# Manuscript Details

| | |
|---|---|
| **Manuscript number** | EVOLHUMBEHAV_2017_26 |
| **Title** | Humans reciprocate intentional harm by discriminating against group peers |
| **Article type** | Research paper |

**Abstract**

Cycles of intergroup revenge appear in large scale conflicts. We experimentally test the hypothesis that humans practice group-based reciprocity: if someone harms or helps them, they harm or help other members of that person's group. Subjects played a trust game, then allocated money between other people. Senders whose partners returned more in the trust game gave more to that partner's group members. The effect was about half as large as the effect of direct reciprocity. Receivers' allocations to group members were not affected by their partners' play in the trust game, suggesting that group reciprocity was only triggered when the partner's intentions were unequivocal.

| | |
|---|---|
| **Keywords** | Upstream reciprocity; group identity; intergroup conflict. |

| | |
|---|---|
| **Corresponding Author** | Ro'i Zultan |
| **Corresponding Author's Institution** | Ben-Gurion University |
| **Order of Authors** | David Hugh-Jones, Itay Ron, Ro'i Zultan |

| | |
|---|---|
| **Suggested reviewers** | Simon Gächter, Yan Chen, Moses Shayo, Carsten De Dreu, Peter Richerson, Urs Fischbacher, Shaul Shalvi |

## Submission Files Included in this PDF

**File Name [File Type]**

just_cover.pdf [Cover Letter]

Cover_EHB.pdf [Response to Reviewers]

gr_EHB-ER.pdf [Manuscript (without Author Details)]

title page.pdf [Title Page (with Author Details)]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

April 7, 2017

Dr. Daniel Hruschka
Editor
*Evolution and Human Behavior*

Re: Manuscript EVOLHUMBEHAV_2017_26

Dear Dr. Hruschka,

Thank you for the opportunity to revise our manuscript. In the following, we list our responses to your comments. Comments are presented in italics, and quotes from the new manuscript in frames.

Sincerely,

David Hugh-Jones, Itay Ron and Ro'i Zultan

1. *Why would group-based upstream reciprocity be important for the evolution of human behavior? How does this contribute to current debates in the literature about the evolution of human behavior?*

We expanded on the relevance of group-based upstream reciprocity to human evolution:

> Intergroup reciprocity could be important for human evolution. First, it may structure intergroup conflicts, just as individual reciprocity structures inter-individual conflict. Existing work suggests that intergroup conflict may be important for the development of (parochial) altruism, since it increases intergroup variation in fitness. But this explanation is incomplete without an understanding of what regulates groups' decisions to initiate or cease conflict. Warfare is costly and dangerous, but pacifist groups risk being victimized by others. Groups that practice group reciprocity can balance the risks of conflict against the risk of not responding to aggression.
> Second, group reciprocity could provide an evolutionary basis for outgroup stereotyping. Upstream group reciprocity has different cognitive requirements from related phenomena. While ingroup altruism and group-based downstream reciprocity require people to differentiate their own group from outsiders—"us" from "them"—upstream group reciprocity requires them to differentiate between different outgroups—between "them and them"—and to keep a mental account of outgroups' reputation.

In the discussion, we relate to recent work on the evolution of intergroup conflict:

> Our results show that upstream reciprocity is moderated by social boundaries. Humans respond to harms from outgroup members by discriminating against others in that specific outgroup. This supports the argument of Pietraszewski (2016) that group identity can modify the cost-benefit calculus of individuals deciding whether to extend a conflict. Unlike parochial altruism and within-group reciprocity, group reciprocity requires humans to differentiate between outgroups, possibly providing a cognitive basis for intergroup stereotyping and prejudice.

2. *Please explain clearly in the methods (before getting to the results) what the the key outcomes are and how they are calculated (and why they are important). The "discrimination" measure especially needs some more justification/explanation.*

We added the following paragraphs:

The key outcomes in this design are based on the allocation decisions made in the third stage. Direct and group reciprocity can be both positive and negative, and therefore are not hypothesized to have a systematic effect on the the amount allocated to either the TG partner or to his team mates. Nonetheless—while there is arguably no reason to discriminate between two neutral players—we hypothesize that direct and group reciprocity will lead the allocator to discriminate either for or against the TG partner or his team mates. Consequently, we predict that the absolute difference between the two allocations will be larger in all treatments compared to the baseline. This difference is measured in our 'Discrimination' outcome.

We measure reciprocity directly by looking at the effect of the TG experience in the second stage on allocations made in the third stage. We define the experience with the TG partner in two ways. For responders, this is the amount sent to them by their partner. For senders, we calculate the amount returned to them by their partner as a fraction of the money available to the responder. Thus, an equal split of the pie implies a value of 1/2, and compensating the sender for his investment implies a value of 1/3. We subsequently define (direct or group) reciprocity as the slope of the allocation made to the TG partner or his team mates on the TG experience.

3. *In the results, it is difficult to understand why you are analyzing "discrimination". It should be presented in its own paragraph, with a clearer interpretation of what the results mean.*

We rephrased this part of the results thus:

The first column in Table 2 presents the mean allocations. Participants gave significantly more to members of their own team at the expense of the neutral recipient ($z = 3.63, p < 0.001$ for senders, $z = 3.59, p < 0.001$ for responders), establishing that our group formation manipulation was successful in inducing group identity and triggering ingroup favouritism. Allocations to the TG partner and his team mates were not significantly different to the baseline 35 ($p > 0.47$ for all comparisons). This result suggests that the experience with the TG partner is, on average, neutral, such that positive and negative experiences balance each other overall.

Nonetheless, both positive and negative treatment of the TG partner or his team mates increase the absolute difference between the two allocations. Indeed, column of Table 2 shows that allocators discriminated significantly more than in the baseline both when interacting with their TG partner ($z = 9.08, p < 0.001$) and with his team mates ($z = 3.93, p < 0.001$). This effect was not significantly different between TG senders and receivers (F test $0.50, p = 0.68$).

4. *Avoid interpreting amount sent and received in terms of intentions ("kindness"). This is confusing and potentially inaccurate. A sender might send a lot because they "trust" the responder to send a bunch back. A responder might send a lot back because of "gratitude." It's not at all clear that the behaviors reflect "kindness."*

We followed the convention established in psychological game theory, that "Given the belief of player $i$ about the strategy choice of the other player $j$, $i$ is kind to the extent that he believes he gives $j$ a (relatively) high material payoff." (Dufwenberg and Kirchsteiger, 2004, following Rabin, 1993). We have now changed the terminology to talk more generally about the subjects' experience in the trust game (TG experience).

5. *The findings should be qualified with a discussion of potential crosscultural variation in how perceived intentions affect moral judgment (see Barrett et al. 2016). Do you expect the difference in group reciprocity between senders and responders to hold across all human groups?*

We added the following to the discussion:

> Since our study was conducted with students from a rich industrialized democracy, results may not generalize to all cultures (Henrich et al., 2010). In particular, the link between intentions and moral judgment may vary across cultures (Barrett et al., 2016), and this could affect how group reciprocity plays out in different societies.

6. *You should very clearly state your theoretical expectations for the analyses in the introduction (including the analysis of social value orientation). And if you are going to present the SVO results, then you should devote more time to why you expected SVO would matter, and also how SVO is related to the different behaviors and outcomes in the game.*

We address this in the following:

> Our expectations were as follows. First, in Direct Reciprocity rounds, individuals' allocations to their TG partner should positively covary with the amount the partner sent (or returned) in the Trust Game. This simply comes from from the well-known theory of direct reciprocity. Second, if group reciprocity is present, then allocations to the TG partner's group member, in Group Reciprocity rounds, should also covary with the amount sent or returned by the TG partner. We also measured participants' social value orientation (Van Lange, 1999). It is plausible that willingness to group-reciprocate should be linked to other social preferences. We were not certain *a priori* whether group reciprocity would be stronger among selfish or among prosocial types. On the one hand, both prosociality and group reciprocity can be seen as actions that benefit the group, by providing support to in-group members or protecting it from outgroups. On the other hand, negative reciprocity in general may be linked to spite (Johnstone and Bshary, 2004). So we test a non-directional hypothesis here.

7. *As a point of comparison, it would be helpful to include the same figures as in Figure 2, but for allocations in the direct reciprocity situation.*

Done.

## References


Barrett, H. C., A. Bolyanatz, A. N. Crittenden, D. M. Fessler, S. Fitzpatrick, M. Gurven, J. Henrich, M. Kanovsky, G. Kushnick, A. Pisor, et al. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences 113*(17), 4688–4693.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Game Econ Behav 47*(2), 268–298.

Henrich, J., S. J. Heine, and A. Norenzayan (2010). Most people are not weird. *Nature 466*(7302), 29–29.

Johnstone, R. A. and R. Bshary (2004). Evolution of spite through indirect reciprocity. *Proceedings of the Royal Society of London B: Biological Sciences 271*(1551), 1917–1922.

Pietraszewski, D. (2016). How the mind sees coalitional and group conflict: the evolutionary invariances of n-person conflict dynamics. *Evolution and Human Behavior 37*(6), 470 – 480.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *Am Econ Rev 83*(5), 1281–1302.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of personality and social psychology 77*(2), 337.

# Humans reciprocate intentional harm by discriminating against group peers

AUTHORS UNDISCLOSED

**Abstract**

Cycles of intergroup revenge appear in large scale conflicts. We experimentally test the hypothesis that humans practice group-based reciprocity: if someone harms or helps them, they harm or help other members of that person's group. Subjects played a trust game, then allocated money between other people. Senders whose partners returned more in the trust game gave more to that partner's group members. The effect was about half as large as the effect of direct reciprocity. Receivers' allocations to group members were not affected by their partners' play in the trust game, suggesting that group reciprocity was only triggered when the partner's intentions were unequivocal.

*Keywords:* Upstream reciprocity, group identity, intergroup conflict.
Word count: 3129

# 1 Introduction

Human society is organized in groups, including families, clans, firms and nations. This structure is reflected in individual behaviour and cognition. Humans identify with their ingroup and are altruistic and prosocial towards ingroup members; towards outgroup members, they display stereotyping and prejudice (Balliet, Wu, and De Dreu, 2014; Chen and Chen, 2011; Chen and Li, 2009; De Dreu, Balliet, and Halevy, 2014; Tajfel and Turner, 1979; Yamagishi and Kiyonari, 2000). Group structure provides the backdrop for intergroup conflict—from economic and political competition to inter-ethnic violence and war—which is pervasive in the species (World Bank, 2011).

Intergroup conflicts often follow a tit-for-tat logic, in which one group's violence leads to revenge from the other side (Chagnon, 1988; Haushofer, Biletzki, and Kanwisher, 2010; Horowitz, 1985; Horowitz, 2001; Shayo and Zussman, 2010). This suggests that humans practice intergroup *reciprocity*. Reciprocity is a well-known mechanism that may underlie the evolution of cooperation (Nowak, 2006, 2012). While in direct reciprocity, individuals help those who have helped them in the past (and similarly for harm), in indirect reciprocity, individuals help or harm other people than those who have helped them. Indirect reciprocity comes in two flavours: *downstream* reciprocity follows the maxim 'do unto thy neighbour as they have done to others', whereas *upstream* reciprocity follows the maxim 'do unto thy neighbour as others have done unto you'.

Compared to downstream reciprocity, upstream reciprocity is cognitively easier to implement, as it does not require tracking individual reputations, but is more difficult to understand from an evolutionary point of view (Boyd and Richerson, 1989; Nowak and Sigmund, 2005). Nonetheless, upstream reciprocity can co-evolve with direct or spatial reciprocity (Nowak and Roch, 2007). Furthermore, laboratory experiments provide positive evidence for upstream reciprocity: individuals are more generous to others if a third party was generous to them (Dufwenberg, Gneezy, Güth, and van Damme, 2001; Greiner and Levati, 2005; Güth, Königstein, Marchand, and Nehring, 2001), and the mere possibility of being harmed by a third party reduces cooperation in a social

dilemma (Weisel and Zultan, 2016).

In this paper we examine group-based upstream reciprocity, or *group reciprocity*. That is, an individual who is harmed (helped) by a member of an outgroup becomes more likely to harm (help) others from that group. Whereas group-based downstream reciprocity (Bernhard, Fehr, and Fischbacher, 2006; Bernhard, Fischbacher, and Fehr, 2006) follows the maxim 'do unto others as they have done to members of *my* tribe', group-based upstream reciprocity follows the maxim 'do unto others as members of *their* tribe have done to me' (Figure 1). Both up- and downstream group reciprocity can expand the scope of conflict, from individual level to group level. While (group-based) downstream reciprocity can bring a victim's groupmates into a conflict as new aggressors, upstream reciprocity can bring in an aggressor's groupmates, as new victims (Pietraszewski, 2016).

Intergroup reciprocity could be important for human evolution. First, it may structure intergroup conflicts, just as individual reciprocity structures interindividual conflict. Existing work suggests that intergroup conflict may be important for the development of (parochial) altruism, since it increases intergroup variation in fitness. But this explanation is incomplete without an understanding of what regulates groups' decisions to initiate or cease conflict. Warfare is costly and dangerous, but pacifist groups risk being victimized by others. Groups that practice group reciprocity can balance the risks of conflict against the risk of not responding to aggression.

Second, group reciprocity could provide an evolutionary basis for outgroup stereotyping. Upstream group reciprocity has different cognitive requirements from related phenomena. While ingroup altruism and group-based downstream reciprocity require people to differentiate their own group from outsiders—"us" from "them"—upstream group reciprocity requires them to differentiate between different outgroups—between "them and them"—and to keep a mental account of outgroups' reputation.

We ran a laboratory experiment to test the hypothesis that people reciprocate towards groups. Although field observations from conflict are highly suggestive, they are loaded with individual and group context and history. Observing group reciprocity under controlled laboratory conditions with artifi-

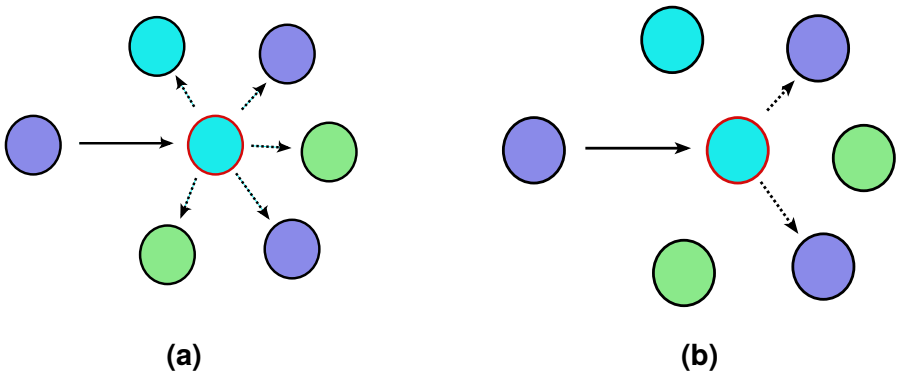



**Figure 1:** Upstream reciprocity. (*a*) Someone who was helped or harmed becomes more likely to help or harm others. (*b*) Upstream group reciprocity targets people who belong to the same group as the initial partner.

cal groups identifies group reciprocity as an innate human tendency. Cleanly identifying group reciprocity requires controlling for three confounds: individual level reciprocity, e.g. if subjects' actions affect an entire group including the original actor who helped or harmed them; generalized reciprocity, where subjects reciprocate not specifically towards the original actor's group, but towards other people in general; and strategic interactions, where apparent reciprocity is driven by reputation-building. our experiment fulfils all three: subjects can differentiate the original actor from his or her group members, they interact both with these group members and with members of other groups, and we minimize strategic concerns by not giving feedback about subjects' actions.

While previous studies looked at retaliation towards groups, this retaliation does not necessarily reflect group reciprocity as defined here. Gaertner, Iuzzini, and O'Mara (2008) found that rejection by one group member leads to more hostility towards the group when the group is perceived as a unified entity. Since hostility was directed towards the whole group, individual and group level reciprocity were confounded. More importantly, manipulating the entitativity of the group creates a context in which the initial rejection can be perceived as a group action. Similarly, Stenstrom, Lickel, Denson, and Miller (2008) manipulated entitativity by making the original perpetrator (a political

analyst) an official affiliate of the group (a presidential campaign). Thus, holding the group accountable for its member's action is justified without resorting to group reciprocity. In contrast, we look at how people reciprocate a clear individual act by one group member to an unrelated other group member, where group structure is minimal and free of existing social context.

Our experimental set up was the following. After an initial group-formation stage, participants interacted in two strategic stages. The upstream action, in which the individual could be helped or harmed by another person, was represented by a Trust Game (TG) (Berg, Dickhaut, and McCabe, 1995). In this game, the Sender (S) receives 150 money-equivalent tokens, and chooses how many of them to send to the Responder (R). The amount sent is multiplied by a factor of 3, so that R receives between 0 and 450 tokens, of which he can send any number back to S. The TG enables us to model two types of interactions. Whereas R is clearly kind when returning money (and nasty when exploiting a generous proposer by keeping the received amount), S's intentions are equivocal. Sending money can be driven by selfish expectations of reciprocity, while not sending can be driven by caution. Thus, while all subjects experience helpful or harmful actions, only senders experience actions that clearly reflect their counterpart's preferences and intentions (Gunnthorsdottir, McCabe, and Smith, 2002; Kimbrough and Vostroknutov, 2015).

The upstream action was followed by the reciprocal action, in which the individual could help others. We implemented this as an Allocation Game in which subjects divided a fixed amount between two recipients. In Direct Reciprocity rounds, the recipients included the TG partner; in Group Reciprocity rounds, a member of the TG partner's group; and in Ingroup Favoritism rounds, a member of the allocator's group. The other recipient was always a member of a third, neutral, group. Baseline rounds included two neutral recipients, to test whether the TG experience leads to arbitrary discrimination in the absence of any reciprocal or group motivations.

Our expectations were as follows. First, in Direct Reciprocity rounds, individuals' allocations to their TG partner should positively covary with the amount the partner sent (or returned) in the Trust Game. This simply comes from from the well-known theory of direct reciprocity. Second, if group reciprocity is

present, then allocations to the TG partner's group member, in Group Reciprocity rounds, should also covary with the amount sent or returned by the TG partner. We also measured participants' social value orientation (Van Lange, 1999). It is plausible that willingness to group-reciprocate should be linked to other social preferences. We were not certain *a priori* whether group reciprocity would be stronger among selfish or among prosocial types. On the one hand, both prosociality and group reciprocity can be seen as actions that benefit the group, by providing support to ingroup members or protecting it from outgroups. On the other hand, negative reciprocity in general may be linked to spite (Johnstone and Bshary, 2004). So we test a non-directional hypothesis here.

## 2 Material and methods

Each session consisted of 24 participants, randomly allocated into six *teams* of four. Each participant was identified throughout the experiment by team colour and individual number (1–4) within the team. At the beginning of the experiment, participants were informed that the experiment had five distinct stages, and that they might interact with the same people in different stages. Specific instructions for each stage were distributed and read aloud at the beginning of the stage. The five stages were a group formation stage, the TG stage, the Allocation Game stage, a social value orientation elicitation stage (Murphy, Ackermann, and Handgraaf, 2011) and a collectivism scale measurement stage (adapted from the horizontal collectivism scale in Singelis, Triandis, Bhawuk, and Gelfand, 1995).

Following (Chen and Li, 2009), we created group identity in the first stage by allowing participants to consult each other by anonymous chat while solving a simple task. Participants solved five Raven matrices (see supplementary material). Each matrix was presented on screen for 120 seconds, during which each participant could both send written messages to the team and update her own answer. The final answer submitted at the end of the 120 seconds determined payoffs, with 10 tokens paid for each correct answer. To further boost

group identity through a common goal, team members each earned an additional bonus of 5 tokens if all four team members answered correctly.

Next, participants were rematched into pairs to play the one-shot TG. To facilitate understanding, participants played five practice rounds, in which they entered decisions both as S and as R. In the actual interaction, participants could see their TG partner's team colour and individual number.

The third stage Allocation Game consisted of six rounds. In each round, participants interacted in groups of three. Individuals in each group were identified to each other by team colour and number. Each round consisted of a random dictator game, as follows. Each player in the group of three had to allocate 100 tokens within the group. The allocator received a fixed 30 tokens, and could freely allocate the remaining 70 tokens between the other two players. Previous research has found that people do not harm, but refrain from helping negatively perceived outgroups (Weisel and Böhm, 2015). Accordingly, we set the parameters of the game so that, compared to the reference point of the allocator's own share, an equal division benefits both other players. Table 1 shows the matching scheme over the six rouds. Each participant was in the same group of three in one of the six rounds with a member of her own team (*ingroup* condition), in one round with her TG partner (*direct reciprocity* condition), and in two rounds with other members of the TG partner's team (*group reciprocity* condition). The remaining two rounds served as the baseline condition. No feedback was provided between rounds. Stage payoffs were determined by one randomly chosen round of the six rounds, and the allocation decision of one randomly chosen player in each group. Note that the matching is not independent. For example, if one player is in the direct reciprocity condition, then one other player is in the direct reciprocity condition and the third player is in either the baseline or group reciprocity condition.

The fourth stage implemented the slider measure of social value orientation (Crosetto, Weisel, and Winter, 2012; Murphy, Ackermann, and Handgraaf, 2011), in which participants choose nine allocations between themselves and another person. For consistency with the previous stages, the team identity of the partner was known. To keep the decision independent of previous experience with the different teams, we matched participants within teams. There-

**Table 1:** Matching example

| Round | Allocates to | | | Treatment |
|---|---|---|---|---|
| 1 | Red 1 | / | Yellow 1 | Group reciprocity (GR) |
| 2 | Yellow 4 | / | Brown 2 | Group reciprocity (GR) |
| 3 | Green 3 | / | Yellow 2 | Direct reciprocity (DR) |
| 4 | Red 1 | / | Brown 1 | Baseline (B) |
| 5 | Brown 2 | / | Brown 4 | Baseline (B) |
| 6 | Blue 3 | / | Green 2 | Ingroup (IG) |

Note: Example treatments shown for player Blue 2, who played the TG with Yellow 2 (see the supplementary material for the full matching scheme).

fore, this measure captures within-group social value orientation. Payoffs were determined by one randomly chosen decision of the nine decisions made by one randomly chosen player in each dyad. The decisions yielded a social orientation angle for each participant, with $0°$ corresponding to selfishness, $45°$ to pure altruism, and negative angles to spitefulness.

After the fifth and final stage (a non-strategic and non-incentivised collectivism measurement), participants learned their cumulative payoff in tokens and were paid in private. One hundred and ninety two participants, recruited using ORSEE (Greiner, 2015) participated in eight sessions conducted between June 2014 and January 2015. The experiment was programmed in z-Tree (Fischbacher, 2007).

The key outcomes in this design are based on the allocation decisions made in the third stage. Direct and group reciprocity can be both positive and negative, and therefore are not hypothesized to have a systematic effect on the the amount allocated to either the TG partner or to his team mates. Nonetheless—while there is arguably no reason to discriminate between two neutral players—we hypothesize that direct and group reciprocity will lead the allocator to discriminate either for or against the TG partner or his team mates. Consequently, we predict that the absolute difference between the two allocations will be larger in all treatments compared to the baseline. This difference is measured in our 'Discrimination' outcome.

**Table 2:** Allocations and Discrimination

| | Allocation | Discrimination | Reciprocity |
|---|---|---|---|
| **Senders** | | | |
| Baseline | 35.00 (—) | 4.15 (0.97) | — |
| Direct Reciprocity | 33.98 (2.30) | 22.00 (1.51) *** | 15.64 (5.12)** |
| Group Reciprocity | 34.39 (0.77) | 8.08 (1.61) *** | 7.78 (2.37)** |
| In-Group | 38.98 (1.11) *** | 15.46 (2.99) *** | 0.20 (5.50) |
| **Responders** | | | |
| Baseline | 35.00 (—) | 2.25 (0.51) | — |
| Direct Reciprocity | 35.38 (1.08) | 22.17 (2.30) *** | 20.87 (6.04)*** |
| Group Reciprocity | 34.79 (0.62) | 6.12 (1.51) ** | 1.20 (2.08) |
| In-Group | 42.13 (1.99) *** | 17.20 (3.40) *** | 4.72 (7.62) |

Mean allocation, mean discrimination, and reciprocity (marginal effect of TG partner's kindness on allocation) by condition. Robust standard errors clustered on sessions. Significance of comparison to Baseline is marked.*, **, and *** indicate $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively.

We measure reciprocity directly by looking at the effect of the TG experience in the second stage on allocations made in the third stage. We define the experience with the TG partner in two ways. For responders, this is the amount sent to them by their partner. For senders, we calculate the amount returned to them by their partner as a fraction of the money available to the responder. Thus, an equal split of the pie implies a value of 1/2, and compensating the sender for his investment implies a value of 1/3. We subsequently define (direct or group) reciprocity as the slope of the allocation made to the TG partner or his team mates on the TG experience.

# 3 Results

We report results on allocations, discrimination between recipients (measured as the absolute difference between the two recipients' allocations), and direct and group reciprocity. All reported statistical tests are based on mixed-effects regressions with bootstrapped standard errors clustered on subjects.

The first column in Table 2 presents the mean allocations. Participants gave significantly more to members of their own team at the expense of the neu-

tral recipient ($z = 3.63, p < 0.001$ for senders, $z = 3.59, p < 0.001$ for responders), establishing that our group formation manipulation was successful in inducing group identity and triggering ingroup favouritism. Allocations to the TG partner and his team mates were not significantly different to the baseline 35 ($p > 0.47$ for all comparisons). This result suggests that the experience with the TG partner is, on average, neutral, such that positive and negative experiences balance each other overall.

Nonetheless, both positive and negative treatment of the TG partner or his team mates increase the absolute difference between the two allocationa. Indeed, column of Table 2 shows that allocators discriminated significantly more than in the baseline both when interacting with their TG partner ($z = 9.08, p < 0.001$) and with his team mates ($z = 3.93, p < 0.001$). This effect was not significantly different between TG senders and receivers (F test $0.50, p = 0.68$).

## *3.1 Direct and group reciprocity*

The third column of Table 2, *Reciprocity*, reports the slope of allocations regressed on the subjects' experience with their TG partners. The responder's experience with the sender is measured as the share of the endowment that the sender chose to send. The sender's experience with the responder is measured as the share of the received amount that the responder chose to send back. The sender's experience was not defined for the six (out of 96) senders who did not send any money. There is strong direct reciprocity: allocations to the TG partners increase with the TG experience both for senders ($z = 3.06, p < 0.01$) and for responders ($z = 3.46, p < 0.001$).

Group reciprocity, however, is only observed for senders, who allocate less to team mates of a responder who returned less—an intentionally harmful action. Responders, on the other hand, although directly reciprocating the TG partner's action, do not systematically discriminate against team mates of a sender who sent little—a harmful action that does not unequivocally signal a bad intention. The regression analysis shows no significant effect of the responder's TG experience on her allocation to the sender's team mates ($z = 0.58, p = 0.56$). The sender's TG experience, on the other hand, significantly in-

creases the allocations made to the responder's team mates ($z = 3.29, p < 0.01$). The estimated ratio of the group and direct reciprocity coefficients is 50%, so that for every allocation dollar a responder loses due to an unkind action in the TG, his team mates lose 50 cents. This relationship is presented graphically in Figure 2 (the corresponding figure for direct reciprocity is included in the supplementary material).

Senders' group reciprocity was related to their social value orientation. The slope of the effect of the TG experience on allocations was 15.97 for those with less than median SVO, and -1.06 for those with median or greater SVO (interaction, $p = 0.061$).These results should be interpreted cautiously, since both scores were affected by the TG experience.

# 4 Discussion

Our results show that upstream reciprocity is moderated by social boundaries. Humans respond to harms from outgroup members by discriminating against others in that specific outgroup. This supports the argument of Pietraszewski, 2016 that group identity can modify the cost-benefit calculus of individuals deciding whether to extend a conflict. Unlike parochial altruism and within-group reciprocity, group reciprocity requires humans to differentiate between outgroups, possibly providing a cognitive basis for intergroup stereotyping and prejudice.

We distinguish between reciprocity towards harm and towards intentional harm (Stanca, Bruni, and Corazzini, 2009). People discriminate against others who harm them even if the harmful action does not necessarily indicate bad intentions. However, they only generalize to the perpetrator's group members if the intentions behind the harmful actions are unequivocally bad.

This observation raises new questions regarding the nature of reciprocity and the role of intentions (or perceptions thereof). One possible interpretation stems from the distinction between intention-based and outcome-based motives in reciprocal behaviour (Falk and Fischbacher, 2006). It is possible that humans generalize intentions across group members more than they gener-

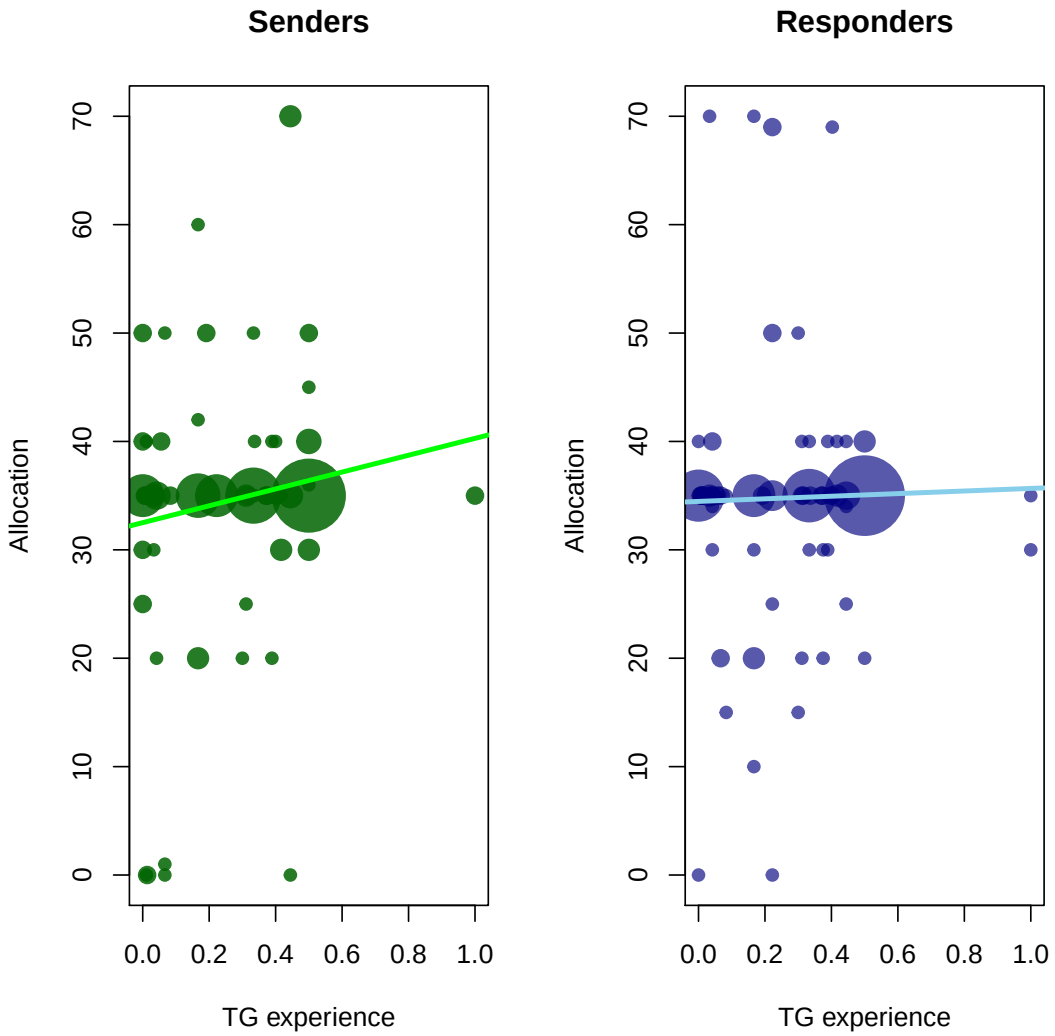



**Figure 2:** Allocations in the Group Reciprocity condition versus the TG experience. Circles show individual data points (circle size proportional to number of observations). Lines show linear regressions.

alize actions across group members. So, if (e.g.) group member 1 wishes to harm them, they are prone to infer that group member 2 also wishes to do so; but if group member 1 takes an action that harms them, they do not necessarily infer that group member 2 would also have done so. Indeed, the conjecture 'One member of the Blue group is a bad person, therefore all Blue members are bad' is plausible. The conjecture 'One member of the Blue group did not send any money, therefore other Blue members did not send money' is not—as, given subjects' knowledge, the other Blue members were not even necessarily senders.

Since our study was conducted with students from a rich industrialized democracy, results may not generalize to all cultures (Henrich, Heine, and Norenzayan, 2010). In particular, the link between intentions and moral judgment may vary across cultures (Barrett, Bolyanatz, Crittenden, Fessler, Fitzpatrick, Gurven, Henrich, Kanovsky, Kushnick, Pisor, et al., 2016), and this could affect how group reciprocity plays out in different societies.

Upstream reciprocity is notoriously difficult to understand in evolutionary terms (Boyd and Richerson, 1989; Nowak and Roch, 2007). Group reciprocity may provide another piece of the puzzle, as it provides two new channels by which upstream reciprocity may evolve. First, group members are interdependent, especially in the small groups that were the norm during most of human evolutionary history. Punishing a perpetrator's group member therefore indirectly harms the perpetrator, who is dependent on his peers for, e.g., public goods provision. Thus, group reciprocity may bridge upstream indirect reciprocity and direct reciprocity.

Second, the evolution of indirect reciprocity acts by way of chains of reciprocal actions, which return with some probability to the original instigator of the chain (Nowak and Roch, 2007). In a population organised in groups, such that people interact more frequently with their own group members, group reciprocity may increase the likelihood of successful reciprocal chains, facilitating the evolution of upstream reciprocity. These ideas could be formalized in future work.

# References


[1] Daniel Balliet, Junhui Wu, and Carsten KW De Dreu. "Ingroup favoritism in cooperation: A meta-analysis." In: *Psychol Bull* 140.6 (2014), pp. 1556–1581.

[2] H Clark Barrett, Alexander Bolyanatz, Alyssa N Crittenden, Daniel MT Fessler, Simon Fitzpatrick, Michael Gurven, Joseph Henrich, Martin Kanovsky, Geoff Kushnick, Anne Pisor, et al. "Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment". In: *Proceedings of the National Academy of Sciences* 113.17 (2016), pp. 4688–4693.

[3] Joyce Berg, John Dickhaut, and Kevin McCabe. "Trust, reciprocity, and social history". In: *Game Econ Behav* 10.1 (1995), pp. 122–142.

[4] Helen Bernhard, Ernst Fehr, and Urs Fischbacher. "Group affiliation and altruistic norm enforcement". In: *Am Econ Rev* 96.2 (2006), pp. 217–221.

[5] Helen Bernhard, Urs Fischbacher, and Ernst Fehr. "Parochial altruism in humans". In: *Nature* 442.7105 (2006), pp. 912–915.

[6] Robert Boyd and Peter J. Richerson. "The evolution of indirect reciprocity". In: *Soc Networks* 11.3 (1989), pp. 213–236.

[7] Napoleon A. Chagnon. "Life Histories, Blood Revenge, and Warfare in a Tribal Population". en. In: *Science* 239.4843 (Feb. 1988), pp. 985–992.

[8] Roy Chen and Yan Chen. "The Potential of Social Identity for Equilibrium Selection". In: *Am Econ Rev* 101.6 (2011), pp. 2562–2589.

[9] Yan Chen and Sherry X. Li. "Group identity and social preferences". In: *Am Econ Rev* 99.1 (2009), pp. 431–457.

[10] Paolo Crosetto, Ori Weisel, and Fabian Winter. "A Flexible z-Tree Implementation of the Social Value Orientation Slider Measure (Murphy et al. 2011): Manual". In: *Jena Economic Research Papers* 2012-062 (2012). Friedrich-Schiller-University Jena, Max-Planck-Institute of Economics.

[11] Carsten K.W. De Dreu, Daniel Balliet, and Nir Halevy. “Chapter One – Parochial Cooperation in Humans: Forms and Functions of Self-Sacrifice in Intergroup Conflict”. In: *Advances in Motivation Science*. Ed. by Andrew J. Elliot. Vol. 1. Elsevier, 2014. Chap. 1, pp. 1–47.

[12] Martin Dufwenberg, Uri Gneezy, Werner Güth, and Eric E. C. van Damme. “Direct versus indirect reciprocity: An experiment”. In: *Homo Oeconomicus* 18.1/2 (2001), pp. 19–30.

[13] Armin Falk and Urs Fischbacher. “A theory of reciprocity”. In: *Game Econ Behav* 54.2 (2006), pp. 293–315.

[14] Urs Fischbacher. “z-Tree: Zurich toolbox for ready-made economic experiments”. In: *Exp Econ* 10.2 (2007), pp. 171–178.

[15] Lowell Gaertner, Jonathan Iuzzini, and Erin M. O’Mara. “When rejection by one fosters aggression against many: Multiple-victim aggression as a consequence of social rejection and perceived groupness”. In: *J Exp Soc Psychol* 44.4 (July 2008), pp. 958–970.

[16] Ben Greiner. “Subject pool recruitment procedures: organizing experiments with ORSEE”. English. In: *Journal of the Economic Science Association* 1.1 (2015), pp. 114–125.

[17] Ben Greiner and M. Vittoria Levati. “Indirect reciprocity in cyclical networks: An experimental study”. In: *J Econ Psychol* 26.5 (2005), pp. 711–731.

[18] Anna Gunnthorsdottir, Kevin McCabe, and Vernon Smith. “Using the Machiavellianism instrument to predict trustworthiness in a bargaining game”. In: *J Econ Psychol* 23.1 (2002), pp. 49–66.

[19] W. Güth, M. Königstein, N. Marchand, and K.D. Nehring. “Trust and Reciprocity in the Investment Game with Indirect Reward”. In: *Homo Oeconomicus* 18 (2001), pp. 241–262.

[20] Johannes Haushofer, Anat Biletzki, and Nancy Kanwisher. “Both sides retaliate in the Israeli–Palestinian conflict”. In: *P Natl Acad Sci Usa* 107.42 (2010), pp. 17927–17932.

[21] Joseph Henrich, Steven J Heine, and Ara Norenzayan. “Most people are not WEIRD”. In: *Nature* 466.7302 (2010), pp. 29–29.

[22] D. L Horowitz. *Ethnic Groups in Conflict*. Berkeley: University of California Press, 1985.

[23] D. L. Horowitz. *The deadly ethnic riot*. University of California Press, 2001.

[24] Rufus A Johnstone and Redouan Bshary. “Evolution of spite through indirect reciprocity”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 271.1551 (2004), pp. 1917–1922.

[25] Erik O Kimbrough and Alexander Vostroknutov. “Norms make preferences social”. In: *J Eur Econ Assoc* 14.3 (2015), pp. 608–638.

[26] Ryan O. Murphy, Kurt A. Ackermann, and Michel J. J. Handgraaf. “Measuring Social Value Orientation”. In: *Judgm Decis Mak* 6.8 (2011), pp. 771–781.

[27] Martin A. Nowak. “Five rules for the evolution of cooperation”. In: *Science* 314.5805 (2006), pp. 1560–1563.

[28] Martin A. Nowak. “Evolving cooperation”. In: *J Theor Biol* 299 (2012), pp. 1–8.

[29] Martin A. Nowak and Sébastian Roch. “Upstream reciprocity and the evolution of gratitude”. In: *P Roy Soc B-biol Sci* 274.1610 (2007), pp. 605–610.

[30] Martin A. Nowak and Karl Sigmund. “Evolution of indirect reciprocity”. In: *Nature* 437.7063 (2005), pp. 1291–1298.

[31] David Pietraszewski. “How the mind sees coalitional and group conflict: the evolutionary invariances of n-person conflict dynamics”. In: *Evolution and Human Behavior* 37.6 (2016), pp. 470–480.

[32] Moses Shayo and Asaf Zussman. “Judicial ingroup bias in the shadow of terrorism”. In: *Q J Econ* (2010).

[33] Theodore M. Singelis, Harry C. Triandis, Dharm P. S. Bhawuk, and Michele J. Gelfand. “Horizontal and vertical dimensions of individualism and collectivism: A theoretical and measurement refinement”. In: *Cross-Cult Res* 29.3 (1995), pp. 240–275.

[34] Luca Stanca, Luigino Bruni, and Luca Corazzini. “Testing theories of reciprocity: Do motivations matter?” In: *J Econ Behav Organ* 71.2 (2009), pp. 233–245.

[35] Douglas M. Stenstrom, Brian Lickel, Thomas F. Denson, and Norman Miller. “The Roles of Ingroup Identification and Outgroup Entitativity in Intergroup Retribution”. en. In: *Pers Soc Psychol B* (Aug. 2008).

[36] Henri Tajfel and John C. Turner. “An integrative theory of intergroup conflict”. In: *The Social Psychology of Intergroup Relations*. Ed. by William G. Austin and Stephen Worchel. Monterey, CA: Brookes/Coole, 1979. Chap. 3, pp. 33–47.

[37] Paul AM Van Lange. “The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation.” In: *Journal of personality and social psychology* 77.2 (1999), p. 337.

[38] Ori Weisel and Robert Böhm. ““Ingroup love” and “outgroup hate” in intergroup conflict between natural groups”. In: *J Exp Soc Psychol* 60 (2015), pp. 110–120.

[39] Ori Weisel and Ro’i Zultan. “Social Motives in Intergroup Conflict: Group Identity and Perceived Target of Threat”. In: *Eur Econ Rev* (2016).

[40] World Bank. *World Development Report 2011: Conflict, Security, and Development*. World Bank, 2011.

[41] Toshio Yamagishi and Toko Kiyonari. “The Group as the Container of Generalized Reciprocity”. In: *Soc Psychol Quart* 63.2 (2000). contains references to literature on in-group favoritism in 2 person PDs, pp. 116–132.

# Appendix A: Complete matching scheme

| Period | Group | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | Blue 2 (GR) | Blue 1 (GR) | Green 4 (GR) | Blue 3 (B) | Red 2 (DR) | Blue 4 (B) | Green 1 (IG) | Red 4 (B) |
| | Red 1 (B) | Yellow 2 (GR) | Brown 4 (B) | Green 3 (GR) | Brown 2 (DR) | Red 3 (DR) | Green 2 (IG) | Yellow 3 (IG) |
| | Yellow 1 (GR) | Purple 2 (B) | Purple 3 (GR) | Purple 4 (GR) | Purple 1 (B) | Brown 3 (DR) | Brown 1 (B) | Yellow 4 (IG) |
| 2 | Green 3 (GR) | Red 3 (B) | Blue 4 (GR) | Blue 2 (GR) | Blue 3 (DR) | Green 2 (DR) | Blue 1 (B) | Red 2 (IG) |
| | Yellow 1 (B) | Green 1 (GR) | Green 4 (B) | Yellow 4 (GR) | Red 1 (B) | Brown 4 (B) | Brown 1 (IG) | Red 4 (IG) |
| | Purple 1 (GR) | Purple 3 (GR) | Yellow 2 (GR) | Brown 2 (B) | Yellow 3 (DR) | Purple 2 (DR) | Brown 3 (IG) | Purple 4 (B) |
| 3 | Red 1 (GR) | Red 4 (GR) | Blue 3 (B) | Red 3 (GR) | Green 4 (DR) | Blue 2 (DR) | Blue 1 (IG) | Yellow 3 (B) |
| | Brown 4 (GR) | Yellow 4 (B) | Red 2 (GR) | Green 2 (B) | Yellow 1 (B) | Green 3 (B) | Blue 4 (IG) | Purple 2 (IG) |
| | Purple 1 (B) | Brown 1 (GR) | Brown 3 (GR) | Brown 2 (GR) | Purple 4 (DR) | Yellow 2 (DR) | Green 1 (B) | Purple 3 (IG) |
| 4 | Blue 4 (GR) | Blue 3 (GR) | Green 2 (GR) | Blue 1 (B) | Red 4 (DR) | Blue 2 (B) | Green 3 (IG) | Red 2 (B) |
| | Red 3 (B) | Yellow 4 (GR) | Brown 2 (B) | Green 1 (GR) | Brown 4 (DR) | Red 1 (DR) | Green 4 (IG) | Yellow 1 (IG) |
| | Yellow 3 (GR) | Purple 4 (B) | Purple 1 (GR) | Purple 2 (GR) | Purple 3 (B) | Brown 1 (DR) | Brown 3 (B) | Yellow 2 (IG) |
| 5 | Green 4 (GR) | Red 4 (B) | Blue 3 (GR) | Blue 1 (GR) | Blue 4 (DR) | Green 1 (DR) | Blue 2 (B) | Red 1 (IG) |
| | Yellow 2 (B) | Green 2 (GR) | Green 3 (B) | Yellow 3 (GR) | Red 2 (B) | Brown 3 (B) | Brown 2 (IG) | Red 3 (IG) |
| | Purple 2 (GR) | Purple 4 (GR) | Yellow 1 (GR) | Brown 1 (B) | Yellow 4 (DR) | Purple 1 (DR) | Brown 4 (IG) | Purple 3 (B) |
| 6 | Red 2 (GR) | Red 3 (GR) | Blue 4 (B) | Red 4 (GR) | Green 3 (DR) | Blue 1 (DR) | Blue 2 (IG) | Yellow 4 (B) |
| | Brown 3 (GR) | Yellow 3 (B) | Red 1 (GR) | Green 1 (B) | Yellow 2 (B) | Green 4 (B) | Blue 3 (IG) | Purple 1 (IG) |
| | Purple 2 (B) | Brown 2 (GR) | Brown 4 (GR) | Brown 1 (GR) | Purple 3 (DR) | Yellow 1 (DR) | Green 2 (B) | Purple 4 (IG) |

# Appendix B: Allocations in the DR condition

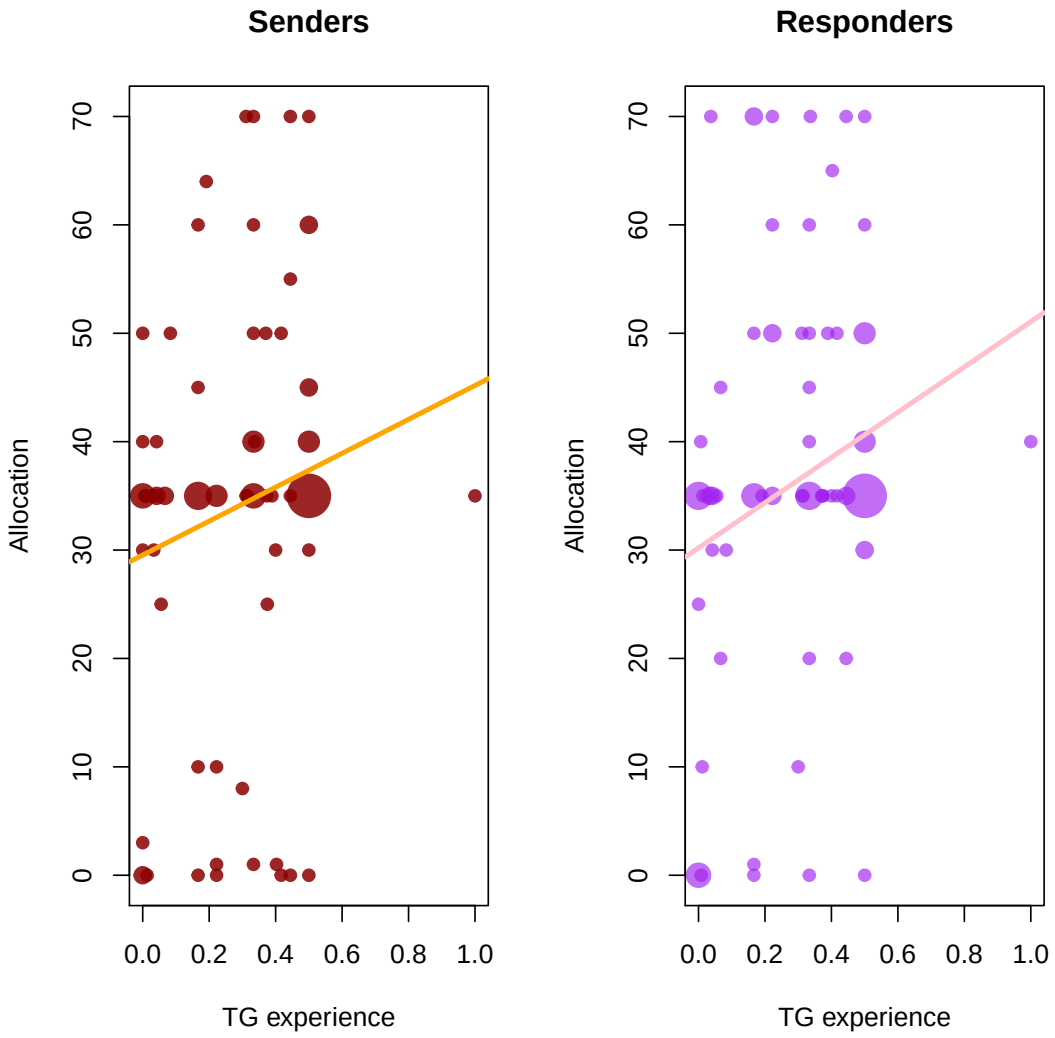



**Figure B.1:** Allocations in the Direct Reciprocity condition versus the TG experience. Circles show individual data points (circle size proportional to number of observations). Lines show linear regressions.

# Appendix C: Experimental instructions

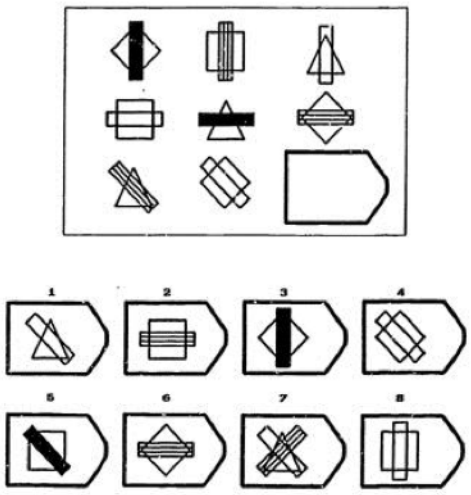
*Instructions for the experiment*

*<Presented as a pdf document and available throughout the experiment>*

**These instructions are identical to all the participants.**
**The experiment is composed of five separate and different phases.** At the beginning of the experiment, all participants will be allocated into **teams of four**. Each team has a unique **colour**. These teams will remain fixed throughout the experiment.
Before each part, we will distribute and read the relevant instructions for that part. In each part the participants will be reallocated into groups. The number of participants in a group can change from part to part. The payments in the part will be determined according to the decisions of the participants in the team. It is possible, but not necessary, that another participant will be in the same group as you in two different parts. In each part of the experiment you will be able to know which team each of the participants in your group belongs to.
**Your final payment in the experiment will be the total of your gain in all of the parts**.
At the end of the experiment, you will be presented with the payments in each part and your total payment, in points and in shekels. Please remain seated until the experimenter calls you for payment.
**If you have any questions, please raise your hand now and the experimenter will come to you**.

*Experiments for the first part*

In this part, you and the members of your team perform a pattern completion task. The computer will present you with five questions. Each question is comprised of eight pictures, and the team members wil be asked to choose a ninth picture out of eight possible pictures to complete the pattern. For example:

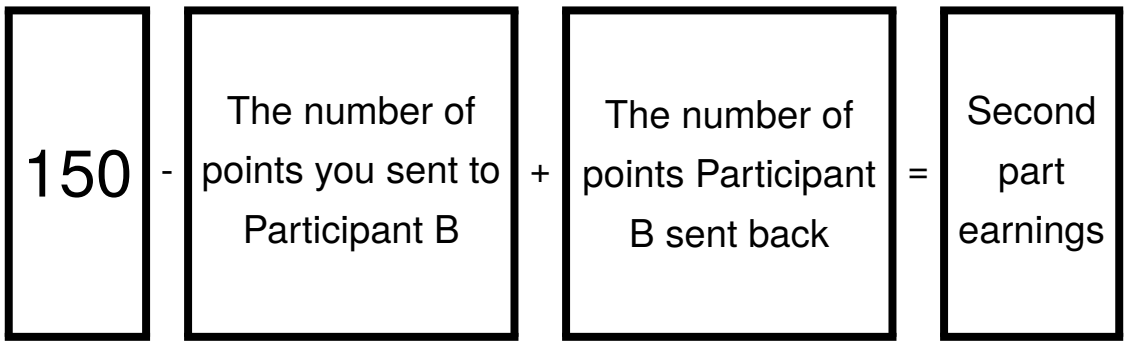


Each team member must answer all of the questions. For each correct answer, the team member will receive **10 points**. Additionally, if all of the team members answer correctly, the whole team will receive a **team bonus of 20 points, to be equally divided among the team members**.
**Each question will be allocated two minutes**. During this time, the team members can **consult each other** using electronic chat. Enter your answer and click Confirm. You can change your answer and click Confirm again at any point during the two minutes. The last answer to be entered is the final answer.
**Attention:** Do not reveal any identifying information. If any participant in the session identifies themselves, we will stop the experiment and release all participants with only the showup fee.
**If you have any questions, please raise your hand now and the experimenter will come to you**.

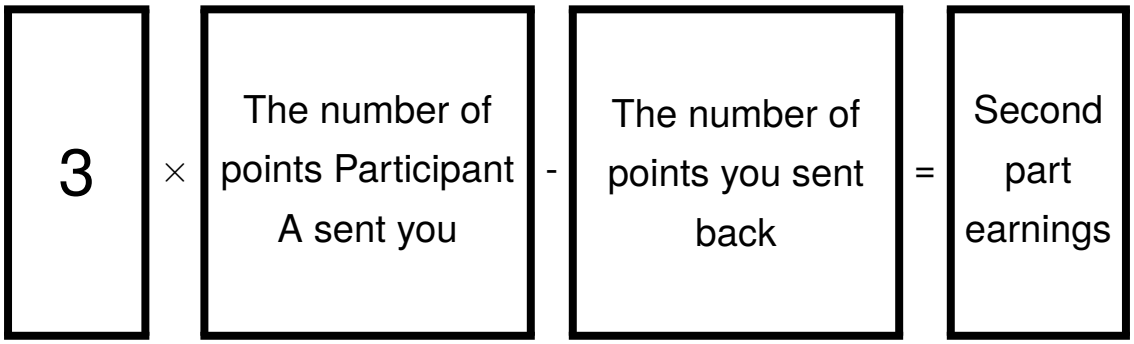
*Instructions for the second part*

In this part participants will be matched in **pairs**. In each pair, one participant will be in role A and the other participant in role B. Participant A receives an allocation of **150 points** and decides how many of the 150 points to **send to Participant B**. The amount is **tripled**. Next, Participant B will decide how many points out of the points received to **send back to to Participant A**. These points will not be multiplied.

If you are allocated to role A, your payment in this part will be:




If you are allocated to role B, your payment in this part will be:




Before making your decision, you will be able to test the payment calculation in a **practice phase**, in which you will be able to make decisions as both **Participant A** and as **Participant B**. In this stage, you will enter decisions in both roles, and see the final payments. The practice will repeat five times.

**If you have any questions, please raise your hand now and the experimenter will come to you**.

## *Instructions for the third part*

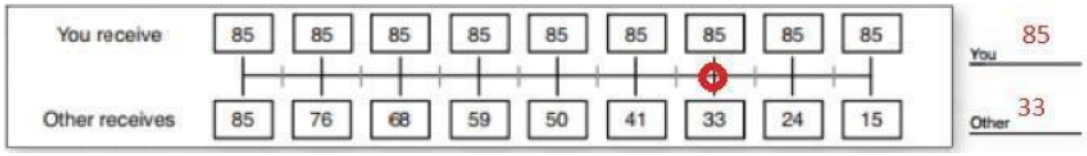
In the third part, all participants will be matched in **groups of three**. Each of the three participants in the group will choose how to **divide 100 points** between the three group members, such that he himself receives **30 points**, and **freely allocates** the remaining **70 points** between the other two group members. This stage has **6 rounds**, and you will be **rematched in a new group**.

## *Payment calculation in the part*

At the end of the experiment, the computer will randomly choose one of the six rounds, and one participant in each group. The payment for this part will be determined according to the decision of the randomly chosen participant in the randomly chosen round.
**If you have any questions, please raise your hand now and the experimenter will come to you**.

## *Instructions for the fourth part*

In this part, participant will be matched in **pairs**.
Each participant will be presented with **6 rulers** that include nine possible allocations of money to the two participants. The amount you chose to **keep for yourself** is indicated above each ruler, and the amount you choose to **give to the other participant** is indicated below the ruler. You are to choose your preferred allocation of the nine possible allocations. For example,




You can choose any point on the ruler. For example, assume you chose the point marked in red. You will receive 85 points and the other participant will receive 33 points.
At the end of the part, the computer will randomly choose on of the two participants in the pair and one of the nine rulers. your payment in this part will be determined by the decision of the randomly chosen participant for the randomly chosen ruler.
**If you have any questions, please raise your hand now and the experimenter will come to you**.

## *Instructions for the fifth part*

In this part you will be asked to answer several questions. The questions have to do with the way one sees himself and his surroundings in different situations. Your task is to indicate how much you agree or disagree with each statement, using the following scale:

1. Strongly disagree.
2. Disagree.
3. Neither agree nor disagree.
4. Agree.
5. Strongly agree.

Note: there are no right and wrong answers. Please indicate the answer that best reflects your character with respect to the statement. Take your time and think about your answer.

# Humans reciprocate intentional harm by discriminating against group peers

David Hugh-Jones[*] Itay Ron[†] Ro'i Zultan[‡]

**Abstract**

Cycles of intergroup revenge appear in large scale conflicts. We experimentally test the hypothesis that humans practice group-based reciprocity: if someone harms or helps them, they harm or help other members of that person's group. Subjects played a trust game, then allocated money between other people. Senders whose partners returned more in the trust game gave more to that partner's group members. The effect was about half as large as the effect of direct reciprocity. Receivers' allocations to group members were not affected by their partners' play in the trust game, suggesting that group reciprocity was only triggered when the partner's intentions were unequivocal.

*Keywords:* Upstream reciprocity, group identity, intergroup conflict.

## Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 214/13).

[*]School of Economics, University of East Anglia. E-mail: d.hugh-jones@uea.ac.uk.
[†]E-mail: itayron@gmail.com
[‡]Corresponding author, Department of Economics, Ben-Gurion University of the Negev. E-mail: zultan@bgu.ac.il.