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

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

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

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

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

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

We added the following paragraphs:

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

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

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

We rephrased this part of the results thus:

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

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

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

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

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

We added the following to the discussion:

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

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

We address this in the following:

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

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

Done.

References

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

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

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

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

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

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

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