Cognitive and neural bases of collateral damage during intergroup conflict

Xiaochun Han[1,2,3], Shuai Zhou[4], Nardine Fahoum[5], Taoyu Wu[1,2,3], Tianyu Gao[1,2,3], Simone Shamay-Tsoory[5], Michele J. Gelfand[6], Xinhuai Wu[4*], Shihui Han[1,2,3*]

[1]School of Psychological and Cognitive Sciences, Peking University, Beijing, China
[2]PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing, China
[3]Beijing Key Laboratory of Behavior and Mental Health, Peking University, China
[4]Department of Radiology, The 7th Medical Center of PLA General Hospital, Beijing, China
[5]Department of Psychology, University of Haifa, Haifa, Israel
[6]Department of Psychology, University of Maryland, College Park, The United States of America

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

Abstract

Civilian casualties occur during military attacks. Such collateral damage is prohibited by international laws but increases with substantial consequences when intergroup conflict escalates. We investigate cognitive and neural bases of collateral damage by integrating brain imaging and statistical modeling of punishment decision-making during intergroup conflict. We test two Chinese groups manipulated in a laboratory setting and two ethnic groups (Israeli-Jews and Arab-Palestinians) in Israel. The results dissociate two psychological constructs, i.e., harm preference and harm avoidance, which respectively characterize punishment decision-making related to outgroup combatants and noncombatants during intergroup conflict. In particular, individuals show decreased avoidance of harming outgroup noncombatants when conflict escalates. Moreover, the decreased harm avoidance is predicted by inhibition of the left middle frontal activity during the selection of punishment decisions. Our findings provide a new perspective on the cognitive and neural bases of collateral damage and have important implications for understanding civilian casualties during the war.

Main

Collateral damage refers to civilian or noncombatant casualties and property damage due to aggressive actions during intergroup conflict such as a military operation during war[1]. Although conflicting with humanitarian interests and being prohibited by international laws[2], collateral damage occurs massively in human societies[3,4] and raises serious moral and ethical concerns. For example, hundreds of civilians were killed and thousands of civilians were wounded due to NATO's bombing intervention in the Kosovo conflict[5]. The ratio of civilian to military deaths during Iraq war is estimated to range between 3:1 and 8:1[4]. During the 2014 Israel-Gaza conflict over two thousand individuals were killed and 67% of the victims were civilians[6]. Given the substantial social consequences of collateral damage, understanding its psychological constructs and brain underpinnings is of both theoretical and practical importance.

Collateral damage is usually regarded as a side effect of aggressive acts following punishment decisions toward outgroup combatants during intergroup conflict. Punishment decision-making engages complicated cognitive and neural processes[7,8], including detection and generation of aversive experiences in the salience network (e.g., the anterior midcingulate and insula)[9], integration of harm and intent into an assessment of blame in the theory-of-mind network (e.g., the medial prefrontal cortex and temporoparietal conjunction)[10], and action selection/control in the central executive network (e.g., the lateral prefrontal cortex)[11]. In previous research, punishment decisions in the context of intergroup conflict were usually made to treat outgroup members indiscriminately as a whole[12-16]. This leaves an open issue related to collateral damage, i.e., how to separate consequences of decision-making pertaining to outgroup combatants and noncombatants so as to disentangle the

cognitive/neural mechanisms underlying decision-making that causes collateral damage. To address this issue would allow us to unravel psychological constructs characterizing the unique pattern of decision-making related to collateral damage during intergroup conflict. It is now known that a critical feature of real-world collateral damage is that it increases when intergroup conflict escalates[17]. This suggests an approach to the understanding of cognitive and neural bases of collateral damage by examining psychological processes that vary along escalation of intergroup conflict.

Here, we propose two psychological constructs involved in punishment decision-making pertaining to outgroup members during intergroup conflict. Because the goal of combat during intergroup conflict is to punish or destroy outgroup combatants but to maintain minimal noncombatant casualties, as required by morality and international laws[2], we hypothesize that harm preference and harm avoidance are two distinct psychological constructs that differentially characterize punishment decisions pertaining to outgroup combatants and noncombatants during intergroup conflict (Hypothesis 1). Moreover, collateral damage arising from intergroup conflict escalation may result from reduced harm avoidance pertaining to outgroup noncombatants (Hypothesis 2). That is, harm avoidance that inhibits damage to noncombatants during intergroup conflict is predicted to decrease when intergroup conflict escalates.

To test these hypotheses, we developed a new behavioral paradigm in which participants were asked to make a decision by choosing electric shocks for punishing an outgroup member who was directly engaged in the conflict (an outgroup combatant) with an ingroup member (an ingroup combatant) when they distributed a bonus as additional payment. This punishment decision was accompanied by harm to another

outgroup member who was not involved in the conflict (an outgroup noncombatant). These manipulations in a laboratory setting characterize the nature of real-world collateral damage 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 groups[18]. 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 collateral damage to outgroup noncombatants escalated when the level of intergroup conflict increased, similar to what happens during real-world intergroup conflicts[17].

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., Israeli-Jews and Arab-Palestinians 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 collateral damage, 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

111 harm avoidance. In particular, we estimated neural underpinnings of collateral
112 damage by examining brain activity associated with the decreased harm avoidance
113 pertaining to outgroup noncombatants as intergroup conflict escalates.

114 **Results**

115 **Decreased harm avoidance during escalation of intergroup conflict in a**
116 **laboratory setting**

117 To disentangle the psychological constructs related to collateral damage from
118 nonessential cognitive/affective processes engaged in punishment decision-making
119 during intergroup conflict, in Experiment 1, we randomly assigned Chinese adults to a
120 group conflict condition (N = 120) or an individual conflict (control) condition (N =
121 120). In the group conflict condition four participants of the same gender, who were
122 strangers, were recruited simultaneously and tested in the same session. To build
123 ingroup/outgroup relationships, the four participants were divided into two mini-
124 groups (N = 2 in each mini-group) by wearing T-shirts of two different colors and
125 playing a 1-hour card game (Phase 1 in Figure 1). This game required the participants
126 to help ingroup members but to interrupt outgroup members so that one's own group
127 would reach a destination in advance (see Methods). After group formation in Phase 1,
128 the participants completed a modified version of the Inclusion of Other in the Self
129 Scale[19] to assess their feelings of closeness between themselves and other participants.
130 The same procedure was applied to the participants in the individual conflict
131 condition except that there was no mini-group assignment and each participant
132 competed against the other three in the game. Thus, no one shared group identity with
133 others in the individual conflict condition, and the comparison of behavioral
134 performance in group and individual conflict conditions allowed us to disentangle the
135 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 ± 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) and the one left 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 unfairness of the bonus payment distribution 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.

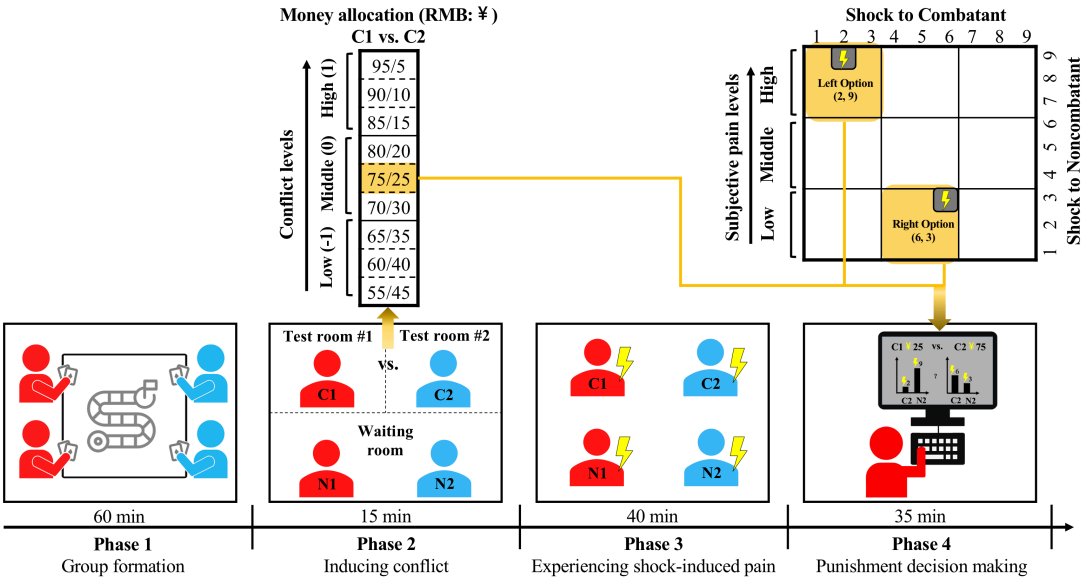


Figure 1. Illustration of the experimental procedure in Experiment 1. Group formation was manipulated in Phase 1 in which four participants were assigned to two groups to play a card game against an outgroup (the group conflict condition) or each participant played a card game against all the other three (the individual conflict condition). Group conflict was induced in Phase 2 in which C1 and C2 made decisions on the distribution of the bonus payment between C1 and C2. Three levels of conflict were defined based on unfairness of money distribution and were coded as low (-1), middle (0), and high (1) in data analyses. In Phase 3 the participants underwent a pain threshold test, respectively, and reported their subjective feelings of pain corresponding to different shock intensities. In Phase 4 each of the four participants was asked to make a

punishment decision in a private room by choosing one from two options shown on the left and right side of a screen. Each option consisted of two electric shocks as punishment on the outgroup combatant and collateral damage to the outgroup noncombatants in the group conflict condition or punishment on a combatant and harm to a bystander in the individual conflict condition. Three levels of electric shocks were defined based on painful feelings corresponding to the stimuli (low, middle, high), resulting in nine combinations of shocks applied to combatant/noncombatant in the group conflict condition or combatant/bystander in the individual conflict condition.

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 outgroup noncombatant as a collateral damage. There were three levels (i.e., 1-3, 4-6, or 7-9 corresponding to the low, middle, or high level of pain) of electric shocks. In either the left or right punishment option there were nine combinations of the two electric shocks applied to the outgroup combatant and noncombatant (i.e., 3 (low, middle, or high) levels of shocks to the combatant x 3 (low, middle, or high) levels of shocks to the noncombatant). These were the same in the individual conflict condition except that one shock would be applied to the combatant as punishment and the other

to a bystander. The participants were told that the bonus and electric shock in one of their decisions would be randomly selected to apply to the target combatant and noncombatant after the experiment.

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

Participants' choices of electric shocks were subject to logistic regression analyses[20]. Because we hypothesized that outgroup combatants and noncombatants are treated differently in terms of harm preference and harm avoidance during intergroup conflict, we included four independent variables in the regression model to 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, Extended Data Fig. 1 and 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:

$$\text{logit} = \ln(p/(1-p)) = \Delta V \qquad (1)$$

$$\Delta V = \boldsymbol{h_c} * \Delta S_c + \boldsymbol{h_n} * \Delta S_n + \boldsymbol{h_{c\text{-}L}} * \Delta S_c * L + \boldsymbol{h_{n\text{-}L}} * \Delta S_n * L \quad (2)$$

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

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 $\boldsymbol{h_c}$ or $\boldsymbol{h_n}$ value indicates harm preference — a propensity to select a stronger electric shock from two options to punish a target. A negative $\boldsymbol{h_c}$ or $\boldsymbol{h_n}$ value indicates harm avoidance — a propensity to select a weaker electric shock from two options to punish a target. $\boldsymbol{h_{c\text{-}L}}$ (or $\boldsymbol{h_{n\text{-}L}}$) manifests how $\boldsymbol{h_c}$ (or $\boldsymbol{h_n}$) varies along escalated levels of conflict and positive (or negative) $\boldsymbol{h_{c\text{-}L}}$ or $\boldsymbol{h_{n\text{-}L}}$ values indicate increase (or decrease) of $\boldsymbol{h_c}$ or $\boldsymbol{h_n}$. 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 analyses[23] 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).

If harm preference and harm avoidance characterize punishment decisions pertaining to outgroup combatants and noncombatants, respectively, we would expect a positive value of $\boldsymbol{h_c}$ but a negative value of $\boldsymbol{h_n}$. In addition, if collateral damage during escalation of intergroup conflict arises from reduced harm avoidance pertaining to outgroup noncombatants, we predicted a positive value of $\boldsymbol{h_{n\text{-}L}}$.

As expected, the results of one sample t-tests (all two-tailed) showed that $\boldsymbol{h_c}$ was significantly larger than zero in both the group conflict condition (mean ± s.d. = 0.18

± 0.80, t(119) = 2.47, $P = 0.015$, FDR corrected, Cohen's $d = 0.23$, 95% CI = 0.04, 0.32; $BF_{10} = 1.84$, only FDR corrected results were reported in the following) and the individual conflict condition (0.31±0.68, t(119) = 4.95, $P < 0.001$, Cohen's $d = 0.45$, 95% CI = 0.19, 0.43; $BF_{10} > 150$, Figure 2, Extended Data Fig. 2.), but did not differ significantly between the two conditions (two sample test: t(238) = -1.35, $P = 0.358$, Cohen's $d = 0.17$, 95% CI = -0.32, 0.06; $BF_{01} = 3.00$, see Figure 2). However, $\boldsymbol{h_n}$ 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 in the 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; $BF_{10} > 150$), but did not differ significantly between the two conditions (t(238) = 0.55, $P = 0.584$, Cohen's $d = 0.07$, 95% CI = -0.09, 0.17; $BF_{01} = 6.14$). These results provide evidence for harm preference for combatants (e.g., positive $\boldsymbol{h_c}$) but harm avoidance related to noncombatants (e.g., negative $\boldsymbol{h_n}$) on average across the three conflict levels.

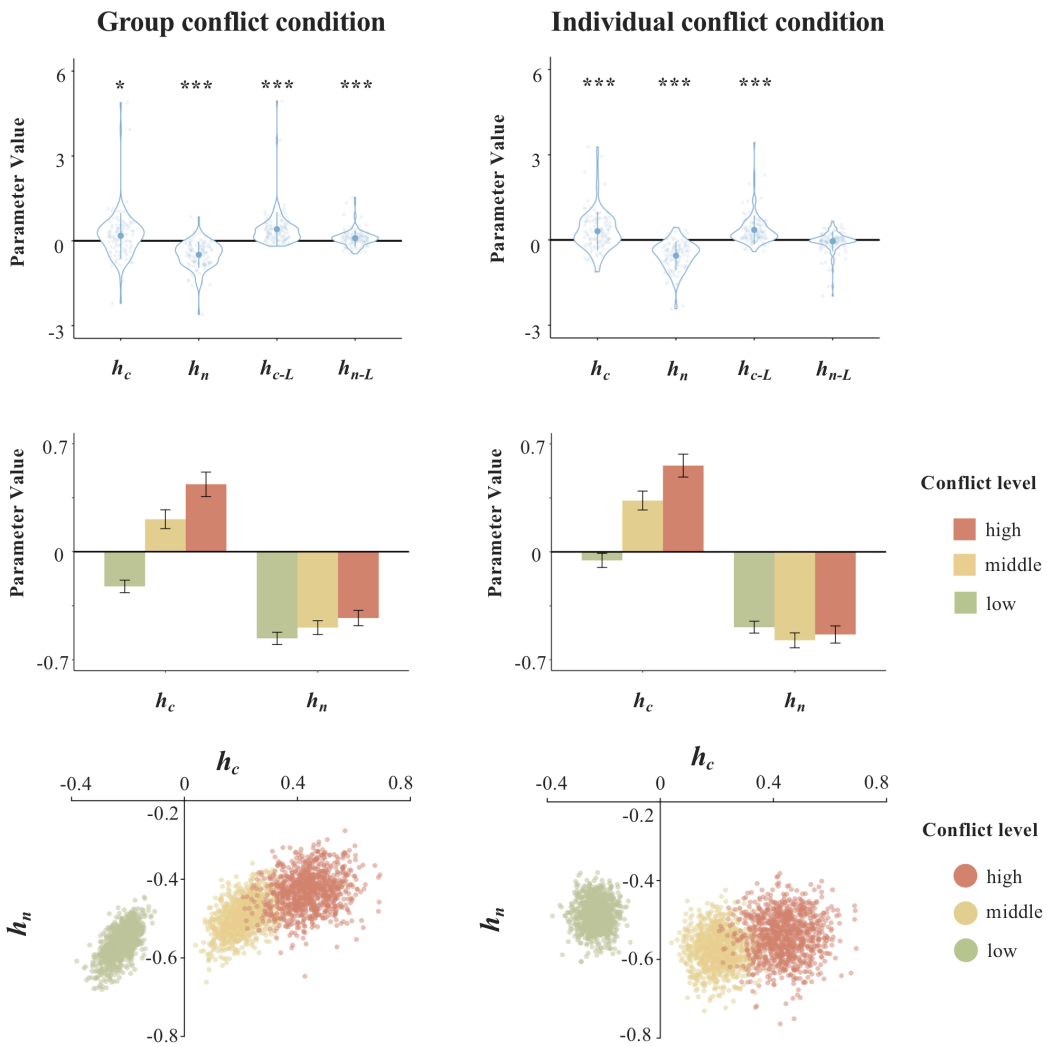


Figure 2. The results of logistic regression parameter analyses in the group and individual conflict conditions in Experiment 1. (**A**) Violin plots show means (big dots), s.d. (bars), and distributions of parameter values. * $P < 0.05$; *** $P < 0.001$, FDR corrected. Each small dot represents an individual participant. (**B**) Bar charts illustrate variations of $\boldsymbol{h_c}$ and $\boldsymbol{h_n}$ that were calculated at each level of conflict and in different conditions. $\boldsymbol{h_n}$ 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; $BF_{10} > 150$). $\boldsymbol{h_{n\text{-}L}}$ 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 $\boldsymbol{h_c}$ and $\boldsymbol{h_n}$ 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 $\boldsymbol{h_c}$ and $\boldsymbol{h_n}$. The procedure was replicated for 1000 times to generate an empirical approximation of the sampling distribution.

The analyses of $\boldsymbol{h_{c\text{-}L}}$ showed similar patterns in the group and individual conflict conditions whereas the analyses of $\boldsymbol{h_{n\text{-}L}}$ revealed different patterns between the two conditions. $\boldsymbol{h_{c\text{-}L}}$ 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 towards combatants increased when conflict escalated. Most interestingly, $\boldsymbol{h_{n\text{-}L}}$ 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 $\boldsymbol{h_{n\text{-}L}}$ 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., $\boldsymbol{h_n}$ 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 collateral damage 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 ± 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 Extended Data Fig. 1). The results indicate that the logistic regression model (Model 1) fits well with the participants' actual performances during decision making.

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

The results of Experiment 1 suggest decreased harm avoidance toward outgroup noncombatants during intergroup conflict escalation in a laboratory setting. In Experiment 2 we examined whether this account applies to collateral damage during intergroup conflict in a real-world context by testing Israeli-Jews (N = 106) and Arab-Palestinians (N = 106) in Israel. The Israeli-Palestinian conflict is one of the most challenging intergroup conflicts in the world and threatens numerous civilians' lives[6,24-26]. To test these two samples allowed us to further understand the role of decreased harm avoidance in collateral damage in a real-world intergroup conflict. Experiment 2 was carried out during the year 2018 when repeated clashes with numerous casualties occurred between the two sides. Because of the difference in national and religious identity between the two ethnic groups, we divided 2 Israeli-Jews and 2 Arab-Palestinians (all were of the same gender and strangers to each other) into two mini-groups on each testing day. The participants were blind to what the study was about during the recruitment. To highlight participants' national identity, after introducing one's own nationality, all participants were presented with images of items or celebrities/holy places that remind them of the nationality of Israel or Palestine[27]. The participants from each mini-group were asked to jointly observe each

picture and rate their feelings of picture-induced belonging and excitement (1 = very little, 9 = very much). The rating scores were high for both ethnic groups and higher in Arab-Palestinians than in Israeli-Jews (see Supplementary Results 1 and 2, Supplementary Table 1 and 2, and Supplementary Figure 1).

Thereafter, participants completed the money distribution task, pain threshold test, and punishment decision-making task. The procedures of these tasks were the same as those in Experiment 1 except that the participants were paid with a corresponding amount of Shekels. Similarly, a logistic regression model was constructed to estimate each participant's performance during punishment decision-making. We first assessed harm preference and harm avoidance characterizing the whole sample by pooling together the two ethnic groups' data and then examined possible group differences. One sample t-tests conducted on the whole sample showed that $\boldsymbol{h_c}$ did not differ significantly from zero (mean ± s.d. = -0.07±0.55, t(211) = -1.76, P = 0.081, Cohen's d = 0.12, 95% CI = -0.14, 0.01; BF_{01} = 2.88), whereas $\boldsymbol{h_n}$ 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} > 150, Figure 3, Extended Data Fig. 3 and 4). Moreover, these effects were significantly smaller in Israeli-Jews than in Arab-Palestinians ($\boldsymbol{h_c}$: -0.19±0.55 vs. 0.06±0.51, t(210) = -3.49, P = 0.001, Cohen's d = 0.57, 95% CI = -0.38, -0.14; BF_{10} > 150; $\boldsymbol{h_n}$: -0.63±0.49 vs. -0.37±0.39, t(210) = -4.17, P < 0.001, Cohen's d = 0.57, 95% CI = -0.38, -0.14; BF_{10} > 150). The results of one sample t-tests conducted on the whole sample revealed that both $\boldsymbol{h_{c\text{-}L}}$ and $\boldsymbol{h_{n\text{-}L}}$ were significantly larger than zero ($\boldsymbol{h_{c\text{-}L}}$: 0.19±0.28, t(211) = 9.84, P < 0.001, Cohen's d = 0.68, 95% CI = 0.15, 0.23; BF_{10} > 150; $\boldsymbol{h_{n\text{-}L}}$: 0.06±0.28, t(211) = 3.04, P = 0.003, Cohen's d = 0.21, 95% CI = 0.02, 0.10) ; BF_{10} = 6.66). These effects did not differ significantly between Israeli-Jews and in Arab-Palestinians ($\boldsymbol{h_{c\text{-}L}}$: 0.19±0.30 vs. 0.19±0.26, t(210) = 0.14, P = 0.888,

Cohen's d = 0.02, 95% CI = -0.07, 0.08; BF_{01} = 6.62; $\boldsymbol{h_{n\text{-}L}}$: 0.05±0.34 vs. 0.07±0.20, t(210) = -0.57, P = 0.756, Cohen's d = 0.08, 95% CI = -0.10, 0.05; BF_{01} = 5.72).

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

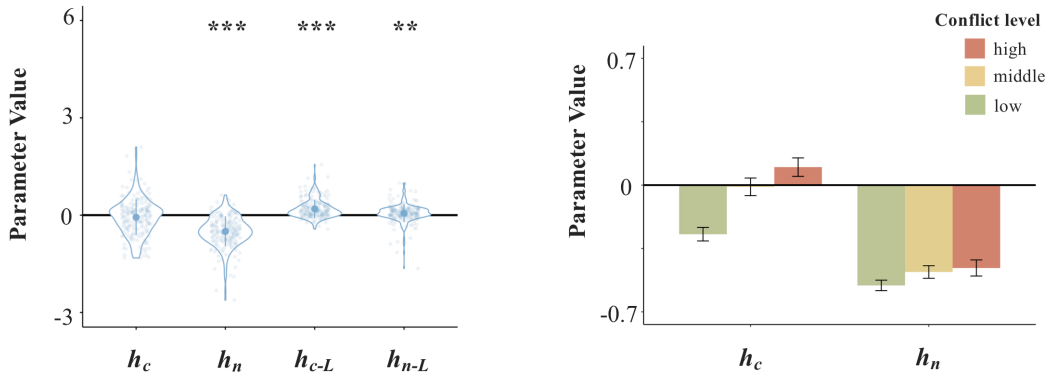


Figure 3. The results of logistic regression parameter analyses of the whole sample in Experiment 2. The left panel shows violin plots of mean (big dots), s.d. (bars), and distributions of parameter values. ** P < 0.005; *** P < 0.001, FDR corrected. Each small dot represents an individual participant. The right panel illustrates variations of $\boldsymbol{h_c}$ and $\boldsymbol{h_n}$ along increased conflict levels. $\boldsymbol{h_n}$ 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} > 150) and $\boldsymbol{h_{n\text{-}L}}$ was significantly larger than zero (0.06±0.28, t(211) = 3.04, P = 0.003, Cohen's d = 0.21, 95% CI = 0.02, 0.10) ; BF_{10} = 6.66).

Together, the results in Experiments 1 and 2 provide evidence for decreased harm avoidance toward noncombatants when intergroup conflicts escalate in both a laboratory setting and a real social context. Moreover, the same logistic regression model fits the empirical data of punishment decision-making during intergroup conflict in both experiments. These results suggest that decreased harm avoidance provides a common cognitive basis of collateral damage in a laboratory setting and a

real-world social context. Notably, our participants in a laboratory setting but not in a real-world situation showed harm preference and Israeli-Jews even exhibited harm avoidance toward to outgroup combatants (see Extended Data Fig. 3). Punishment decisions seemed to be made more carefully in a real social context of intergroup conflict than in a laboratory setting possibly due to higher stakes in the latter situation. The Israeli side of the conflict in the real-world situation is the stronger party with power resources that widen its range of strategies in managing the conflict, whereas the Palestinian side is relatively weaker and has rather limited options to react to the conflict[28]. In addition, victims in such asymmetric conflicts seek to restore their sense of power whereas perpetrators strive to restore their public moral image[29]. Therefore, in our collateral damage paradigm, the weaker side of the conflict (the Arab-Palestinians) may have been more willing to punish, in order to repair their percieved weakness. By contrast, the Israeli-Jews may have been more concerned about presenting a moral image. Future research should examine these distinct possibilities and replicate them with other samples embedded in real world conflicts.

Neural correlates of decreased harm avoidance during intergroup conflict escalation

In Experiment 3, we further investigated neural correlates of the decreased harm avoidance during intergroup conflict by testing an independent sample of Chinese 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 Results 3, Extended Data Table 2 and Fig. 5 for details).

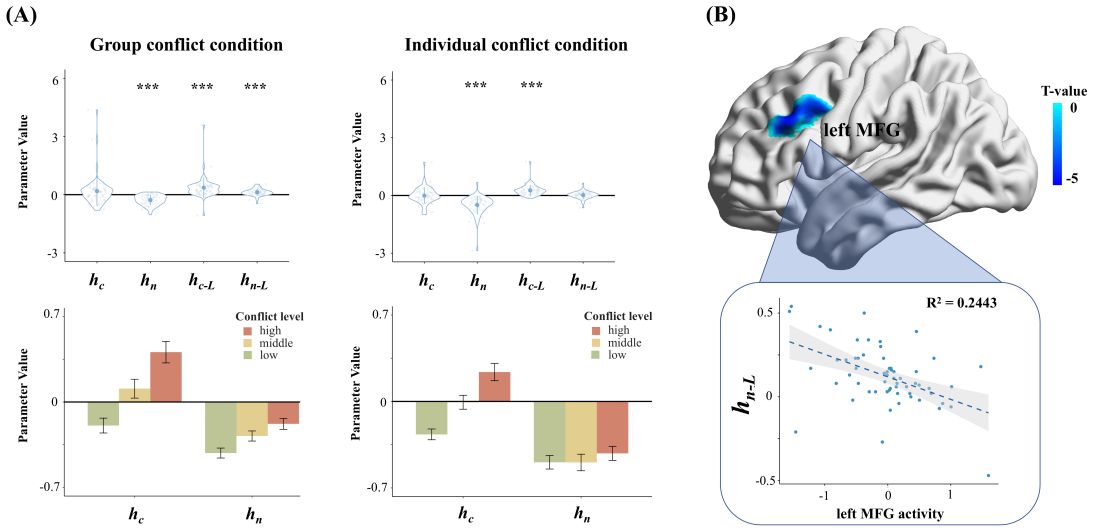


Figure 4. The results in Experiment 3. (**A**) The results of logistic regression parameter analyses in the group and individual conflict conditions, respectively. The top panels show violin plots of mean (big dots), s.d. (bars), and distributions of parameter values. *** $P < 0.001$, FDR corrected.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

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, ΔS_c * L, and ΔS_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 $\boldsymbol{h_c}$), ΔS_n (with the covariate $\boldsymbol{h_n}$), ΔS_c * L (with the covariate $\boldsymbol{h_{c\text{-}L}}$), and ΔS_n * L (with the covariate $\boldsymbol{h_{n\text{-}L}}$) of each participant. Most interestingly, the second-level analyses showed that a larger $\boldsymbol{h_{n\text{-}L}}$ 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 $\boldsymbol{h_{n\text{-}L}}$ was statistically significant when directly comparing the group and individual conflict conditions, we conducted a moderation analysis of the association between $\boldsymbol{h_{n\text{-}L}}$ and the left MFG activity. This analysis searched for regions in the left frontal lobe defined by a mask in the Brainnetome Atlas[30] in which the association between $\boldsymbol{h_{n\text{-}L}}$ 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 $\boldsymbol{h_{n\text{-}L}}$ 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 $\boldsymbol{h_n}$ showed that $\boldsymbol{h_n}$ 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 $\boldsymbol{h_c}$ and $\boldsymbol{h_{c\text{-}L}}$ between the group and individual conflict conditions, we pooled fMRI data in the two conditions to examine neural correlates of $\boldsymbol{h_c}$ and $\boldsymbol{h_{c\text{-}L}}$, respectively. The results of the second-level whole-brain regression analyses showed that a larger $\boldsymbol{h_c}$ 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 $\boldsymbol{h_{c\text{-}L}}$ 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.

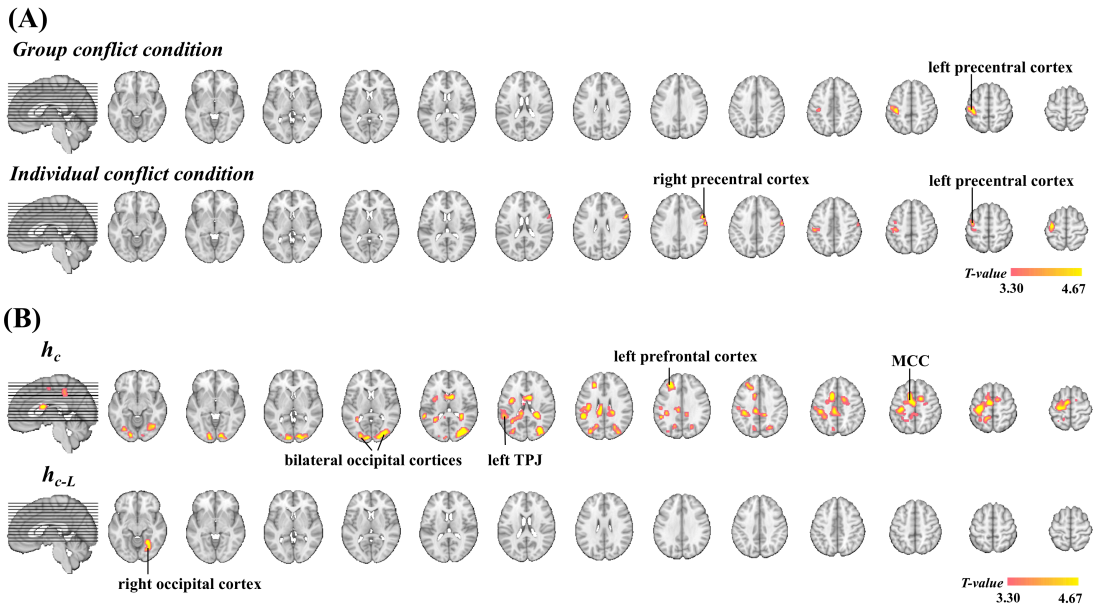


Figure 5. fMRI results of Experiment 3. (A) Neural correlates of $\boldsymbol{h_n}$ in the group and individual conflict groups, respectively. (B) Neural correlates of $\boldsymbol{h_c}$ and $\boldsymbol{h_{c\text{-}L}}$ when combining group and individual conflict conditions. Brain activations were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

Discussion

The present study combined functional brain imaging and statistical modeling of punishment decision-making to examine the cognitive and neural bases of collateral damage during intergroup conflict. Unlike previous research of intergroup conflict that measured punishment decisions upon outgroup members as a whole[12-16], 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 collateral damage when intergroup conflict escalated. Our fMRI results further uncovered the neural correlate of this psychological construct in the left MFG.

498 Our statistical modeling revealed a significantly negative value of the parameter
499 h_n as an index of harm avoidance during the selection of electric shocks that were
500 supposed to be possibly applied to noncombatants. This finding indicates that people
501 tended to give weaker electric shocks to or to avoid harming others who were not
502 directly engaged in aggressive actions during conflict. This was true during conflicts
503 between two groups or between two individuals and is consistent with morality and
504 international laws in human societies[31,32]. Nevertheless, the statistical modeling
505 revealed a positive value of the parameter $h_{n\text{-}L}$ in the situation of intergroup conflict
506 across three experiments. The positive $h_{n\text{-}L}$ suggests that harm avoidance that inhibits
507 damage to noncombatants during intergroup conflict tends to decrease when
508 intergroup conflict escalates.

509 It has been speculated that, as conflict escalates, decision makers are less likely to
510 differentiate between combatants and noncombatants due to depleted cognitive
511 resources and have to deal with the incidental cognitive dissonance[17]. These
512 hypothesized cognitive mechanisms underlying collateral damage, however, have not
513 been tested empirically. Our empirical findings uncover a psychological construct, i.e.,
514 harm avoidance, which varies across levels of intergroup conflict in a fashion similar
515 to collateral damage. Furthermore, we showed that decreased harm avoidance during
516 conflict escalation was evident in individuals facing conflict between an ingroup
517 member and an outgroup member but not in those facing conflicts between two
518 individuals who had no ingroup/outgroup relationships with the decision maker. This
519 provides further evidence that, rather than reflecting changes of cognitive processes
520 that are general to any conflict, the decreased harm avoidance specifically
521 characterizes punishment decision-making that leads to collateral damage during
522 escalation of intergroup conflict. In addition, our results demonstrated a positive value

523 of $h_{n\text{-}L}$ in both a laboratory setting in China and in a natural conflict context in Israel.
524 Thus a similar psychological construct may underlie collateral damage during
525 escalation of intergroup conflict in both a laboratory setting and a real-world context.
526 Our statistical modeling also revealed a significantly positive value of the
527 parameter h_c as an index of harm preference for outgroup combatants during selection
528 of electric shocks in a laboratory setting but not in a real social context of intergroup
529 conflict. However, the statistical modeling showed a significant positive value of the
530 parameter $h_{c\text{-}L}$ across the three experiments in both the group and individual conflict
531 conditions. The positive $h_{c\text{-}L}$ suggests that our participants tended to choose stronger
532 electric shocks to punish an individual if this person had proposed a more unfair
533 money distribution plan. The increased harm preference as a function of conflict
534 escalation might have different psychological causes in the group and individual
535 conflict conditions. In the group conflict condition, when facing a more unfair money
536 distribution plan, noncombatants might choose stronger electric shocks to protect
537 ingroup members' benefit, similar to previous findings of ingroup favoritism in
538 decision-making[33-35]. In the individual conflict condition, however, unfair money
539 distribution between two individuals did not directly affect a bystander's payoff and
540 punishment decision-making was possibly made to protect social norms, similar to
541 third-party punishment[36]. Our data analyses did not show consistent evidence for
542 differences in parameter estimates of harm preference and harm avoidance between
543 combatants and noncombatants as decision makers. It is likely that the motives to help
544 ingroup and to protect social norms dominated punishment decision-making whereas
545 one’s own payment was much less taken into consideration in our participants. Future
546 research requires a new behavioral paradigm that helps to reveal how combatants and
547 noncombatants make punishment decisions differently.

548 We integrated fMRI and statistical modeling of punishment decision-making to
549 examine neural correlates of harm avoidance and harm preference during intergroup
550 conflict. Our fMRI results showed that $\boldsymbol{h_c}$ and $\boldsymbol{h_n}$ were linked to distinct neural circuits,
551 i.e., $\boldsymbol{h_n}$ was associated with increasing activity in the precentral cortex whereas $\boldsymbol{h_c}$ was
552 associated with increased activity in the left temporoparietal junction, left prefrontal
553 cortex, and midcingulate. These results provide further neuroimaging evidence that
554 harm avoidance and harm preference represent two distinct psychological processes
555 supported by different neural underpinnings. Most importantly, we showed that the
556 decreased left MFG activity predicted a larger $\boldsymbol{h_{n\text{-}L}}$ and this effect was specific to the
557 participants engaged in intergroup conflict. This finding suggests a key functional role
558 of the left MFG in controlling harm avoidance related to collateral damage to
559 noncombatants during intergroup conflict. Specifically, inhibition of the left MFG
560 might lead to failure to avoid harm to noncombatants when making punishment
561 decisions during high intergroup conflict. The functional role of the left MFG during
562 intergroup conflict is consistent with previous fMRI findings. For example, the left
563 MFG activity increases with the complexity of different cognitive tasks and lesions of
564 this brain region result in performance deficits in different cognitive tasks[37]. The
565 lateral frontal cortex also represents the value of employing cognitive control and
566 plays an important role in selection/control of a specific punishment decision[8,11] and
567 in value-based learning and decision-making[38]. These findings agree upon a general
568 function of the left MFG in executive control of decision-making and behavioral
569 responses that is independent of task domains.

570 Our fMRI results revealed increasing activation in the precentral cortex as neural
571 correlates of harm avoidance related to noncombatants as indexed by $\boldsymbol{h_n}$. This is
572 consistent with the functional role of this brain region in general behavioral

573 inhibition[39]. Harm preference toward combatants as indexed by $\boldsymbol{h_c}$ was associated
574 with increased activity in the left temporoparietal junction, left prefrontal cortex, and
575 midcingulate. These brain regions are involved in punishment decision-making to
576 detect aversive experiences and to infer others' mental states[7-10]. Similar mental
577 processes may support harm preference regardless of when conflicts occur between
578 two groups or between two individuals. However, although the motives to punish
579 outgroup combatants and to avoid harming to noncombatants may coexist, inhibition
580 of the left MFG activity is specifically related to the decreased harm avoidance as a
581 psychological construct underlying collateral damage to noncombatants when
582 intergroup conflict escalates.

583 We also considered revenge as a potential motivator of punishment in this study.
584 Revenge is usually manifested in taking actions of harming someone in retaliation for
585 an injury[40,41]. Our recent fMRI research showed that the medial prefrontal activity in
586 response to ingroup pain caused by outgroup members predicted the propensity to
587 give painful electric shocks to outgroup members[42]. In the current study the
588 participants made choices by considering shocks to both the outgroup combatant and
589 noncombatant. While our statistical modeling of punishment decision-making
590 suggested harm preference toward the outgroup combatant, the results also suggest
591 harm avoidance toward the outgroup noncombatant. It is thus unlikely that the
592 participants made their choices only by considering retaliation upon the outgroup
593 combatant. Moreover, revenge cannot explain why harm avoidance related to
594 outgroup noncombatants decreased when intergroup conflict escalated since the
595 outgroup noncombatants did not bring any harm to the decision maker in any case.
596 We also assessed the possibility that the participants made their choices by
597 considering retaliation upon the outgroup combatant and noncombatant as a whole (or

group revenge) by constructing a different model (i.e., Model 3 reported in Extended Data Fig. 1 and Table 1). In this model the sum of shock intensities to the combatant and the noncombatant (i.e., group revenge) was used in a regression model to predict the participants choices. The results suggest that this 'group revenge' model is not as good as the model reported in the main text in predicting the participants' decisions. It appears that revenge cannot fully explain punishment decision making in a context when a punishment decision is related to both outgroup combatants and noncombatants during intergroup conflict.

Previous research on mechanisms of intergroup conflict usually took all outgroup members as one unit. Perceiving the outgroup as a competitor for limited resources[43] or a threat to the ingroup[44] may trigger intergroup conflict. Dehumanization of outgroup members[45] may further drive discrimination and conflict. Humans may evolve ingroup favoritism in implicit attitudes[46] and empathy[47-52] and aggressiveness against outgroups as genetic or cultural traits[53]. While these traits help individuals to survive, humans as a rational species are confronted with the contradiction between harming outgroup combatants as enemies and protecting outgroup noncombatants as innocents during intergroup conflict. Our findings provide a new neurocognitive account of decision-making that underlies collateral damage during intergroup conflict.

While our findings contribute to the understanding of neurocognitive underpinnings of collateral damage in real-world situations, several limitations of our findings should be acknowledged. First, collateral damages in real-world situations usually arise from armed attacks toward outgroup combatants and such attacks cost lives of and bring persistent physical/mental harms to outgroup combatants and noncombatants. The harm that intended to apply to the outgroup in our paradigm,

however, was much weaker than those in real-world situations due to ethical requirements. Nevertheless, despite the differences in degree of harm, our paradigm captures the key features of collateral damages in real-world situations (i.e., increasing along conflict escalation). Besides, previous research has shown that conflict expenditures of groups can be substantially larger than those of conflict between individuals and allowing group members to punish each other results in even larger conflict expenditures[54]. In our paradigm, however, conflict expenditures (distributed money between the two combatants) were controlled in the group and individual conflict conditions as an irrelevant variable. This leaves an open question for future research, i.e., how cognitive and neural mechanisms underlying collateral damage during intergroup conflict vary along escalated conflict expenditures.

Second, collateral damage in real-world situations usually takes place during intergroup conflict with a history. People make decisions to attack the outgroup by considering not only the consequence of the upcoming attack on the outgroup but also potential feedback/revenge from the outgroup that may occur to the ingroup thereafter. Recent theories of motivations indeed emphasize key functional roles of updated beliefs, beliefs of others, and plans of action in decision making[55]. Our paradigm modeled an initial decision making in response to intergroup conflict without providing the participants with the information of potential feedback/revenge from the outgroup. While there are indeed similar situations in real-world intergroup conflict, future research on collateral damage during intergroup conflict should consider decision making in complicated situations when decision makers receive feedback or potential revenge from the outgroup after each decision making.

Third, our work recruited three cultural samples (i.e., Chinese, Israeli-Jews, and Arab-Palestinians) who showed similar decreased harm avoidance related to outgroup

noncombatants as conflicts escalated but difference in harm preference for outgroup combatants. It is important to test other cultural samples to verify whether collateral damage arising from intergroup conflict has similar cognitive and neural underpinnings in other sociocultural contexts. This is necessary given broad sociocultural differences in[56-58] and potential interactions of biological and cultural factors[59-61] 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 collateral damage 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 stimulation[62].

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 and outgroup noncombatants, respectively. 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 new light on the neurocognitive underpinnings of collateral damage during intergroup conflict.

Methods

Participants

673 The sample size was pre-estimated using G*Power[63]. In Experiment 1, we aimed
674 to conducted one-sample t-tests for each parameter in the logistic regression model
675 and two-sample t-tests to compared parameters between the group and individual
676 conflict conditions. Based on G*Power estimation, a sample size of 54 participants
677 was required to obtain a medium effect size of 0.5 with an error probability of 0.05
678 and power of 0.95 in two-tailed one-sample t-tests. For the two-tailed two-sample t-
679 tests, a sample size of 105 participants for both the group and individual conflict
680 conditions were required to obtain a medium effect size of 0.5 with an error
681 probability of 0.05 and power of 0.95. The sample size of Israeli-Jews and Arab-
682 Palestinians in Experiment 2 was determined to be similar with those in the group and
683 individual conflict in Experiment 1. The sample size in Experiment 3 was smaller
684 than that in Experiment 1 by considering the cost of fMRI scanning. Based on
685 G*Power estimation, a sample size of 34 participants was required to obtain a
686 medium effect size of 0.5 with an error probability of 0.05 and power of 0.80 in two-
687 tailed one-sample t-tests. For the two-tailed two-sample t-tests, a sample size of 64
688 participants for both the group and individual conflict conditions were required to
689 obtain a medium effect size of 0.5 with an error probability of 0.05 and power of 0.80.
690 In Experiment 1 we recruited 120 Chinese university students (half male, mean
691 age ± s.d. = 22.85 ± 2.46 years) in the group conflict condition and 120 Chinese
692 university students (half male, 21.68 ± 2.43 years) in the individual conflict condition.
693 In Experiment 2 we recruited 106 Israeli-Jew university students (46 males, 24.97 ±
694 3.78 years). To assure that only Arabs who clearly identify themselves as Palestinians
695 participated in our study, we originally recruited 152 Arab-Palestinians who
696 completed a scale in which they rated their national identification from 0% Palestinian
697 to 100% Palestinian. Only participants who rated themselves as 100% Palestinians

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 measurements. Both Israeli-Jews and Arab-Palestinians were Israeli citizens.

In Experiment 3 we recruited 64 Chinese university students (half males, 23.13 ± 1.82 years) in the group conflict condition. Four participants were excluded due to excessive head motion during scanning, leaving 29 combatants (14 males, 22.93 ± 1.56 years) and 31 noncombatants (16 males, 23.19 ± 2.06 years) for analyses of data in the group conflict condition. An independent group of 64 Chinese university students (half males, 22.48 ± 2.42 years) were recruited in the individual conflict condition. Five participants were excluded due to excessive head motion during scanning, leaving 29 combatants (14 males, 22.76 ± 2.84 years) and 30 noncombatants (15 males, 22.23 ± 2.08 years) for analyses of data in the individual conflict condition. All participants had normal or corrected-to-normal vision and reported no neurological or psychiatric history. Informed consent was obtained prior to the experiment. All participants were paid for their participation. Experiments 1 and 3 were approved by the local ethics committee at the School of Psychological and Cognitive Sciences, Peking University. Experiment 2 was approved by the local ethics committee at the Rambam Medical Center, Haifa University.

Experiment Procedure

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

Phase 1: Group formation

723 In the group conflict condition, participants were randomly assigned to two
724 groups of 2 persons by choosing a card. Participants from each group were asked to
725 wear T-shirts of the same color (red or blue). Participant introduced their own names,
726 nicknames, majors, and hobbies to get familiar with each other. They then started to
727 play the Saboteur card game (http://www.annarbor.com/entertainment/saboteur-card-
728 game-review/). During this game ingroup members played cards to build a tunnel to a
729 gold location or to block the tunnel to prevent outgroup members from reaching the
730 location. This game required ingroup members to cooperate with each other but to
731 interrupt outgroup members so as to reach the gold location before the outgroup. The
732 intergroup relationship was built by playing this game for 50 minutes. To check the
733 effectiveness of the group manipulation, after the game, participants were asked to
734 complete a modified version of the Inclusion of Other in the Self Scale[19] to assess
735 their feelings of closeness between oneself and ingroup members and between oneself
736 and outgroup members. The individual conflict condition was the same as the group
737 conflict condition except that participants were not divided into groups and they
738 played the Saboteur card game against the other three participants. After the game,
739 participants were asked to assess their feelings of closeness between oneself and the
740 other three participants with a modified version of the Inclusion of Other in the Self
741 Scale[19]. Phase 1 lasted for 60 minutes.

742 *Phase 2: Inducing conflict*

743 For the group conflict condition one participant from each group was randomly
744 assigned as a combatant who allotted 100 RMB (in Experiment 1; 100 Shekels in
745 Experiment 2; and 200 RMB in Experiment 3) between the two combatants by giving
746 different weights to a number of plans. The other participant from each group waited,
747 as a noncombatant, while combatants made their decisions (Phase 2 in Figure 1). This

procedure was the same for the individual conflict condition except that 2 participants 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

773 participation. Phase 2 lasted for 15 minutes.

774 *Phase 3: Pain threshold measurement*

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

785 *Phase 4: Punishment decision making*

786 In this phase, the participants were informed that the 108 money allocation plans
787 from one combatant would be presented one by one. There were 108 trials with equal
788 numbers of the nine money distribution plans which were actually created by the
789 computer program and randomly presented to participants. On each trial, a money
790 distribution plan was first presented for 4 s on a screen. Thereafter, two options of
791 electric shocks were shown on the left and right side of the screen, respectively (Phase
792 4 in Figure 1). Each option (either the left or right) consisted of two vertical bars that
793 indicated intensities of two electric shocks. The shock intensities were indexed by
794 rating scores of subjective painful feelings (from 1 = slightly pain to 9 = extremely
795 painful) that each participant experienced in Phase 3. In the group conflict condition,
796 one shock would be applied to the outgroup combatant as punishment and the other to
797 the outgroup noncombatant as a collateral damage. In the individual conflict condition,

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

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

According to the unfairness of the money allocation plan, all 108 trials were coded as being in one of three conflict levels. In low, middle and high conflict levels, the target combatant would keep 55 – 65 RMB, 70 – 80 RMB and 85 – 95 RMB for

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

The experimental procedures of Experiment 2 in Israeli-Jews and Palestinian-Arab were the same as those in Study 1 except the following. There was no group formation phase at the start. Instead, two Israeli and two Palestinians (all were of the same gender and were strangers to each other) were recruited each time and they introduced their nationality to each other. Furthermore, in order to highlight the nationality of each participant, participants of the same nationality performed a priming task together which contained images of items that relate to one's nationality. Such as the symbols of each group's nation (e.g. flags, emblems, anthems), which was found to increase the feeling of identification with one's own nation[27], as well as photos of famous people (e.g., Yasser Arafat vs. Shimon Peres) and holy places (e.g., Al-Aqsa Mosque vs. Hakotel). The participants from each group were asked to jointly rate the feelings of belonging and excitement that each image evoked (1 = very little, 9 = very much).

The experimental procedures of Study 3 in the neuroimaging lab were the same

848 as those in Study 1 except the following. The Saboteur card game lasted 90 minutes
849 during Phase 1 to strengthen the intergroup relationship since the whole experiment
850 lasted for a longer time in the Experiment 3. Each participant was scanned using
851 fMRI while making punishment choices during Phase 4. There were 108 trials for
852 each participant which were equally distributed into six scans. In each scan, there
853 were 18 trials which included 6 low conflict trials, 6 middle conflict trials, and 6 high
854 conflict trials. On each trial, a jitter fixation was first presented for 2, 4, 6, 8, or 10 s.
855 A money distribution option given by the target combatant was then presented for 2 s.
856 A following jitter fixation was presented for 2, 4, 6, 8, or 10 s. Thereafter, two options
857 of electric shocks towards the target combatant and noncombatant were presented and
858 participants were asked to make a choice between the options within 5 seconds. Each
859 scan started with a 12-s fixation to get a baseline for blood-oxygen-level-dependent
860 (BOLD) signals. The last trial of each scan was followed by a 12-s fixation.

861 **Model comparisons**

862 We tested additional two models (Models 2 and 3) and compared these models
863 with the one we reported in the main text (Model 1). Model 2 only included intercept
864 in the logistic regression model to regress the inherent tendency to choose left or right
865 option of punishment decisions for each participant (ΔV = intercept; the participants
866 tended to choose one (the left or right) option regardless of shock intensities and
867 monetary distribution plans). This is a baseline control model compared with other
868 models. Model 3 hypothesized that outgroup combatants and noncombatants are
869 treated in the same way during intergroup conflict. We included two independent
870 variables in the regression model to account for the probability that participants would
871 select the left (p) or right options (1-p) in each trial. These included the difference in
872 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:

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

$$\Delta V = \boldsymbol{h} * \Delta S + \boldsymbol{h_L} * \Delta S * L \quad (2)$$

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

A larger p is associated with higher possibility of selection of the left option. A positive $\boldsymbol{h}$ 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 $\boldsymbol{h}$ 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. $\boldsymbol{h_L}$ manifest how $\boldsymbol{h}$ varies along escalated levels of conflict (positive and negative $\boldsymbol{h_L}$ value indicates increase or decrease of $\boldsymbol{h}$, 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 AUC is the preferred model.

fMRI data acquisition and preprocessing

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

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

Whole brain parametric analyses

The first level general linear model (GLM) was constructed to identify brain 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 (ΔS_c * L); (iv) the interaction between conflict level and difference of electric shocks towards noncombatant (ΔS_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 ($\boldsymbol{h_c}$), harm avoidance ($\boldsymbol{h_n}$), variation of harm tendency with conflict levels ($\boldsymbol{h_{c\text{-}L}}$), and variation of harm avoidance with conflict levels ($\boldsymbol{h_{n\text{-}L}}$). The first regression analysis included the contrast images of parametric modulation effects of ΔS_c with $\boldsymbol{h_c}$ as the covariate. The second regression analysis included the contrast images of parametric modulation effects of ΔS_n with $\boldsymbol{h_n}$ as the covariate. The third regression analysis included the contrast images of parametric modulation effects of ΔS_c * L with

948 $\boldsymbol{h_{c\text{-}L}}$ as the covariate. The fourth regression analysis included the contrast images of
949 parametric modulation effects of ΔS_n * L with $\boldsymbol{h_{n\text{-}L}}$ as the covariate. The brain regions
950 associated with individual differences in parameters were estimated by calculating
951 one-sample t-test using the contrasts of [0 1] or [0 -1] to assess the positive or
952 negative associations. Brain activations were defined by combining a whole-brain
953 voxel-level threshold of $P < 0.001$ and a cluster-level threshold at $P < 0.05$ FWE
954 corrected.

955 ***Moderation analyses***

956 To test whether the associations between $\boldsymbol{h_{n\text{-}L}}$ and the left MFG activity were
957 significantly different between the group conflict and individual conflict conditions,
958 we conducted a second-level whole-brain two-sample t-test with $\boldsymbol{h_{n\text{-}L}}$ in these two
959 conditions as covariates. The design matrix for this t-test included the contrast images
960 of the parametric modulation effect of ΔS_n * L in each condition and $\boldsymbol{h_{n\text{-}L}}$ in the two
961 conditions as the first and second covariates. The moderation effect was estimated by
962 calculating two-sample t-test using the contrasts of [0 0 -1 1] to assess the enhanced
963 coupling between $\boldsymbol{h_{n\text{-}L}}$ and brain activity that was stronger in the group conflict
964 condition than in the individual conflict condition. Small volume corrections within
965 the mask of the left MFG defined using the Brainnetome Atlas (Region 22:
966 MFG_L_7_4_9 (46v)[30] were applied to search for significant results with the
967 threshold of $p<0.05$, FWE corrected for each voxel.

968 ***Bayesian factor analyses***

969 Bayesian factor analyses were used to compare the likelihood of the data fitting
970 under the null hypothesis with the likelihood of fitting under the alternative
971 hypothesis[23]. Compared with widely used p-values, the Bayes factor allows the
972 researcher to make a statement about the alternative hypothesis, rather than just the

null hypothesis. It also provides a clearer estimate of the amount of evidence present in the data. In the current paper, Bayes factor was calculated using the r toolbox "Bayesian factor" (https://richarddmorey.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})[64].

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 Neurosci.* **8**, 300-311 (2007).

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

9. Patil, I., Calò, M., Fornasier, F., Cushman, F. & Silani, G. The behavioral and neural basis of empathic blame. *Sci Rep.* **7**, 1-14 (2017).

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

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

12. Halevy, N., Bornstein, G. & Sagiv, L. "In-group love" and "out-group hate" as motives for individual participation in intergroup conflict: A new game paradigm. *Psychol Sci.* **19**, 405-411 (2008).

13. De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., ... & Feith, S. W. The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* **328**, 1408-1411 (2010).

14. De Dreu, C. K., Gross, J., Méder, Z., Giffin, M., Prochazkova, E., Krikeb, J. & Columbus, S. In-group defense, out-group aggression, and coordination failures in intergroup conflict. *Proc Natl Acad Sci USA.* **113**, 10524-10529 (2016).

15. Sääksvuori, L., Mappes, T. & Puurtinen, M. Costly punishment prevails in intergroup conflict. *Proc Biol Sci B*. **278**, 3428-3436 (2011).

16. Zhang, H., Gross, J., De Dreu, C. & Ma, Y. Oxytocin promotes coordinated out-

group attack during intergroup conflict in humans. *Elife* **8**, e40698 (2019).

17. Wolfe, R.J. & Darley, J.M. Protracted asymmetrical conflict erodes standards for avoiding civilian casualties, *Peace Conflict: J Peace Psychol.* **11**, 55–61 (2005).

18. Fehr, E. & Gächter, S. Fairness and retaliation: The economics of reciprocity. *J Econ Perspect.* **14**, 159-181 (2000).

19. Aron, A., Aron, E. N. & Smollan, D. Inclusion of other in the self scale and the structure of interpersonal closeness. *J Pers Soc Psychol*. **63**, 596-612 (1992).

20. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *J Edu Res.* **96**, 3-14 (2002).

21. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).

22. Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R. & Phelps, E. A. The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychol Sci.* **24**, 2498-2504 (2013).

23. Jarosz, A. F. & Wiley, J. What are the odds? A practical guide to computing and reporting Bayes factors. *J Prob Solving* **7**, 2-9 (2014).

24, Rouhana, N. N. & Bar-Tal, D. Psychological dynamics of intractable ethnonational conflicts: The Israeli–Palestinian case. *Ame Psychol.* **53**, 761-770. (1998).

25. Bekerman, Z. & Horenczyk, G. Arab‐Jewish bilingual coeducation in Israel: A long‐term approach to intergroup conflict resolution. *J Soc Issues* **60**, 389-404 (2004).

26. Shamay-Tsoory, S. G., Abu-Akel, A., Palgi, S., Sulieman, R., Fischer-Shofty, M., Levkovitz, Y. & Decety, J. Giving peace a chance: oxytocin increases empathy to pain in the context of the Israeli–Palestinian conflict. *Psychoneuroendocrinology* **38**,

1048 3139-3144 (2013).
1049 27. Feshbach, S. & Sakano, N. The structure and correlates of attitudes toward one's
1050 nation in samples of United States and Japanese college students: A comparative
1051 study. *Patriotism in the Lives of Individuals and Nations* 91-107 (1997).
1052 28. Aggestam, K. Mediating asymmetrical conflict. *Mediterranean Politics* **7**, 69-91
1053 (2002).
1054 29. Shnabel, N. & Nadler, A. A needs-based model of reconciliation: satisfying the
1055 differential emotional needs of victim and perpetrator as a key to promoting
1056 reconciliation. *J Pers Soc Psychol.* **94**, 116-132 (2008).
1057 30. Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S.,
1058 Laird, A.R. & Fox, P.T. The human brainnetome atlas: a new brain atlas based on
1059 connectional architecture. *Cereb Cortex* **26**, 3508-3526 (2016).
1060 31. Fullinwider, R. K. War and innocence. *Phi Pub Affairs* 90-97 (1975).
1061 32. Coates, A. J. *The Ethics of War*. Manchester, U.K. (Manchester Univer. Press,
1062 1997).
1063 33. Schiller, B., Baumgartner, T. & Knoch, D. Intergroup bias in third-party
1064 punishment stems from both ingroup favoritism and outgroup discrimination. *Evol*
1065 *Hum Behav.* **35**, 169-175 (2014).
1066 34. Bernhard, H., Fischbacher, U. & Fehr, E. Parochial altruism in humans. *Nature*
1067 **442**, 912-915 (2006).
1068 35. Delton, A. W. & Krasnow, M. M. The psychology of deterrence explains why
1069 group membership matters for third-party punishment. *Evol Hum Behav.* **38**, 734-743
1070 (2017).
1071 36. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol Hum*
1072 *Behav*. **25**, 63-87 (2004).

37. Volle, E., Kinkingnéhun, S., Pochon, J.B., Mondon, K., Thiebaut de Schotten, M., Seassau, M., Duffau, H., Samson, Y., Dubois, B. & Levy, R. The functional architecture of the left posterior and lateral prefrontal cortex in humans. *Cereb Cortex* **18**, 2460-2469 (2008).

38. Dixon, M. L. & Christoff, K. The lateral prefrontal cortex and complex value-based learning and decision making. *Neurosci Biobehav Rev.* **45**, 9-18 (2014).

39. Skunde, M., Walther, S., Simon, J. J., Wu, M., Bendszus, M., Herzog, W. & Friederich, H. C. Neural signature of behavioural inhibition in women with bulimia nervosa. *J Psychia Neurosci. JPN* **41**, E69. (2016).

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

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

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

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

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

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

46. Amodio, D. M. & Mendoza, S. A. Implicit intergroup bias: Cognitive, affective, and motivational underpinnings. In B. Gawronski & B. K. Payne (Eds.), Handbook of implicit social cognition: Measurement, theory, and applications. pp. 353–374 (The

Guilford Press, 2010).

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

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

49. Molenberghs, P. The neuroscience of in-group bias. *Neurosci Biobehav Rev.* **37**, 1530-1536 (2013).

50. Cikara, M., & Van Bavel, J. J. The neuroscience of intergroup relations: An integrative review. *Perspec Psychol Sci.* **9**, 245-274 (2014).

51. Levy, J., Goldstein, A., Influs, M., Masalha, S., Zagoory-Sharon, O. & Feldman, R. Adolescents growing up amidst intractable conflict attenuate brain response to pain of outgroup. *Proc Natl Acad Sci USA.* **113**, 13696-13701 (2016).

52. Han, S. Neurocognitive basis of racial ingroup bias in empathy. *Trends Cogn Sci.* **22**, 400-421 (2018).

53. Rusch, H. The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension. *Proc Biol Sci B.* **281**, 20141539 (2014).

54. Abbink, K., Brandts, J., Herrmann, B. & Orzen, H. Intergroup conflict and intra-group punishment in an experimental contest game. *Ame Econ Rev.* **100**, 420-47 (2010).

55. Battigalli, P. & Dufwenberg, M. Dynamic psychological games. *J Econ Theory* **144**, 1-35 (2009).

56. Nisbett, R. E. & Masuda, T. Culture and point of view. *Proc Natl Acad Sci USA.* **100**, 11163-11170 (2003).

1123 57. Han, S., & Northoff, G. Culture-sensitive neural substrates of human cognition: A
1124 transcultural neuroimaging approach. *Nat Rev Neurosci* **9**, 646-654 (2008).
1125 58. Han, S. *The sociocultural brain: A cultural neuroscience approach to human*
1126 *nature.* (Oxford Univ. Press, 2017).
1127 59. Han, S., & Ma, Y. A culture–behavior–brain loop model of human development.
1128 *Trends Cogn Sci.* **19**, 666-676 (2015).
1129 60. Ma, Y., Wang, C., Li, B., Zhang, W., Rao, Y. & Han, S. Does self-construal
1130 predict activity in the social brain network? A genetic moderation effect. *Soc Cogn*
1131 *Affect Neurosci.* **9**, 1360-1367 (2014).
1132 61. Luo, S., Ma, Y., Liu, Y., Li, B., Wang, C., Shi, Z. ... & Han, S. Interaction
1133 between oxytocin receptor polymorphism and interdependent culture values on
1134 human empathy. *Soc Cogn Affect Neurosci.* **10**, 1273-1281 (2015).
1135 62. Buckholtz, J. W., Martin, J. W., Treadway, M. T., Jan, K., Zald, D. H., Jones, O.
1136 & Marois, R. From blame to punishment: disrupting prefrontal cortex activity reveals
1137 norm enforcement mechanisms. *Neuron* **87**, 1369-1380 (2015).
1138 63. Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. Statistical power analyses
1139 using G* Power 3.1: Tests for correlation and regression analyses. *Behav Res*
1140 *Methods* **41**, 1149-1160 (2009).
1141 64. Raftery, A. E. Bayesian model selection in social research. *Sociol Meth.* **25**, 111-
1142 164 (1995).
1143 65. Maoz, I. & McCauley, C. Threat perceptions and feelings as predictors of Jewish-
1144 Israeli support for compromise with Palestinians. *J Peace Res.* **46**, 525-539 (2009).
1145 66. Halperin, E., Porat, R., Tamir, M. & Gross, J. J. Can emotion regulation change
1146 political attitudes in intractable conflicts? From the laboratory to the field. *Psychol Sci.*
1147 **24**, 106-111 (2013).

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

Data availability

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

Acknowledgements

This research was supported by the National Natural Science Foundation of China (projects 31871134, 31421003, and 31661143039) and the Ministry of Science and Technology of China (2019YFA0707103). 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

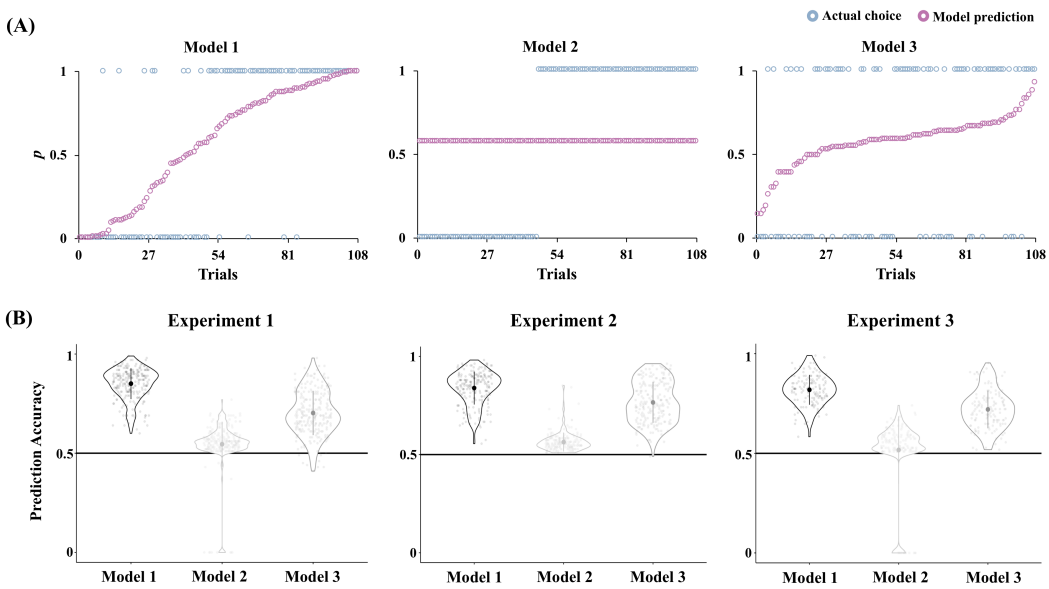
X. H. and S.H. conceived the research programme and designed the experiments. X. H., S. Z., N. F., T. W., T. G., S. S., X. W., S. H. carried out the experiments. X. H., N. F., and S. H. analyzed the data. X. H., S. H., and S. S. wrote the paper. M.J.G reviewed and edited the paper. S.H. supervised the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

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper.

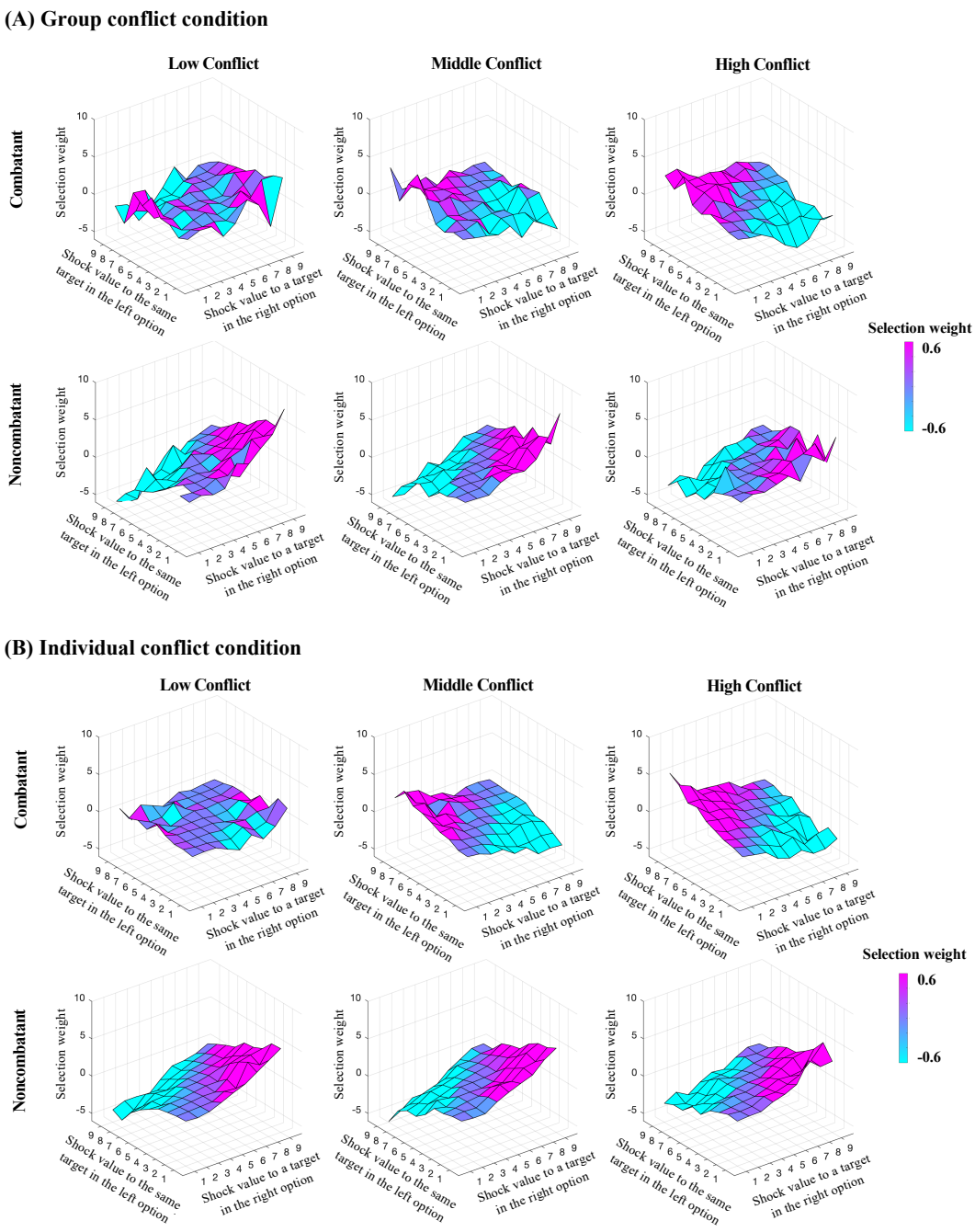


Extended Data Fig. 1. Prediction accuracies of different models of participants' choices of electric shocks. (A) An example data from a participant in the group conflict condition in Experiment 1. *p* represents probability to choose the left option. If $p > 0.5$, the model predicts selection of the left option; otherwise, the model predicts selection of the right option. The violet circle represents the probability of choosing the left option by the corresponding models. The blue circle represents a participant's real choice in that trial. A blue circle with a value of zero along y-axis indicates a choice of the right option whereas a value of 1 indicates a choice of the left option. (B) Prediction accuracies of the three testing models. The prediction accuracy of Model 1 is the best, and Model 2 is the worst. Violin plots show the distributions of prediction accuracy with the large dots showing the means of the prediction accuracy and a vertical bar indicating standard deviation of the prediction accuracy. Each small gray dot represents one participant. see Methods for details of the three testing models.

Extended Data Table 1. Results of model comparisons in three experiments.

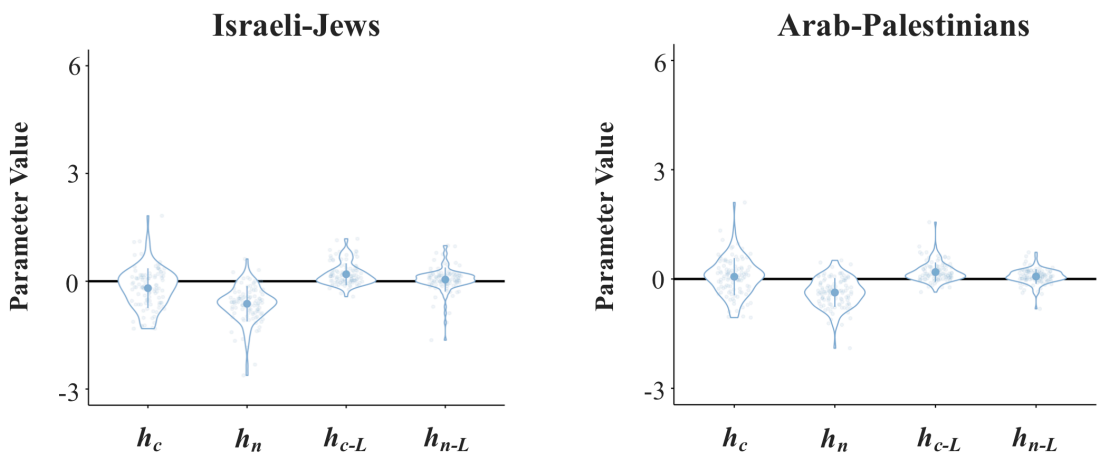
Experiment	Model	BIC	Prediction accuracy (mean±s.d.)	AUC (mean±s.d.)
1	Model 1	18159.58	0.85±0.08	0.91±0.08
	Model 2	35584.88	0.54±0.11	0.50±0.00
	Model 3	27061.69	0.70±0.11	0.77±0.12
2	Model 1	17076.65	0.84±0.08	0.89±0.09
	Model 2	31442.07	0.56±0.05	0.50±0.00
	Model 3	21669.25	0.77±0.10	0.81±0.12
3	Model 1	10456.98	0.82±0.08	0.88±0.07
	Model 2	20170.1	0.52±0.17	0.50±0.00
	Model 3	15231.57	0.72±0.09	0.78±0.11

Notes. All the results of BIC, prediction accuracy, and AUC support that Model 1 is the best model of the participants' punishment decisions across three Experiments. Smaller BIC but larger prediction accuracy and AUC indicate better model fit. BIC=Bayesian Information Criterion; AUC=area under the curve.

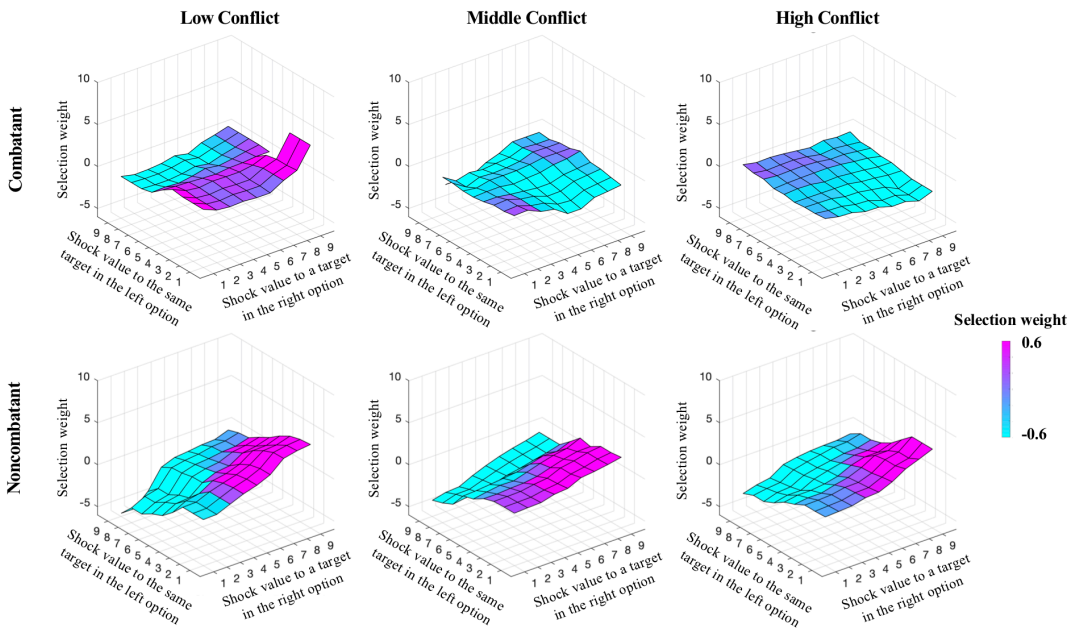


Extended Data Fig. 2. Illustration of selection weight of the left and right options corresponding to specific shock intensities shown in Phase 4 in Experiment 1. Selection weight was calculated as: Selection weight $= \frac{1}{n} * \sum_{i=1}^{n} \text{Choice}_i * \text{abs}(\Delta S_i)$; n = number of participants; ΔS_i = shock value to a target in the left option minus shock value to the same target in the right option of the participant *i*; Choice_i = 1 if the participant *i* choosing the left option of punishment on one trial; Choice_i = -1 if the participant *i* choosing the right option of punishment on one trial; abs = absolute value. Selection weight represents tendencies to select the left or right options when considering the difference in shock intensity related to combatant (or noncombatant) between the left and right options across all participants. (A) Selection weight in the group conflict condition. Figures

illustrate a sharp pattern of selection of the left or right choice in correspondence with the magnitudes of shocks to outgroup combatants particularly when conflict was high. By contrast, there is a sharp pattern of selection of the left or right choice in correspondence with the magnitudes of shocks to outgroup noncombatants particularly when conflict was low. (B) Selection weight in the individual conflict condition. Similarly, figures illustrate a sharp pattern of selection of the left or right choice in correspondence with the magnitudes of shocks to outgroup combatants particularly when conflict was high. However, there are similar sharp patterns of selection of the left or right choice in correspondence with the magnitudes of shocks to outgroup noncombatants regardless conflict levels.



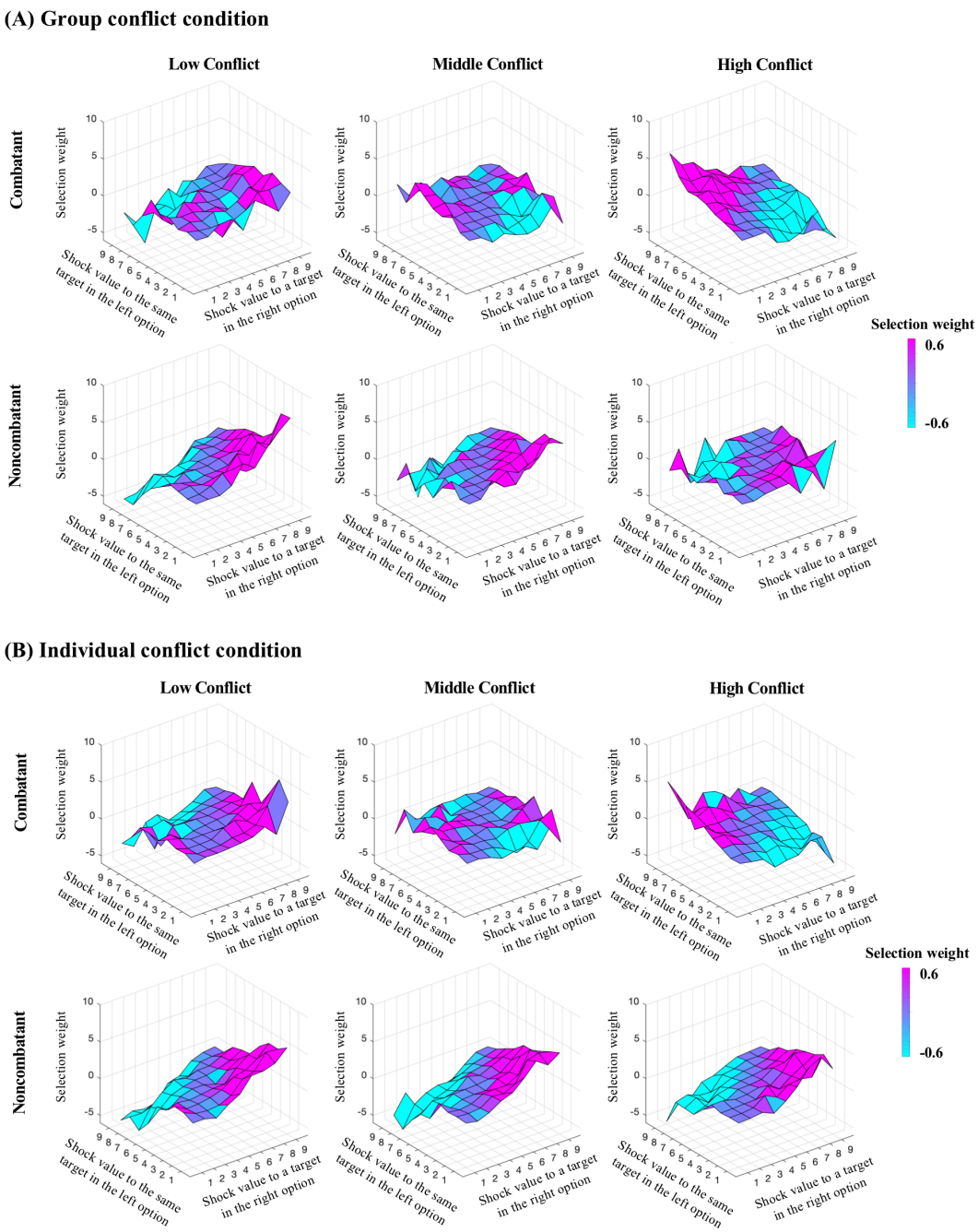
Extended Data Fig. 3. Separate logistic regression parameters in Israeli-Jews and Arab-Palestinians in Experiment 2. Violin plots show means (big dots), s.d. (bars), and distributions of parameter values. Each dot represents an individual participant.



Extended Data Fig. 4. Illustration of selection weight of the left and right options corresponding to specific shock intensities shown in Phase 4 in Experiment 2. Selection weight is calculated in the same way as that in Experiment 1. Missing selection weights due to random generation of options of shock intensities, which did not influence the logistic regression modeling, were calculated as the average of the surrounding selection weights.

Extended Data Table 2. Group-level statistical details of logistic regression parameters in Experiment 3.

	mean(s.d.)	t	*P*	Cohen's *d*	95% CI	BF_{10}
Group conflict condition						
h_c	0.18(0.77)	1.79	0.078	0.23	-0.02, 0.38	0.63
h_n	-0.28(0.27)	-8.04	< 0.001	1.04	-0.35, -0.21	>150
$h_{c\text{-}L}$	0.36(0.54)	5.20	< 0.001	0.67	0.22, 0.50	>150
$h_{n\text{-}L}$	0.12(0.18)	5.36	< 0.001	0.69	0.08, 0.17	>150
Individual conflict condition						
h_c	-0.01(0.43)	-0.10	0.919	0.01	-0.12, 0.11	0.14
h_n	-0.50(0.49)	-7.83	< 0.001	1.02	-0.62, -0.37	>150
$h_{c\text{-}L}$	0.27(0.28)	7.25	< 0.001	0.94	0.19, 0.34	>150
$h_{n\text{-}L}$	0.02±0.19	0.94	0.467	0.12	-0.03, 0.07	0.22
Differences between group and individual conflict conditions						
h_c		1.61	0.148	0.29	-0.04, 0.41	0.62
h_n		3.06	0.008	0.56	-0.36, -0.08	12.19
$h_{c\text{-}L}$		1.19	0.236	0.22	-0.06, 0.25	0.37
$h_{n\text{-}L}$		2.97	0.008	0.54	0.03, 0.17	9.53



Extended Data Fig. 5. Illustration of selection weight of the left and right options corresponding to specific shock intensities shown in Phase 4 in Experiment 3. Selection weight is calculated in the same way as that in Experiment 1. (A) Selection weight in the group conflict condition. (B) Selection weight in the individual conflict condition.