1 **Cognitive and neural bases of decision-making causing civilian casualties during**
2 **intergroup conflict**

3
4 Xiaochun Han^{1,2,3}, Shuai Zhou⁴, Nardine Fahoum⁵, Taoyu Wu^{1,2,3}, Tianyu Gao^{1,2,3},
5 Simone Shamay-Tsoory⁵, Michele J. Gelfand⁶, Xinhui Wu^{4*}, Shihui Han^{1,2,3*}

6
7 ¹School of Psychological and Cognitive Sciences, Peking University, Beijing, China

8 ²PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China

9 ³Beijing Key Laboratory of Behavior and Mental Health, Peking University, China

10 ⁴Department of Radiology, The 7th Medical Center of PLA General Hospital, Beijing,
11 China

12 ⁵Department of Psychology, University of Haifa, Haifa, Israel

13 ⁶Department of Psychology, University of Maryland, College Park, The United States
14 of America

15

16

17

18 * Corresponding authors: Shihui Han (shan@pku.edu.cn, ORCID ID: 0000-0003-
19 3350-5104); Xinhui Wu (bei925@sina.com, ORCID: 0000-0002-4134-3507)

20

21 **Abstract**

22
 23 Civilian casualties occur during military attacks. Such “collateral damage” is
 24 prohibited by international laws but increases with substantial consequences when
 25 intergroup conflict escalates. Here, we investigate cognitive and neural bases of
 26 decision-making processes resulting in civilian harm, using a task that simulates
 27 punishment decision-making during intergroup conflict. We test two groups of
 28 Chinese participants in a laboratory setting, and two ethnic groups (Jewish and
 29 Palestinian) in Israel. The results dissociate two psychological constructs, harm
 30 preference and harm avoidance, which respectively characterize punishment decision-
 31 making related to outgroup combatants and outgroup noncombatants during
 32 intergroup conflict. In particular, individuals show decreased avoidance of harming
 33 outgroup noncombatants when conflict escalates. Brain imaging (functional magnetic
 34 resonance imaging) reveals that decreased harm avoidance is predicted by inhibition
 35 of the left middle frontal activity during selection of punishment decisions. Our
 36 findings provide insight into the cognitive and neural bases of decision-making
 37 involving civilian harm during intergroup conflict.

38

39 Introduction

40 “Collateral damage” refers to civilian or noncombatant casualties and property
 41 damage due to aggressive actions during intergroup conflict such as a military
 42 operation during war¹. Although conflicting with humanitarian interests and being
 43 prohibited by international laws², “collateral damage” occurs massively in human
 44 societies^{3,4} and raises serious moral and ethical concerns. For example, hundreds of
 45 civilians were killed and thousands of civilians were wounded due to NATO's
 46 bombing intervention in the Kosovo conflict⁵. The ratio of civilian to military deaths
 47 during Iraq war is estimated to range between 3:1 and 8:1⁴. During the 2014 Israel-
 48 Gaza conflict over two thousand individuals were killed and 67% of the victims were
 49 civilians⁶. Given the substantial social consequences of civilian casualties,
 50 understanding its psychological constructs and brain underpinnings is of both
 51 theoretical and practical importance.

52 Civilian casualties are usually regarded as a side effect of aggressive acts
 53 following punishment decisions toward outgroup combatants during intergroup
 54 conflict. Punishment decision-making engages complicated cognitive and neural
 55 processes^{7,8}, including detection and generation of aversive experiences in the
 56 salience network (e.g., the anterior midcingulate and insula)⁹, integration of harm and
 57 intent into an assessment of blame in the theory-of-mind network (e.g., the medial
 58 prefrontal cortex and temporoparietal conjunction)¹⁰, and action selection/control in
 59 the central executive network (e.g., the lateral prefrontal cortex)¹¹. In previous
 60 research, punishment decisions were made towards outgroup members
 61 indiscriminately as a whole in the context of intergroup conflict¹²⁻¹⁶. This leaves an
 62 open issue related to civilian casualties, i.e., how to separate consequences of
 63 decision-making pertaining to outgroup combatants versus noncombatants so as to

64 disentangle the cognitive/neural mechanisms underlying decision-making that causes
65 harm to civilians. To address this issue would allow us to unravel psychological
66 constructs characterizing the unique pattern of decision-making related to “collateral
67 damage” during intergroup conflict. It is now known that a critical feature of real-
68 world civilian casualties is that it increases when intergroup conflict escalates¹⁷. This
69 suggests an approach to the understanding of cognitive and neural bases of
70 punishment decision-making resulting in civilian casualties by examining
71 psychological processes that vary along escalation of intergroup conflict.

72 Here, we propose two psychological constructs involved in punishment decision-
73 making pertaining to outgroup members during intergroup conflict. Because the goal
74 of combat during intergroup conflict is to punish or destroy outgroup combatants but
75 to maintain minimal noncombatant casualties, as required by morality and
76 international laws², we hypothesize that harm preference and harm avoidance are two
77 distinct psychological constructs that differentially characterize punishment decisions
78 pertaining to outgroup combatants versus noncombatants during intergroup conflict
79 (Hypothesis 1). Moreover, civilian casualties arising from intergroup conflict
80 escalation may result from reduced harm avoidance pertaining to outgroup
81 noncombatants (Hypothesis 2). That is, harm avoidance that inhibits harm to
82 noncombatants during intergroup conflict is predicted to decrease when intergroup
83 conflict escalates.

84 To test these hypotheses, we developed a behavioral paradigm in which
85 participants were asked to make a decision by choosing electric shocks for punishing
86 an outgroup member who was directly engaged in a conflict (an outgroup combatant)
87 involving the distribution of resources. This punishment of the combatant was
88 accompanied by varying degrees of harm to another outgroup member who was not

involved in the conflict (an outgroup noncombatant). These manipulations in a laboratory setting simulate the nature of real-world civilian casualties during intergroup conflict in several aspects. First, the conflict between two groups was initiated by an unfair distribution of limited resources, which is consistent with research that shows that unfair money distributions induce conflict between two individuals or two groups¹⁸. Second, the participants had to make decisions to punish outgroup combatants, and these decisions also brought harm to outgroup noncombatants who were not involved in the conflict. Third, intergroup conflict was manipulated to vary across three (low, middle, or high) levels in our design. This design enables us not only to dissociate harm preference and harm avoidance by conducting statistical modeling of choices of electric shocks related to outgroup combatants and noncombatants, respectively, but also to estimate whether harm to outgroup noncombatants escalated when the level of intergroup conflict increased, similar to what happens during real-world intergroup conflicts¹⁷.

We examined how harm preference and harm avoidance varied along escalation of intergroup conflict in both minimal groups manipulated in a laboratory setting and two ethnic groups with conflicts in a real-world situation (i.e., Jewish and Palestinian groups during the Israeli-Palestinian conflict). Our statistical modeling of behavioral choices of punishment decisions showed consistent evidence for decreased harm avoidance pertaining to outgroup noncombatants during escalation of intergroup conflict. To further improve our understanding of neural underpinnings of punishment decision-making causing civilian casualties, we conducted an additional functional magnetic resonance imaging (fMRI) experiment to examine which neural network, e.g., the salience network, theory-of-mind network, or central executive network that contribute to punishment decision-making^{7,8}, is associated with harm preference or

114 harm avoidance. In particular, we estimated neural underpinnings of punishment
115 decision-making resulting in civilian casualties by examining brain activity associated
116 with the decreased harm avoidance pertaining to outgroup noncombatants as
117 intergroup conflict escalates.

118 **Results**

119 **Decreased harm avoidance during escalation of intergroup conflict in a** 120 **laboratory setting**

121 To disentangle the psychological constructs related to civilian casualties from
122 nonessential cognitive/affective processes engaged in punishment decision-making
123 during intergroup conflict, in Experiment 1, we randomly assigned Chinese adults to a
124 group conflict condition (N = 120) or an individual conflict (control) condition (N =
125 120). In the group conflict condition four participants of the same gender, who were
126 strangers, were recruited simultaneously and tested in the same session. To build
127 ingroup/outgroup relationships, the four participants were divided into two mini-
128 groups (N = 2 in each mini-group) by wearing T-shirts of two different colors and
129 playing a 1-hour card game (Phase 1 in Figure 1). This game required the participants
130 to help ingroup members but to interrupt outgroup members so that one's own group
131 would reach a destination in advance (see Methods). After group formation in Phase 1,
132 the participants completed a modified version of the Inclusion of Other in the Self
133 Scale¹⁹ to assess their feelings of closeness between themselves and other participants.
134 The same procedure was applied to the participants in the individual conflict
135 condition except that there was no mini-group assignment and each participant
136 competed against the other three in the game. Thus, no one shared group identity with
137 others in the individual conflict condition, and the comparison of behavioral
138 performance in group and individual conflict conditions allowed us to disentangle the

effect specific to intergroup conflict in the group conflict condition by controlling other irrelevant factors. The analyses of the mean rating scores showed that the participants in the group conflict condition reported greater feelings of closeness to ingroup compared to outgroup members (mean $\pm$ s.d. = 4.92 ± 0.99 vs. 2.95 ± 1.14 ; $F(1,116) = 232.40$, $P < 0.001$, $\eta^2_p = 0.667$, 95% CI = 1.59, 2.61, $BF_{10} > 150$). The feelings of closeness to others in the individual conflict condition were lower than that pertaining to ingroup members in the group conflict condition (3.13 ± 1.03 vs. 4.92 ± 0.99 ; $F(1,232) = 187.73$, $P < 0.001$, $\eta^2_p = 0.447$, 95% CI = -2.05, -1.53, $BF_{10} > 150$) but did not differ significantly from that related to outgroup members in the group conflict condition (3.13 ± 1.03 vs. 2.95 ± 1.14 ; $F(1,232) = 1.63$, $P = 0.203$, $\eta^2_p = 0.01$, 95% CI = -0.10, 0.45, $BF_{01} = 23.61$). These results indicate that group formation increased the affective link between ingroup members rather than decreased the affective link to outgroup members relative to the control condition.

Thereafter, in Phase 2 shown in Figure 1, one participant was randomly selected from each mini-group to play a combatant role in the group conflict condition (called C1 and C2, i.e. a combatant from group 1 and a combatant from group 2) and the remaining one in each group played the role of a noncombatant (called N1 and N2). In the individual conflict condition two participants were randomly selected as combatants (also called C1 and C2) with the left two being bystanders (also called N1 and N2). Thus C1 and N1 (or C2 and N2) shared group identity in the group conflict condition but not in the individual conflict condition. The following procedure was the same in the group and individual conflict conditions. Each noncombatant was informed of a 100 RMB payment for their participation. However, each combatant was told that, in addition to a basic payment of 50 RMB, he/she would share a 100 RMB bonus with the outgroup combatant as an additional payment and how much

each person would obtain depends on the final distribution plan. C1 and C2 were then asked to make decisions privately on how to allot the bonus payment between C1 and C2 by choosing one of the distribution plans. Selection of a distribution plan of the bonus payment did not affect the payment for N1 and N2 so that N1 and N2 were not involved in the conflict and their noncombatant roles were not influenced by C1 and C2's distribution plans. Selection of a distribution plan was repeated 108 times to make the participants believe various distribution plans used in the following punishment decision-making task. We defined three levels of conflict (low, middle, high) based on the distribution of the bonus payment between C1 and C2. In Phase 3 all participants went through a pain threshold test to rate their painful feelings (1 = slightly painful, 9 = extremely painful) corresponding to different intensities of electric shocks.

During the following punishment decision-making task (Phase 4 in Figure 1) each participant was tested in a private room to minimize social influences on their decision-making. Each trial began with the presentation of a plan of an unfair distribution (e.g., C1 and C2 receive 25 and 75 RMB, respectively), which was supposed to be given by an outgroup combatant (e.g., given by C2 in the case illustrated in Phase 4 in Figure 1) but was actually generated by a computer program. In the group conflict condition, the participant was instructed that, given the perceived distribution plan that was unfair to the ingroup combatant, they had to make a decision to punish the outgroup combatant. They were shown two punishment options, one on the left side and one on the right side of a screen, respectively (see illustration of Phase 4 in Figure 1). Each option consisted of two vertical bars indicating intensities (1 = slightly painful, 9 = extremely painful) of two electric shocks. One shock would be applied to the outgroup combatant as punishment and the other to the

189 outgroup noncombatant as a collateral damage. There were three levels (i.e., 1-3, 4-6,
190 or 7-9 corresponding to the low, middle, or high level of pain) of electric shocks. In
191 either the left or right punishment option there were nine combinations of the two
192 electric shocks applied to the outgroup combatant and noncombatant (i.e., 3 (low,
193 middle, or high) levels of shocks to the combatant x 3 (low, middle, or high) levels of
194 shocks to the noncombatant). These were the same in the individual conflict condition
195 except that one shock would be applied to the combatant as punishment and the other
196 to a bystander. The participants were told that the bonus and electric shock in one of
197 their decisions would be randomly selected to apply to the target combatant and
198 noncombatant after the experiment.

199 The distribution plan and the options of electric shocks were manipulated to
200 balance conflict levels and punishment options. Each participant completed 108 trials
201 of punishment decision-making with 3 levels of conflict based on the given
202 distribution plan (i.e., leaving 55-65, 70-80, or 85-95 RMB for oneself,
203 corresponding to low, middle, or high conflict) and three (low, middle, or high) levels
204 of electric shocks. The intensities of the two electric shocks applied to outgroup
205 combatants and noncombatants in the group conflict condition (or to combatants and
206 bystanders in the individual conflict condition) varied independently across trials.
207 There were 108 different combinations of conflict level and electric shock intensity
208 that were presented to each participant in a random order.

209 Participants' choices of electric shocks were subject to logistic regression
210 analyses²⁰. Because we hypothesized that outgroup combatants and noncombatants
211 are treated differently in terms of harm preference and harm avoidance during
212 intergroup conflict, we included four independent variables in the regression model to
213 account for the probability that participants would select the left option (p) or the

probability that participants would select the right option ($1-p$) in each trial (see Method, Supplementary Fig. 1 and Supplementary Table 1) for the methods and results of model comparisons). These included the difference in intensities of electric shocks between left and right options applied to the combatant (ΔS_c) and the noncombatant (ΔS_n), the interactions between ΔS_c and level of conflict (L) and between ΔS_n and L. The logistic regression model was defined as following:

$$\text{logit} = \ln \frac{p}{1-p} = \Delta V \quad (1)$$

$$\Delta V = \mathbf{h}_c \times \Delta S_c + \mathbf{h}_n \times \Delta S_n + \mathbf{h}_{c-L} \times \Delta S_c \times L + \mathbf{h}_{n-L} \times \Delta S_n \times L \quad (2)$$

$$p = \frac{1}{1+e^{-\Delta V}} \quad (3)$$

A larger p is associated with higher possibility of selection of the left option.

The regression coefficients of these independent variables in this logistic regression model are associated with different psychological constructs. A positive $\mathbf{h}_c$ or $\mathbf{h}_n$ value indicates harm preference — a propensity to select a stronger electric shock from two options to punish a target. A negative $\mathbf{h}_c$ or $\mathbf{h}_n$ value indicates harm avoidance — a propensity to select a weaker electric shock from two options to punish a target. $\mathbf{h}_{c-L}$ (or $\mathbf{h}_{n-L}$) manifests how $\mathbf{h}_c$ (or $\mathbf{h}_n$) varies along escalated levels of conflict and positive (or negative) $\mathbf{h}_{c-L}$ or $\mathbf{h}_{n-L}$ values indicate increase (or decrease) of $\mathbf{h}_c$ or $\mathbf{h}_n$. To reduce potential biases in estimation and to allow for finite parameter estimates for data with separation, we performed logistic function estimation using Firth's penalized-likelihood logistic regression (*logistf* function in R)^{21,22}. The logistic regression model was estimated and the four regression coefficients were extracted for each participant for the second group-level t-tests. We also conducted Bayesian factor analyses²³ to assess the likelihood of data driven alternative hypothesis over the likelihood of data driven null hypothesis (BF_{10}) or the reverse (BF_{01}) (see Methods for details).

239 If harm preference and harm avoidance characterize punishment decisions
240 pertaining to outgroup combatants and noncombatants, respectively, we would expect
241 a positive value of h_c (harm preference to combatants) but a negative value of h_n
242 (harm avoidance related to noncombatants). In addition, if collateral damage during
243 escalation of intergroup conflict arises from reduced harm avoidance pertaining to
244 outgroup noncombatants, we predicted a positive value of h_{n-L} .

245 As expected, the results of one sample t-tests (all two-tailed) showed that h_c was
246 significantly larger than zero in both the group conflict condition (mean $\pm$ s.d. = 0.18
247 $\pm$ 0.80, $t(119) = 2.47$, $P = 0.015$, FDR corrected, Cohen's $d = 0.23$, 95% CI = 0.04,
248 0.32; $BF_{10} = 1.84$, only FDR corrected results were reported in the following) and the
249 individual conflict condition (0.31 $\pm$ 0.68, $t(119) = 4.95$, $P < 0.001$, Cohen's $d = 0.45$,
250 95% CI = 0.19, 0.43; $BF_{10} > 150$, Figure 2, Supplementary Fig. 2), but did not differ
251 significantly between the two conditions (two sample test: $t(238) = -1.35$, $P = 0.358$,
252 Cohen's $d = 0.17$, 95% CI = -0.32, 0.06; $BF_{01} = 3.00$, see Figure 2). However, h_n was
253 significantly smaller than zero in both the group conflict condition (-0.52 $\pm$ 0.53, $t(119)$
254 = -10.71, $P < 0.001$, Cohen's $d = 0.98$, 95% CI = -0.61, -0.42; $BF_{10} > 150$) and in the
255 individual conflict condition (-0.55 $\pm$ 0.50, $t(119) = -12.14$, $P < 0.001$, Cohen's $d =$
256 1.11, 95% CI = -0.64, -0.46; $BF_{10} > 150$), but did not differ significantly between the
257 two conditions ($t(238) = 0.55$, $P = 0.584$, Cohen's $d = 0.07$, 95% CI = -0.09, 0.17;
258 $BF_{01} = 6.14$). These results provide evidence for harm preference for combatants
259 (indicated by a positive h_c) but harm avoidance related to noncombatants (indicated
260 by a negative h_n) on average across the three conflict levels.

261 The analyses of h_{c-L} showed similar patterns in the group and individual conflict
262 conditions whereas the analyses of h_{n-L} revealed different patterns between the two
263 conditions. h_{c-L} was significantly larger than zero in both the group conflict condition

(0.41±0.61, $t(119) = 7.33$, $P < 0.001$, Cohen's $d = 0.67$, 95% CI = 0.30, 0.52; $BF_{10} > 150$) and individual conflict condition (0.35±0.51, $t(119) = 7.58$, $P < 0.001$, Cohen's $d = 0.69$, 95% CI = 0.26, 0.44; $BF_{10} > 150$), but did not differ significantly between the two conditions ($t(238) = 0.79$, $P = 0.576$, Cohen's $d = 0.10$, 95% CI = -0.09, 0.20; $BF_{01} = 5.28$). The results suggest that, for both group conflict and individual conflict conditions, harm preference (indicated by h_c) towards combatants increased when conflict escalated. Most interestingly, h_{n-L} was significantly larger than zero in the group conflict condition (0.09±0.29, $t(119) = 3.56$, $P < 0.001$, Cohen's $d = 0.33$, 95% CI = 0.04, 0.15; $BF_{10} = 36.87$) but not in the individual conflict condition (-0.05±0.35, $t(119) = -1.50$, $P = 0.135$, Cohen's $d = 0.14$, 95% CI = -0.11, 0.02; $BF_{01} = 3.30$). A two-sample t-test further confirmed the significant difference in h_{n-L} between the group and individual conflict conditions ($t(238) = 3.42$, $P = 0.003$, Cohen's $d = 0.44$, 95% CI = 0.06, 0.22; $BF_{10} = 32.28$). These results provide evidence that harm avoidance toward noncombatants decreased (i.e., h_n (harm avoidance related to noncombatants) shifted toward less negative) along escalated conflict in the group conflict condition but not in the individual conflict condition. This finding supports the idea that decreased harm avoidance functions as a psychological construct related to civilian casualties that occurs specifically during intergroup conflict.

To test how well the logistic regression model could predict individuals' punishment decisions, we compared participants' choices and the model prediction of a choice in each trial (e.g., probability to choose left or right option with 0.5 as the threshold). The results showed that the logistic regression model correctly predicted $85.03 \pm 7.77\%$ of choices, which is significantly higher than the chance level ($t(239) = 69.85$, $P < 0.001$, Cohen's $d = 4.51$, 95% CI = 0.84, 0.86; $BF_{10} > 150$, see

288 Supplementary Fig. 1). The results indicate that the logistic regression model (Model
289 1) fits well with the participants' actual performances during decision making.

290 **Decreased harm avoidance during escalation of intergroup conflict in a real-**
291 **world context**

292 The results of Experiment 1 suggest decreased harm avoidance toward outgroup
293 noncombatants during intergroup conflict escalation in a laboratory setting. In
294 Experiment 2 we examined whether this account applies to collateral damage during
295 intergroup conflict in a real-world context by testing Jewish (N = 106) and Palestinian
296 participants (N = 106) in Israel. The Israeli-Palestinian conflict is one of the most
297 challenging intergroup conflicts in the world and threatens numerous civilians'
298 lives^{6,24-26}. The study of participants from these two groups allowed us to further
299 understand the role of decreased harm avoidance in civilian casualties in a real-world
300 intergroup conflict. Experiment 2 was carried out during the year 2018 when repeated
301 clashes with numerous casualties occurred between the two sides. Because of the
302 difference in national and religious identity between the two ethnic groups, we
303 divided 2 Jewish participants and 2 Palestinian participants (all were of the same
304 gender and strangers to each other) into two mini-groups on each testing day. The
305 participants were blind to what the study was about during the recruitment. To
306 highlight participants' ethnic identity, after introducing one's own ethnicity, all
307 participants were presented with images of items or celebrities/holy places that
308 remind them of the ethnicity-specific cultural elements²⁷. The participants from each
309 mini-group were asked to jointly observe each picture and rate their feelings of
310 picture-induced belonging and excitement (1 = very little, 9 = very much). The rating
311 scores were high for both ethnic groups and higher in Palestinian participants than in
312 Jewish participants (see Supplementary information, Results 1 and 2, Supplementary

313 Tables 2 and 3, and Supplementary Fig. 3).

314 Thereafter, participants completed the money distribution task, pain threshold test,
315 and punishment decision-making task. The procedures of these tasks were the same as
316 those in Experiment 1 except that the participants were paid with a corresponding
317 amount of Shekels. Similarly, a logistic regression model was constructed to estimate
318 each participant's performance during punishment decision-making. We first assessed
319 harm preference and harm avoidance characterizing the whole sample by pooling
320 together the two ethnic groups' data and then examined possible group differences.
321 One sample t-tests conducted on the whole sample showed that h_c (harm preference to
322 combatants) did not differ significantly from zero (mean $\pm$ s.d. = -0.07 ± 0.55 , $t(211) =$
323 -1.76 , $P = 0.081$, Cohen's $d = 0.12$, 95% CI = $-0.14, 0.01$; $BF_{01} = 2.88$), whereas h_n
324 (harm avoidance related to noncombatants) was significantly smaller than zero ($-$
325 0.50 ± 0.46 , $t(211) = -15.67$, $P < 0.001$, Cohen's $d = 1.08$, 95% CI = $-0.56, -0.44$; $BF_{10} >$
326 150 , Figure 3, Supplementary Fig. 4 and 5). Moreover, these effects were
327 significantly smaller in Jewish than in Palestinian participants (h_c : -0.19 ± 0.55 vs.
328 0.06 ± 0.51 , $t(210) = -3.49$, $P = 0.001$, Cohen's $d = 0.57$, 95% CI = $-0.38, -0.14$; $BF_{10} >$
329 150 ; h_n : -0.63 ± 0.49 vs. -0.37 ± 0.39 , $t(210) = -4.17$, $P < 0.001$, Cohen's $d = 0.57$, 95%
330 CI = $-0.38, -0.14$; $BF_{10} > 150$). The results of one sample t-tests conducted on the
331 whole sample revealed that both h_{c-L} and h_{n-L} were significantly larger than zero (h_{c-L} :
332 0.19 ± 0.28 , $t(211) = 9.84$, $P < 0.001$, Cohen's $d = 0.68$, 95% CI = $0.15, 0.23$; $BF_{10} >$
333 150 ; h_{n-L} : 0.06 ± 0.28 , $t(211) = 3.04$, $P = 0.003$, Cohen's $d = 0.21$, 95% CI = $0.02,$
334 0.10) ; $BF_{10} = 6.66$). These effects did not differ significantly between Jewish and
335 Palestinian participants (h_{c-L} : 0.19 ± 0.30 vs. 0.19 ± 0.26 , $t(210) = 0.14$, $P = 0.888$,
336 Cohen's $d = 0.02$, 95% CI = $-0.07, 0.08$; $BF_{01} = 6.62$; h_{n-L} : 0.05 ± 0.34 vs. 0.07 ± 0.20 ,
337 $t(210) = -0.57$, $P = 0.756$, Cohen's $d = 0.08$, 95% CI = $-0.10, 0.05$; $BF_{01} = 5.72$).

338 Similarly, the logistic regression model predicted a high proportion of punishment
339 decisions across trials that was significantly higher than the chance level
340 ($83.89\% \pm 8.39\%$, $t(211) = 58.79$, $P < 0.001$, Cohen's $d = 4.04$, $95\% \text{ CI} = 82.75\%$,
341 85.02% ; $\text{BF}_{10} > 150$, Supplementary Fig. 1), indicating that the logistic regression
342 model fits well with actual performance in the sample tested in a real-world situation.

343 Together, the results in Experiments 1 and 2 provide evidence for decreased harm
344 avoidance toward noncombatants when intergroup conflicts escalate in both a
345 laboratory setting and a real social context. Moreover, the same logistic regression
346 model fits the empirical data of punishment decision-making during intergroup
347 conflict in both experiments. These results suggest that decreased harm avoidance
348 provides a common cognitive basis of decision-making involving civilian casualties in
349 a laboratory setting and a real-world social context. Notably, our participants in a
350 laboratory setting but not in a real-world situation showed harm preference and
351 Jewish participants even exhibited harm avoidance toward outgroup combatants (see
352 Supplementary Fig. 4). Punishment decisions seemed to be made more carefully in a
353 real social context of intergroup conflict than in a laboratory setting possibly due to
354 higher stakes in the latter situation. It is also possible that the procedure used in our
355 study created demand characteristics that resulted in social desirability pressures to
356 appear moral. Future research should examine these distinct possibilities and replicate
357 them with other samples embedded in real world conflicts.

358 **Neural correlates of decreased harm avoidance during intergroup conflict** 359 **escalation**

360 In Experiment 3, we further investigated neural correlates of the decreased harm
361 avoidance during intergroup conflict by testing an independent sample of Chinese
362 adults. The experimental procedures were the same as those in Experiment 1 except

that participants were scanned using fMRI during the punishment decision task and the amounts of the basic payment and additional bonus payment were doubled. Participants were randomly assigned to the group conflict condition ($N = 64$) and the individual conflict condition ($N = 64$). Logistic regression analyses of the behavioral data replicated the findings of decreased harm avoidance in punishment decision-making related to outgroup noncombatants during conflict escalation in the group conflict condition (Figure 4A, see Supplementary information, Result 3, Supplementary Table 4 and Supplementary Fig. 6 for details).

The first-level whole-brain general linear model (GLM) analyses of fMRI data estimated brain activities associated with the parametric modulators including ΔS_c , ΔS_n , $\Delta S_c * L$, and $\Delta S_n * L$ in each participant. We conducted four second-level whole-brain regression analyses to test brain regions associated with individual differences in parameters estimated using the logistic regression modelling. The second-level regression analyses included the contrast images of parametric modulation effects of ΔS_c (with the covariate $\mathbf{h}_c$), ΔS_n (with the covariate $\mathbf{h}_n$), $\Delta S_c * L$ (with the covariate $\mathbf{h}_{c-L}$), and $\Delta S_n * L$ (with the covariate $\mathbf{h}_{n-L}$) of each participant. Most interestingly, the second-level analyses showed that a larger $\mathbf{h}_{n-L}$ was significantly correlated with decreased activity in the left middle frontal gyrus (MFG, MNI coordinates $x/y/z = -48/28/32$, a voxel-level threshold $P < 0.001$, cluster-level threshold $P < 0.05$, FWE corrected, Figure 4B) in the group conflict condition but not in the individual conflict condition, suggesting a functional role of the left MFG in controlling harm avoidance pertaining to outgroup noncombatants when intergroup conflict escalated.

To further confirm that the difference in the association between the left MFG activity and $\mathbf{h}_{n-L}$ was statistically significant when directly comparing the group and

individual conflict conditions, we conducted a moderation analysis of the association between h_{n-L} and the left MFG activity. This analysis searched for regions in the left frontal lobe defined by a mask in the Brainnetome Atlas²⁸ in which the association between h_{n-L} and the left MFG activity differed significantly between the group and individual conflict conditions. The result revealed a significant activation in a region in which the voxel with peak activation was located at -48/28/30 ($z = 3.65$, voxel-level threshold $P < 0.05$ with small volume FWE correction). This region overlaps with the left MFG shown in the second-level whole-brain regression analyses and provides evidence that intergroup conflict significantly enhanced the coupling between h_{n-L} and the left MFG activity compared to the control condition.

The second-level whole-brain regression analysis across the participants using the contrast image of parametric modulation effects of ΔS_n and h_n showed that h_n (harm avoidance related to noncombatants) was associated with increasing activity in the left precentral cortex in the group conflict condition (-36/-30/56) and in the bilateral precentral cortices in the individual conflict condition (-36/-14/64 and 60/8/32, a voxel-level threshold $P < 0.001$, cluster-level threshold $P < 0.05$, FWE corrected, Figure 5A), but these activations did not differ significantly between the group and individual conflict conditions. Because there was no significant difference in h_c (harm preference to combatants) and h_{c-L} between the group and individual conflict conditions, we pooled fMRI data in the two conditions to examine neural correlates of h_c and h_{c-L} , respectively. The results of the second-level whole-brain regression analyses showed that a larger h_c was associated with increased activity in the left temporoparietal junction (-50/-40/26), left prefrontal cortex (-28/36/32), midcingulate (14/2/48), and bilateral occipital cortex (-34/-76/26 and 34/-74/16), whereas increased activity linked to h_{c-L} was observed only in the right occipital cortex (24/-54/-10,

voxel-level threshold $P < 0.001$, cluster-level threshold $P < 0.05$, FWE corrected, Figure 5B). Apparently, these activations do not overlap with the left MFG which seems to be specifically associated with decreased harm avoidance related to outgroup noncombatants during intergroup conflict. Taken together, these neuroimaging results uncovered brain activity that was specifically associated with collateral damage during intergroup conflict (i.e., the left MFG activity). Our neuroimaging results also revealed brain activities underlying punishment decision making that, however, were similarly observed in the group and individuals conflict conditions.

Discussion

The present study combined functional brain imaging and statistical modeling of punishment decision-making to examine the cognitive and neural bases of decision-making resulting in civilian casualties during intergroup conflict. Unlike previous research of intergroup conflict that measured punishment decisions upon outgroup members as a whole¹²⁻¹⁶, our design distinguished between outgroup combatants and noncombatants in punishment decision making. By comparing individuals' behavioral performance in group-conflict and individual-conflict conditions, we highlighted a psychological construct, i.e., decreased harm avoidance, which was specifically related to selection of punishment decisions that led to civilian casualties when intergroup conflict escalated. Our fMRI results further uncovered the neural correlate of this psychological construct in the left MFG.

Our statistical modeling revealed a significantly negative value of the parameter h_n as an index of harm avoidance during the selection of electric shocks that were supposed to be possibly applied to noncombatants. This finding indicates that people tended to give weaker electric shocks to or to avoid harming others who were not directly engaged in aggressive actions during conflict. This was true during conflicts

between two groups or between two individuals and is consistent with moral imperatives and international laws in human societies^{29,30}. Nevertheless, the statistical modeling revealed a positive value of the parameter h_{n-L} in the situation of intergroup conflict across three experiments. The positive h_{n-L} suggests that harm avoidance that inhibits harm to noncombatants during intergroup conflict tends to decrease when intergroup conflict escalates.

It has been speculated that, as conflict escalates, decision makers are less likely to differentiate between combatants and noncombatants due to depleted cognitive resources and have to deal with the incidental cognitive dissonance¹⁷. These hypothesized cognitive mechanisms underlying civilian casualties, however, have not been tested empirically. Our empirical findings uncover a psychological construct, i.e., harm avoidance, which varies across levels of intergroup conflict in a fashion similar to actual civilian casualties. Furthermore, we showed that decreased harm avoidance during conflict escalation was evident in individuals facing conflict between an ingroup member and an outgroup member but not in those facing conflicts between two individuals who had no ingroup/outgroup relationships with the decision maker. This provides further evidence that, rather than reflecting changes of cognitive processes that are general to any conflict, the decreased harm avoidance specifically characterizes punishment decision-making that leads to civilian casualties during escalation of intergroup conflict. In addition, our results demonstrated a positive value of h_{n-L} in both a laboratory setting in China and in a natural conflict context in Israel. Thus a similar psychological construct may underlie civilian casualties during escalation of intergroup conflict in both a laboratory setting and a real-world context.

Our statistical modeling also revealed a significantly positive value of the parameter h_c as an index of harm preference for outgroup combatants during selection

of electric shocks in a laboratory setting but not in a real social context of intergroup conflict. However, the statistical modeling showed a significant positive value of the parameter h_{c-L} across the three experiments in both the group and individual conflict conditions. The positive h_{c-L} suggests that our participants tended to choose stronger electric shocks to punish an individual if this person had proposed a more unfair money distribution plan. The increased harm preference as a function of conflict escalation might have different psychological causes in the group and individual conflict conditions. In the group conflict condition, when facing a more unfair money distribution plan, noncombatants might choose stronger electric shocks to protect ingroup members' benefit, similar to previous findings of ingroup favoritism in decision-making³¹⁻³³. In the individual conflict condition, however, unfair money distribution between two individuals did not directly affect a bystander's payoff and punishment decision-making was possibly made to protect social norms, similar to third-party punishment³⁴. Our data analyses did not show consistent evidence for differences in parameter estimates of harm preference and harm avoidance between combatants and noncombatants as decision makers. It is likely that the motives to help ingroup and to protect social norms dominated punishment decision-making whereas one's own payment was much less taken into consideration in our participants. Future research requires a new behavioral paradigm that helps to reveal how combatants and noncombatants make punishment decisions differently.

We integrated fMRI and statistical modeling of punishment decision-making to examine neural correlates of harm avoidance and harm preference during intergroup conflict. Our fMRI results showed that h_c and h_n were linked to distinct neural circuits, i.e., h_n was associated with increasing activity in the precentral cortex whereas h_c was associated with increased activity in the left temporoparietal junction, left prefrontal

cortex, and midcingulate. These results provide further neuroimaging evidence that harm avoidance and harm preference represent two distinct psychological processes supported by different neural underpinnings. Most importantly, we showed that the decreased left MFG activity predicted a larger h_{n-L} and this effect was specific to the participants engaged in intergroup conflict. This finding suggests a key functional role of the left MFG in controlling harm avoidance related to civilian casualties during intergroup conflict. Specifically, inhibition of the left MFG might lead to failure to avoid harm to noncombatants when making punishment decisions during high intergroup conflict. The functional role of the left MFG during intergroup conflict is consistent with previous fMRI findings. For example, the left MFG activity increases with the complexity of different cognitive tasks and lesions of this brain region result in performance deficits in different cognitive tasks³⁵. The lateral frontal cortex also represents the value of employing cognitive control and plays an important role in selection/control of a specific punishment decision^{8,11} and in value-based learning and decision-making³⁶. These findings agree upon a general function of the left MFG in executive control of decision-making and behavioral responses that is independent of task domains.

Our fMRI results revealed increasing activation in the precentral cortex as neural correlates of harm avoidance related to noncombatants as indexed by h_n . This is consistent with the functional role of this brain region in general behavioral inhibition³⁷. Harm preference toward combatants as indexed by h_c was associated with increased activity in the left temporoparietal junction, left prefrontal cortex, and midcingulate. These brain regions are involved in punishment decision-making to detect aversive experiences and to infer others' mental states⁷⁻¹⁰. Similar mental processes may support harm preference regardless of when conflicts occur between

513 two groups or between two individuals. However, although the motives to punish
514 outgroup combatants and to avoid harming to noncombatants may coexist, inhibition
515 of the left MFG activity is specifically related to the decreased harm avoidance as a
516 psychological construct underlying civilian casualties when intergroup conflict
517 escalates.

518 We also considered revenge as a potential motivator of punishment in this study.
519 Revenge is usually manifested in taking actions of harming someone in retaliation for
520 an injury^{38,39}. Our recent fMRI research showed that the medial prefrontal activity in
521 response to ingroup pain caused by outgroup members predicted the propensity to
522 give painful electric shocks to outgroup members⁴⁰. In the current study the
523 participants made choices by considering shocks to both the outgroup combatant and
524 noncombatant. While our statistical modeling of punishment decision-making
525 suggested harm preference toward the outgroup combatant, the results also suggest
526 harm avoidance toward the outgroup noncombatant. It is thus unlikely that the
527 participants made their choices only by considering retaliation upon the outgroup
528 combatant. Moreover, revenge cannot explain why harm avoidance related to
529 outgroup noncombatants decreased when intergroup conflict escalated since the
530 outgroup noncombatants did not bring any harm to the decision maker in any case.
531 We also assessed the possibility that the participants made their choices by
532 considering retaliation upon the outgroup combatant and noncombatant as a whole (or
533 group revenge) by constructing a different model (i.e., Model 3 reported in
534 Supplementary Fig. 1 and Supplementary Table 1). In this model the sum of shock
535 intensities to the combatant and the noncombatant (i.e., group revenge) was used in a
536 regression model to predict the participants choices. The results suggest that this
537 'group revenge' model is not as good as the model reported in the main text in

538 predicting the participants' decisions. It appears that revenge cannot fully explain
539 punishment decision making in a context when a punishment decision is related to
540 both outgroup combatants and noncombatants during intergroup conflict.

541 Previous research on mechanisms of intergroup conflict usually took all outgroup
542 members as one unit. Perceiving the outgroup as a competitor for limited resources⁴¹
543 or a threat to the ingroup⁴² may trigger intergroup conflict. Dehumanization of
544 outgroup members⁴³ may further drive discrimination and conflict. Humans may
545 evolve ingroup favoritism in implicit attitudes⁴⁴ and empathy⁴⁵⁻⁵⁰ and aggressiveness
546 against outgroups as genetic or cultural traits⁵¹. While these traits help individuals to
547 survive, humans as a rational species are confronted with the contradiction between
548 harming outgroup combatants as enemies and protecting outgroup noncombatants as
549 innocents during intergroup conflict. Our findings provide a neurocognitive account
550 of decision-making that underlies civilian casualties during intergroup conflict.

551 While our findings contribute to the understanding of neurocognitive
552 underpinnings of decisions involving harm to noncombatants in real-world situations,
553 several limitations of our findings should be acknowledged. First, civilian casualties
554 in real-world situations usually arise from armed attacks toward outgroup combatants
555 and such attacks cost lives of and bring long-term physical/mental harms to outgroup
556 combatants and noncombatants. The harm that participants intended to apply to the
557 outgroup in our paradigm, however, was much weaker than those in real-world
558 situations due to ethical requirements. Nevertheless, despite the differences in degree
559 of harm, our paradigm captures the key features of collateral damages in real-world
560 situations (i.e., increasing along conflict escalation). Besides, previous research has
561 shown that conflict expenditures of groups can be substantially larger than those of
562 conflict between individuals and allowing group members to punish each other results

563 in even larger conflict expenditures⁵². In our paradigm, however, conflict
 564 expenditures (distributed money between the two combatants) were controlled in the
 565 group and individual conflict conditions as an irrelevant variable. This leaves an open
 566 question for future research, i.e., how cognitive and neural mechanisms underlying
 567 decision-making resulting in civilian casualties during intergroup conflict vary along
 568 escalated conflict expenditures.

569 Second, civilian casualties in real-world situations usually take place during
 570 intergroup conflict with a history. People make decisions to attack the outgroup by
 571 considering not only the consequence of the upcoming attack on the outgroup but also
 572 potential feedback/revenge from the outgroup that may occur to the ingroup thereafter.
 573 Recent theories of motivations indeed emphasize key functional roles of updated
 574 beliefs, beliefs of others, and plans of action in decision making⁵³. Our paradigm
 575 modeled an initial decision making in response to intergroup conflict without
 576 providing the participants with the information of potential feedback/revenge from the
 577 outgroup. While there are indeed similar situations in real-world intergroup conflict,
 578 future research on civilian casualties during intergroup conflict should consider
 579 decision making in complicated situations when decision makers receive feedback or
 580 potential revenge from the outgroup after each decision making.

581 Third, our work recruited three cultural samples (i.e., Chinese, Jewish and
 582 Palestinian participants) who showed similar decreased harm avoidance related to
 583 outgroup noncombatants as conflicts escalated but difference in harm preference for
 584 outgroup combatants. It is important to test other cultural samples to verify whether
 585 collateral damage arising from intergroup conflict has similar cognitive and neural
 586 underpinnings in other sociocultural contexts. This is necessary given broad

sociocultural differences in⁵⁴⁻⁵⁶ and potential interactions of biological and cultural factors⁵⁷⁻⁵⁹ on cognition and brain underpinnings of human behavior.

Finally, even though we showed evidence for an association between decreased harm avoidance and inhibition of the left MFG activity during escalation of intergroup conflict, it remains unclear whether the left MFG activity is necessary for controlling civilian casualties during intergroup conflict. This can be clarified in future research by integrating the paradigm developed in our work with the technique for interrupting brain activity such as transcranial magnetic stimulation⁶⁰.

In conclusion, by integrating brain imaging and statistical modeling of punishment decision-making during intergroup conflict in three cultural samples, the current work disentangled harm preference and harm avoidance as two psychological constructs underlying punishment decision making related to outgroup combatants versus outgroup noncombatants. In addition, our fMRI results suggest different neural circuits supporting these two psychological constructs. Most important, we showed evidence that avoidance of harming outgroup noncombatants decreased when intergroup conflict escalated and that the decreased harm avoidance was associated with inhibition of the left middle frontal activity during selection of punishment decisions. These findings shed light on the neurocognitive underpinnings of decision-making involving civilian casualties during intergroup conflict.

Methods

Participants

Sample sizes were pre-determined using G*Power⁶¹. In Experiment 1, we aimed to conducted one-sample t-tests for each parameter in the logistic regression model and two-sample t-tests to compared parameters between the group and individual conflict conditions. Based on G*Power estimation, a sample size of 54 participants

was required to obtain a medium effect size of 0.5 with an error probability of 0.05 and power of 0.95 in two-tailed one-sample t-tests. For the two-tailed two-sample t-tests, a sample size of 105 participants for both the group and individual conflict conditions were required to obtain a medium effect size of 0.5 with an error probability of 0.05 and power of 0.95. The sample size of the Jewish and Palestinian groups in Experiment 2 was determined to be similar with those in the group and individual conflict in Experiment 1. The sample size in Experiment 3 was smaller than that in Experiment 1 by considering the cost of fMRI scanning. Based on G*Power estimation, a sample size of 34 participants was required to obtain a medium effect size of 0.5 with an error probability of 0.05 and power of 0.80 in two-tailed one-sample t-tests. For the two-tailed two-sample t-tests, a sample size of 64 participants for both the group and individual conflict conditions were required to obtain a medium effect size of 0.5 with an error probability of 0.05 and power of 0.80.

In Experiment 1 we recruited 120 Chinese university students (half male, mean age $\pm$ s.d. = 22.85 ± 2.46 years) in the group conflict condition and 120 Chinese university students (half male, 21.68 ± 2.43 years) in the individual conflict condition. In Experiment 2 we recruited 106 Israeli-Jewish university students (46 males, 24.97 ± 3.78 years). To assure that only Arab participants who clearly identify themselves as Palestinian participated in our study, we originally recruited 152 Arab-Palestinian participants who completed a scale in which they rated their ethnic identification from 0% Palestinian to 100% Palestinian. Only participants who rated themselves as 100% Palestinian were further invited to participate in Experiment 2 (N=106, 46 males, 21.82 ± 2.23 years). During recruitment we did not reveal to participants that our study examined questions related to intergroup conflict so as to minimize any effects of participants' prior attitudes about the Israeli-Palestinian conflict on behavioral

637 measurements. All participants in Experiment 2 (i.e., both Jewish and Palestinian
638 participants) were Israeli citizens.

639 In Experiment 3 we recruited 64 Chinese university students (half males, $23.13 \pm$
640 1.82 years) in the group conflict condition. Four participants were excluded due to
641 excessive head motion during scanning, leaving 29 combatants (14 males, $22.93 \pm$
642 1.56 years) and 31 noncombatants (16 males, 23.19 ± 2.06 years) for analyses of data
643 in the group conflict condition. An independent group of 64 Chinese university
644 students (half males, 22.48 ± 2.42 years) were recruited in the individual conflict
645 condition. Five participants were excluded due to excessive head motion during
646 scanning, leaving 29 combatants (14 males, 22.76 ± 2.84 years) and 30
647 noncombatants (15 males, 22.23 ± 2.08 years) for analyses of data in the individual
648 conflict condition. All participants had normal or corrected-to-normal vision and
649 reported no neurological or psychiatric history. Informed consent was obtained prior
650 to the experiment. All participants were paid for their participation. Experiments 1
651 and 3 were approved by the local ethics committee at the School of Psychological and
652 Cognitive Sciences (Reference number: 2105-12-04), Peking University. Experiment
653 2 was approved by the local ethics committee at the Rambam Medical Center, Haifa
654 University (Reference number: 044/17).

655 ***Experiment Procedure***

656 On each testing day, 4 participants of the same gender were recruited for a testing
657 group who had not known each other before their participation. The experimental
658 procedures consisted of four phases.

659 *Phase 1: Group formation*

660 In the group conflict condition, participants were randomly assigned to two
661 groups of 2 persons by choosing a card. Participants from each group were asked to

662 wear T-shirts of the same color (red or blue). Participant introduced their own names,
 663 nicknames, majors, and hobbies to get familiar with each other. They then started to
 664 play the Saboteur card game ([http://www.annarbor.com/entertainment/saboteur-card-](http://www.annarbor.com/entertainment/saboteur-card-game-review/)
 665 [game-review/](http://www.annarbor.com/entertainment/saboteur-card-game-review/)). During this game ingroup members played cards to build a tunnel to a
 666 gold location or to block the tunnel to prevent outgroup members from reaching the
 667 location. This game required ingroup members to cooperate with each other but to
 668 interrupt outgroup members so as to reach the gold location before the outgroup. The
 669 intergroup relationship was built by playing this game for 50 minutes. To check the
 670 effectiveness of the group manipulation, after the game, participants were asked to
 671 complete a modified version of the Inclusion of Other in the Self Scale¹⁹ to assess
 672 their feelings of closeness between oneself and ingroup members and between oneself
 673 and outgroup members. The individual conflict condition was the same as the group
 674 conflict condition except that participants were not divided into groups and they
 675 played the Saboteur card game against the other three participants. After the game,
 676 participants were asked to assess their feelings of closeness between oneself and the
 677 other three participants with a modified version of the Inclusion of Other in the Self
 678 Scale¹⁹. Phase 1 lasted for 60 minutes.

679 *Phase 2: Inducing conflict*

680 For the group conflict condition one participant from each group was randomly
 681 assigned as a combatant who allotted 100 RMB (in Experiment 1; 100 Shekels in
 682 Experiment 2; and 200 RMB in Experiment 3) between the two combatants by giving
 683 different weights to a number of plans. The other participant from each group waited,
 684 as a noncombatant, while combatants made their decisions (Phase 2 in Figure 1). This
 685 procedure was the same for the individual conflict condition except that 2 participants
 686 were randomly assigned as combatants and 2 as noncombatants.

There were nine different money distribution plans, which included Self(55)/Rival(45), Self(60)/Rival(40), Self(65)/Rival(35), Self(70)/Rival(30), Self(75)/Rival(25), Self(80)/Rival(20), Self(85)/Rival(15), Self(90)/Rival(10), Self(95)/Rival(5) (The ratios of money allocation plans were the same across three experiments). Each combatant was given 108 cards and had to allot these cards to the nine plans (there must be at least 1 card for each plan). Both combatants and noncombatants were informed that a computer would randomly select one out of 108 cards, which corresponded to one money allocation plan from each combatant. That is, a plan with more cards allocated would be more likely to be selected by the computer. After the computer selected a plan from each combatant, the money would be finally divided between the two combatants according to the plan with a larger difference between Self and Rival. If two plans from the two combatants were the same, the computer would randomly select one plan to decide the final payment. There were typically two strategies to allocate the cards. Allotting more cards to Self(55)/Rival(45) plan represents a fairness strategy whereas more cards to Self(95)/Rival(5) represents a selfish strategy. All plans allot more money to self and the rules benefited the combatant who employed a more selfish strategy. These designs induced conflict between two combatants. The two combatants finished the money allocation task in two private rooms so that the other participants would not know their decisions. All the participants were informed that each noncombatant would receive 100 RMB for their participation. Each combatant, however, would received a basic payment of 50 RMB and share a 100 RMB bonus with the other combatant based on the final money allocation plan. On average each combatant would received 100 RMB for his/her participation. Phase 2 lasted for 15 minutes.

Phase 3: Pain threshold measurement

712 To help participants to have an intuitive feeling of different levels of electric
713 shocks, each participant underwent a pain titration procedure with a Digitimer DS7
714 electric stimulator individually (Phase 3 in Figure 1). Following a brief overview of
715 the equipment and titration process, two electrodes were attached to the back of the
716 left hand. The titration procedure began with a low-voltage electric shock (0.5 mA). A
717 participant was then asked to rate his/her subjective feelings of pain on an 11-point
718 scale (0 = no pain, 10 = intolerable). The initial rating was followed by a series of
719 shocks increasing with a step of 0.5 mA. Subjective ratings of pain were collected
720 after each shock until a rating of 10 was reached. The same procedure was repeated
721 twice for each participant. Phase 3 lasted for 40 minutes.

722 *Phase 4: Punishment decision making*

723 In this phase, the participants were informed that the 108 money allocation plans
724 from one combatant would be presented one by one. There were 108 trials with equal
725 numbers of the nine money distribution plans which were actually created by the
726 computer program and randomly presented to participants. On each trial, a money
727 distribution plan was first presented for 4 s on a screen. Thereafter, two options of
728 electric shocks were shown on the left and right side of the screen, respectively (Phase
729 4 in Figure 1). Each option (either the left or right) consisted of two vertical bars that
730 indicated intensities of two electric shocks. The shock intensities were indexed by
731 rating scores of subjective painful feelings (from 1 = slightly pain to 9 = extremely
732 painful) that each participant experienced in Phase 3. In the group conflict condition,
733 one shock would be applied to the outgroup combatant as punishment and the other to
734 the outgroup noncombatant as a “collateral damage”. In the individual conflict
735 condition, one shock would be applied to the combatant as punishment and the other
736 to a bystander. The intensities of two electric shocks in one option for punishing the

737 combatant and harming the noncombatant, the intensities of two electric shocks in the
 738 left and right options, and the intensities of electric shocks varied independently
 739 across trials. Non-independent variations between the shock intensities might lead to
 740 high collinearity between two variables and result in unstable estimation of the
 741 regression parameters. The participants were asked to make a choice between the left
 742 and right options by pressing one of two buttons using the right index and middle
 743 fingers. A fixation of 2 s was presented after a button press when the next trial started.
 744 They were also informed that electric shocks in one of their decisions would be
 745 randomly chosen to be applied to the target combatant and noncombatant after the
 746 experiment, which, though were not really applied to the participants who were
 747 debriefed. All the four participants in each session knew that the two outgroup
 748 members would make decisions that might lead to them receiving electric shocks.
 749 C1/C2 and N1/N2 performed the decision making task simultaneously in private
 750 rooms during Phase 4.

751 Task instructions in Phases 2-4 were given at the beginning of Phase 2 to all the
 752 participants simultaneously. The combatants and the noncombatants knew each
 753 other's tasks. That means, N1/N2 knew that they would be subject to electric shocks
 754 when the other participants made decisions to punish the combatants. All the
 755 participants in Phase 4 knew that other participants were doing the same decision-
 756 making tasks.

757 According to the unfairness of the money allocation plan, all 108 trials were
 758 coded as being in one of three conflict levels. In low, middle and high conflict levels,
 759 the target combatant would keep 55 – 65 RMB, 70 – 80 RMB and 85 – 95 RMB for
 760 oneself, respectively. In each conflict level, there were 36 trials with 36 combinations
 761 of two options of electric shocks. Specifically, nine levels of electric shocks were

762 divided into three pain levels (e.g., low, middle and high pain). These three pain
763 levels towards combatant and noncombatant would create 9 different combinations of
764 options. In each trial, two out of nine options were selected and presented. This
765 created 36 different combinations of two options. To this end, the difference of
766 electric shocks between two options will vary in each trial and vary independently
767 between combatant and noncombatant, which were entered into the following logistic
768 regression modelling. Finally, the experiment design and purpose were explained in
769 debriefing and all participants received 100 RMB for participation. Phase 4 lasted for
770 35 minutes.

771 The experimental procedures of Experiment 2 in Jewish and Palestinian
772 participants were the same as those in Study 1 except the following. There was no
773 group formation phase at the start. Instead, two Jewish participants and two
774 Palestinians (all were of the same gender and were strangers to each other) were
775 recruited each time and they introduced their nationality to each other. Furthermore,
776 in order to highlight the nationality of each participant, participants of the same
777 nationality performed a priming task together which contained images of items that
778 relate to one's nationality. Such as the symbols of each group's nation (e.g. flags,
779 emblems, anthems), which was found to increase the feeling of identification with
780 one's own nation²⁷, as well as photos of famous people (e.g., Yasser Arafat vs.
781 Shimon Peres) and holy places (e.g., Al-Aqsa Mosque vs. Hakotel). The participants
782 from each group were asked to jointly rate the feelings of belonging and excitement
783 that each image evoked (1 = very little, 9 = very much).

784 The experimental procedures of Study 3 in the neuroimaging lab were the same
785 as those in Study 1 except the following. The Saboteur card game lasted 90 minutes
786 during Phase 1 to strengthen the intergroup relationship since the whole experiment

lasted for a longer time in the Experiment 3. Each participant was scanned using fMRI while making punishment choices during Phase 4. There were 108 trials for each participant which were equally distributed into six scans. In each scan, there were 18 trials which included 6 low conflict trials, 6 middle conflict trials, and 6 high conflict trials. On each trial, a jitter fixation was first presented for 2, 4, 6, 8, or 10 s. A money distribution option given by the target combatant was then presented for 2 s. A following jitter fixation was presented for 2, 4, 6, 8, or 10 s. Thereafter, two options of electric shocks towards the target combatant and noncombatant were presented and participants were asked to make a choice between the options within 5 seconds. Each scan started with a 12-s fixation to get a baseline for blood-oxygen-level-dependent (BOLD) signals. The last trial of each scan was followed by a 12-s fixation.

Model comparisons

We tested additional two models (Models 2 and 3) and compared these models with the one we reported in the main text (Model 1). Model 2 only included intercept in the logistic regression model to regress the inherent tendency to choose left or right option of punishment decisions for each participant ($\Delta V = \text{intercept}$; the participants tended to choose one (the left or right) option regardless of shock intensities and monetary distribution plans). This is a baseline control model compared with other models. Model 3 hypothesized that outgroup combatants and noncombatants are treated in the same way during intergroup conflict. We included two independent variables in the regression model to account for the probability that participants would select the left (p) or right options ($1-p$) in each trial. These included the difference in the sum of shock intensities to the combatant and the noncombatant between left and right options ($\Delta S = \Delta S_c + \Delta S_n$) and the interaction between ΔS and level of conflict (L). The logistic regression model was defined as following:

$$\text{logit} = \ln \frac{p}{1-p} = \Delta V \quad (1)$$

$$\Delta V = \mathbf{h} \times \Delta S + \mathbf{h}_L \times \Delta S \times L \quad (2)$$

$$p = \frac{1}{1+e^{-\Delta V}} \quad (3)$$

A larger p is associated with higher possibility of selection of the left option.

A positive $\mathbf{h}$ value indicates harm preference towards both combatants and noncombatants — a propensity to select an electric shock option with a greater sum of intensities from two options. A negative $\mathbf{h}$ value indicates harm avoidance towards both combatants and noncombatants — a propensity to select an electric shock option with a smaller sum of intensities from two options. $\mathbf{h}_L$ manifest how $\mathbf{h}$ varies along escalated levels of conflict (positive and negative $\mathbf{h}_L$ value indicates increase or decrease of $\mathbf{h}$, respectively).

We compared these models using Bayesian Information Criterion (BIC), prediction accuracy, and area under the curve (AUC). BIC penalizes models with a greater number of parameters, and the model with the lowest group BIC is the preferred model. We computed BIC scores for each model fit in each individual participant and summed the BIC scores across subjects to obtain a group BIC score. We also compared predicted choice probability of each model with the categorization threshold of 0.5 and participants' choices in all trials to calculate the prediction accuracy of each model. If the predicted choice probability is larger than 0.5 in one trial, the model will predict that participant choose left option in that trial. Otherwise, participant will choose right option. We calculated mean prediction accuracy across all participants for each model. The model with the highest prediction accuracy is the preferred model. AUC was calculated as the area under the Receiver Operator Characteristic curve, which provides a simple way to summarize all the confusion matrices by setting different categorization thresholds. The model with the largest

837 AUC is the preferred model.

838 *fMRI data acquisition and preprocessing*

839 Brain images were acquired using a 3.0T GE Signa MR750 scanner (GE
840 Healthcare; Waukesha, WI) with a standard 8 channel head coil. Functional images
841 were acquired by using T2-weighted, gradient-echo, echo-planar imaging (EPI)
842 sequences sensitive to BOLD signals ($64 \times 64 \times 32$ matrix with $3.75 \times 3.75 \times 5 \text{ mm}^3$
843 spatial resolution, repetition time = 2000 ms, echo time = 30 ms, flip angle = 90° ,
844 field of view = $24 \times 24 \text{ cm}^2$). A high-resolution T1-weighted structural image ($512 \times$
845 512×180 matrix with a spatial resolution of $0.47 \times 0.47 \times 1.0 \text{ mm}^3$, repetition time =
846 8.204ms, echo time = 3.22ms, flip angle = 12°) was acquired after the first three
847 functional scans. Padded clamps were used to minimize head motion and earplugs
848 were used to attenuate scanner noise. The stimuli were projected onto a screen at the
849 head of the magnet bore using Presentation. Participants viewed the screen through a
850 mirror attached to the head coil.

851 Functional images were preprocessed using SPM12 software (the Wellcome
852 Trust Centre for Neuroimaging, London, UK). Head movements were corrected
853 within each run and six movement parameters (translation; x, y, z and rotation; pitch,
854 roll, yaw) were extracted for further analysis in the statistical model. The anatomical
855 image was coregistered with the mean realigned functional image. The functional
856 images were resampled to $3 \times 3 \times 3 \text{ mm}^3$ voxels, normalized to Montreal
857 Neurological Institute (MNI) template space and then spatially smoothed using an
858 isotropic of 8 mm full-width half-maximum (FWHM) Gaussian kernel.

859 *Whole brain parametric analyses*

860 The first level general linear model (GLM) was constructed to identify brain
861 regions responding parametrically at choice onset to the difference of electric shocks

between chosen and unchosen options in each participant. The GLM contained three sets of event regressors describing: (1) the onsets of the money allocation plans, the corresponding duration of 1 TR, and a parametric modulator of different money allocation plans representing three different levels of conflict (L); (2) the onsets of two options of shocks, the corresponding duration of participants' reaction time on each trial, and four parametric modulators: (i) the difference of electric shocks between chosen and unchosen options towards combatant (ΔS_c); (ii) the difference of electric shocks between chosen and unchosen options towards noncombatant (ΔS_n); (iii) the interaction between conflict level and difference of electric shocks towards combatant ($\Delta S_c * L$); (iv) the interaction between conflict level and difference of electric shocks towards noncombatant ($\Delta S_n * L$); (3) the onsets of the button press for punishment decision making and a duration of zero. Besides, the GLM also included the realignment parameters to account for any residual movement-related effect.

Whole brain regression analyses

We conducted four second-level whole-brain one-sample t-tests with parameters estimated by the logistic regression modeling as covariates. These tests were constructed to examine brain regions associated with individual differences in harm tendency (h_c), harm avoidance (h_n), variation of harm tendency with conflict levels (h_{c-L}), and variation of harm avoidance with conflict levels (h_{n-L}). The first regression analysis included the contrast images of parametric modulation effects of ΔS_c with h_c as the covariate. The second regression analysis included the contrast images of parametric modulation effects of ΔS_n with h_n as the covariate. The third regression analysis included the contrast images of parametric modulation effects of $\Delta S_c * L$ with h_{c-L} as the covariate. The fourth regression analysis included the contrast images of parametric modulation effects of $\Delta S_n * L$ with h_{n-L} as the covariate. The brain regions

887 associated with individual differences in parameters were estimated by calculating
888 one-sample t-test using the contrasts of [0 1] or [0 -1] to assess the positive or
889 negative associations. Brain activations were defined by combining a whole-brain
890 voxel-level threshold of $P < 0.001$ and a cluster-level threshold at $P < 0.05$ FWE
891 corrected.

892 ***Moderation analyses***

893 To test whether the associations between h_{n-L} and the left MFG activity were
894 significantly different between the group conflict and individual conflict conditions,
895 we conducted a second-level whole-brain two-sample t-test with h_{n-L} in these two
896 conditions as covariates. The design matrix for this t-test included the contrast images
897 of the parametric modulation effect of $\Delta S_n * L$ in each condition and h_{n-L} in the two
898 conditions as the first and second covariates. The moderation effect was estimated by
899 calculating two-sample t-test using the contrasts of [0 0 -1 1] to assess the enhanced
900 coupling between h_{n-L} and brain activity that was stronger in the group conflict
901 condition than in the individual conflict condition. Small volume corrections within
902 the mask of the left MFG defined using the Brainnetome Atlas (Region 22:
903 MFG_L_7_4_9 (46v)²⁸ were applied to search for significant results with the
904 threshold of $p < 0.05$, FWE corrected for each voxel.

905 ***Bayesian factor analyses***

906 Bayesian factor analyses were used to compare the likelihood of the data fitting
907 under the null hypothesis with the likelihood of fitting under the alternative
908 hypothesis²³. Compared with widely used p-values, the Bayes factor allows the
909 researcher to make a statement about the alternative hypothesis, rather than just the
910 null hypothesis. It also provides a clearer estimate of the amount of evidence present
911 in the data. In the current paper, Bayes factor was calculated using the r toolbox

“Bayesian factor” (<https://richarddmoney.github.io/BayesFactor/>). BF_{01} represents the ratio that contrasts the likelihood of the data fitting under the null hypothesis with the likelihood of fitting under the alternative hypothesis. On the contrary, BF_{10} represents the ratio that contrasts the likelihood of the data fitting under the alternative hypothesis with the likelihood of fitting under the null hypothesis. Therefore, as BF_{01} increases, there is more evidence in support of the null hypothesis. While as BF_{10} increases, there is more evidence in support of the alternative hypothesis. A Bayesian factor falling between 1 to 3 is regarded as weak evidence, between 3 to 20 is regarded as positive evidence, between 20 to 150 is regarded as strong evidence and larger than 150 is regarded as very strong evidence in support of null (BF_{01}) or alternative hypothesis (BF_{10})⁶².

References

1. Reynolds, J. D. Collateral damage on the 21st century battlefield: Enemy exploitation of the law of armed conflict, and the struggle for a moral high ground. *Air Force Law Rev.* **56**, 1,101 (2005).
2. *The 1977 Additional Protocol I of the Fourth Geneva Convention of 1949.*
3. Eckhardt, W. Civilian deaths in wartime. *Bull. Peace Proposals* **20**, 89-98 (1989).
4. Roberts, A. Lives and statistics: Are 90% of war victims civilians? *Survival* **52**, 115-136 (2010).
5. Voon, T. Pointing the finger: Civilian casualties of NATO bombing in the Kosovo conflict. *Am. U Int'l. L. Rev.* **16**, 1083-1113 (2001).
6. *United Nations Office for the Coordination of Humanitarian Affairs–Occupied Palestinian Territory (UNOCHA OPT).* (2015). *Fragmented Lives: Humanitarian Overview* 2014.
7. Seymour, B., Singer, T. & Dolan, R. The neurobiology of punishment. *Nat. Rev.*

- 937 *Neurosci.* **8**, 300-311 (2007).
- 938 8. Krueger, F., & Hoffman, M. The emerging neuroscience of third-party punishment.
- 939 *Trends Neurosci.* **39**, 499-501 (2016).
- 940 9. Patil, I., Calò, M., Fornasier, F., Cushman, F. & Silani, G. The behavioral and
- 941 neural basis of empathic blame. *Sci. Rep.* **7**, 1-14 (2017).
- 942 10. Ginther, M. R., Bonnie, R. J., Hoffman, M. B., Shen, F. X., Simons, K. W., Jones,
- 943 O. D. & Marois, R. Parsing the behavioral and brain mechanisms of third-party
- 944 punishment. *J. Neurosci.* **36**, 9420-9434 (2016).
- 945 11. Buckholz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D.
- 946 & Marois, R. The neural correlates of third-party punishment. *Neuron* **60**, 930-940
- 947 (2008).
- 948 12. Halevy, N., Bornstein, G. & Sagiv, L. “In-group love” and “out-group hate” as
- 949 motives for individual participation in intergroup conflict: A new game paradigm.
- 950 *Psychol. Sci.* **19**, 405-411 (2008).
- 951 13. De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas,
- 952 M., ... & Feith, S. W. The neuropeptide oxytocin regulates parochial altruism in
- 953 intergroup conflict among humans. *Science* **328**, 1408-1411 (2010).
- 954 14. De Dreu, C. K., Gross, J., Méder, Z., Giffin, M., Prochazkova, E., Kriek, J. &
- 955 Columbus, S. In-group defense, out-group aggression, and coordination failures in
- 956 intergroup conflict. *Proc. Natl. Acad. Sci. USA.* **113**, 10524-10529 (2016).
- 957 15. Sääksvuori, L., Mappes, T. & Puurtinen, M. Costly punishment prevails in
- 958 intergroup conflict. *Proc. Biol. Sci. B.* **278**, 3428-3436 (2011).
- 959 16. Zhang, H., Gross, J., De Dreu, C. & Ma, Y. Oxytocin promotes coordinated out-
- 960 group attack during intergroup conflict in humans. *Elife* **8**, e40698 (2019).
- 961 17. Wolfe, R.J. & Darley, J.M. Protracted asymmetrical conflict erodes standards for

- 962 avoiding civilian casualties, *Peace Conflict: J. Peace Psychol.* **11**, 55–61 (2005).
- 963 18. Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. *J.*
964 *Econ. Perspect.* **14**, 159-181 (2000).
- 965 19. Aron, A., Aron, E. N. & Smollan, D. Inclusion of other in the self scale and the
966 structure of interpersonal closeness. *J. Pers. Soc. Psychol.* **63**, 596-612 (1992).
- 967 20. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression
968 analysis and reporting. *J. Edu. Res.* **96**, 3-14 (2002).
- 969 21. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38
970 (1993).
- 971 22. Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R. & Phelps, E. A. The price of
972 racial bias: Intergroup negotiations in the ultimatum game. *Psychol. Sci.* **24**, 2498-
973 2504 (2013).
- 974 23. Jarosz, A. F. & Wiley, J. What are the odds? A practical guide to computing and
975 reporting Bayes factors. *J. Prob. Solving* **7**, 2-9 (2014).
- 976 24. Rouhana, N. N. & Bar-Tal, D. Psychological dynamics of intractable
977 ethnonational conflicts: The Israeli–Palestinian case. *Ame. Psychol.* **53**, 761-770.
978 (1998).
- 979 25. Bekerman, Z. & Horenczyk, G. Arab - Jewish bilingual coeducation in Israel: A
980 long - term approach to intergroup conflict resolution. *J. Soc. Issues* **60**, 389-404
981 (2004).
- 982 26. Shamay-Tsoory, S. G., Abu-Akel, A., Palgi, S., Sulieman, R., Fischer-Shofty, M.,
983 Levkovitz, Y. & Decety, J. Giving peace a chance: oxytocin increases empathy to
984 pain in the context of the Israeli–Palestinian conflict. *Psychoneuroendocrinology* **38**,
985 3139-3144 (2013).
- 986 27. Feshbach, S. & Sakano, N. The structure and correlates of attitudes toward one's

- 987 nation in samples of United States and Japanese college students: A comparative
988 study. *Patriotism in the Lives of Individuals and Nations* 91-107 (1997).
- 989 28. Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S.,
990 Laird, A.R. & Fox, P.T. The human brainnetome atlas: a new brain atlas based on
991 connectional architecture. *Cereb. Cortex* **26**, 3508-3526 (2016).
- 992 29. Fullinwider, R. K. War and innocence. *Phi Pub Affairs* 90-97 (1975).
- 993 30. Coates, A. J. *The Ethics of War*. Manchester, U.K. (Manchester Univer. Press,
994 1997).
- 995 31. Schiller, B., Baumgartner, T. & Knoch, D. Intergroup bias in third-party
996 punishment stems from both ingroup favoritism and outgroup discrimination. *Evol.*
997 *Hum. Behav.* **35**, 169-175 (2014).
- 998 32. Bernhard, H., Fischbacher, U. & Fehr, E. Parochial altruism in humans. *Nature*
999 **442**, 912-915 (2006).
- 1000 33. Delton, A. W. & Krasnow, M. M. The psychology of deterrence explains why
1001 group membership matters for third-party punishment. *Evol. Hum. Behav.* **38**, 734-
1002 743 (2017).
- 1003 34. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum.*
1004 *Behav.* **25**, 63-87 (2004).
- 1005 35. Volle, E., Kinkingnéhun, S., Pochon, J.B., Mondon, K., Thiebaut de Schotten, M.,
1006 Seassau, M., Duffau, H., Samson, Y., Dubois, B. & Levy, R. The functional
1007 architecture of the left posterior and lateral prefrontal cortex in humans. *Cereb. Cortex*
1008 **18**, 2460-2469 (2008).
- 1009 36. Dixon, M. L. & Christoff, K. The lateral prefrontal cortex and complex value-
1010 based learning and decision making. *Neurosci. Biobehav. Rev.* **45**, 9-18 (2014).
- 1011 37. Skunde, M., Walther, S., Simon, J. J., Wu, M., Bendszus, M., Herzog, W. &

1012 Friederich, H. C. Neural signature of behavioural inhibition in women with bulimia
1013 nervosa. *J. Psychia. Neurosci. JPN* **41**, E69. (2016).

1014 38. Elshout, M., NELISSEN, R. M. & Van Beest, I. A prototype analysis of
1015 vengeance. *Pers. Relat.* **22**, 502-523 (2015).

1016 39. Jackson, J. C., Choi, V. K. & Gelfand, M. J. Revenge: A multilevel review and
1017 synthesis. *Ann. Rev. Psychol.* **70**, 319-345 (2019).

1018 40. Han, X., Gelfand, M. J., Wu, B., Zhang, T., Li, W., Gao, T., Pang, C., Wu, T.,
1019 Zhou, Y., Zhou, S., Wu, X. & Han, S. A neurobiological association of revenge
1020 propensity during intergroup conflict. *Elife* **9**, e52014 (2020).

1021 41. Sherif, M. *Group conflict and cooperation*. (London: Routledge, 1966).

1022 42. Stephan, W. G., & Stephan, C. W. Intergroup threat theory. *Interna. Encyclo.*
1023 *Intercul. Commu.* 1-12 (2017).

1024 43. Kteily, N., Hodson, G. & Bruneau, E. They see us as less than human:
1025 Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *J.*
1026 *Pers. Soc. Psychol.* **110**, 343-370 (2016).

1027 44. Amodio, D. M. & Mendoza, S. A. Implicit intergroup bias: Cognitive, affective,
1028 and motivational underpinnings. In B. Gawronski & B. K. Payne (Eds.), *Handbook of*
1029 *implicit social cognition: Measurement, theory, and applications*. pp. 353–374 (The
1030 Guilford Press, 2010).

1031 45. Xu, X., Zuo, X., Wang, X. & Han, S. Do you feel my pain? Racial group
1032 membership modulates empathic neural responses. *J. Neurosci.* **29**, 8525-8529 (2009).

1033 46. Hein, G., Silani, G., Preuschoff, K., Batson, C. D. & Singer, T. Neural responses
1034 to ingroup and outgroup members' suffering predict individual differences in costly
1035 helping. *Neuron* **68**, 149-160 (2010).

1036 47. Molenberghs, P. The neuroscience of in-group bias. *Neurosci. Biobehav. Rev.* **37**,

- 1037 1530-1536 (2013).
- 1038 48. Cikara, M., & Van Bavel, J. J. The neuroscience of intergroup relations: An
1039 integrative review. *Perspec. Psychol. Sci.* **9**, 245-274 (2014).
- 1040 49. Levy, J., Goldstein, A., Inliss, M., Masalha, S., Zagoory-Sharon, O. & Feldman,
1041 R. Adolescents growing up amidst intractable conflict attenuate brain response to pain
1042 of outgroup. *Proc. Natl. Acad. Sci. USA.* **113**, 13696-13701 (2016).
- 1043 50. Han, S. Neurocognitive basis of racial ingroup bias in empathy. *Trends Cogn. Sci.*
1044 **22**, 400-421 (2018).
- 1045 51. Rusch, H. The evolutionary interplay of intergroup conflict and altruism in
1046 humans: a review of parochial altruism theory and prospects for its extension. *Proc.*
1047 *Biol. Sci. B.* **281**, 20141539 (2014).
- 1048 52. Abbink, K., Brandts, J., Herrmann, B. & Orzen, H. Intergroup conflict and intra-
1049 group punishment in an experimental contest game. *Ame. Econ. Rev.* **100**, 420-47
1050 (2010).
- 1051 53. Battigalli, P. & Dufwenberg, M. Dynamic psychological games. *J. Econ. Theory*
1052 **144**, 1-35 (2009).
- 1053 54. Nisbett, R. E. & Masuda, T. Culture and point of view. *Proc. Natl. Acad. Sci. USA.*
1054 **100**, 11163-11170 (2003).
- 1055 55. Han, S., & Northoff, G. Culture-sensitive neural substrates of human cognition: A
1056 transcultural neuroimaging approach. *Nat. Rev. Neurosci.* **9**, 646-654 (2008).
- 1057 56. Han, S. *The sociocultural brain: A cultural neuroscience approach to human*
1058 *nature.* (Oxford Univ. Press, 2017).
- 1059 57. Han, S., & Ma, Y. A culture–behavior–brain loop model of human development.
1060 *Trends Cogn. Sci.* **19**, 666-676 (2015).
- 1061 58. Ma, Y., Wang, C., Li, B., Zhang, W., Rao, Y. & Han, S. Does self-construal

- 1062 predict activity in the social brain network? A genetic moderation effect. *Soc. Cogn.*
 1063 *Affect. Neurosci.* **9**, 1360-1367 (2014).
- 1064 59. Luo, S., Ma, Y., Liu, Y., Li, B., Wang, C., Shi, Z. ... & Han, S. Interaction
 1065 between oxytocin receptor polymorphism and interdependent culture values on
 1066 human empathy. *Soc. Cogn. Affect. Neurosci.* **10**, 1273-1281 (2015).
- 1067 60. Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O.
 1068 & Marois, R. From blame to punishment: disrupting prefrontal cortex activity reveals
 1069 norm enforcement mechanisms. *Neuron* **87**, 1369-1380 (2015).
- 1070 61. Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. Statistical power analyses
 1071 using G* Power 3.1: Tests for correlation and regression analyses. *Behav. Res.*
 1072 *Methods* **41**, 1149-1160 (2009).
- 1073 62. Raftery, A. E. Bayesian model selection in social research. *Sociol. Meth.* **25**, 111-
 1074 164 (1995).
- 1075 63. Maoz, I. & McCauley, C. Threat perceptions and feelings as predictors of Jewish-
 1076 Israeli support for compromise with Palestinians. *J. Peace. Res.* **46**, 525-539 (2009).
- 1077 64. Halperin, E., Porat, R., Tamir, M. & Gross, J. J. Can emotion regulation change
 1078 political attitudes in intractable conflicts? From the laboratory to the field. *Psychol.*
 1079 *Sci.* **24**, 106-111 (2013).

1080 **Data availability**

1081 The data that support the findings of this study are available from the corresponding
 1082 author upon reasonable request.

1083 **Code availability.** The codes used to analyze the data that support the findings of
 1084 this study are available from the corresponding author upon reasonable request.

1085 **Acknowledgements**

1086 This research was supported by the National Natural Science Foundation of China

(projects 31421003, 31871134, and 31661143039), the Ministry of Science and Technology of China (2019YFA0707103), and the Israel Science Foundation (2510/16). We thank the National Center for Protein Sciences at Peking University for assistance with this study. The authors thank L. Zhu, J. Li, and Y. Ma for insightful discussion about experimental design and data analyses, R. Hampton for proof reading, and S. Wang for modification of Figure 1. The funder had no role in the conceptualization, design, data collection, analysis, decision to publish or preparation of the manuscript.

Author contributions

X. H. and S.H. conceived the research programme and designed the experiments. X. H., S. Z., N. F., T. W., T. G., S. S., X. W., S. H. carried out the experiments. X. H., N. F., and S. H. analyzed the data. X. H., S. H., and S. S. wrote the paper. M.J.G reviewed and edited the paper. S.H. supervised the entire work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper.

1105 **Figure legends**

1106 **Figure 1.** Illustration of the experimental procedure in Experiment 1. Group
 1107 formation was manipulated in Phase 1 in which four participants were assigned to two
 1108 groups to play a card game against an outgroup (the group conflict condition) or each
 1109 participant played a card game against all the other three (the individual conflict
 1110 condition). Group conflict was induced in Phase 2 in which C1 and C2 made
 1111 decisions on the distribution of the bonus payment between C1 and C2. Three levels
 1112 of conflict were defined based on unfairness of money distribution and were coded as
 1113 low (-1), middle (0), and high (1) in data analyses. In Phase 3 the participants
 1114 underwent a pain threshold test, respectively, and reported their subjective feelings of
 1115 pain corresponding to different shock intensities. In Phase 4 each of the four
 1116 participants was asked to make a punishment decision in a private room by choosing
 1117 one from two options shown on the left and right side of a screen. Each option
 1118 consisted of two electric shocks as punishment on the outgroup combatant and
 1119 “collateral damage” to the outgroup noncombatants in the group conflict condition or
 1120 punishment on a combatant and harm to a bystander in the individual conflict
 1121 condition. Three levels of electric shocks were defined based on painful feelings
 1122 corresponding to the stimuli (low, middle, high), resulting in nine combinations of
 1123 shocks applied to combatant/noncombatant in the group conflict condition or
 1124 combatant/bystander in the individual conflict condition.

1125
 1126 **Figure 2.** Results of logistic regression parameter analyses in the group and
 1127 individual conflict conditions in Experiment 1. **(A)** Violin plots show means (big
 1128 dots), s.d. (bars), and distributions of parameter values. * $P < 0.05$; *** $P < 0.001$,
 1129 FDR corrected. Each small dot represents an individual participant. **(B)** Bar charts
 1130 illustrate variations of h_c and h_n that were calculated at each level of conflict and in

different conditions. h_n was significantly smaller than zero in both the group conflict condition (-0.52 ± 0.53 , $t(119) = -10.71$, $P < 0.001$, Cohen's $d = 0.98$, 95% CI = -0.61 , -0.42 ; $BF_{10} > 150$) and individual conflict condition (-0.55 ± 0.50 , $t(119) = -12.14$, $P < 0.001$, Cohen's $d = 1.11$, 95% CI = -0.64 , -0.46 ; $BF_{10} > 150$). h_{n-L} was significantly larger than zero in the group conflict condition (0.09 ± 0.29 , $t(119) = 3.56$, $P < 0.001$, Cohen's $d = 0.33$, 95% CI = 0.04 , 0.15 ; $BF_{10} = 36.87$) but not in the individual conflict condition (-0.05 ± 0.35 , $t(119) = -1.50$, $P = 0.135$, Cohen's $d = 0.14$, 95% CI = -0.11 , 0.02 ; $BF_{01} = 3.30$). (C) The results of bootstrapping analyses illustrate variations of sample distributions of h_c and h_n along three levels of conflict in the group and individual conflict conditions. We conducted repeated random resampling with replacement from the available data and the mean of this bootstrapped sample was then calculated and plotted in a two-dimensional space defined by h_c and h_n . The procedure was replicated for 1000 times to generate an empirical approximation of the sampling distribution. h_c , h_n , h_{c-L} and h_{n-L} are regression coefficients in the logistic regression model. A positive h_c or h_n value indicates harm preference whereas a negative h_c or h_n value indicates harm avoidance. h_{c-L} (or h_{n-L}) manifests how h_c (or h_n) varies along escalated levels of conflict. Positive (or negative) h_{c-L} or h_{n-L} values indicate increase (or decrease) of h_c or h_n .

Figure 3. Results of logistic regression parameter analyses of the whole sample in Experiment 2. The left panel shows violin plots of mean (big dots), s.d. (bars), and distributions of parameter values. ** $P < 0.005$; *** $P < 0.001$, FDR corrected. Each small dot represents an individual participant. The right panel illustrates variations of h_c and h_n along increased conflict levels. h_n was significantly smaller than zero (-0.50 ± 0.46 , $t(211) = -15.67$, $P < 0.001$, Cohen's $d = 1.08$, 95% CI = -0.56 , -0.44 ; $BF_{10} >$

1156 150) and h_{n-L} was significantly larger than zero (0.06 ± 0.28 , $t(211) = 3.04$, $P = 0.003$,
 1157 Cohen's $d = 0.21$, 95% CI = 0.02, 0.10) ; $BF_{10} = 6.66$).

1158 **Figure 4.** Behavioral and fMRI results in Experiment 3. (A) The results of logistic
 1159 regression parameter analyses in the group and individual conflict conditions,
 1160 respectively. The top panels show violin plots of mean (big dots), s.d. (bars), and
 1161 distributions of parameter values. *** $P < 0.001$, FDR corrected. Each small dot
 1162 represents an individual participant. The bottom panels illustrate variations of h_c and
 1163 h_n along increased conflict levels. h_n was significantly smaller than zero in both the
 1164 group conflict condition (-0.28 ± 0.27 , $t(59) = -8.04$, $P < 0.001$, Cohen's $d = 1.04$, 95%
 1165 CI = -0.35, -0.21; $BF_{10} > 150$) and individual conflict condition (-0.50 ± 0.49 , $t(58) = -$
 1166 7.83 , $P < 0.001$, Cohen's $d = 1.02$, 95% CI = -0.62, -0.37; $BF_{10} > 150$). h_{n-L} was
 1167 significantly larger than zero in the group conflict condition (0.12 ± 0.18 , $t(59) = 5.36$,
 1168 $P < 0.001$, Cohen's $d = 0.69$, 95% CI = 0.08, 0.17; $BF_{10} = 36.87$) but not in the
 1169 individual conflict condition (0.02 ± 0.19 , $t(58) = 0.94$, $P = 0.467$, Cohen's $d = 0.12$, 95%
 1170 CI = -0.03, 0.07; $BF_{01} = 0.22$). (B) The decreased left MFG activity during selection of
 1171 punishment decisions predicted larger h_{n-L} in the group conflict condition (a voxel-
 1172 level threshold $P < 0.001$, cluster-level threshold $P < 0.05$, FWE corrected). The top
 1173 panel illustrates the location of the left MFG. The bottom panel illustrates variation of
 1174 the left MFG activity across individuals' h_{n-L} .

1176 **Figure 5.** fMRI results of Experiment 3. (A) Neural correlates of h_n in the group and
 1177 individual conflict groups, respectively. (B) Neural correlates of h_c and h_{c-L} when
 1178 combining group and individual conflict groups. Brain activations were defined by
 1179 combining a whole-brain voxel-level threshold of $P < 0.001$ and a cluster-level
 1180 threshold at $P < 0.05$ FWE corrected. MCC=midcingulate cortex;

1182 TPJ=temporoparietal junction.

1183

A**Group conflict condition** **Individual conflict condition** **B** **Conflict level**

- high
- middle
- low

C **Conflict level**

- high
- middle
- low

(A)

(B)

(A)

Group conflict condition

Individual conflict condition

(B)

h_c

h_{c-L}