Editor

1. *The reviewers provide numerous constructive comments that will be important address before submitting a revision. Reviewer 2 in particular points out key concerns about theoretical framing, experimental setup, and interpretation which deserve careful consideration. Also, regarding the simulation, I agree with R1 that they don't contribute much to the paper, and that the real potential of the paper is in the empirical ndings.*

Thank you for the opportunity to revise our manuscript. We did our best to address the comments and suggestions of the referees. In line with your comment, we moved the simulations to the supplementary material, and provide a short summary in the text.

Reviewer #1

1. *This is a well-designed study on an interesting and under-theorized question: does group-based reciprocity drive harm towards outgroups. The manuscript is well-written with the experimental design clearly articulated and it fits within a larger body of experimental work using versions of the DG and TG to explore and test hypotheses about real-world intergroup relationships. I see the key contribution of this manuscript being that it advances a novel and important line of theorizing about the origins of intergroup relationships. In this respect, I am optimistic it will simulate more theory on this topic.*

I recommend publication pending minor revisions and have a few general and specific comments below.

Thank you for the encouraging assessment.

2. *I find the simulations not very convincing and think they could be safely removed from the ms without losing much of substance. But it's also fine if they remain in.*

In line with your and the editor's comments, we kept a summary of the simulations, and moved the more detailed description to the supplementary material.

3. *Their results find group based reciprocity only towards receivers, not senders, which weakens the interpretation from their experimental results. They provide one plausible explanation but it might also be the case that upstream reciprocity doesn't shape these types of intergroup interactions or the experimental design doesn't adequately capture real-world phenomena in this context. Nonetheless, I don't see this as a serious flaw of their paper.*

We believe that the difference between sender and responder behavior makes the results non-trivial, and therefore interesting and pointing at new research directions and insights. We argue this in the following:

> We observed group reciprocity only towards receivers, not senders. On the one hand, we find group reciprocity towards receivers, confirming that the experiment was successful in setting up the type of group interactions that triggers group reciprocity. On the other hand, we find *direct* reciprocity towards senders, indicating that responders perceived the TG interaction as meaningful and relevant for the later allocation decisions. We therefore conclude that it is some characteristic of the responder decision, not shared with the sender decision, that triggers group reciprocity.

We agree that that the empirical findings are only a first step to understanding this phenomenon, and that the explanations we put forward in the paper are somewhat tentative at this point. We extended the discussion to allow for varying interpretations, and removed the words "intentional harm" from the title:

> One possible interpretation for this difference between senders and responders stems from the distinction between intention-based and outcome-based motives in reciprocal behaviour (Falk and Fischbacher, 2006; Stanca, Bruni, and Corazzini, 2009). In this sense, senders' intentions are more ambiguous, as they do not know what the responder will do. Responders who do not return money, in contrast, are clearly intentionally harming the senders. It is possible that humans generalize *intentions* across group members. That is, if group member 1 takes an action that deliberately harms them, they predict that group member 2 wishes to harm them also. If not returning money is seen as deliberately harmful, while not sending money can be explained by caution or mistrust, then this would generate the difference in group reciprocity that we observe.
> Another distinction made in the literature between trust (sender behavior) and trustworthiness (responder behavior) is based on norms and rules of conduct. In their analylsis of Adam Smith's *A Theory of Moral Sentiments*, Wilson and Smith (2017) argue that trust is benificient act, while breaking trust is misconduct. Accordingly, Wilson and Smith (2017) found that people punish responders but not senders. Similarly, Kimbrough and Vostroknutov (2015) found that 'rule followers' are more trustworthy, but not more trusting, than other individuals. We view these interpretations of our results as tentative. Further research will be necessary to map and understand the boundaries of the group reciprocity phenomenon.

4. *I would have found the claim of "intentional" harm more compelling if they included an unintentional harm as a control. However, I don't see this as serious enough to warrant running more experiments for the sake of this paper though.*

We use the term "intentional harm" to mean an act that conveys unambiguous intentions to harm. In this sense, senders' decisions in the trust game are not intentional. We clarified in the passage quoted above.

5. *Page 4, Line 199: Typo: "reciprocity" is spelled incorrectly.*

Corrected.

6. *Page 15, lines 842-846: can you unpack this sentence? It's a little too opaque.*

We rephrased thus:

> Group reciprocity may provide another piece of the puzzle. Group reicprocity allows individuals to use reciprocal strategies based on group reputation. Consequently, upstream reciprocity can direct group-level selection in ways parallel to those by which direct reciprocity direct individual-level selection.

7. *Line 84, page 2. Tit-for-tat conflict doesn't imply general reciprocity including both harm and help but rather reciprocity for harm.*

We now acknowledge this point:

Tit-for-tat conflict looks like negative group reciprocity.

8. *Page 3, line 168: chimpanzees do not live in "bands", which refers to a type of human social organization. They live in "communities", which is the term for the chimpanzee unit of social organization. Many human hunter-gatherers and some horticulturalists lived in "bands" but not all of them.*

Thank you. We have corrected the terminology.

9. *Page 16:: "well-defined" institutions? What does 'well-defined' mean? Is this a term Kelly uses? If not, I don't think it's an accurate assessment of war/ peace systems and would strongly suggest deleting it.*

We were thinking of e.g. the payment of blood money or Court of Good Men described in Boehm (1984). Kelly (p. 119) describes the Andaman islanders as having a "war/peace" system and describes the "peace dance" by which they tried to reconcile with European convicts. We have modified our claim:

Some of these societies also have "war/peace systems" featuring institutions for ending conflict as well as beginning it, such as the Andamanese Peace Dance, or the Montenegrin Court of Good Men for ending feuds (Boehm 1984).

10. *Page 14, line 780: "the argument of Pietraszewski" is introduced out of the blue as if this is if his ideas were what was designed to be tested in this study. I would suggest rephrasing it or introducing Pietraszewski's paper earlier in the manuscript.*

We have deleted this sentence.

11. *Page 14, lines 814–815. Whether HGs had "generally peaceful" relationships is the subject of intense dispute and the authors don't adequately support this claim and cite work elsewhere in this manuscript that argues to the contrary. This debate can be safely dodged by rephrasing this sentence to say "…group reciprocity can help explain why some groups have relatively peaceful intergroup relationships." Or "…why hunter-gatherers have lower rates of intergroup violence than horticulturalists or pastoralists". Neither of these are contentious. For a reference that directly compares rates of death due to warfare between subsistence styles you can cite: Wrangham, Richard W., Michael L. Wilson, and Martin N. Muller. "Comparative rates of violence in chimpanzees and humans." Primates 47.1 (2006): 14-26.*

We have gone along with the reviewer's first suggestion. We think group reciprocity is better able to explain differences between humans and chimps, than betwen hunter-gatherers and other subsistence styles. We now say:

We have argued that group reciprocity could help explain why some groups have relatively peaceful intergroup relations. It may also provide a step from the "chimpanzee model" of conflict towards the large-scale, organized intergroup conflicts observed in tribal and state-level societies. For example, Wrangham et al. (2006) provide evidence that hunter-gatherers and farmers have similar levels of lethal violence to chimpanzees but much less non-lethal violence. This could be because the threat of high-level violence can contain low-level violence.

12. *I have several suggestions about references you might consider:*

P.2. Line 7: the citation to World Bank is strange. You might consider citing a review on group based conflict, these two below cover this terrain more comprehensively:

Esteban, J., Mayoral, L., & Ray, D. (2012). Ethnicity and conflict: Theory and facts. science, 336(6083), 858-865.

Or

Glowacki, Luke, Michael L. Wilson, and Richard W. Wrangham. "The evolutionary anthropology of war." Journal of Economic Behavior & Organization (2017).

The role of parochial altruism and intergroup conflict has been written about more widely than just Choi and Bowles. See for example: Rusch, Hannes. "The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension." Proceedings of the Royal Society of London B: Biological Sciences 281.1794 (2014): 20141539.

Citations for tit-for-tat logic would do well to include a review or theory paper consider:

Boehm, Christopher. Blood revenge: The enactment and management of conflict in Montenegro and other tribal societies. University of Pennsylvania Press, 1984.

Boehm, Christopher. "Retaliatory violence in human prehistory." The British Journal of Criminology 51.3 (2011): 518-534.

Page 156-157: Fry and Soderberg don't actually show that intergroup conflict among HGs is rare. They only report number of people killed with no denominator. They found 148 killings occurred and were present in all but 3 societies and made up 34% of total killings. So I think it's a bit off hand on the basis of this to state that conflict appears to have been rare, especially when numerous other reviews find evidence to the contrary, including the one you cite (Wrangham and Glowacki 2012). What you can safely claim is that HGs had lower rates of deaths from warfare than farmers and pastoralists. For support for this you can see Wrangham, Richard W., Michael L. Wilson, and Martin N. Muller. "Comparative rates of violence in chimpanzees and humans." Primates 47.1 (2006): 14-26.

Thank you for the very helpful suggestions! We updated the references. We've deleted the claim that HG conflict is rare; we kept the World Bank report in to make a point about the economic costs of conflict. New citations have been added.

Reviewer #2

The research questions are important and likely of interest to the readers of E&HB. I value the idea of the the experimental design in general and the results (partly) support the authors' hypotheses. Hence, there are several things to like about the manuscript. That said, however, there are also a number of weaknesses in the theoretical overview, the experiment, and the interpretation of results. I describe my concerns and suggestions in detail below in the order of appearance in the paper.

Thank you for the positive assessment and for the helpful suggestions.

1. *The introduction is well written. However, this overview misses some important work. In my view, the authors need to discuss the theory of group-bounded reciprocity by Yamagishi and colleagues, which is closely related to the present work. The authors already cite some of this work, but do not discuss it. There is also recent empirical research supporting and extending this theory, which might be helpful for interpreting the present results and putting them into context.*

We agree that Yamagishi's work is important. As we understand it, his key argument is that altruism within a group is related to an expectation of reciprocity within the "container" of the group. Our analysis complements this by relating between-group behaviour to reciprocity. We now mention this in Footnote 4:

> This argument is a between-group parallel to Yamagishi and Kiyonari (2000), which argues that expectations of generalized reciprocity lie behind altruism within a group.

2. *The simulation results appear quite important and relevant for the present paper. I strongly suggest including them in the main text. The results provide important insights for the hypotheses of the experiment. However, in the current form, the simulation and its parameters are not discussed with regard to other simulation work in the field of direct/indirect reciprocity. The authors should rewrite this section to help the less informed reader understanding and interpreting their results. Of course, this requires substantial rewriting.*

Following the other reviewer and editor's comments, we have moved the results to an online appendix. We now discuss them more briefly, and we hope more clearly, as follows:

While group reciprocity can benefit the group, to evolve it must increase individual fitness. In the supplementary materials, we report on a series of simulations tracking the evolution of group reciprocators under different environmental parameters. The key result is that when the relative cost of helping (or failing to harm) a target is low enough, relative to the benefit supplied to the target and to the group size, group reciprocity can evolve to fixation in a group. The mechanism is that groups with a high share of group reciprocators learn to cooperate with each other, while not helping groups that have a high share of selfish types. Selfish types in cooperative groups free ride on the group reputation, exploiting members of other cooperating groups, thereby gaining higher fitness then the group reciprocators in their group. Nonetheless, if the share of group reciprocators within the group varies sufficiently between groups, group reciprocators are overrepresented in the cooperative groups, and therefore have higher fitness overall. The high benefit/cost ratio required for group reciprocity to evolve fits the "chimpanzee model" of conflict where an attack is only launched if it is low risk. It may also explain why most salient examples of group reciprocity are seen in the negative domain of harm and conflict: when the benefit/cost ratio is high, hurting or refraining from helping is a "nasty" action, since it imposes a large loss on the target for a small gain to oneself.

3. *The dictator game is not well explained. From the methods section and the instructions I guess the allocator had to allocate the remaining 70 tokens and could nothing keep for him/herself? In this case, this is a third-party dictator game (with some fixed endowment for the allocator) and allocations have nothing to do with prosociality, i.e., they are not costly to the allocator. This is an important limitation and the authors should at least discuss this.*

The aim of this design is to best capture group reciprocity, rather then to study costly prosociality. As can be seen in our response to comment #10 below, this design choice eliminates potential confounds. We clarify in the following:

Each player in the group of three had to allocate 100 tokens within the group. The allocator always received exactly 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 an equal division between the other two players provides them with 35 tokens each, more than the allocator's own share.[2]

[2] The allocator's decision is not costly, which might have introduced additional confounding considerations. As our aim in this paper is to identify and study the qualitative characteristics of group reciprocity, we accept the limitations that this choice imposes on the ability to estimate the *magnitude* of preferences for group reciprocity.

4. *Potentially even more critically, it is explained: "Each player in the group of three had to allocate 100 tokens within the group." Did participants know about that? This introduces the confound that all allocations may not be based on own preferences based on the trust game allocations (e.g., reciprocity concerns) but could also be based on participants reciprocity to their belief about others' reciprocity based on the trust game. In other words, I might give more to a*

person because I expect this person give more to me because of direct/group-based reciprocity or group-based preferences. In this case, all behavior may be based on direct (expected) reciprocity based on participant's beliefs. In my view, this undermines the whole idea of the otherwise well-designed experiment.

An important feature of the design is that only one dictator allocation is implemented. Therefore, each participant makes a decision conditional on being the only allocator. Also, participants in the allocation stage do not know which treatment their partners are in, nor which role they played in the preceding trust game. That is, when A is group-reciprocating to B, it is likely that A is a neutral player from B's viewpoint. We elaborate on these points in the following:

> One round of the six rounds was randomly chosen for payment. In that round, the payoffs of the members of each group were determined by the allocation decision of one randomly chosen player in the group. No feedback was provided between rounds. At the end of the stage, players learned the payoff round, whether their allocation was chosen to determine payoffs, and their payoff for the round. Thus, all allocation decisions were completely independent of each other, both within and between participants.

5. *The authors do not explain how the payment of participants worked. From the instructions I conclude that decisions were incentivized. How big were the stakes and what were the average payments?*

We added details on the payments:

> The average payment was approximately $18) for a duration of 70 minutes. The lowest and highest payments were approximately $6 and $32, respectively.

6. *Was the SVO slider measure incentivized? Given it was not and considering the additional weakness that it was assessed at the end of the experiment and could be affected by participants' experiences in the previous parts, I do not see how this can be interpreted at all. As a last point, SVO was assessed with in-group recipients, which differs from the usual approach where recipients are unknown others (although there is some evidence that the results are equivalent, e.g., Böhm, Fleiss & Rybnicek, 2017). Although the authors are cautious regarding the interpretation of their results in relation to SVO, I think these results should be dropped completely from the results section. This does not mean that I suggest to drop the information from the methods section that this variable was assessed, but it should be clearly explained why this data could not be used in any meaningful analyses. This is really a pity because it would have been interesting, particularly with regard to parochial altruism, how SVO relates to the other results.*

We clarify in the revised manuscript that all decisions were incentivized. We further qualify the results with the SVO partners being in-group members. Despite the shortcomings, we believe that most readers would be curious to see a brief (two sentences) report of the SVO results.

7. *How was the mixed-effects analysis specified? Did you consider random intercepts, random slopes, or both among subjects to consider the within-subjects homogeneity in error terms?*

The description of our statistics referred to a previous specification. We corrected it to read:

> All reported statistical tests are based on linear regressions with standard errors clustered by session.

I.e., we ran simple linear regressions, we clustered errors by session rather than by subject, and they are not bootstrap clusters. Clustering errors at the highest level is recommended by Cameron, Gelbach, and Miller (2011). We could also have included per-subject random intercepts. Below, we show the results if we do so:

	Allocation	Discrimination	Reciprocity
Senders			
Baseline	35.00 (—)	4.15 (1.29)	—
Direct Reciprocity	33.98 (1.06)	21.94 (1.84)	15.64 (5.14)
Group Reciprocity	34.40 (0.76)	8.09 (1.49)	7.77 (3.65)
In-Group	38.98 (1.06)	15.46 (1.83)	0.20 (5.11)
Responders			
Baseline	35.00 (—)	2.25 (1.00)	—
Direct Reciprocity	35.38 (0.95)	22.17 (1.64) ***	20.87 (4.76)***
Group Reciprocity	34.79 (0.69)	6.10 (1.25) **	1.20 (3.46)
In-Group	42.13 (0.95) ***	17.17 (1.65) ***	4.72 (4.76)

Estimates are virtually the same as in Table 2 in the paper. In particular, the Group Reciprocity parameter remains significant for senders but not receivers. We also tried to run models with per-individual reciprocity slopes, but these failed to converge. Although we would like to get estimates of subject heterogeneity in reciprocity, that is probably asking too much of these data, which have only six decisions per subject.

8. *The is a typo on p.11: It should be THE SECOND column of Table 2.*

Corrected to state "the first data column".

9. *I am puzzled about the difference regarding senders and responders in group-based reciprocity. Although the authors suggest one possible explanation in the Discussion, I think there could be an alternative explanation. That is, participants who play in the later position of the trust game, i.e., responders, may perceive their back-transfer already as reciprocity and see no need to reciprocate again in the later allocation game.*

This explanation can be ruled out, as responders do reciprocate directly towards their TG partner. We address this point more generally in the discussion:

On the one hand, we find group reciprocity towards receivers, confirming that the experiment was successful in setting up the type of group interactions that triggers group reciprocity. On the other hand, we find *direct* reciprocity towards senders, indicating that responders perceived the TG interaction as meaningful and relevant for the later allocation decisions.

10. *Related to the point above, the behavior(s) in the trust game are insufficiently controlled in the analyses of the allocation game. The trust game is quite complex. Responders' back-transfers may be influenced by the absolute amount sent in the first place. For instance, if a responder received only a small amount by the sender, he/she might be less willing to make a fair back-transfer compared to a sender who received a larger, "fair" amount. These "level differences" need to be considered, and therefore, the whole analyses presented could be biased.*

The amount sent is indeed correlated with the proportion returned. So mean senders are "treated" with a lower proportion returned, creating a potential confound. However, recall that in the Allocation Games, senders have to allocate a fixed amount between two others. While sender behavior in the Trust Game might correlate with pro-sociality, There is no reason that it would correlate ex ante with preference for one group rather than another. We tested this argument by rerunning the Sender group reciprocity regression, interacting amount sent with share returned. The coefficient on receiver's kindness remained large and significant at 5%; coefficients on amount sent, and on its interaction with share returned, were small and insignificant:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	31.580	2.260	13.972	$< 0.001^{***}$
kindness	9.882	4.593	2.151	0.033^{*}
sent (demeaned)	-0.029	0.058	-0.505	0.614
kindness × sent (demeaned)	0.067	0.135	0.491	0.624

We address this in new Footnote 3:

Responders who receive higher amounts also return a higher share. As a result, senders with a more positive TG experience are, on average, those who sent more in the TG, creating a potential confound. There is no reason, however, why different senders should systematically discriminate between groups in a non-costly way. Indeed, the results hold when we control in the regression for the amount sent and its interaction with the share returned.

11. *In the introduction, the authors mention differences between positive and negative reciprocity but rarely relate to this differentiation when presenting their results. This requires controlling (e.g., mean center) for the baseline allocations, both in the analyses and when displaying the results (i.e., in Figure 2).*

We agree that it would be good to differentiate positive and negative group reciprocity, and originally we included some statistical analysis along these lines. We dropped this because our statistical power is not really high enough for the purpose. We now mention this as a limitation in the discussion.

12. *More generally, it is not clear what are the consequences of the mechanism the author suggest and discuss at the end. Given that people display group-based reciprocity, should they seek group-based markers? Should groups enforce group-based signals, potentially more so when they are prosocial rather when they are aggressive/non-cooperative across group boundaries? I think the discussion could be improved a lot by suggesting some testable hypotheses for future research.*

Yes, that is a potential implication. We now mention it in the discussion as follows:

> Furthermore, by making group reputation a valuable asset, group reciprocity could encourage groups to differentiate themselves symbolically from others, and to police their members' behaviour towards outgroups—both behaviours that we indeed observe in humans (Fearon and Laitin, 1996).