk that punishment might not occur at all? In their conceptual framework El Zein et al. (2019) propose that collective decisions have several advantages over sole decisions. They allow minimizing the material and psychological burden of an individual's responsibility and shield collective decision-makers from the consequences of negative outcomes. This conceptual framework fits our case. Conditional punishers are part of the norm-enforcement and they help the other punisher(s), while sharing the blame and costs for punishing if punishment is implemented. If no punishment is implemented, conditional punishers can keep up a self-image of behaving altruistically by showing a general willingness to participate in the punishment of unfair behavior, but also do not spend points and avoid being the sole punisher. Conditional punishers' answers to the motives for collective actions and shared responsibility are in line with these assumptions, that is, conditional punishers indicated that they wanted to help the independent punishers, but avoided being the sole punisher. In conclusion, it seems that conditional punishment has a strong strategic component, which optimizes the punishment choice for all potential situations. Interestingly, there is strong evidence from task-dependent and task-independent studies (structure and baseline activation) that the DLPFC is involved in strategic decision-making (e.g., Crone & Steinbeis, 2017; Gianotti et al., 2018; Ruff, Ugazio, & Fehr, 2013; Soutschek et al., 2015; Spitzer et al., 2007; Steinbeis et al., 2012; Strang et al., 2014), and that the functional role of the DLPFC in strategic decision-making involves aspects of self-control, that is, a strategic decision often involves some kind of sacrifice, for example, money or time. Thus, we speculate that the larger volume in the right DLPFC in conditional punishers (compared to the other two types) might indicate that their decision is more influenced by strategic considerations, possibly due to an increased capacity for self-control processes and strategic reasoning. This increased capacity might allow them to implement strategies of behavior that minimize the material burden (costs) as well as the psychological burden (feeling bad if not participating in punishment).

Since we explored a novel punishment typology, we also investigated the out-of-sample classification accuracy of the structural brain characteristics and motives. To this end, we used discriminant analyses, first by classifying based upon either brain characteristics or motives, and then combining both data. Each dataset, on its own, is able to capture two of the three punishment types with an accuracy of at least 60%, but neither is better than chance at discriminating the remaining type. Combining the two data sets leads to correct classification of 63.7% of all subjects, and each behavioral type is classified with an accuracy of at least 60%, which is nearly double as high as a naïve baseline (33%). Thus, with the help of structural brain characteristics and psychometric motives, we were able to correctly classify individual subjects into distinct behavioral types, going beyond the exploration of average differences between behavioral types.

To corroborate the interpretation of the anatomical brain findings and to allow more specific and mechanistic conclusions, future research could additionally acquire neuropsychological test batteries (e.g., cognitive abilities and strategic reasoning) or social and affective decision-making measures (e.g., mentalizing, self-other distinction, and reaction to rewards).

Summing up, our findings reveal a typology of punishers that can be characterized by structural brain characteristics associated with social cognition, reward processing, behavioral control, and strategic reasoning. In order to increase participation in third-party punishment (or to decrease the second-order free-rider problem) it would be necessary to shift some of the considerably numerous nonpunishers to be conditional or independent punishers. Obvious ways for shifting nonpunishers would include making punishment less harmful or costly. Other ways could be to use specific behavioral trainings (e.g., mentalizing training, meditation, and working memory training) or neuro-modulation techniques (e.g., brain stimulation and neurofeedback). These trainings and techniques have been demonstrated to improve social cognition (e.g., Santiesteban et al., 2012; Santiesteban, Banissy, Catmur, & Bird, 2012) and strategic reasoning/behavioral control capacities (e.g., Anguera et al., 2013; Houben,

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Dassen, & Jansen, 2016; Kouijzer, de Moor, Gerrits, Congedo, & van Schie, 2009) and to change underlying structural brain characteristics in the TPJ and DLPFC (e.g., Jausovec & Jausovec, 2012; Klimecki et al., 2019; Valk et al., 2017). Thus, it is conceivable that behavioral trainings and neuro-modulation techniques that target the brain areas involved in social cognition and strategic reasoning could help to increase the number of punishers, be it conditional or independent, thereby promoting the enforcements of social norms via third-party punishment.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

Jan Hausfeld, Miguel dos Santos, and Daria Knoch conceived and designed the study. Thomas Baumgartner, Jan Hausfeld, and Miguel dos Santos performed research. Thomas Baumgartner and Jan Hausfeld analyzed the behavioral/psychometric data, and Thomas Baumgartner analyzed the brain data. Thomas Baumgartner, Jan Hausfeld, and Daria Knoch wrote and Miguel dos Santos commented the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Thomas Baumgartner https://orcid.org/0000-0001-5966-7377
Jan Hausfeld https://orcid.org/0000-0001-5171-4829
Miguel dos Santos https://orcid.org/0000-0002-2198-1560
Daria Knoch https://orcid.org/0000-0003-1935-053X

REFERENCES

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., … Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*, 97–101.

Arsalidou, M., Vijayarajah, S., & Sharaev, M. (2020). Basal ganglia lateralization in different types of reward. *Brain Imaging and Behavior*, *14*, 2618–2646.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage*, *38*, 95–113.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. *NeuroImage*, *11*, 805–821.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*, 839–851.

Balafoutas, L., Grechenig, K., & Nikiforakis, N. (2014). Third-party punishment and counter-punishment in one-shot interactions. *Economics Letters*, *122*, 308–310.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*, 594–615.

Baumgartner, T., Dahinden, F. M., Gianotti, L. R., & Knoch, D. (2019). Neural traits characterize unconditional cooperators, conditional cooperators, and noncooperators in group-based cooperation. *Human Brain Mapping*, *40*, 4508–4517.

Baumgartner, T., Gotte, L., Gugler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping*, *33*, 1452–1469.

Baumgartner, T., Schiller, B., Rieskamp, J., Gianotti, L. R., & Knoch, D. (2013). Diminishing parochialism in intergroup conflict by disrupting the right temporo-parietal junction. *Social Cognitive and Affective Neuroscience*, *9*, 653–660.

Bellucci, G., Camilleri, J. A., Iyengar, V., Eickhoff, S. B., & Krueger, F. (2020). The emerging neuroscience of social punishment: Meta-analytic evidence. *Neuroscience and Biobehavioral Reviews*, *113*, 426–439.

Bjork, J. M., Smith, A. R., Chen, G., & Hommer, D. W. (2010). Adolescents, adults and rewards: Comparing motivational neurocircuitry recruitment using fMRI. *PLoS One*, *5*, e11440.

Brañas-Garza, P., Espín, A. M., Exadaktylos, F., & Herrmann, B. (2014). Fair and unfair punishers coexist in the ultimatum game. *Scientific Reports*, *4*, 6025.

Brassen, S., Gamer, M., Peters, J., Gluth, S., & Buchel, C. (2012). Don't look back in anger! Responsiveness to missed chances in successful and nonsuccessful aging. *Science*, *336*, 612–614.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*, 930–940.

Buckholtz, J. W., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*, 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*, 1369–1380.

Carpenter, J. P., & Matthews, P. H. (2012). Norm enforcement: Anger, indignation, or reciprocity? *Journal of the European Economic Association*, *10*, 555–572.

Carter, R. M., & Huettel, S. A. (2013). A nexus model of the temporal–parietal junction. *Trends in Cognitive Sciences*, *17*, 328–336.

Civai, C., Crescentini, C., Rustichini, A., & Rumiati, R. I. (2012). Equality versus self-interest in the brain: Differential roles of anterior insula and medial prefrontal cortex. *NeuroImage*, *62*, 102–112.

Crone, E. A., & Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends in Cognitive Sciences*, *21*, 205–215.

Dahnke, R., Yotter, R. A., & Gaser, C. (2013). Cortical thickness and central surface estimation. *NeuroImage*, *65*, 336–348.

De Quervain, D. J., Fischbacher, U., Treyer, V., & Schellhammer, M. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258.

Delgado, M. R. (2007). Reward-related responses in the human striatum. *Annals of the New York Academy of Sciences*, *1104*, 70–88.

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. *Nature Human Behaviour*, *3*, 554–559.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*, 2017–2030.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*, 171–178.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*, 397–404.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313.

Gerfo, E. L., Gallucci, A., Morese, R., Vergallito, A., Ottone, S., Ponzano, F., … Lauro, L. J. R. (2019). The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. *NeuroImage*, *200*, 501–510.

Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. (2018). Effect size and power in fMRI group analysis. *BioRxiv*, pp. 1–23.

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gianotti, L. R., Dahinden, F. M., Baumgartner, T., & Knoch, D. (2019). Understanding individual differences in domain-general prosociality: A resting EEG study. *Brain Topography*, *32*, 118–126.

Gianotti, L. R., Nash, K., Baumgartner, T., Dahinden, F. M., & Knoch, D. (2018). Neural signatures of different behavioral types in fairness norm compliance. *Scientific Reports*, *8*, 10513.

Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones, O. D., & Marois, R. (2016). Parsing the behavioral and brain mechanisms of third-party punishment. *Journal of Neuroscience*, *36*, 9420–9434.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*, 114–125.

Gu, R., Huang, W., Camilleri, J., Xu, P., Wei, P., Eickhoff, S. B., & Feng, C. (2019). Love is analogous to money in human brain: Coordinate-based and functional connectivity meta-analyses of social and monetary reward anticipation. *Neuroscience and Biobehavioral Reviews*, *100*, 108–128.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, *316*, 1622–1625.

Hardin, M. G., Pine, D. S., & Ernst, M. (2009). The influence of context valence in the neural coding of monetary outcomes. *NeuroImage*, *48*, 249–257.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience*, *30*, 583–590.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., … Henrich, N. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*, 1480–1484.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Henrich, N. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367.

Hosking, J. G., Kastman, E. K., Dorfman, H. M., Samanez-Larkin, G. R., Baskin-Sommers, A., Kiehl, K. A., … Buckholtz, J. W. (2017). Disrupted prefrontal regulation of striatal subjective value signals in psychopathy. *Neuron*, *95*, 221–231 e4.

Houben, K., Dassen, F. C., & Jansen, A. (2016). Taking control: Working memory training in overweight individuals increases self-regulation of food intake. *Appetite*, *105*, 567–574.

Hu, Y., Strang, S., & Weber, B. (2015). Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Frontiers in Behavioral Neuroscience*, *9*, 24.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*, 451–462.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*, 284–294.

Jausovec, N., & Jausovec, K. (2012). Working memory training: Improving intelligence—Changing brain activity. *Brain and Cognition*, *79*, 96–106.

Kamei, K. (2014). Conditional punishment. *Economics Letters*, *124*, 199–202.

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*, 231–242.

Klimecki, O., Marchant, N. L., Lutz, A., Poisnel, G., Chetelat, G., & Collette, F. (2019). The impact of meditation on healthy ageing - the current state of knowledge and a roadmap to future directions. *Current Opinion in Psychology*, *28*, 223–228.

Kouijzer, M. E. J., de Moor, J. M. H., Gerrits, B. J. L., Congedo, M., & van Schie, H. T. (2009). Neurofeedback improves executive functioning in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *3*, 145–162.

Krueger, F., & Hoffman, M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*, 499–501.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*, 75–84.

Lamm, C., Bukowski, H., & Silani, G. (2016). From shared to distinct self–other representations in empathy: Evidence from neurotypical function and socio-cognitive disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150083.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, *19*, 1233–1239.

Mathew, S., & Boyd, R. (2011). Punishment sustains large-scale cooperation in prestate warfare. *Proceedings of the National Academy of Sciences*, *108*, 11375–11380.

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., & Grafman, J. (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences*, *103*, 15623–15628.

Molleman, L., Kölle, F., Starmer, C., & Gächter, S. (2019). People prefer coordinated punishment in cooperative interactions. *Nature Human Behaviour*, *3*, 1145–1153.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, *75*, 73–79.

Nash, K., Gianotti, L. R., & Knoch, D. (2015). A neural trait approach to exploring individual differences in social preferences. *Frontiers in Behavioral Neuroscience*, *8*, 458.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

O'Gorman, R., Henrich, J., & Vugt, M. V. (2009). Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *276*, 323–329.

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*, 499–502.

Park, S. Q., Kahnt, T., Dogan, A., Strang, S., Fehr, E., & Tobler, P. N. (2017). A neural link between generosity and happiness. *Nature Communications*, *8*, 15964.

Rademacher, L., Schulte-Rüther, M., Hanewald, B., & Lammertz, S. (2015). Reward: From basic reinforcers to anticipation of social cues. In *Social behavior from rodents to humans* (pp. 207–221). Cham, Switzerland: Springer.

Rajapakse, J. C., Giedd, J. N., & Rapoport, J. L. (1997). Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Transactions on Medical Imaging*, *16*, 176–186.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*, 395–405.

Ruff, C. C., Ugazio, G., & Fehr, E. (2013). Changing social norm compliance with noninvasive brain stimulation. *Science*, *342*, 482–484.

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology*, *22*, 2274–2277.

Santiesteban, I., White, S., Cook, J., Gilbert, S. J., Heyes, C., & Bird, G. (2012). Training social cognition: From imitation to theory of mind. *Cognition*, *122*, 228–235.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*, 83–98.

Soutschek, A., Sauter, M., & Schubert, T. (2015). The importance of the lateral prefrontal cortex for strategic decision making in the prisoner's dilemma. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 854–860.

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Spitzer, M., Fischbacher, U., Herrnberger, B., Gron, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, *56*, 185–196.

Steinbeis, N. (2016). The role of self–other distinction in understanding others' mental and emotional states: Neurocognitive mechanisms in children and adults. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150074.

Steinbeis, N., Bernhardt, B. C., & Singer, T. (2012). Impulse control and underlying functions of the left DLPFC mediate age-related and age-independent individual differences in strategic social behavior. *Neuron*, *73*, 1040–1051.

Strang, S., Gross, J., Schuhmann, T., Riedl, A., Weber, B., & Sack, A. T. (2014). Be nice if you have to—The neurobiological roots of strategic fairness. *Social Cognitive and Affective Neuroscience*, *10*, 790–796.

Strobel, A., Zimmermann, J., Schmitz, A., Reuter, M., Lis, S., Windmann, S., & Kirsch, P. (2011). Beyond revenge: Neural and genetic bases of altruistic punishment. *NeuroImage*, *54*, 671–680.

Tabibnia, G., Satpute, A. B., & Lieberman, M. D. (2008). The sunny side of fairness: Preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). *Psychological Science*, *19*, 339–347.

Tohka, J., Zijdenbos, A., & Evans, A. (2004). Fast and robust parameter estimation for statistical partial volume models in brain MRI. *NeuroImage*, *23*, 84–97.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*(62), 1–10.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*, 273–289.

Valizadeh, S. A., Liem, F., Merillat, S., Hanggi, J., & Jancke, L. (2018). Identification of individual subjects on the basis of their brain anatomical features. *Scientific Reports*, *8*, 5611.

Valk, S. L., Bernhardt, B. C., Trautwein, F. M., Bockler, A., Kanske, P., Guizard, N., … Singer, T. (2017). Structural plasticity of the social brain: Differential change after socio-affective and cognitive mental training. *Science Advances*, *3*, e1700489.

van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, *30*, 829–858.

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*, 1558–1564.

Yamagishi, T., Takagishi, H., Fermin Ade, S., Kanai, R., Li, Y., & Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 5582–5587.

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage*, *129*, 95–104.

Zhou, Y., Jiao, P., & Zhang, Q. (2017). Second-party and third-party punishment in a public goods experiment. *Applied Economics Letters*, *24*, 54–57.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Baumgartner, T., Hausfeld, J., dos Santos, M., & Knoch, D. (2021). Who initiates punishment, who joins punishment? Disentangling types of third-party punishers by neural traits. *Human Brain Mapping*, *42*(17), 5703–5717. https://doi.org/10.1002/hbm.25648

10970193, 2021, 17, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/hbm.25648 by University of Amsterdam, Wiley Online Library on [27/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License