THE ROYAL SOCIETY PUBLISHING

PROCEEDINGS B

Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait

Maxwell N. Burton-Chellew and Claire Guérin

Article citation details
Proc. R. Soc. B **288**: 20211611.
http://dx.doi.org/10.1098/rspb.2021.1611

Review timeline
Original submission: 16 July 2021
Revised submission: 11 October 2021
Final acceptance: 18 October 2021

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2021-1611.R0 (Original submission)

Review form: Reviewer 1

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Acceptable

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This experimental paper looks at the conditions under which people will punish in a repeated public goods game. To get at this the authors consider three experimental scenarios, 1) in which all players can mutually punish each other, 2) in which one individual is immune from being punished (but all players can punish and all others can be punished) and 3) in which only one sole player can punish. They find that individuals who are immune from punishment (scenarios 2-3) reduce their contributions to the public good over time, and punish less overall compared to non-immune individuals. The interpretation of this result is that cooperation-punishment do not form a single altruistically motivated trait, which contradicts the "strong reciprocity" hypothesis and (in contrast to previous results which confounded a cooperative environment with a fear of punishment), but supports the idea that players learn over the course of an experiment. I find these results interesting and valuable, and I support publication of the paper.

The experimental methodology is fully explained, and the statistical analysis is satisfactory. The main difficulty I have in reading the paper is the motivation for and interpretation of the results, which at times is a bit confused and jargon heavy. The paper is focused on the debate over strong reciprocity and provides clear evidence that cooperation and punishment can decouple. But the idea of cooperation-punishment as a single altruistic linked trait needs to be explained more clearly in the context of the experiment, and, ideally, in intuitive terms. In particular the authors need to spell out what exactly it means, in humans, for cooperation and punishment to "stem from one conjoined, altruistic, trait (dubbed 'strong reciprocity')" (line 60). Is the idea that cooperation and punishment are somehow intrinsically linked by a genetic architecture which uniquely determines behavior, so that punishment could never occur without the cooperation and vice versa? Or is it rather that the two behaviors co-evolved as a single successful strategy, which could be decoupled if incentives change? And in the latter case, what is our expectation for players' behavior in this experiment? In previous studies, as the authors show, the apparent evidence for strong reciprocity can be explained as an artifact arising from fear of punishment. In this experiment we see confused learning, in which punishment and cooperation decouple over time. However I am unclear what we would expect a confused learner under the strong reciprocity hypothesis to look like except under the very strong assumption that cooperation and punishment cannot decouple for mechanistic reasons. I see four potential categories of conditional cooperator in this experiment i) {strong reciprocity, no learning}, ii) {no strong reciprocity, no learning}, iii) {strong reciprocity, learning} and iv) {no strong reciprocity, learning}. The conclusion of the paper is that we see iv) but i'm not clear how we would distinguish iii) and iv). If the authors can explore this distinction, the importance of the paper will be much clearer.

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Excellent

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
In this paper, the authors report the results of a cleverly designed public goods game that decouples cooperation and punishment to stringently test the altruistic punishment/strong reciprocity hypothesis. The paper is clear, very well written, the analyses are sophisticated and appear correctly executed, and the argumentation is compelling. I do a lot of research in this area and can confidently say this will be a very well-cited and influential paper, and I think these results shed some very important clarifying light on the field. Frankly, I think the paper could be published as is. I don't have many substantive comments to give, but here are some minor points that I noted.

- In lines 89-94, the authors allude to the confused learner hypothesis. I'm personally familiar with this work and think it's a very relevant counterpoint to the altruistic punishment hypothesis's interpretation of PGG results, but I'm not sure it's common knowledge in the field yet (it should be!). I think the authors would do well to have an additional few sentences setting up that work by briefly describing what some of findings are and how it contradicts the AP

hypothesis. Particularly since this is ultimately what the authors endorse in the discussion, I think elaboration is needed.

- On lines 149 and 151 the authors mention that the maximum MUs were different in sessions 1 and 2 than in later sessions. I think this warrants either a footnote or a reference to the supplement explaining why.

- At the point the authors introduce the classified "conditional cooperators' on line 230, they have only briefly introduced how they categorized people at the end of the methods section and it's hard to follow exactly what went into that (e.g., line 129 says "approximately match the mean average contribution of their groupmates"). Some more clarification here would be good.

- I don't think it undercuts the point made in the "immune individuals punished less" section (lines 268-286, but what do the authors make of the level of punishment in the sole punisher condition, which appears to be in the ballpark of non-immune players? They make the comparison directly to immune punishers, but it does seem interesting that sole punishers appear equivalent to mutual punishers, which seems to me predicted by both the AP and CL hypotheses so it's not particularly informative. My take, in line with how the authors discuss it, would be that the direct comparison in the immune punisher condition is most relevant for the free riding question.

- Line 319: "..the instances of social punishment" I think a "pro-" is missing.

- In the discussion on line 367 the authors mention the inconsistency in punishment. It might be worth referring here again to the work on the confused learner hypothesis: some of this might be explained by participants simply not quite grasping how to maximize their payoffs in the game and are either testing different strategies or responding somewhat randomly.

Decision letter (RSPB-2021-1611.R0)

08-Sep-2021

Dear Dr Burton-Chellew:

I have now received comments on your manuscript from two peer reviewers and an Associate Editor. I have also read your paper, which I enjoyed; your experiment is clever and I think an important addition to the literature on human altruism and punishment. However, I also agree with the reviewers that while the experiment is strong, the framing needs revision to improve its impact. The most important issue is how strong reciprocity is treated, as discussed by Reviewer 1, and I also agree with Reviewer 2 that the paper is somewhat jargon-heavy in places and will not be particularly accessible outside of the field. This needs to be addressed as this is a topic that should have broad appeal. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. Of course, please be sure to address each of their comments fully.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with

Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" - in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article. Please see our Data Sharing Policies (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the

accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
Two reviewers have provided feedback on this paper. Both commend the experimental design and the important contribution of this study to the literature. However, both reviewers also propose that the article could be enhanced considerably via some careful reframing of the study aims and conclusions. I agree that these suggestions are valuable, especially given the broad readership of this journal. Decreasing jargon where possible, providing real world examples, and perhaps discussing how this fits within a broader framework of cooperation seen outside of experimental contexts and/or in different species would increase the accessibility and reach of this article.

Additionally, I have some minor questions and requests for clarification
1. I see that demographic information about the participants is provided in the supplementary materials, but this could be more explicitly referenced in the main article.
2. line 130 please define the "Other" category
3. line 197 please describe more clearly what you mean by "original result" and provide a citation if this refers to a previous study (is this [7] that is referenced on line 261?)
4. line 352 does the "they" here refer to the punisher? I assume so, but please clarify.
5. From looking at Figure 2, it seems that the is much greater variance in the "can be punished" participants' responses in the immune and sole punisher conditions (i.e., in their rate of contribution change over time). this variation seems worth discussing.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
This experimental paper looks at the conditions under which people will punish in a repeated public goods game. To get at this the authors consider three experimental scenarios, 1) in which all players can mutually punish each other, 2) in which one individual is immune from being punished (but all players can punish and all others can be punished) and 3) in which only one sole player can punish. They find that individuals who are immune from punishment (scenarios 2-3) reduce their contributions to the public good over time, and punish less overall compared to non-immune individuals. The interpretation of this result is that cooperation-punishment do not form a single altruistically motivated trait, which contradicts the “strong reciprocity” hypothesis

and (in contrast to previous results which confounded a cooperative environment with a fear of punishment), but supports the idea that players learn over the course of an experiment. I find these results interesting and valuable, and I support publication of the paper.

The experimental methodology is fully explained, and the statistical analysis is satisfactory. The main difficulty I have in reading the paper is the motivation for and interpretation of the results, which at times is a bit confused and jargon heavy. The paper is focused on the debate over strong reciprocity and provides clear evidence that cooperation and punishment can decouple. But the idea of cooperation-punishment as a single altruistic linked trait needs to be explained more clearly in the context of the experiment, and, ideally, in intuitive terms. In particular the authors need to spell out what exactly it means, in humans, for cooperation and punishment to "stem from one conjoined, altruistic, trait (dubbed 'strong reciprocity')" (line 60). Is the idea that cooperation and punishment are somehow intrinsically linked by a genetic architecture which uniquely determines behavior, so that punishment could never occur without the cooperation and vice versa? Or is it rather that the two behaviors co-evolved as a single successful strategy, which could be decoupled if incentives change? And in the latter case, what is our expectation for players' behavior in this experiment? In previous studies, as the authors show, the apparent evidence for strong reciprocity can be explained as an artifact arising from fear of punishment. In this experiment we see confused learning, in which punishment and cooperation decouple over time. However I am unclear what we would expect a confused learner under the strong reciprocity hypothesis to look like except under the very strong assumption that cooperation and punishment cannot decouple for mechanistic reasons. I see four potential categories of conditional cooperator in this experiment i) {strong reciprocity, no learning}, ii) {no strong reciprocity, no learning}, iii) {strong reciprocity, learning} and iv) {no strong reciprocity, learning}. The conclusion of the paper is that we see iv) but i'm not clear how we would distinguish iii) and iv). If the authors can explore this distinction, the importance of the paper will be much clearer.

Referee: 2
Comments to the Author(s)
In this paper, the authors report the results of a cleverly designed public goods game that decouples cooperation and punishment to stringently test the altruistic punishment/strong reciprocity hypothesis. The paper is clear, very well written, the analyses are sophisticated and appear correctly executed, and the argumentation is compelling. I do a lot of research in this area and can confidently say this will be a very well-cited and influential paper, and I think these results shed some very important clarifying light on the field. Frankly, I think the paper could be published as is. I don't have many substantive comments to give, but here are some minor points that I noted.

- In lines 89-94, the authors allude to the confused learner hypothesis. I'm personally familiar with this work and think it's a very relevant counterpoint to the altruistic punishment hypothesis's interpretation of PGG results, but I'm not sure it's common knowledge in the field yet (it should be!). I think the authors would do well to have an additional few sentences setting up that work by briefly describing what some of findings are and how it contradicts the AP hypothesis. Particularly since this is ultimately what the authors endorse in the discussion, I think elaboration is needed.

- On lines 149 and 151 the authors mention that the maximum MUs were different in sessions 1 and 2 than in later sessions. I think this warrants either a footnote or a reference to the supplement explaining why.

- At the point the authors introduce the classified "conditional cooperators' on line 230, they have only briefly introduced how they categorized people at the end of the methods section and it's hard to follow exactly what went into that (e.g., line 129 says "approximately match the mean average contribution of their groupmates"). Some more clarification here would be good.

- I don't think it undercuts the point made in the "immune individuals punished less" section (lines 268-286, but what do the authors make of the level of punishment in the sole punisher condition, which appears to be in the ballpark of non-immune players? They make the comparison directly to immune punishers, but it does seem interesting that sole punishers appear equivalent to mutual punishers, which seems to me predicted by both the AP and CL hypotheses so it's not particularly informative. My take, in line with how the authors discuss it, would be that the direct comparison in the immune punisher condition is most relevant for the free riding question.

- Line 319: "..the instances of social punishment" I think a "pro-" is missing.

- In the discussion on line 367 the authors mention the inconsistency in punishment. It might be worth referring here again to the work on the confused learner hypothesis: some of this might be explained by participants simply not quite grasping how to maximize their payoffs in the game and are either testing different strategies or responding somewhat randomly.

Author's Response to Decision Letter for (RSPB-2021-1611.R0)

See Appendix A.

Decision letter (RSPB-2021-1611.R1)

18-Oct-2021

Dear Dr Burton-Chellew

I am pleased to inform you that your manuscript entitled "Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Data Accessibility section
Please remember to make any data sets live prior to publication, and update any links as needed when you receive a proof to check. It is good practice to also add data sets to your reference list.

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions

(http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 10 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,
Dr Sarah Brosnan
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
Thank you very much for your careful and thorough revisions. I think the clarity and accessibility of your work is now much enhanced. The introduction in particular much more clearly sets up the background literature and rationale for your work in a manner that will be engaging to a broad audience. I think this is an important and interesting contribution to the field.

Appendix A

08-Sep-2021

Dear Dr Burton-Chellew:

I have now received comments on your manuscript from two peer reviewers and an Associate Editor. I have also read your paper, which I enjoyed; your experiment is clever and I think an important addition to the literature on human altruism and punishment. However, I also agree with the reviewers that while the experiment is strong, the framing needs revision to improve its impact. The most important issue is how strong reciprocity is treated, as discussed by Reviewer 1, and I also agree with Reviewer 2 that the paper is somewhat jargon-heavy in places and will not be particularly accessible outside of the field. This needs to be addressed as this is a topic that should have broad appeal. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. Of course, please be sure to address each of their comments fully.

Best wishes,

Dr Sarah Brosnan
Editor, Proceedings B

RESPONSE: Thank you for your kind comments and the opportunity to resubmit our manuscript! We have responded to all the comments and improved our manuscript by reducing jargon, using more familiar language and examples, and taking more time to explain key concepts. We have also expanded our discussion and included more non-human citations.

Associate Editor
Board Member: 1
Comments to Author:
Two reviewers have provided feedback on this paper. Both commend the experimental design and the important contribution of this study to the literature. However, both reviewers also propose that the article could be enhanced considerably via some careful reframing of the study aims and conclusions. I agree that these suggestions are valuable, especially given the broad readership of this journal. Decreasing jargon where possible, providing real world examples, and perhaps discussing how this fits within a broader framework of cooperation seen

outside of experimental contexts and/or in different species would increase the accessibility and reach of this article.

RESPONSE: we have added a paragraph to our introduction that is more accessible and provides real world hypothetical human examples.

"Can human cooperation, often considered biologically unique, be explained by a phenomenon of altruistic punishment, whereby individuals punish non-cooperators for the good of society? Do people, for example, if they see someone littering in public, harming others, or using public transport without paying, punish them? Economic experiments have attempted to model these situations. Key results suggest that human cooperation in such situations is reliant on the presence of 'altruistic punishers' who police a minority of non-cooperators even though they have nothing to gain."

We have then decreased jargon by removing the terms 'Conditional Cooperation' and 'Free Riders' from the opening, where they were not yet needed, and moved them to the Methods. Instead, we simply refer to cooperators and non-cooperators at this stage.

We have also added two paragraphs to our discussion to discuss the broader framework of cooperation and punishment in and outside the laboratory, including many non-human citations.

"We suggest that punishment and cooperation are social behaviours better understood through the evolutionary benefits they potentially offer to the actors. For example, punishment can provide reputational benefits or lead to more cooperation in long-term partners. The altruistic punishment paradigm has tested a severely restricted behavioural interaction, but outside the laboratory, behaviour is more open ended, meaning that would-be punishers face more benefits, but also more potential costs, such as from retaliation ('counter punishment') or feuds.

The laboratory evidence for altruistic punishment also suffers from other findings. Specifically, the fact that that costs of punishment tend to erode any collective gains from cooperation in such experiments (Supplementary Figure 5), challenges the idea that altruistic punishment could even be favoured by group selection. Although some evolutionary models work on the assumption that punishment will be sufficiently rare when altruistic punishers are common, meaning altruistic punishers will not be at too large a disadvantage. However, our results and other experiments show that the laboratory behaviours taken as evidence for altruistic punishment in cooperative societies are frequent and easily triggered."

Additionally, I have some minor questions and requests for clarification

1. I see that demographic information about the participants is provided in the supplementary materials, but this could be more explicitly referenced in the main article.

RESPONSE: we have now included demographic data in the main text Methods, specifically,

"*Participants were mostly students enrolled at either UNIL or the Swiss Federal Polytechnic School (EPFL). We had a near equal gender ratio (202 Female, 215 Male, 2 Other, and 1 declined to answer) and most of our participants were under 26 years of age (134 aged under 20, 257 aged 20-25, 23 aged 26-30, 2 aged 30-35, 3 x Over 35, and 1 declined to answer).*"

2. line 130 please define the "Other" category

RESPONSE: we have now defined the Other category,

*"This allowed us to categorize individuals into different types: Conditional Cooperators that approximately match the mean average contribution of their groupmates; Free Riders that never contribute regardless of what their groupmates contribute, and Other/**Unclassified**, **who satisfied neither of these criteria."***

3. line 197 please describe more clearly what you mean by "original result" and provide a citation if this refers to a previous study (is this [7] that is referenced on line 261?)

RESPONSE: Yes it was reference 7, Fehr & Gachter 2002. We have now made this clearer,

*"When all four groupmembers could punish (Mutual-Punishers scenario), mean contributions were stable across the five rounds between 42-45%, showing no significant decline over time**,** replicating the original **~~result~~ findings of stable contributions under punishment reported by Fehr & Gachter [7]."***

4. line 352 does the "they" here refer to the punisher? I assume so, but please clarify.

RESPONSE: Yes you are correct, our apologies, it referred to the immune punishers. We have now made this clear in the text.

*"~~They~~ **Immune individuals** also often continued to punish intermediate contributors, but not the top contributors, even though ~~they~~ **immune individuals often** hypocritically contributed less (Figure 5, Tables 1 and 2)."*

5. From looking at Figure 2, it seems that the is much greater variance in the "can be punished" participants' responses in the immune and sole punisher conditions

(i.e., in their rate of contribution change over time). this variation seems worth discussing.

RESPONSE: This was interesting but we have tested the data, in R with var.test(), and found no significant difference in the variances between immune and non-immune players in either scenario, or combining scenarios (analysing all rounds of the game together). Nor among non-immune players in the Mutual Punishers scenario versus the other scenarios. We repeated the variance tests for just round 1 of the game and again found no significant differences. In total, the lowest p-value we obtained from 8 tests was 0.245 (comparing Mutual Punishers versus non-immune players in the other two scenarios combined). Variance is difficult in public goods games because the contributions are bounded between 0-100%, and so when the mean is closer to 50% the variance tends to be bigger anyway.

Reviewer(s)' Comments to Author:

Referee: 1

Comments to the Author(s)
This experimental paper looks at the conditions under which people will punish in a repeated public goods game. To get at this the authors consider three experimental scenarios, 1) in which all players can mutually punish each other, 2) in which one individual is immune from being punished (but all players can punish and all others can be punished) and 3) in which only one sole player can punish. They find that individuals who are immune from punishment (scenarios 2-3) reduce their contributions to the public good over time, and punish less overall compared to non-immune individuals. The interpretation of this result is that cooperation-punishment do not form a single altruistically motivated trait, which contradicts the "strong reciprocity" hypothesis and (in contrast to previous results which confounded a cooperative environment with a fear of punishment), but supports the idea that players learn over the course of an experiment. I find these results interesting and valuable, and I support publication of the paper.

RESPONSE: thank you for your valuable time as a reviewer and your kind comments!

The experimental methodology is fully explained, and the statistical analysis is satisfactory. The main difficulty I have in reading the paper is the motivation for and interpretation of the results, which at times is a bit confused and jargon heavy.

RESPONSE: Our apologies for too much jargon. We have now reduced our level of jargon and hopefully the paper is now clearer in response to peer review.

The paper is focused on the debate over strong reciprocity and provides clear evidence that cooperation and punishment can decouple. But the idea of cooperation-punishment as a single altruistic linked trait needs to be explained more clearly in the context of the experiment, and, ideally, in intuitive terms. In particular the authors need to spell out what exactly it means, in humans, for cooperation and punishment to "stem from one conjoined, altruistic, trait (dubbed 'strong reciprocity')" (line 60). Is the idea that cooperation and punishment are somehow intrinsically linked by a genetic architecture which uniquely determines behavior, so that punishment could never occur without the cooperation and vice versa? Or is it rather that the two behaviors co-evolved as a single successful strategy, which could be decoupled if incentives change? And in the latter case, what is our expectation for players' behavior in this experiment?

RESPONSE: The idea would be that the two behaviours are at least stably correlated, at the phenotypic level, and that is what we are testing here. We mention this in our Introduction, "*This 'Altruistic Punishment' hypothesis (also known as the 'Strong Reciprocity' hypothesis) posits that "**Strong reciprocators bear the cost of rewarding or punishing even if they gain no individual economic benefit whatsoever from their acts.**", and has been supported by experiments showing that punishment is mostly directed towards below average, or relatively lower, contributors. This suggests that altruistic cooperation and punishment are indeed correlated"*

If the trait is the result of genetical evolution then cooperation and punishment would be genetically linked yes. There would be many ways this could happen mechanistically, for example a common altruistic motivation could underpin and activate both traits depending on how different contexts are perceived. However the trait could be the result of cultural evolution and more labile. Previous research has posited the idea of 3 types, 1) cooperators who don't punish, 2) non-cooperators that also don't punish, and 3) cooperators who also punish (altruistic punishers). We are pointing out this omits the 4th possible type from combining 2 traits, a non-cooperator that punishes, which we observe here.

We now write in our introduction,

"However, Natural Selection, or individual learning of cultural traits, could often favour the decoupling of cooperation and punishment, making Strong Reciprocity unstable."

We had written more but in the end had to remove it due to space constraints.

In previous studies, as the authors show, the apparent evidence for strong reciprocity can be explained as an artifact arising from fear of punishment. In this experiment we see confused learning, in which punishment and cooperation decouple over time. However I am unclear what we would expect a confused learner under the strong reciprocity hypothesis to look like except under the very strong assumption that cooperation and punishment cannot decouple for mechanistic reasons. I see four potential categories of conditional cooperator in this experiment i) {strong reciprocity, no learning}, ii) {no strong reciprocity, no learning}, iii) {strong reciprocity, learning} and iv) {no strong reciprocity, learning}. The conclusion of the paper is that we see iv) but i'm not clear how we would distinguish iii) and iv). If the authors can explore this distinction, the importance of the paper will be much clearer.

RESPONSE: While it is possible for Strong Reciprocators (SRs) to learn, an interpretation of the data relying on SRs learning would invalidate the very data often used to infer their existence. Previous studies have assumed individuals perfectly understood that their decisions were costly and benefited others, and hence were altruistically motivated. In this context, learning among SRs would be oxymoronic.

We now discuss this near the start of our new Discussion,

"*These results still applied to individuals previously classified as Conditional Cooperators (Supplementary Figures 3 & 4) and show that cooperation and punishment are not linked traits, as often assumed.*

Instead, our results are consistent with confused individuals initially contributing and then learning to reduce their contributions. Declining contributions are often attributed to frustration among impotent strong reciprocators deprived of the ability to punish, but that explanation is not possible here. This is because our immune individuals were (1) surrounded by a stable level of contributions, and (2) even had the power to punish non-cooperators. While it is possible that strong reciprocators also learn, their very existence has been inferred from previous results that required assuming individuals fully understood the consequences of their decisions."

Referee: 2

Comments to the Author(s)

In this paper, the authors report the results of a cleverly designed public goods game that decouples cooperation and punishment to stringently test the altruistic punishment/strong reciprocity hypothesis. The paper is clear, very well written, the

analyses are sophisticated and appear correctly executed, and the argumentation is compelling. I do a lot of research in this area and can confidently say this will be a very well-cited and influential paper, and I think these results shed some very important clarifying light on the field. Frankly, I think the paper could be published as is. I don't have many substantive comments to give, but here are some minor points that I noted.

RESPONSE: thank you for your valuable time as a reviewer and your kind comments!

- In lines 89-94, the authors allude to the confused learner hypothesis. I'm personally familiar with this work and think it's a very relevant counterpoint to the altruistic punishment hypothesis's interpretation of PGG results, but I'm not sure it's common knowledge in the field yet (it should be!). I think the authors would do well to have an additional few sentences setting up that work by briefly describing what some of findings are and how it contradicts the AP hypothesis. Particularly since this is ultimately what the authors endorse in the discussion, I think elaboration is needed.

RESPONSE: You are right, we have added the following explainer (**new in bold**),

"*In contrast,* ***there is increasing evidence that participants are initially confused in public goods games, but learn from experience, and that levels of altruistic contributions have been over estimated ('Confused Learners' hypothesis).*** *If immune individuals are instead motivated by personal gain but require experience to learn how to play the game, then they will learn to (1) reduce their contributions despite high levels of contributions among their groupmates, and yet (2) they may still punish others, even if they themselves are hypocritically contributing even less, demonstrating that cooperation (contributing) and punishment are not linked traits."*

We had also written more here, but in the end had to remove it due to space constraints (we are really at the limit!).

- On lines 149 and 151 the authors mention that the maximum MUs were different in sessions 1 and 2 than in later sessions. I think this warrants either a footnote or a reference to the supplement explaining why.

RESPONSE: we have now explained why this was the case in our Supplementary Methods and directed the reader towards there and 3 points, the two points you mention, and in a new subsection called Financial Incentives.

"***Financial incentives***

Each MU was worth 0.04 CHF, so 20 MU was worth 0.8 CHF ***(see Supplementary Methods for details of exceptions in Sessions 1-3)****. All earnings were rounded up to the nearest CHF, and the mean average payment was 22.60 CHF (this includes*

the 10 CHF show up fee), and ranged from 18 CHF to 31 CHF, with a median and a mode of 22 CHF."

We then explain fully in the ESM,
*"**Exceptions to exchange rate and punishment budget in sessions 1-3***
The exceptions were that in the first three sessions, the exchange rate was 1 MU = 0.05 CHF, not 0.04 CHF, and in the first two sessions individuals had a smaller budget to spend on punishment (18 MU instead of 30 MU per round, so the total endowment = 230 MU, worth 11.50 CHF). We increased the punishment budget after session 2 from 6 to 10 MU per groupmate for theoretical reasons. We wanted to enable a full contributor (20 MU) motivated by inequity-aversion to be able to equalize the payoff between themselves and a complete free-rider (contributes 0 MU). A punishment of 10 MU would close the gap in income by 20 MU. Increasing the punishment budget did not increase mean spending on punishment (mean spending in Sessions 1 & 2 = 2.1 MU; Remaining sessions using same scenarios (Immune Punisher or Sole Punisher) = 2.0 MU). Upon increasing the punishment budget, we realized mean earnings were consequently higher than we could afford, so we reduced the exchange rate from 0.05 to 0.04 CHF / MU."

- At the point the authors introduce the classified "conditional cooperators' on line 230, they have only briefly introduced how they categorized people at the end of the methods section and it's hard to follow exactly what went into that (e.g., line 129 says "approximately match the mean average contribution of their groupmates"). Some more clarification here would be good.

RESPONSE: We replicated the method of classification from Thoni & Volk 2018. We have now made this clear in our Methods and included what the precise criteria were.

"The method categorises individuals as either Conditional Cooperators, who either perfectly, or at least approximately, match their groupmate's mean contribution (Pearson correlation > 0.5 and the amount they contribute when their groupmates contribute fully is greater than their mean contribution for all 21 scenarios (Thoni & Volk, 2018), or Free Riders, who never cooperate regardless (contribute 0 MU for every possible scenario); or Other/Unclassified, who satisfied neither of these criteria."

- I don't think it undercuts the point made in the "immune individuals punished less" section (lines 268-286, but what do the authors make of the level of punishment in the sole punisher condition, which appears to be in the ballpark of non-immune players? They make the comparison directly to immune punishers,

but it does seem interesting that sole punishers appear equivalent to mutual punishers, which seems to me predicted by both the AP and CL hypotheses so it's not particularly informative. My take, in line with how the authors discuss it, would be that the direct comparison in the immune punisher condition is most relevant for the free riding question.

RESPONSE: Yes, it is interesting, however, the Sole Punisher was actually a bit lower than in the Mutual Punishers scenario (2.7 versus 3.2 MU), and only the mutual punishers had the 'revenge' motive. We agree with you that the most relevant comparison is between the sole punisher and the immune punisher, as both could not be punished, but only one had the sole 'responsibility' for punishing, leading to potential free riding on punishment (which is a second order public good). Studying levels of punishment can be difficult because they tend to be higher when groups are less cooperative, and in repeated games there is a circular relationship between contributions and punishment, making causal inferences difficult. The Sole Punishers were unable to stop the typical decline in contributions, had sole responsibility for punishing, and were likely under the most experimenter demand to punish, all of which could inflate levels of their punishment.

Our thinking is that if an individual believes punishing will lead to higher returns in the future via increased cooperation, but is costly, then they will be more likely to do it when they have sole responsibility. In the immune punisher scenario, it is more akin to a volunteer's dilemma, so free riding is possible. However we also suspect that the Sole Punisher condition creates quite a strong experimenter demand to be a punisher, it is like the teacher assigning you a role of responsibility, and there is no hiding in the mix, you know the experimenter is interested in how you punish.

We have added to our text the value of this comparison between the Sole Punishers and the Immune Punishers,

"*This comparison between Sole Punishers and Immune Punishers also has the advantage that both types were immune from punishment and thus could not be motivated by 'revenge', although we suspect Sole Punishers may have felt more pressure from the experimenter to punish."*

- Line 319: "..the instances of social punishment" I think a "pro-" is missing.

RESPONSE: you are correct, we have fixed the error thank you.

- In the discussion on line 367 the authors mention the inconsistency in punishment. It might be worth referring here again to the work on the confused learner hypothesis: some of this might be explained by participants simply not

quite grasping how to maximize their payoffs in the game and are either testing different strategies or responding somewhat randomly.

RESPONSE: we have added the following,

"While variation in social strategies or motivations is likely ('heterogenous preferences'), we think our also demonstrate that experiments which offer participants multiple behavioural possibilities are likely to find multiple behaviours, and that there was no clear preference for altruistic punishment. Such exploratory behaviour is perhaps more consistent with confused learners than rational punishers."

We have also added to the final paragraph, "*Regarding punishment, it is hard to rationalize hypocritical punishment. Perhaps some individuals* ***were confused and*** *thought they could somehow gain from the punishment, either directly, or indirectly, via a chain of interactions."*

****