July 24, 2018

Professor Daniel Hruschka
Editor
Evolution and Human Behavior

Re: Resubmission of Manuscript EVOLHUMBEHAV_2017_270

Dear Professor Hruschka,

We hereby submit a revised version of our manuscript titled "Human reciprocate intentional harm by discriminating against group peers." We hope that you will find this version suitable for publication in *Evolution and Human Behavior*.

In the following letter, we list and respond to to the comments of the reviewers. For ease of reading, the original comments are presented in italics, and quotes from the new manuscript appear in frames. We leave some of the minor comments made in the pdf annotations out of this response letter, but provide a full mark up of our revision for reference.

Sincerely,

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

Editor

1. *The reviewers include one of the original reviewers who feel you adequately addressed that reviewers' specific comments and another new reviewer who raises a number of important continuing issues with the paper that I agree need to be addressed. These include tempering frequent strong and unsubstantiated claims and a number of concerns about analysis, presentation and interpretation.*

Also, please remove the discussion of the simulations. Upon further reflection, they add very little to the paper and should contribute to an independently reviewer paper. You might consider submitting that paper to the Journal of Theoretical Biology.

Thank you for the additional opportunity to revise our manuscript. We did our best to address the comments and suggestions of the reviewers. Following your suggestion, we removed the simulations from the manuscript and are indeed considering developing them into an independent paper.

Reviewer #1

1. *The authors have adequately addressed the points I raised in my initial review. As the authors point out in the discussion, this study is a first step and I think it will contribute to new theorizing on the basis of the questions they raise.*

In my opinion, this paper should be accepted pending light edits detailed below.

Thank you for the positive evaluation.

2. *Page 12, above section 3.1. "indeed, column of Table …" there should be a column name after the world "column".*

Done.

3. *Page 7, second line in 1st full paragraph: "in which the individual could help others". Here you should specify which individual it is as it's not immediately clear.*

We rephrased thus:

> The upstream action was followed by the reciprocal action, in which each individual who participated in the TG could help others.

4. *Page 7–8: it's not clear to me whether initially the participants are interacting in person or they're all at terminals in a room and instructions are then read to the participants: "instructions read aloud".*

The new version states that:

> Each session consisted of 24 participants, sitting at isolated computer terminals.

5. *Page 8: "each player … had to allocate 100 tokens within the group. The allocator always received exactly 30 tokens and could freely allocate the remaining 70…" Doesn't this mean that each player had to allocate only 70 tokens, not 100? Because 30 were designated for their personal retention? Please reword on revisions.*

We rephrased thus:

> Each round consisted of a random dictator game, in which each player chose how to divide 100 tokens between the three group members if he or she is chosen to be the allocator. The allocator always received exactly 30 tokens, and could freely allocate the remaining 70 tokens between the other two players.

6. *Page 11 "his team mates" were there only male players? You can use "their teammates". 'Their' is acceptable in English as a gender neutral single pronoun.*

Changed.

7. *I believe "team mates" is usually spelled as one word as in 'teammates'*

Changed.

8. *New work in Nature Human Behaviour recently came out that is relevant to discussion of whether aggression is aimed at retaliation (in this case—defense) shows that much of what may appear to be offensive aggression is actually motivated by defensive strategizing. It doesn't undermine any of the claims in this paper but is relevant to the literature and the discussion of intentionality (Spoils division rules shape aggression between natural groups. Dogan et al.).*

Our use of the allocation game precludes using discrimination as defensive strategizing. Given the large number of papers studying intergroup contests, we don't think this paper's relevance is enough to discuss it in the paper.

Reviewer #3

1. *This manuscript offers a glimpse into the positive and negative group reciprocity, in that individuals may reward or punish out-group members for the behavior of an out-group target. For addressing this question, the combination of the allocation and trust games seem a good choice. However, as written, the paper is too strident about its results, relying extensively on claims about universality that are not supported either by the results it describes or the papers it cites. Other major issues include lack of precision in the prose and the need for random intercepts by participant, nested within session. I'll walk through these comments here, but note that I have left more detailed comments in the attached PDF.*

The authors begin the manuscript – from the title onwards – with big claims about humans and their behavior towards out-groups. Humans do not universally display stereotyping and prejudice towards out-groups (as implied by paragraph one), a universal propensity to group-reciprocate is claimed but not demonstrated, and the title makes too strong of a statement about the results of the paper. The discussion does a much better job of illustrating the plasticity that characterizes human intergroup relationships: individuals cognizing out-groups in different ways. If the rest of the paper was as moderate as this section, it would do a more accurate job of characterizing humans.

I agree that humans do exhibit evidence of group reciprocity, per vengeance behavior and the generalization from one to many documented by work on stereotypes (both positive and negative) and intergroup contact. The paper and its results could still be strong with a more moderate framing: that humans can use group reciprocity, not that it's universally the case. Further, as the results are ultimately about both positive and negative group reciprocity, given the experimental design, an introduction that acknowledges that behavior towards out-groups can be positive or negative would better foreground the study.

On this topic, and elsewhere in the paper, big claims are made that need citations. I've made some suggestions in the PDF on which papers support the claims made, or where these claims should be tempered.

Throughout the text, and most especially in the introduction, there's a lot of imprecision in the prose. Some sections are hard to follow, and assumptions are made about the benefits of behaving aggressively toward out-group members (what are those benefits?), that the mind represents groups as entities (this comes up as an aside, with no cognitive psych support, even though it's out there (see Tooby, Cosmides, & Price 2006)), that reputation is a confound with respect to the predictions at hand (how so?). These assumptions are shaky and need more theoretical support given the audience of E&HB. Further, it cannot be assumed that the E&HB audience knows the psych terms included in the manuscript, given how diverse the audience is. I've flagged a number of places where more clarity is needed in the attached PDF, with explanations for why it is needed.

Thank you for the detailed suggestions. We revised the manuscript accordingly, qualifying claims of universality, and adding clarifications and references where needed. Below we list the changes in detail. Note that we skip the comments regarding the simulations, which we took out of the paper.

One small point: Reviewer 3 sent comments in a PDF file. Some of these comments were empty, and the PDF software revealed the reviewer's name.

2. *I think the model with random intercepts for each subject, nested within random intercepts for session, should be the model reported in the paper, not relegated to the response to reviewers. There are multiple decisions per participants being analyzed in a single model. Clustering by participant is needed.*

Regressions with nested random intercepts for subjects and sessions either failed to converge, or gave results mathematically identical to ones with random intercepts for subjects alone. This is probably because estimating random intercepts from 8 sessions (i.e. estimating the mean and variance of a normal distribution for the intercepts, simultaneously with the distribution of per-subject intercepts) is too demanding. We therefore ran regressions with random intercepts per subject, and fixed effect dummies for sessions (and clustered errors per subject). This is less efficient than random errors per-session, but robust to correlation between sessions and other independent variables. Results are in the paper and are qualitatively unchanged.

3. *I agree with Reviewer 2 that you should drop the SVO: its measurement following the games is problematic; you provide no a priori reason about why it should affect group reciprocity (as you note explicitly) and not direct reciprocity or in-group favoritism, and it doesn't add anything to the manuscript. After you mention that you collected the SVO and collectivism data, mention that they are not included in this paper as you had no a priori predictions for why they should affect behavior. (On that note, it seems to me that the collectivism thing does have the possibility to be predictive, although whether it should be more predictive of in-group favoritism or group reciprocity is unclear to me…please address why it was not included in the models in the manuscript.)*

We removed the results from the SVO, as suggested. We note briefly when describing the procedure that:

> The results from the social value orientation and collectivism measures did not reveal any systematic and interpretable pattern, and are therefore not included in this paper.

4. *Contingent giving is not analyzed as a potential shortcoming of the repeat appearance of certain players in the Allocation Game. Though participants were told only one decision would be randomly selected as the basis for payouts, it is very unlikely the mind is good at representing each allocation decision as independent of the others. (Which is another reason why you need random intercepts by participant in the paper.) So, taking the example from Table 1, Brown 2 appears in two allocation decisions for Blue 2. One could imagine that if Blue 2 gives Brown 2 more in the first allocation decision, Blue 2 might give Brown 2 less in the second, thinking "I already gave him/her some." This is something that should be checked: that decisions in the allocation game are not contingent on previous decisions made for targets that are encountered twice.*

Note that if participants give less (or more) to a recipient who was previously encountered, this might bias the estimate of in-group favoritism. Direct and group reciprocity, in contrast,

are measured as the slope on the TG experience, which is orthogonal to repeated encounters. Therefore, such effects would only inject noise and bias our measures downwards. Nevertheless, we tested our results for robustness to repeated encounters effects in two ways. First we reran the regressions of table 2, including a dummy variable for having encountered the same participant before. This was significant for regressions on amount given by senders to recipients only. The main results were qualitatively unchanged, as follows:

	Allocation	Discrimination	Reciprocity
Senders			
Baseline	35.00 (—)	6.90 (5.11)	—
Direct Reciprocity	32.55 (2.02)	25.03 (5.56)***	17.02 (3.72)***
Group Reciprocity	33.27 (2.21)*	10.97 (5.52)**	9.08 (4.55)*
In-Group	37.56 (3.82)	18.56 (5.64)***	1.63 (8.39)
Responders			
Baseline	35.00 (—)	3.94 (1.75)	—
Direct Reciprocity	37.17 (1.62)	23.88 (3.06)***	20.80 (7.76)**
Group Reciprocity	36.66 (1.89)	7.79 (2.39)***	1.10 (2.38)
In-Group	43.93 (1.43)***	18.88 (1.67)***	4.65 (4.87)

We also reran the base regressions including only allocations to participants who had not been encountered before. Again, the results were qualitatively unaffected:

	Allocation	Discrimination	Reciprocity
Senders			
Baseline	35.00 (—)	6.41 (5.20)	—
Direct Reciprocity	32.84 (2.62)	24.85 (5.63)***	17.00 (3.68)***
Group Reciprocity	33.55 (2.98)*	11.10 (5.57) **	10.57 (4.53)*
In-Group	37.84 (4.18)	18.38 (5.67)***	1.61 (8.44)
Responders			
Baseline	35.00 (—)	3.97 (2.18)	—
Direct Reciprocity	37.66 (1.72)	24.18 (3.44)***	20.93 (7.62)**
Group Reciprocity	37.43 (2.11)	8.41 (3.02)***	1.62 (2.86)
In-Group	44.41 (1.44)***	19.18 (1.88)***	4.78 (4.77)

This is acknowledged in new Footnote 3:

> Some participants are matched with each other in two different rounds, as Blue 1 and Brown 2 in the example provided in Table 1. This should not matter, as only one round is actually paid. The results are robust both to controlling for and to removing repeated encounters from the regressions.

5. *I have no strong feelings about whether the simulations described in the manuscript stay or go, and I see that Reviewers 1 and 2 didn't agree on this either. As written, some of the take-home points were not clear (again, issues with the precision of the prose), and it's not*

immediately apparent to me that the simulations add anything to the paper. If the simulations were excluded, I don't think it would detract from the paper, as by the end of the manuscript, the simulations were ultimately forgettable (and not discussed in the Discussion at all).

We removed the simulations following the editor's request.

6. *The Discussion gets a little redundant, goes off topic, and is a paragraph or two longer than it needs to be. I've made some suggestions for where it can be condensed.*

See our responses to the specific suggestions below.

Comments made in the PDF file

7. *Title feels too strong: too universal on several fronts.*

We prefer to keep the title, and trust the readers to understand that the 'under certain circumstances' caveat always applies.

8. *Abstract: "group reciprocity was only triggered when the partner's intentions were unequivocal": This is confusing.*

We don't go into details due to the space limitations in the abstract. We rephrased to clarify:

> Receivers' allocations to group members were not affected by their partners' play in the trust game, suggesting that group reciprocity was only triggered by strong norm violations.

9. *Page 2: This is a broad overgeneralization. Humans are very flexible in their behavior towards out-group members. See Hruschka & Henrich 2013 and Pisor & Surbeck PeerJ preprint for relevant reviews of the literature.*

We agree that humans flexibly employ behavioral strategies in intergroup relations. This does not contradict our statement that humans display stereotyping and prejudice towards out-groups. To avoid making a too-strong statement, we changed the text to say that humans *may* display stereotyping and prejudice towards out-groups.

10. *Page 2: What do you mean by group structure?*

We replaced "Group structure" with "Organization into groups…"

11. *Page 3: If you're going to use the phrase "negative reciprocity" (which is a term that comes with some baggage: see Sahlins' Stone Age Economics), you might as well define it and use it in the second paragraph. I found it jarring to read about reciprocity in reference to reciprocal harms in a parenthetical statement; spend a few more words defining "positive" and "negative" reciprocity in that paragraph, and add a footnote saying you don't mean it in the way that Sahlins uses it.*

We added a clarification:

> …negative group reciprocity, i.e., reciprocity where harm is reciprocated with harm.

12. *Page 4: Word choice: use reciprocity in the first part of the sentence too, to keep your point clear.*

Done.

13. *Page 4: You never established what "does better" means. What is the currency in which chimps are benefitting from attacking?*

We rephrased thus:

> although both groups would do better if neither attack—thus avoiding costly conflict—each group does better by attacking when the odds are good enough, thereby gaining territory, resources, etc.

14. *Page 4: This needs a citation. I also don't follow the claim: population density is not the same thing as population numbers, and I think the latter is what you mean. Populations might actually be more dense if they're avoiding a group with whom they have tension, if they're squeezed into more space because they're avoiding a buffer region.*

…

Fearon & Laitin 1996…really relevant to this sentence and the next.

…

I don't follow how ranging over larger areas increases population density. Again, doesn't that lower it, as people are spreading out? Some more precise language is needed in this paragraph.

In line with your comments, we rephrased as follows. The argument continues the argument due to Wilson and Wrangham (2003) cited earlier in the paragraph.

> Indeed, peaceful unsegmented societies resolve intergroup conflict by avoiding the other group, which entails a loss of access to valuable resources, constricting population expansion.
> The evolution of group reciprocity could deter opportunistic conflict. When there is group reciprocity, someone who harms an outgroup member brings retaliation on his own group, and this gives his group members an incentive to maintain peace (Boehm, 1984; Fearon and Laitin, 1996).

15. *Page 5: When group boundaries are salient. Also, universality is a very strong statement to make without supporting evidence (in the form of citations at least)*

…

Non sequitur? What's the relevance of these things they're loaded with?

We rephrased this paragraph and added citations. Note that we don't claim universality. On the contrary, we argue here that the existing evidence is merely suggestive.

Real world examples of apparent intergroup revenge suggest there may be a human propensity to group-reciprocate (Bauerlein, 2001; Chagnon, 1988; Horowitz, 1985; Horowitz, 2001). That is, individuals sometimes, but not always, discriminate against outgroups. Behavior toward outgroup members varies on the basis of the individuals' experiences with the outgroup. In this paper, we aim to study the existence and form of the proximate psychological mechanism for group reciprocity in modern humans. Although field observations from conflict are highly suggestive, it is hard to identify group-reciprocity motives in naturally occurring data. Actions that may look like group reciprocity may be rooted in other motivations—such as wishing to signal group strength—or reflect centralized group decisions rather than individual tendencies for group reciprocity (Gould, 2000; Mamdani, 2001).

16. *Page 5: Language needs more precision here. Do you mean something like a raid that attacks a small gathering of people that includes the perpetrator, but others are killed because they are encountered first (à la Yanomamo), vs these other group members being specifically targeted because they're not the perpetrator? Also, what do you mean by unrelated? Low r or not involved in the perpetration? Again, precision needed.*

…

The second sentence just makes things less clear. Are you testing mechanisms or the presence of group reciprocity? Pick one of the two statements.

…

If this really is individual level, like you say it is, then what you probably mean is that the subject is trying to reciprocate toward the target but there are positive externalities as a consequence of that that benefit the target's same-group members.

We clarify in the following.

Moreover, in the field, group reciprocity may be conflated with individual level reciprocity, i.e., acts that aim to help or harm the perpetrator, but have side effects on the entire group. We therefore designed a controlled laboratory experiment to test the human tendency for group reciprocity in a clean way.

17. *Page 6: This is the first appearance of reputation in the paper, and it's not clear why this is a confound. If you're arguing that group reciprocity likely originally evolved via individual-level selection, then something like within-group reputation could be really important to getting this kind of system off the ground. If you're spinning off a Nowak image-scoring idea, make that clear.*

We agree that reputation building could be part of the evolutionary basis of group reciprocity. When considering the proximate mechanism, it's important to make the distinction between strategic reputation building, which is forward looking, and group reciprocity, which responds to past actions. We state so explicitly in the new version:

…and actions driven by strategic reputation-building rather than by a motivation to reciprocate past actions.

18. *Page 6: I don't follow. The subject knows who the recipient is in the Direct Reciprocity case, right? Isn't that reputational? Is what you mean that in the non DR case, subjects had no way of knowing what the recipients had done previously?*

While the actions taken as part of the initial interaction are public, the reciprocal action is not. We clarify:

we minimize strategic concerns by not giving feedback about the reciprocator's action.

19. *Page 6: Precision. What I think you mean is, subjects weren't given the opportunity to exhibit hostility towards just one individual from that group; what was instead measured was the presence or absence of hostility toward that group. Right?*

…

No definition given. If you mean "uniified entitiy" again, then just say that.

That is right. We have clarified:

Gaertner, Iuzzini, and O'Mara (2008) found that rejection by one group member leads to more hostility towards the group when "entitativity" is high, i.e. the group is perceived as a unified entity. Since subjects could only display hostility to the whole group (by exposing them to unpleasant noise), individual and group level reciprocity were confounded.

20. *Page 6: I think there's a sentence missing here. What's the relevance of the Strenstrom et al paper to the rest of the paragraph?*

…

What is group structure??

The new version clarifies the limitations of the Stenstrom et al. (2008) study vis-à-vis group reciprocity and the features of our experiment.

Similarly, Böhm, Rusch, and Gürerk (2016) examine intergroup retaliation using the intergroup prisoner's dilemma paradigm, but cannot distinguish between individual and group reciprocity. Stenstrom et al. (2008) manipulated entitativity by making the original perpetrator (a political analyst) "tightly affiliated" with the group (a presidential campaign). Under this manipulation, holding the group accountable for the perpetrator's act can be practically and/or legally justified, without resorting to group reciprocity. In contrast, we look at how people reciprocate a clear individual act by one group member to an uninvolved other group member, where groups are created in the lab and are free of existing social context.

21. *Page 6: Who formed the groups, the experimenters or the subjects?*

We added the following sentence:

> Participants were randomly assigned to groups and collaborated on a task to build group identity.

22. *Page 6: What social norm? In what context? WEIRD samples?*

The cited paper by Kimbrough and Vostroknutov (2016), which uses a similar sample of western students, supports this claim and provides details. We return to this issue in the discussion section.

23. *Page 7: sounds like "strong reciprocity." Pick a different phrase.*

We changed "stronger reciprocity" to "more reciprocity".

24. *Page 7: Unnecessary (you've already established this) and sounds flippant.*

...

Where does the Ingroup Favoritism factor in?

We deleted the first sentence, and added:

> Lastly, Ingroup favoritism rounds checked that we had successfully created group identity among participants.

25. *Page 9: Any effects of contingent giving, then? For example, even though subjects were told only one round would be paid out, one could think "I already gave a good amount to Brown 2. I'll give them less this time."*

See our response to comment #4 above.

26. *Page 10: So if the IG favoritism decision was chosen for payouts (1/6 chance), these participants would potentially have a very different SVO measure than the other 5/6 of participants. Also, if they got a small or no allocation from an out-group member on payoff, they could exhibit more in-group favoritism on the SVO. How did you deal with this?*

The participants did not receive any feedback about dictator game allocations, or the rounds selected for payment, until after the SVO measure at the very end of the experiment.

27. *Page 10: Walk us through it here: this is the difference between the amount allocated to an IG, an OG direct, or an OG indirect versus a neutral, every single time.*

We rephrased to clarify:

> we predict that the absolute difference between the amounts allocated to the two recipients will be larger in all treatments compared to the baseline.

28. *Page 12: Redundant: condense into one sentence*

We think redundancy helps to clarify the results here.

29. *Page 13: A good place to say why, per my comment earlier*

We removed the description of the SVO results.

30. *Page 13: "under certain circumstances." This is the third claim of universality you've made in this paper and none of them have been sufficiently supported.*

We added "under certain circumstances". See our responses above for previous comments.

31. *Page 13: Bingo. This is the subtlety needed in the first paragraph: out-groups can be discriminated against, but are not always, and are discriminated against to different degrees when they are on the basis of an individual's experiences of them. I recommend putting this upfront rather than saving this important distinction for the discussion.*

We incorporated some notions that appear here into the introduction. See our response to comment #15 above.

32. *Page 15: "refraining from benefitting"*

We reworded from "harming" to "exploiting".

33. *Page 15: Cite intergroup contact theory here.*

We added Footnote 6:

> Relately, intergroup contact theory stipulates that contact between groups reduces prejudice by correcting misperceptions of the motives driving outgroup behaviors (Pettigrew, 1998).

34. *Page 15: Lots of trust games out there. This is a good opportunity to cite your claims about how senders interpret recipients' behavior with data, rather than just guessing*

We provide references in the next paragraph.

35. *Page 15: Connect to rest of the paragraph.*

Done.

36. *Page 16: All the ideas from here until the end of the paragraph seem not to belong in this paper: the data don't provide sufficient data to assess this.*

We have moved the discussion of hunter-gatherers to earlier in the paper, with the discussion of war/peace systems and population density. These ideas are not tested here, but set the context for why group reciprocal motivations could matter. We mention third party group reciprocity in the conclusion, as a topic for future research.

37. *Page 16: This needs a cite from Fearon & Laitin.*

Done.

38. *Page 16: This is a new use of the word "norm" in this paper; be more precise.*

Fixed.