PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
NEUROSCIENCE

Neurocomputational evidence that conflicting prosocial motives guide distributive justice

Yue Li[a,b,1], Jie Hu[a,c,1,2], Christian C. Ruff[c], and Xiaolin Zhou[a,b,d,e,2]

Edited by René Marois, Vanderbilt University; received June 13, 2022; accepted October 18, 2022 by Editorial Board Member Michael S. Gazzaniga

In the history of humanity, most conflicts within and between societies have originated from perceived inequality in resource distribution. How humans achieve and maintain distributive justice has therefore been an intensely studied issue. However, most research on the corresponding psychological processes has focused on inequality aversion and has been largely agnostic of other motives that may either align or oppose this behavioral tendency. Here we provide behavioral, computational, and neuroimaging evidence that distribution decisions are guided by three distinct motives—inequality aversion, harm aversion, and rank reversal aversion—that interact with each other and can also deter individuals from pursuing equality. At the neural level, we show that these three motives are encoded by separate neural systems, compete for representation in various brain areas processing equality and harm signals, and are integrated in the striatum, which functions as a crucial hub for translating the motives to behavior. Our findings provide a comprehensive framework for understanding the cognitive and biological processes by which multiple prosocial motives are coordinated in the brain to guide redistribution behaviors. This framework enhances our

from the advantaged to the disadvantaged party (19, 20). Supporting this tendency, people are averse to overturn stable hierarchies in a society even though such preexisting hierarchies may

reduced the inequality level (effect of Δ Inequality with ORE = 1.58, 95% CI [1.37–1.83], $P < 0.001$, Fig. 1*C*, *Left* and *SI Appendix*, Table S2) and when the initial inequality was greater (effect of Δ Initial endowment with ORE = 1.12, 95% CI [1.01–1.24], $P = 0.04$, *SI Appendix*, Fig. S2A and Table S3). However, individuals' probability to choose the more equal offer was lower in the Rank-reversal condition than in the No Rank-reversal condition (ORE = 0.37, 95% CI [0.33 – 0.42], $P_{\text{No Rank-reversal}}$ (Equal) = 0.78 ± 0.03 (MEAN ± SE), $P_{\text{Rank-reversal}}$ (Equal) = 0.38 ± 0.04, $t(56) = 8.88$, $P < 0.001$, Fig. 1*C*, *Right*). Altogether, these results indicate that rank reversal aversion influences choices independently fl30 Tw (t5 Tw 6ality wasec6>6>de tha6>6 (qwhich r in thv)6.1)]TJ-0.012 Tc 04-1.1 Td[(conditmatch inac (t5 Tssank-rt

focused these analyses on the Rank-reversal condition, which in contrast to the No Rank-reversal condition allowed us to differentiate inequality aversion from harm aversion and rank reversal aversion. We describe the principles and rationales of the four model families (M1–M4) in the following section and then report the results of the corresponding analyses. For detailed expositions of all the models and technical details of model selection and estimation, please see *SI Appendix, SI Materials and Methods* and Table 1.

Model Construction. The control model M1 only considered inequality aversion, whereas M2–M4 considered combinations of inequality aversion and the other motives.

The simplest model M1 followed the classical inequality aver-inequa2lo(corr)10 (espond29)30.3 (o RanM3c)Tc T

there was no reliable difference in the accuracy with which choice data generated by M3a were recovered by M4a (0.84 ± 0.02) and M3a (0.82 ± 0.02, $t(53) = 1.66$, $P = 0.104$). Thus, the winning model M4a was indeed able to predict and capture unique aspects of the data compared to the closest alternative model.

Model Parameters. In line with the model-free analyses, model-based analyses confirmed that participants' redistribution behaviors in the Rank-reversal condition were driven by inequality aversion, harm aversion, and rank reversal aversion: Participants weighed the inequality difference between the two alternative offers ($\alpha = 0.51 \pm 0.06$, $t(56) = 8.90$, $P < 0.001$, Cohen's d = 1.18), devalued the more equal offer by the extra harm for the initially advantaged party ($\beta = 0.45 \pm 0.06$, $t(56) = 7.83$, $P < 0.001$, Cohen's d = 1.04), and valued rank reversal negatively ($\delta = 0.96 \pm 0.07$, $t(56) = 13.23$, $P < 0.001$, Cohen's d = 1.75, Fig. 2*B*). In line with expectations, greater inequality aversion (α) was associated with higher probability of more equal choice ($tau = 0.74$, $P < 0.001$, *SI Appendix*, Fig. S4, *Left*). By contrast, greater harm aversion (β, $tau = -0.27$, P.s(tau)Tj/T1 stm(ff41)( )]TJ-0.02 Tc 0.015 Tw 90 0 0 10 36 1802817 Tm[(cank r)10 (ev)6 (ersal av)6 (e

coordinates: [15, 20, −5], voxel-wise p(FWE) = 0.064, t-value = 3.55, k = 76) varied parametrically with equality ($-\Delta F$) in the No Rank-reversal condition (Fig. 3*A*), but not in the Rank-reversal condition. A comparison between conditions confirmed a more positive striatal parametric effect of equality in the No Rank-reversal than Rank-reversal condition (peak MNI coordinates: [6, 14, −5], voxel-wise p(FWE) = 0.032, t-value = 4.01, k = 45, Fig. 3*B* and *SI Appendix*, Fig. S7 for a visualization of this effect). Note that this effect was also confirmed in the subsequent whole-brain analysis (*SI Appendix*, Table S7). The absence of striatum responses to equality in the Rank-reversal condition may be due to interactions between inequality aversion and the other motives that are stronger in this condition, a possibility that we tested explicitly in analyses described later.

Our second ROI analysis showed that VMPFC was not involved in equality processing. However, consistent with prior studies (35, 36), this area (MNI peak coordinates: [3, 56, −14], t-value = 2.76, voxel-wise p (FWE-SVC) = 0.049, k = 30, within VMPFC ROI with 8 mm radius centered on the peak MNI coordinates [0, 52, −8] involved in monetary incentive processing in ref. 35) was involved in representing the model-predicted value of the chosen option. This finding provides neural validation of our computational behavioral model.

Given that striatum was involved in signaling equality in the No Rank-reversal condition, we examined whether activity in this area can bias behavior in line with inequality aversion. A post-hoc correlation analysis showed that greater sensitivity to equality signals (i.e., more positive parametric estimates of $-\Delta F$) in putamen (MNI peak coordinates: [−18, 11, −2], max t-value =2.65, voxel-wise p (FWE-SVC) = 0.043, k = 6, ROI center MNI coordinates [−12, 10, −6]) was indeed associated with a significantly higher probability of more equal choice in the No Rank-reversal condition (Kendall's *tau* = 0.27, P = 0.003, robust regression: b = 7.66, P = 0.002, Fig. 3*C*) but not in the Rank-reversal condition (*SI Appendix*, Fig. S8). Whole-brain analyses revealed that no other region correlated with individuals' choices in either condition.

Taken together, these findings show that, in situations where inequality aversion is the main motive guiding behavior, the striatum plays a critical role in processing equality and biasing redistribution behaviors in line with these concerns.

Cortical Regions Involved in Signaling Harm. In the Rank-reversal condition, whole-brain analyses showed that activity in several brain areas correlated with the harm signals related to the more equal offer. These areas comprised dorsomedial prefrontal cortex/anterior cingulate cortex (DMPFC/ACC), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), TPJ, and inferior temporal gyrus (ITG) (Fig. 3*D* and *SI Appendix*, Table S7). Thus, these areas could either represent the strength of the harm aversion motive, or they could be involved in processing/resolving the conflict between concerns about inequality and harm. The latter interpretation may be in line with previous findings that DMPFC/ACC, IFG, and MFG are often activated during cognitive control, conflict resolution, or behavioral adaptation (37, 38); and that TPJ is involved in mentalizing and perspective taking (39, 40). However, none of the neural effects in these areas were associated with the strength of behavioral harm aversion or inequality aversion, or the probability of more equal choice in the Rank-reversal condition. This motivated us to further examine whether and how the strength of the different motives was represented by interactions between the different neural systems representing harm and equality.

DMPFC, as a Region Signaling Harm, Dampens Neural Sensitivity to Equality in Striatum. We had observed weaker inequality aversion and dampened striatal sensitivity to equality in the Rank-reversal condition. These findings suggest that behaviorally relevant neural equality signals may not be represented invariably across different contexts, but may be modulated in situations where they conflict with harm signals. If this "conflict modulation" scenario held true, we should be able to observe that the reduction in striatal equality in the Rank-reversal condition relates to the strength of neural representations in harm-processing regions.

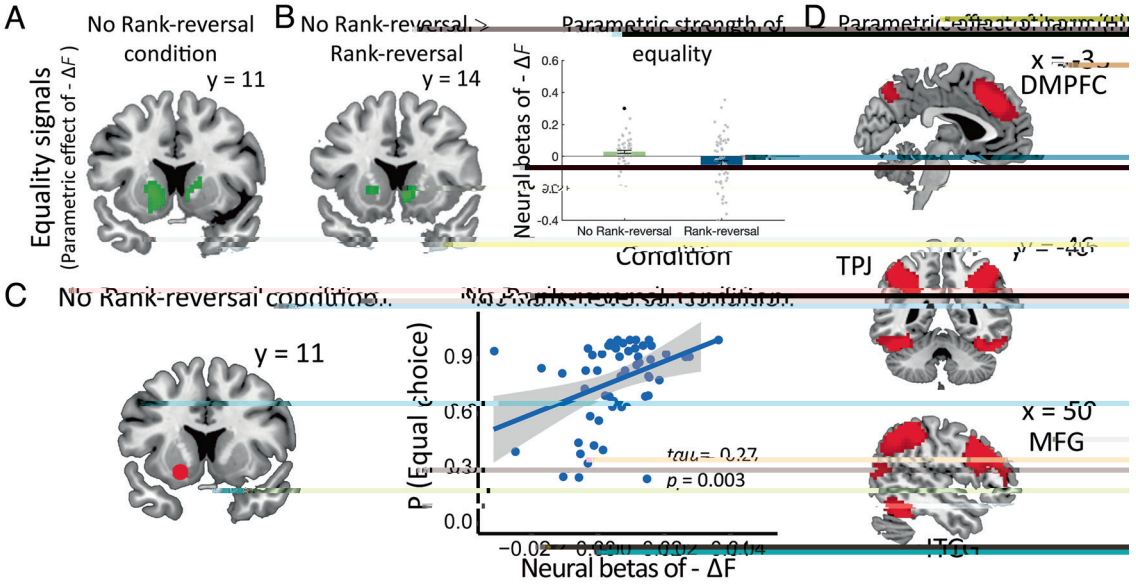

Fig. 3. Neural representations of equality and harm. (*A*) Activity in the striatum was associated with equality signals (−Δ*F*) in the No Rank-reversal condition. (*B*) More positive parametric strength of equality signal in the striatum in the No Rank-reversal than Rank-reversal condition (*Left* panel). For visualization, neural estimates of the significant cluster were extracted from both conditions (*Right* panel). Each dot represents one participant, and error bars indicate the SEMs. •P < 0.05. (*C*) Scatter plot shows a correlation between the parametric strength of equality signal in the striatum (peak MNI coordinates [−18, 11, −2]) and individuals' probability of more equal choice in No Rank-reversal condition, suggesting that people whose striatum is more sensitive to equality have stronger preferences for more equal distribution. (*D*) Parametric effects of harm to the advantaged party in the Rank-reversal condition. Activity in DMPFC, TPJ, MFG, and ITG increased with the extent of harm to the advantaged party, suggesting processing of harm signals in these brain regions. Significant clusters are thresholded at voxel-wise P < 0.001 uncorrected and cluster-wise FWE corrected P < 0.05. Correlation result in (*C*) is thresholded at voxel-wise P < 0.05 FWE, small volume correction.

Downloaded from https://www.pnas.org by Universitaet Zuerich on December 8, 2022 from IP address 89.206.81.235.

To test this hypothesis, we performed PPI analyses examining how interregional functional connectivity varies with inequality levels (GLMs 3 and 4; for ease of visualization −ΔF was split into two bins (high −ΔF vs. low −ΔF), but note that all effects are also present for a parametric regressor of −Δ*F*; for details, see *SI Appendix*, *SI Materials and Methods*). As the seed region for these analyses, we used an unbiased striatum region that was fully independent of the equality results described above (i.e., based on the peak coordinates in the Neurosynth "Striatum" activation map, Fig. 4*A* and *SI Appendix*, *SI Materials and Methods*). The PPI analyses were set up to identify brain regions that change their functional coupling with the striatum in line with how strongly equality concerns are relevant for the current choice. Evidence for this was assessed via the interaction term in the model, which quantifies for each voxel how much the correlation of the BOLD signal with that in the striatum changes as a function of the equality context (i.e., the equality concern triggered by the payoffs on the present trial), while simultaneously controlling for any main effects of (i.e., simple correlations with) the striatum time course and the equality context (41). These analyses revealed that dorsomedial prefrontal cortex (DMPFC, MNI peak coordinates: [0, 47, 40], k = 634, t-value = 4.89, cluster-wise *p* (FWE) = 0.002) was functionally connected with striatum more strongly for high equality contexts (high −ΔF) in the Rank-reversal condition (Fig. 4*C*, *Left*; note that this effect was also present in control PPI analysis containing parametric inequality regressors; see SI Results). Importantly, the DMPFC region identified here largely overlapped with the DMPFC region involved in signaling harm to others (Fig. 4*C*, *Left*). A post-hoc comparison confirmed that this equality effect on DMPFC-Striatum connectivity was stronger in the Rank-reversal than No Rank-reversal condition (peak MNI coordinates: [3, 50, 34], t-value = 3.59, voxel-wise *p* (FWE-SVC) = 0.004, k = 63, ROI center MNI coordinates [0, 47, 40], Fig. 4*C* right, Rank reversal absence vs. presence).

To assess whether the pattern of DMPFC-Striatum connectivity may reflect functional influences on the striatum that change behavioral sensitivity to equality concerns, we tested for the Rank-reversal condition whether across individuals, a stronger effect of equality signals on DMPFC-Striatum connectivity may relate to a weaker striatum response to equality and a dampened tendency for equal choice. To this end, we extracted an index of neural equality sensitivity (Beta (high −Δ*F*) − Beta (low −Δ*F*)) from the independent striatum seed region shown in Fig. 4*A*. As already shown in the initial ROI analyses described above, this index confirmed that the striatum was sensitive to equality in the No Rank-reversal condition, but not in the Rank-reversal condition (Fig. 4*B*). In line with the conjecture that DMPFC may act to dampen these striatal equality representations, the index of DPMFC-Striatum connectivity (Fig. 4*C*) exhibited the opposite pattern: It was stronger during Rank-reversal and weaker during the No Rank-reversal condition. Importantly, this effect was not just present on average but also at an individual level, since the differences between the Rank-reversal and No Rank-reversal condition in equality-related DMPFC-Striatum connectivity correlated negatively with the corresponding differences in neural equality sensitivity in the striatum (*dospondinPFCrolated DMPFC*

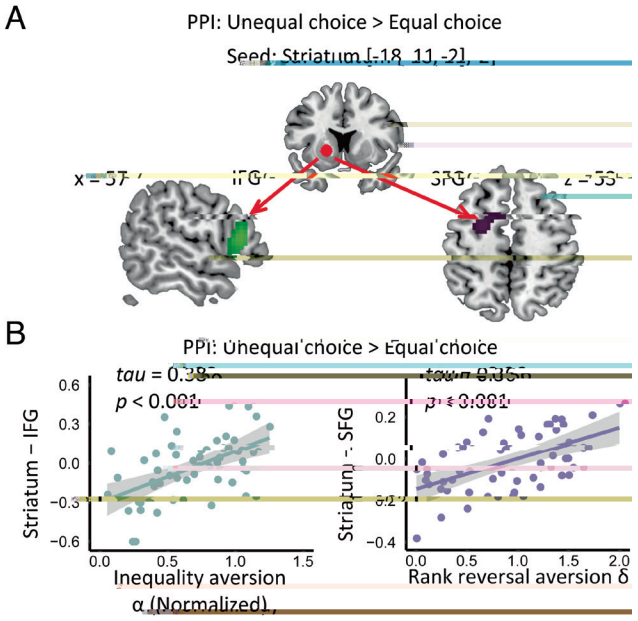

Fig. 6. Neural networks linking different motives to redistribution decisions. (*A*) In the Rank-reversal condition, Striatum-IFG connectivity strength was enhanced when more inequality-averse (higher α) individuals chose more unequal offers vs. more equal offers (*Left* panel), and Striatum-SFG connectivity strength was enhanced when more rank reversal-averse (i.e., higher δ) individuals chose more unequal offers vs. more equal offers (*Right* panel). Neural estimates of the significant clusters were extracted, and scatter plots show the correlation patterns (*B*). Significant clusters were thresholded at voxel-wise $P < 0.001$ uncorrected and cluster-wise FWE corrected $P < 0.05$.

rank reversal aversion) in econometric models, especially since conflicts between these different motives are prevalent in real-life distribution decisions (e.g., taxation policy).

Harm aversion, as a critical type of moral virtue, drives people to achieve a more equal distribution by transferring as little money as possible between two parties. When making moral decisions, people typically conform to the "do-no-harm" principle and prefer not to benefit one party by harming another party (2, 18). Studies of morality suggest that people are not willing to take responsibility for others' bad outcomes when making moral decisions (18, 48), as such moral responsibility will induce individuals' anticipatory guilt emotion which proscribes people from harming others (30, 49). Therefore, taking more money away from others brings not only greater cost for the initially advantaged party but also greater cost of moral responsibility (i.e., harm aversion) for participants which will in turn dampen their motives to seek equality.

Moreover, we suggest that rank reversal aversion is another prosocial motive that discounts the utility of equality during wealth redistribution. A stable hierarchy can provide fitness advantage by satisfying individuals' psychological need for order (50) and enhancing intragroup cooperation and productivity (51). Therefore, it is not surprising that people prefer to preserve rather than reverse preexisting hierarchy (8, 21). In line with these findings, our results suggest that the reversal of initial rankings also contributes to the disutility of equality when rank preserving and equality seeking are in conflict. Together, we demonstrate that in contrast to inequality aversion, harm aversion and rank reversal aversion function as two different third-party prosocial preferences to deter more equal wealth redistribution.

Our neural results first clarified how equality-related information is represented. GLM results support the hypothesis that individuals are sensitive to equality signals in the absence of any conflict but will be less sensitive to equality and base their decisions more heavily on other motives when they conflict with inequality aversion. Although previous studies have proposed that the striatum signals rewarding aspects of equality-related distributions (5–7), it is still unclear which specific aspects of the distributions behavior engage the striatum and trigger the corresponding behavior—does it signal equality or other potentially rewarding aspects, such as efficiency or the other's outcomes? While stronger activity in putamen was related to higher efficiency (i.e., greater overall profits) (5), efficiency cannot account for the pattern of results in the current study since neither of the two alternative offers changed the overall profits of the distributions. An alternative explanation is that striatum activity reflects dopaminergic responses in reward computation of social welfare, as it has been widely observed that stronger striatum activity is associated with charitable giving (52, 53), altruistic punishment to norm violation (23), and more equal wealth distributions (6, 7).

Moreover, striatum has been involved in arousal representations (54). For example, stronger striatal activation was related to greater motivation for norm compliance (55). In the current study, smaller equality difference between the two alternative offers may require participants to base their decisions more heavily on the evidence of equality signals and result in stronger motivation to comply with fairness norms for them, which is manifested by enhanced striatal activity. Together with the finding that greater sensitivity to equality in putamen was related to higher probability of more equal choice, our results suggest that striatum not only reflects processing of equality signals but also promotes fairness norm compliance.

Importantly, representations of equality in striatum were only observed in the No Rank-reversal condition, and this striatal signaling of equality was dampened in the context with conflicts between motives (i.e., Rank-reversal condition). Moreover, stronger DMPFC-Striatum connectivity was associated with lower equality sensitivity in striatum, less equal choice, and higher strength of harm aversion in the Rank-reversal condition. These findings help to clarify the neurocognitive mechanisms of the weighing processes of different motives, by providing a potential neural explanation for the weaker impact of equality on redistribution decisions in the Rank-reversal condition: DMPFC may process harm-related information, convey the harm aversion motive to striatum, interact with striatum, and dampen the tendency for more equal choice. Evidence from two lines of research supports such a modulating role of DMPFC. First, DMPFC, with adjacent regions ACC, is engaged in conflict monitoring, conflict resolution, and action selection in a variety of cognitive tasks (37, 38), which may support the resolution of conflict between different motives in the current paradigm. Second, DMPFC is also thought to be part of the mentalizing system that supports vicarious experiences of others' pain or beliefs (39, 56), which may support harm signals in the current paradigm. In line with our findings, connectivity between prefrontal cortex and striatal value representations was also found to modulate individuals' behaviors in other kinds of social and non-social decision-making (31, 57). However, despite the logical consistency of this interpretation, it is difficult to unambiguously infer the directionality and precise functional contributions of neural interactions from the results of PPI analyses. Future studies with brain stimulation may be needed to establish whether DMPFC influences on striatum are indeed causally involved in guiding redistribution behaviors under circumstances with conflicts between multiple motives.

Our results also provide crucial evidence for frontostriatal circuitry in redistribution decisions. The critical role of frontostriatal circuitry in decision-making has been highlighted in both social and non-social behaviors (31, 55, 57). In general, striatum is

Downloaded from https://www.pnas.org by Universitaet Zuerich on December 8, 2022 from IP address 89.206.81.235.

suggested to receive inputs of goal-related representations from lateral prefrontal cortex and output value signals to guide response selection to maximize reward (58). In line with these suggestions, lateral prefrontal cortices are implicated in either modulating intuitive motivations or value representations that integrate information from different sources for moral and prosocial decision-making (31, 59). Our findings further refine previous accounts of frontostriatal circuitry in moral decision-making by clarifying that different prosocial motives modulate redistribution decisions through differential frontostriatal circuitries. Nevertheless, the specific functional contributions (i.e., inhibitory or modulatory) of these interactions between the striatal and frontal regions still need to be clarified in future studies.

Another critical contribution of our study is to clarify what neural processes underlie the modulations of different prosocial motives on redistribution decisions. Apart from processes involved in arbitrating between motives (i.e., DMPFC-Striatum connectivity), it is also important to identify processes that bias behavior on a trial-by-trial level in line with different motives and which may differ between people with different motive strengths. Activity in both DMPFC and TPJ was stronger when more inequality-averse individuals chose the more unequal offer, and activity in putamen was stronger when more harm-averse individuals chose the more unequal offer. One possibility suggested by the literature is that DMPFC and TPJ may support social cognitive processes such as mentalizing, perspective taking, inference, and learning about others' preferences (39, 56, 60). Recent studies further differentiated the roles of these two regions, by suggesting that while DMPFC is implicated in value-based action selection in a domain general manner (61–63), TPJ may be more specifically involved in processing of context-dependent social information (64, 65). Although our findings cannot provide a clear dissociation between DMPFC and TPJ, among all the regions involved in harm signaling, these two regions may be well-suited to link latent social motives to specific decisions. These findings also parallel the observation of stronger activity in TPJ for unequal choice vs equal choice in the No Rank-reversal condition, which may implicate the role of TPJ in social cognitive processing irrespective of whether there are conflicts between different motives.

In general, our findings may have economic, political, and social implications (66). The endowment effect has been introduced for decades to explain individuals' tendency to increase the subjective value of objects they own already (versus those they want to purchase) (67). Forgoing one's own good is seen as a kind of loss, and loss aversion will make it harder to give up the good (68, 69). In analogy to the endowment effect (70), our study highlights that people are inclined to maintain initial relative rankings and to take less money away from others in wealth redistribution, considering the reversal of initial rankings and others' monetary loss as a kind of third-party loss which proscribes actions to achieve higher equality (8). More generally, our findings may also explain resistance to reform policies that aim to promote social welfare or reduce income inequality (21, 71). For instance, rich people in regions with more equal income distribution, whose advantaged ranks can be more easily reversed, are less supportive of redistribution than those in regions with more unequal income distribution (16). Given that the effects of different motives are scientifically validated in the current study, this may help to develop better taxation policies by taking these motives into account when designing measures to reduce social inequality on the one hand and satisfy people in different income groups who pursue different motives on the other hand.

To conclude, the current study provides a neurocomputational account of the trade-off between multiple prosocial motives underlying resource distribution. Our findings suggest that in addition to inequality aversion, harm aversion and rank reversal aversion work as two separate prosocial motives to modulate individuals' behaviors during wealth redistribution. Moreover, our study offers neural explanations for how different prosocial motives modulate redistribution behaviors, by highlighting a crucial role of striatum in equality processing and modulation of motives on ultimate decisions. Our approach improves our understanding of cognitive and neurobiological mechanisms underlying social preferences and distributive justice and may have implications for development of reform policies to promote fairness norms and social justice.

Materials and Methods

Participants. Sixty-three right-handed healthy adults were recruited in the experiment. Six participants were excluded because of either making the same decision all the time or excessive head movement (> ± 3 mm in translation and/or > ± 3° in rotation). The remaining 57 participants were aged between 19 and 28 y (mean = 21.83 SD = 1.91; 31 female). No participant reported any history of psychiatric, neurological, or cognitive disorders. Informed written consent was obtained from each participant before the experiment. The study was carried out in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Department of Psychology, Peking University.

Experimental Procedure. In the present study, we developed a redistribution task to assess individuals' preferences to redistribute unequal wealth allocations. In this task, participants were first presented with a monetary distribution scheme between two anonymous strangers. The initial endowment of each party was allocated unequally and randomly by computer, and participants had to choose between two redistribution options (i.e., alternative offers) which transferred a certain amount of money from the one with higher initial endowment (advantaged party) to the one with lower initial endowment (disadvantaged party, Fig. 1*A*). In the No Rank-reversal condition, both alternative offers were more equal than the initial offer and kept the same total payoffs and the same relative rankings between the two parties as the initial offer. While in the Rank-reversal condition, participants were presented with the same initial offer and the same more unequal alternative offer as the No Rank-reversal condition, but with a different more equal alternative offer that had the same inequality level as the more equal alternative offer in the No Rank-reversal condition but would reverse the initially relative advantageous/disadvantageous rankings of the two parties (Fig. 1*B*). There were 66 trials in each of the No Rank-reversal and Rank-reversal conditions and 15 trials in each of two filler conditions. The 162 trials were divided into three scanning sessions lasting ~15 min each. After the experiment, each participant received CNY 120 (~ USD 20) for compensation. For further details of the experimental paradigm, see *SI Appendix, SI Materials and Methods*.

Computational Modeling Analyses. To formalize different motives underlying redistribution behaviors, we performed model-based analyses by establishing four families of computational models to examine how inequality aversion, harm aversion, and rank reversal aversion affect individuals' redistribution behaviors in the Rank-reversal condition. For detailed modeling analyses, including model construction, estimation, comparison, and simulation, see *SI Appendix, SI Materials and Methods*.

Neuroimaging Analyses. We collected T2*-weighted echo-planar images using a GE-MR750 3.0 T scanner with a standard head coil at Tongji University, China. The images were acquired in 40 axial slices parallel to the AC-PC line in an interleaved order, with an in-plane resolution of 3 mm × 3 mm, a slice thickness of 4 mm, an inter-slice gap of 4 mm, a repetition time of 2000 ms, an echo time of 30 ms, a flip angle of 90°, and a field of view of 200 mm × 200 mm. We used Statistical Parametric Mapping software SPM12 (Wellcome Trust Department of Cognitive Neurology, London, UK) which was run-through MATLAB (MathWorks) to preprocess the fMRI images, perform GLM analyses and PPI analyses. For detailed neuroimaging analyses, see *SI Appendix, SI Materials and Methods*.

Data, Materials, and Software Availability. Data (behavioral and fMRI) and customized MATLAB and R codes are available online (https://osf.io/zd2tg/).

Downloaded from https://www.pnas.org by Universitaet Zuerich on December 8, 2022 from IP address 89.206.81.235.

ACKNOWLEDGMENTS. This study was supported by grants from the National Natural Science Foundation of China (31630034, 71942001). Dr. J.H. and Dr. C.C.R. also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725355, BRAINCODES), from the UZH URPP AdaBD, and the SNSF (grant 100019L-173248).

Author affiliations: [a]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100871, China; [b]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China; [c]Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Zurich 8006, Switzerland; [d]School of Business and Management, Shanghai International Studies University, Shanghai 200083, China; and [e]Shanghai Key Laboratory of Mental Health and Psychological Crisis Intervention and School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China

1. K. Binmore, *Natural Justice* (Oxford University Press, 2005).
2. J. Baron, Blind justice: Fairness to groups and the do-no-harm principle. *J. Behav. Decis. Mak.* **8**, 71–83 (1995).
3. J. Offer, R. Pinker, Eds., *Social Policy and Welfare Pluralism* (Bristol University Press ed. 1, 2017) 10.2307/j.ctt22p7jvf.
4. J. Rawls, *A Theory of Justice* (Harvard University Press, 1999), 10.2307/j.ctvkjb25m.
5. M. Hsu, C. Anen, S. R. Quartz, The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science* **320**, 1092–1096 (2008).
6. E. Tricomi, A. Rangel, C. F. Camerer, J. P. O'Doherty, Neural evidence for inequality-averse social preferences. *Nature* **463**, 1089–1091 (2010).
7. Y. Hu, S. Strang, B. Weber, Helping or punishing strangers: Neural correlates of altruistic decisions as third-party and of its relation to empathic concern. *Front. Behav. Neurosci.* **9**, 24 (2015).
8. W. Xie, B. Ho, S. Meier, X. Zhou, Rank reversal aversion inhibits redistribution across societies. *Nat. Hum. Behav.* **1**, 0142 (2017).
9. C. Corradi-Dell'Acqua, C. Civai, R. Rumiati, G. Fink, Disentangling self-and fairness-related neural mechanisms involved in the ultimatum game: An fMRI study. *Soc. Cogn. Affect. Neurosci.* **8**, 424–431 (2013).
10. J. J. Jordan, M. Hoffman, P. Bloom, D. G. Rand, Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).
11. E. Lo Gerfo *et al.*, The role of ventromedial prefrontal cortex and temporo-parietal junction in third-party punishment behavior. *Neuroimage* **200**, 501–510 (2019).
12. B. R. House *et al.*, Social norms and cultural diversity in the development of third-party punishment. *Proc. R. Soc. B* **287**, 20192794 (2020).
13. E. Xiao, D. Houser, Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7398–401 (2005).
14. R. Yu, A. J. Calder, D. Mobbs, Overlapping and distinct representations of advantageous and disadvantageous inequality. *Hum. Brain Mapp.* **35**, 3290–3301 (2014).
15. M. Iosifidi, N. Mylonidis, Relative effective taxation and income inequality: Evidence from OECD countries. *J. Eur. Soc. Policy* **27**, 57–76 (2017).
16. M. Dimick, D. Rueda, D. Stegmueller, The Altruistic rich? Inequality and other-regarding preferences for redistribution in the US. *Quart. J. Polit. Sci.* **11**, 385–439 (2016).
17. A. Argentiero, S. Casal, L. Mittone, A. Morreale, Tax evasion and inequality: Some theoretical and empirical insights. *Econ. Gov.* **22**, 309–320 (2021).
18. M. J. Crockett, Z. Kurth-Nelson, J. Z. Siegel, P. Dayan, R. J. Dolan, Harm to others outweighs harm to self in moral decision making. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17320–17325 (2014).
19. M. C. Leliveld, E. van Dijk, I. van Beest, Initial ownership in bargaining: Introducing the giving, splitting, and taking ultimatum bargaining game. *Pers. Soc. Psychol. Bull.* **34**, 1214–1225 (2008).
20. Y. Wu, J. Hu, E. van Dijk, M. C. Leliveld, X. Zhou, Brain activity in fairness consideration during asset distribution: Does the initial ownership play a role? *PLoS One* **7**, e39627 (2012).
21. R. Fernandez, D. Rodrik, Resistance to reform: Status quo bias in the presence of individual-specific uncertainty. *Am. Econ. Rev.* **81**, 1146–155 (2004).
22. E. M. Zitek, L. Z. Tiedens, The fluency of social hierarchy: The ease with which hierarchical relationships are seen, remembered, learned, and liked. *J. Pers. Soc. Psychol.* **102**, 98–115 (2012).
23. D. J. F. De Quervain *et al.*, The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).
24. A. Strobel *et al.*, Beyond revenge: Neural and genetic bases of altruistic punishment. *Neuroimage* **54**, 671–680 (2011).
25. G. Charness, M. Rabin, Understanding social preferences with simple tests. *Q. J. Econ.* **117**, 817–869 (2002).
26. E. Fehr, K. Schmidt, A theory of fairness, competition and cooperation. *Q. J. Econ.* **114**, 817–868 (1999).
27. A. W. Cappelen *et al.*, Equity theory and fair inequality: A neuroeconomic study. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 15368–15372 (2014).
28. D. J. de Quervain, U. Fischbacher, V. Treyer, M. Schellhammer, The neural basis of altruistic punishment. *World* **305**, 1–14 (2004).
29. L. Glass, L. Moody, J. Grafman, F. Krueger, Neural signatures of third-party punishment: Evidence from penetrating traumatic brain injury. *Soc. Cogn. Affect. Neurosci.* **11**, 253–262 (2015).
30. L. J. Chang, A. Smith, M. Dufwenberg, A. G. Sanfey, Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* **70**, 560–572 (2011).
31. M. J. Crockett, J. Z. Siegel, Z. Kurth-Nelson, P. Dayan, R. J. Dolan, Moral transgressions corrupt neural representations of value. *Nat. Neurosci.* **20**, 879–885 (2017).
32. T. Nihonsugi, A. Ihara, M. Haruno, Selective increase of intention-based economic decisions by noninvasive brain stimulation to the dorsolateral prefrontal cortex. *J. Neurosci.* **35**, 3412–3419 (2015).
33. X. Gao *et al.*, Distinguishing neural correlates of context-dependent advantageous- and disadvantageous-inequity aversion. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7680–E7689 (2018).
34. S. Lewandowsky, S. Farrell, *Computational Modeling in Cognition* (2010), 10.1017/CBO9781107415324.004.
35. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage* **76**, 412–427 (2013).
36. J. A. Clithero, A. Rangel, Informatic parcellation of the network involved in the computation of subjective value. *Soc. Cogn. Affect. Neurosci.* **9**, 1289–1302 (2014).
37. C. R. Oehrn *et al.*, Neural communication patterns underlying conflict detection, resolution, and adaptation. *J. Neurosci.* **34**, 10438–10452 (2014).
38. S. F. de Kloet *et al.*, Bi-directional regulation of cognitive control by distinct prefrontal cortical output neurons to thalamus and striatum. *Nat. Commun.* **12**, (2021).
39. F. Van Overwalle, Social cognition and the brain: A meta-analysis. *Hum. Brain Mapp.* **30**, 829–58 (2009).
40. C. A. Hill *et al.*, A causal account of the brain network computations underlying strategic social behavior. *Nat. Neurosci.* **20**, 1142–1149 (2017), 10.1038/nn.4602.
41. K. J. Friston *et al.*, Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).
42. A. C. Loewke, A. R. Minerva, A. B. Nelson, A. C. Kreitzer, L. A. Gunaydin, Frontostriatal projections regulate innate avoidance behavior. *J. Neurosci.* **41**, 5487–5501 (2021).
43. K. E. Pickett, R. G. Wilkinson, Income inequality and health: A causal review. *Soc. Sci. Med.* **128**, 316–326 (2015).
44. B. K. Payne, J. L. Brown-Iannuzzi, J. W. Hannay, Economic inequality increases risk taking. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 4643–4648 (2017).
45. E. Fehr, U. Fischbacher, Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190 (2004).
46. E. Fehr, U. Fischbacher, Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
47. J. Hu *et al.*, Social status modulates the neural response to unfairness. *Soc. Cogn. Affect. Neurosci.* **11**, 1–10 (2015).
48. I. Ritov, J. Baron, Reluctance to vaccinate: Omission bias and ambiguity. *J. Behav. Decis. Making* **3**, 263–277 (1990).
49. H. Yu, J. Hu, L. Hu, X. Zhou, The voice of conscience: Neural bases of interpersonal guilt and compensation. *Soc. Cogn. Affect. Neurosci.* **9**, 1150–1158 (2014).
50. J. P. Friesen, A. C. Kay, R. P. Eibach, A. D. Galinsky, Seeking structure in social organization: Compensatory control and the psychological advantages of hierarchy. *J. Pers. Soc. Psychol.* **106**, 590–609 (2014).
51. N. Halevy, E. Y. Chou, A. D. Galinsky, J. K. Murnighan, When hierarchy wins: Evidence from the national basketball association. *Soc. Psychol. Personal. Sci.* **3**, 398–406 (2012).
52. J. Moll *et al.*, Human fronto-mesolimbic networks guide decisions about charitable donation. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 15623–15628 (2006).
53. W. T. Harbaugh, U. Mayr, D. R. Burghart, Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science* **316**, 1622–1624 (2007).
54. B. Knutson, C. M. Adams, G. W. Fong, D. Hommer, Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.* **21**, 1–5 (2001).
55. M. Spitzer, U. Fischbacher, B. Herrnberger, G. Grön, E. Fehr, The neural signature of social norm compliance. *Neuron* **56**, 185–196 (2007).
56. M. M. Garvert, M. Moutoussis, Z. Kurth-Nelson, T. E. J. Behrens, R. J. Dolan, Learning-Induced plasticity in medial prefrontal cortex predicts preference malleability. *Neuron* **85**, 418–428 (2015).
57. W. van den Bos, C. A. Rodriguez, J. B. Schweitzer, S. M. McClure, Connectivity strength of dissociable striatal tracts predict individual differences in temporal discounting. *J. Neurosci.* **34**, 10298–10310 (2014).
58. J. W. Buckholtz, R. Marois, The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nat. Neurosci.* **15**, 655–61 (2012).
59. J. Hu, Y. Hu, Y. Li, X. Zhou, Computational and neurobiological substrates of cost-benefit integration in altruistic helping decision. *J. Neurosci.* **41**, 3545–3561 (2021).
60. A. Ogawa, T. Kameda, Dissociable roles of left and right temporoparietal junction in strategic competitive interaction. *Soc. Cogn. Affect. Neurosci.* **14**, 1037–1048 (2020).
61. C. K. Kovach *et al.*, Anterior prefrontal cortex contributes to action selection through tracking of recent reward trends. *J. Neurosci.* **32**, 8434–8442 (2012).
62. E. D. Boorman, M. F. Rushworth, T. E. Behrens, Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. *J. Neurosci.* **33**, 2242–2253 (2013).
63. J. X. O'Reilly *et al.*, Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E3660–E3669 (2013).
64. S. M. Lee, G. McCarthy, Functional heterogeneity and convergence in the right temporoparietal junction. *Cereb. Cortex* **26**, 1108–1116 (2016).
65. A. Konovalov, C. Hill, J. Daunizeau, C. C. Ruff, Dissecting functional contributions of the social brain to strategic behavior. *Neuron* **109**, 3323–3337.e5 (2021).
66. B. Irlenbusch, M. C. Villeval, Behavioral ethics: How psychology influenced economics and how economics might inform psychology? *Curr. Opin. Psychol.* **6**, 87–92 (2015).
67. Z. Carmon, D. Ariely, Focusing on the forgone: How value can appear so different to buyers and sellers. *J. Consum. Res.* **27**, 360–370 (2000).
68. D. Kahneman, J. L. Knetsch, R. H. Thaler, Anomalies: The endowment effect, loss aversion, and status quo bias. *J. Econ. Perspect.* **5**, 193–206 (1991).
69. C. K. Morewedge, L. L. Shu, D. T. Gilbert, T. D. Wilson, Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. *J. Exp. Soc. Psychol.* **45**, 947–951 (2009).
70. R. A. Y. Weaver, S. Frederick, A reference price theory of the endowment effect. *J. Mark. Res.* **XLIX**, 696–707 (2012).
71. I. Kuziemko, R. W. Buell, T. Reich, M. I. Norton, Last-place aversion: Evidence and redistributive implications. *Q. J. Econ.* **129**, 105–149 (2014).

Downloaded from https://www.pnas.org by Universitaet Zuerich on December 8, 2022 from IP address 89.206.81.235.