

# Humans reciprocate intentional harm by discriminating against group peers<sup>\*</sup>

David Hugh-Jones<sup>†</sup>    Itay Ron<sup>‡</sup>    Ro'i Zultan<sup>§</sup>

This version: August 3, 2016.

## 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 60 per cent of the size of the direct reciprocity effect. Receivers' allocations to group members did not depend on their partner's play, suggesting that group reciprocity was only triggered when the partner's intentions were unequivocal.

## 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

---

<sup>\*</sup>This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 214/13)

<sup>†</sup>School of Economics, University of East Anglia. E-mail: d.hugh-jones@uea.ac.uk.

<sup>‡</sup>E-mail: itayron@gmail.com

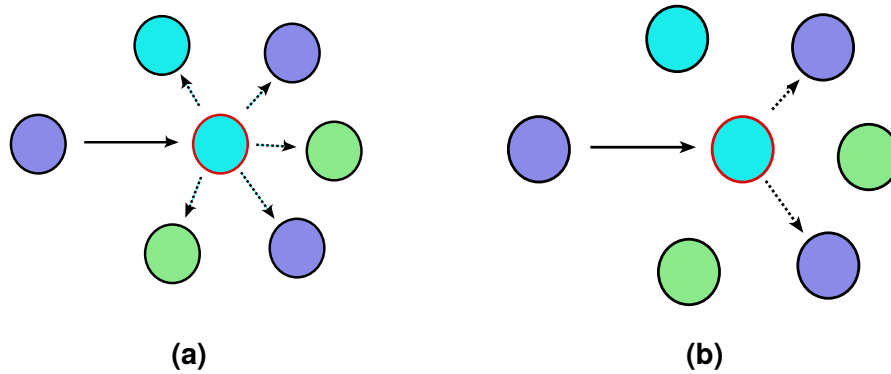
<sup>§</sup>Department of Economics, Ben-Gurion University. E-mail: zultan@bgu.ac.il.

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 inter-group conflict—from economic and political competition to inter-ethnic violence and war—which is pervasive in the species.

Intergroup conflicts often follow a tit-for-tat logic, in which one group's violence leads to revenge from the other side (Chagnon, 1988; 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'. **these harm a member of their own group (Bernhard, Fehr, and Fischbacher, 2006; Bernhard, Fischbacher, and Fehr, 2006).**

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 out-group 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 an individual-level conflict into an inter-group conflict. While (group-based) downstream reciprocity can bring a victim's groupmates in as retaliators, upstream reciprocity makes members of the aggressor's group into potential targets. **C** 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 Rich-





**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.

erson, 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).

Upstream group reciprocity also has different cognitive requirements from related phenomena. While in-group 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. Upstream group reciprocity could thus provide an evolutionary basis for outgroup stereotyping.

We ran a laboratory experiment to test the hypothesis that people reciprocate towards groups. Although field observations from conflict are highly suggestive, 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; general-

ized reciprocity, where subjects reciprocate not just towards the original actor's group but also to members of other groups; and strategic interactions, where apparent reciprocity is driven by reputation-building. While none of the previous work (Gaertner, Iuzzini, and O'Mara, 2008; Hugh-Jones and Leroch, 2013; Stenstrom, Lickel, Denson, and Miller, 2008) has satisfied all three conditions simultaneously, 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. 

After an initial group-formation stage, participants interacted in two strategic stages. The first stage represented the upstream interaction, in which the individual could be helped or harmed by another person. We implemented this as a Trust Game (TG) (Berg, Dickhaut, and McCabe, 1995), in which 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. 

The second stage represented 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 In-Group Favouritism 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.

## 2 Method

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 AG stage, a social value orientation elicitation (Murphy, Ackermann, and Handgraaf, 2011) stage and a collectivism scale measurement (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 kindness of the sender was measured as the share of the endowment sent, and that of the responder as the share of the received amount sent back. Responder's kindness was not defined for six (out of 96) responders whose partner did not send any money.

The third stage 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

**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).

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 found that people do not harm, but refrain from helping negatively perceived out-groups (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 rounds. Each participant was in the same group of three in one of the six rounds with a member of her own team (*in-group* 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. Therefore, 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  $0^\circ$  corresponding to selfishness,  $45^\circ$  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 final payment in New Israeli Shekels. 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 average payment was 72 NIS (approximately \$18) for a duration of 70 minutes.

### 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. See the supplementary material for the full specification and results.

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 in-group favouritism. Allocations to the TG partner and his team mates were not significantly different to the baseline 35 ( $p > 0.47$  for all comparisons). Nonetheless, as the second column of Table 2 shows, allocators discriminated significantly more than in the baseline both when interacting with their TG partner ( $z = 9.08, p < 0.001$ ) and with his

**Table 2:** Allocations and Discrimination

	Allocation	Discrimination	Reciprocity
<b>Senders</b>			
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)
<b>Responders</b>			
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.

team mates ( $z = 3.93, p < 0.001$ ). This effect was not significantly different between TG senders and receivers (F test 0.50,  $p = 0.68$ ).

### *Direct and group reciprocity*

The third column of Table 2, *Reciprocity*, reports the slope of allocations regressed on the *kindness* of subjects' TG partners. There is strong direct reciprocity: allocations to the TG partners increase with that partner's kindness 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 sent less—an intentionally harmful action. Responders, on the other hand, although directly reciprocating the TG partner, 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 sender kindness on responder's allocation to the sender's team mates ( $z = 0.58, p = 0.56$ ). The responder's kindness, on the other hand, significantly increases allocations made to his 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.

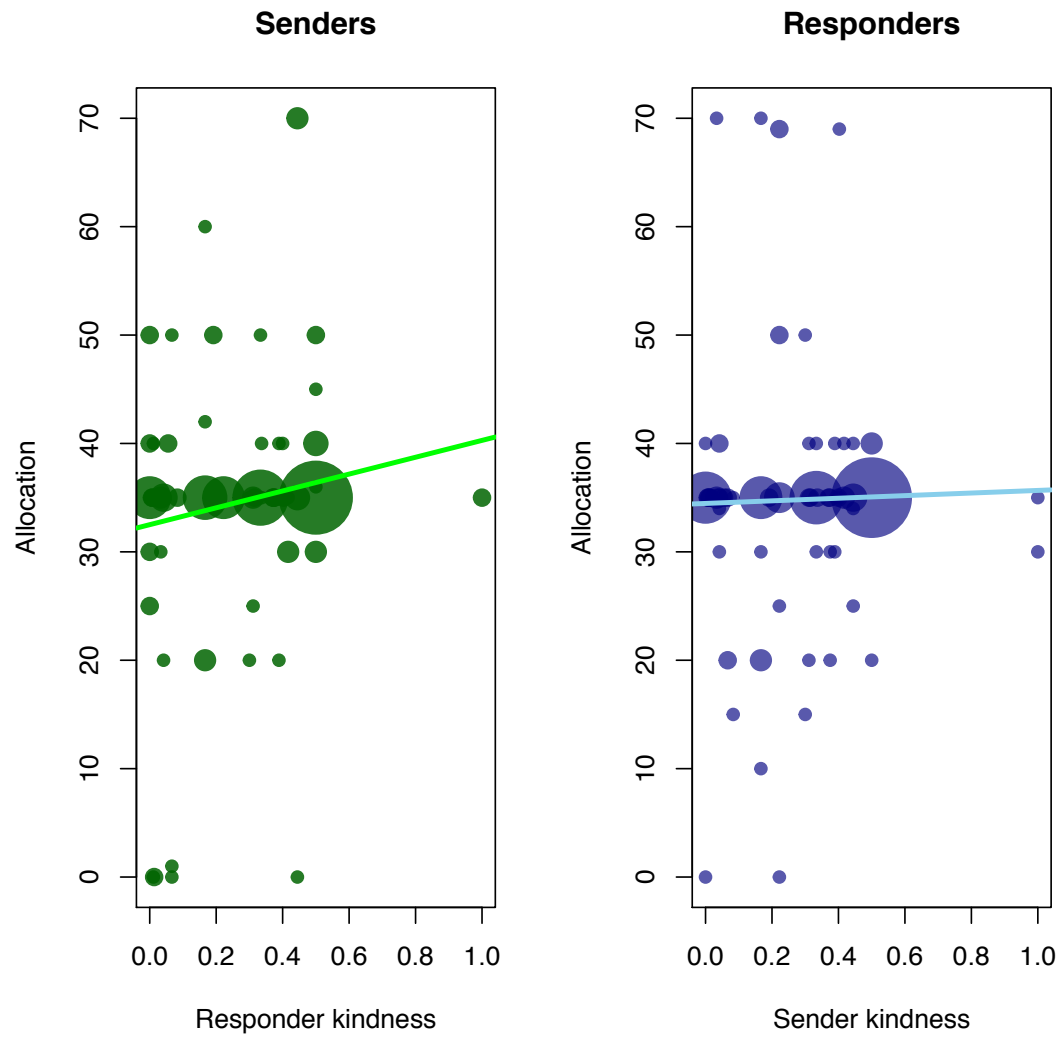
Senders' group reciprocity was related to their social value orientation. The slope of kindness on allocations was 15.97 for those with less than median SVO, and -1.06 for those with median or greater SVO (interaction,  $p = 0.061$ ). These results should be interpreted cautiously, since both scores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

## 4 Discussion

Our results show that upstream reciprocity is moderated by social boundaries. 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 generalize 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.

Upstream reciprocity is notoriously difficult to understand in evolutionary terms (Boyd and Richerson, 1989; Nowak and Roch, 2007) Group reciprocity



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

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. We aim to explore these ideas in future work.

## References

- [1] Daniel Balliet, Junhui Wu, and Carsten KW De Dreu. “Ingroup favoritism in cooperation: A meta-analysis.” In: *Psychological Bulletin* 140.6 (2014), pp. 1556–1581.
- [2] Joyce Berg, John Dickhaut, and Kevin McCabe. “Trust, reciprocity, and social history”. In: *Games and economic behavior* 10.1 (1995), pp. 122–142.
- [3] Helen Bernhard, Ernst Fehr, and Urs Fischbacher. “Group affiliation and altruistic norm enforcement”. In: *American Economic Review* 96.2 (2006), pp. 217–221.
- [4] Helen Bernhard, Urs Fischbacher, and Ernst Fehr. “Parochial altruism in humans”. In: *Nature* 442.7105 (2006), pp. 912–915.
- [5] Robert Boyd and Peter J. Richerson. “The evolution of indirect reciprocity”. In: *Social Networks* 11.3 (1989), pp. 213–236.
- [6] Napoleon A. Chagnon. “Life Histories, Blood Revenge, and Warfare in a Tribal Population”. en. In: *Science* 239.4843 (Feb. 1988), pp. 985–992. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.239.4843.985. URL: <http://www.sciencemag.org/content/239/4843/985> (visited on 10/02/2013).
- [7] Roy Chen and Yan Chen. “The Potential of Social Identity for Equilibrium Selection”. In: *American Economic Review* 101.6 (2011), pp. 2562–2589.
- [8] Yan Chen and Sherry X. Li. “Group identity and social preferences”. In: *American Economic Review* 99.1 (2009), pp. 431–457. ISSN: 0002-8282.
- [9] 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.

- [10] 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. DOI: <http://dx.doi.org/10.1016/bs.adms.2014.08.001>.
- [11] 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.
- [12] Armin Falk and Urs Fischbacher. “A theory of reciprocity”. In: *Games and Economic Behavior* 54.2 (2006), pp. 293–315.
- [13] Urs Fischbacher. “z-Tree: Zurich toolbox for ready-made economic experiments”. In: *Experimental Economics* 10.2 (2007), pp. 171–178.
- [14] 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: *Journal of Experimental Social Psychology* 44.4 (July 2008), pp. 958–970. ISSN: 0022-1031. DOI: 16/j.jesp.2008.02.004. URL: <http://www.sciencedirect.com/science/article/pii/S0022103108000267> (visited on 07/20/2011).
- [15] Ben Greiner. “Subject pool recruitment procedures: organizing experiments with ORSEE”. English. In: *Journal of the Economic Science Association* 1.1 (2015), pp. 114–125. ISSN: 2199-6776. DOI: 10.1007/s40881-015-0004-4.
- [16] Ben Greiner and M. Vittoria Levati. “Indirect reciprocity in cyclical networks: An experimental study”. In: *Journal of Economic Psychology* 26.5 (2005), pp. 711–731.
- [17] 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.
- [18] D. L Horowitz. *Ethnic Groups in Conflict*. Berkeley: University of California Press, 1985.
- [19] D. L. Horowitz. *The deadly ethnic riot*. University of California Press, 2001.

- [20] David Hugh-Jones and Martin A. Leroch. “Intergroup revenge: a laboratory experiment on the causes”. In: *Available at SSRN 2275173* (2013).
- [21] Ryan O. Murphy, Kurt A. Ackermann, and Michel J. J. Handgraaf. “Measuring Social Value Orientation”. In: *Judgment and Decision Making*, 6.8 (2011), pp. 771–781.
- [22] Martin A. Nowak. “Five rules for the evolution of cooperation”. In: *science* 314.5805 (2006), pp. 1560–1563.
- [23] Martin A. Nowak. “Evolving cooperation”. In: *Journal of Theoretical Biology* 299 (2012), pp. 1–8.
- [24] Martin A. Nowak and Sébastien Roch. “Upstream reciprocity and the evolution of gratitude”. In: *Proceedings of the Royal Society B: Biological Sciences* 274.1610 (2007), pp. 605–610.
- [25] Martin A. Nowak and Karl Sigmund. “Evolution of indirect reciprocity”. In: *Nature* 437.7063 (2005), pp. 1291–1298.
- [26] Moses Shayo and Asaf Zussman. “Judicial ingroup bias in the shadow of terrorism”. In: *Quarterly Journal of Economics, Forthcoming* (2010).
- [27] 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-Cultural Research* 29.3 (1995), pp. 240–275.
- [28] Luca Stanca, Luigino Bruni, and Luca Corazzini. “Testing theories of reciprocity: Do motivations matter?” In: *Journal of economic behavior & organization* 71.2 (2009), pp. 233–245.
- [29] Douglas M. Stenstrom, Brian Lickel, Thomas F. Denson, and Norman Miller. “The Roles of Ingroup Identification and Outgroup Entitativity in Intergroup Retribution”. en. In: *Personality and Social Psychology Bulletin* (Aug. 2008). ISSN: 0146-1672, 1552-7433. DOI: 10.1177/0146167208322999. URL: <http://psp.sagepub.com/content/early/2008/08/20/0146167208322999> (visited on 06/10/2015).

- [30] 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.
- [31] Ori Weisel and Robert Böhm. ““Ingroup love” and “outgroup hate” in intergroup conflict between natural groups”. In: *Journal of Experimental Social Psychology* 60 (2015), pp. 110–120.
- [32] Ori Weisel and Ro’i Zultan. “Social Motives in Intergroup Conflict: Group Identity and Perceived Target of Threat”. In: *European Economic Review* (2016). DOI: <http://dx.doi.org/10.1016/j.euroecorev.2016.01.004>.
- [33] T. Yamagishi and T. Kiyonari. “The Group as the Container of Generalized Reciprocity”. In: *Social Psychology Quarterly* 63.2 (2000). contains references to literature on in-group favoritism in 2 person PDs, pp. 116–132.