nature portfolio

# Peer Review File

## Evolutionary dynamics of behavioral motivations for cooperation

Corresponding Author: Dr Qi Su

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The paper provides a model of dynamics of cooperative behavior among agents with specific utility function that can switch between "cooperative" and "selfish" depending on the parameter.

The results presented by the authors are interesting in terms of explaining what dynamics we should expect from the utility function and information structure that they have chosen. However, it is not clear what we should learn from this model. I describe the details in points:

1) while the authors state in the introduction that there are too many models of cooperative mechanisms in the literature, they themselves invent yet another one. There are truly a lot of papers in the last 30 years that consider some dynamics of cooperation taking some utility function and showing some results. It is unclear to me what we can learn from yet another exercise in this vein with some new assumptions on the utility function and the structure of agents' interaction that seem mostly arbitrary. Why should we think that this specification of cooperative dynamics is "more true" than any other dynamics with different assumptions?

2) The authors strangely do not mention social norms in their argument about what drives cooperation. In all social sciences there is a consensus by now that social norms are the main drivers of coopration behavior. How should we think about norms in the framework provided by the authors?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I have read the manuscript "Evolutionary dynamics of behavioral motivations" with great interest. The manuscript describes a framework for studying the evolution of self-evaluative heuristics, which have agents compare their past outcomes against their (subjective) needs before deciding whether to cooperate or defect. The manuscript distinguishes 'philanthropic' types (which cooperate only once their needs are met) and 'aspirational' types (which cooperate as long as their needs are not met), and varies the need threshold (more or less 'demanding'). The manuscript presents a wide range of results including on the evolution of cooperation given fixed motivational types and the evolutionary dynamics of motivations conditional on incentives and social network structure. I found many aspects of the manuscript intriguing and appreciate the careful effort to establish the robustness of the results. At the same time, I struggle somewhat to connect the model to what we know about human motivation.

# Major issues

First, I want to acknowledge that I am not an expert in evolutionary game theory. I read this manuscript as an experimentalist interested in (human) cooperation, albeit one with an interest in evolutionary theory. That said, the modelling work in this manuscript strikes me as very robust: the authors present converging analytic and computational results and explore a wide range of specifications (e.g. six different classes of networks). I hope that other reviewers can comment on technical matters in more detail.

I am, however, less convinced by the framing of the model. The authors explicitly motivate the shape of the behavioural motivations with reference to Maslow's hierarchy of needs. Maslow's model, however, was originally proposed on purely theoretical grounds and has since received only limited empirical support (and much criticism). More specifically on the matter of cooperative behaviour, the manuscripts cites a swath of literature purporting that richer individuals are less cooperative, but the highest-quality evidence does not support this claim (e.g. Andreoni et al. 2021). In the specific case of Piff et al. (2012), a recent study failed to replicate the original findings under highly controlled conditions (Jung et al., 2023). Even if these findings were robust, however, they would be at best tangentially related to the model studied here, which suggests within-person changes in cooperation conditional on past earnings meeting an internal need threshold.

Perhaps the literature on scarcity and cooperative behaviour provides a better context for the models studied here (e.g. Elbaek et al., 2021). The ethnographic literature on this topic is also quite fitting (e.g. Townsend, 2020). However, this literature quite consistently evidences 'philanthropic' behaviour: deprivation (i.e. not meeting one's need threshold) reduces cooperative behaviour. One intriguing result of the present model is that it suggests a flipside to this: conditions under which 'aspirational' motivations prevail. Unfortunately, right now, the empirical literature only makes an appearance in the introduction, and rather superficially so, which made it difficult for this reader at least to connect the models back to what we know about human motivation and cooperative behaviour.

This is related to my other concern: outside of the introduction, the manuscript is very technical, including in the discussion. Part of the reason is surely that it is so full of exciting material. However, this does make it difficult to translate the modelling results back into the broader context. I especially struggled with the implications of the social network structure for the evolutionary dynamics. This is addressed across two paragraphs in the discussion (pp. 16--17), but these remain both technical *and* unspecific ("By considering the population structure..." -- but exact what structures are to be considered is not mentioned here). I think a more high-level discussion which connects the model back to substantive application areas would make the manuscript more accessible and more valuable for many readers.

# Minor issues

p. 9: Is it accurate to speak of 'cooperation frequency' when L > 2?
p. 12: Would be good to plot Figures 5a and 5b on the same coordinates

# References

Andreoni, J., Nikiforakis, N., & Stoop, J. (2021). Higher socioeconomic status does not predict decreased prosocial behavior in a field experiment. Nature Communications, 12(1), 4266.
Elbaek, C., Mitkidis, P., Aarøe, L., & Otterbring, T. (2021). Material scarcity and unethical economic behavior: A systematic review and meta-analysis. Preprint.
Jung, M. H., Smeets, P., Stoop, J., & Vosgerau, J. (2023). Social status and unethical behavior: Two replications of the field studies in Piff et al.(2012). Journal of Experimental Psychology: General, 152(5), 1368.
Townsend, C. (2020). Neither nasty nor brutish. Aeon. https://aeon.co/essays/why-were-the-ik-people-vilified-as-selfish-and-nasty

I always sign my reviews.

Simon Columbus | MIT
simon@simoncolumbus.com

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Summary:

This manuscript studies the evolution of behavioral motivations in the context of prosocial decision-making. The model considers two dimensions of behavioural motivations: the need threshold (whether an individual is more or less needy) and the motivation intensity (whether an individual tends to be more cooperative when personally satisfied or unsatisfied). This study aims at understanding i) the effect of each motivation profile on cooperative behaviour and ii) which of these profiles evolves.

The authors employ several sorts of analysis. They start by deriving the cooperation frequency for populations with a fixed motivation (for weak intensity of selection). They consider several network structures and games (more than two actions and non-linear donation games). Then, the authors analyze the fixation probability of a given motivation profile in a population. Finally, the analytical results are complemented by simulations. The results are robust across the various methods: when the benefit-to-cost ratio of cooperation is above a certain threshold, undemanding philanthropic and demanding aspirational motivations can evolve, and thus, cooperation increases. This study offers 4 key conclusions:

-Undemanding philanthropists and demanding aspirationalists improve cooperation.
-Extending the action space beyond binary options (Cooperate / Defect) will either improve or impede cooperation,

depending on the individuals need threshold.
-There is a critical benefit-to-cost ratio for cooperation, above which behavioral motivations evolve towards either undemanding philanthropists or demanding aspirationalists.
-This critical benefit-to-cost ratio depends on network topology.

Major comments:

We found this study creative and relevant in synthetizing literate on the seemingly contradictory motives for individuals to cooperate. The authors point to previous psychology literature arguing that prosociality is high in both the lower and the higher social classes (which some papers refer to as the U-Shaped charitable giving profile). This paper provides an initial attempt to explain such phenomena through the lens of evolutionary game theory. We appreciate that this work focuses on the indirect evolution of strategies (actions); that is, instead of a direct evolution of actions (e.g., whether to cooperate or defect), the proposed model allows for the evolution of the mechanisms/heuristics/learning/motivations that govern the decision-making process. We agree with the authors that in evolutionary game theory literature, this kind of model remains underexplored. For that reason, this study advances our knowledge about what sort of motivations can evolve and can also inspire more research in this direction. We also value that different levels of needs are explored (which allows for a natural extension of aspiration dynamics to multi-action games).

We do, however, have some concerns and questions. Let's start with the main observations and then refer to minor comments, such as typos.

1) The conclusion relating cooperation with undemanding philanthropists and demanding aspirationalists directly follows from the model assumptions. As the need threshold (alpha) increases for aspirationalists the probability of cooperation directly increases; the same happens by decreasing alpha for philanthropists. Why is this surprising given the model definition? We believe the relevance of this conclusion should be stressed, otherwise the relation between cooperation and undemanding philanthropists / demanding aspirationalists should not be presented as a key conclusion.

2) There is already some work that incorporates the idea of indirect evolution, which could be good to reference. To name a few,
-Akçay et al. "A theory for the evolution of other-regard integrating proximate and ultimate perspectives." PNAS, 2009.
-McNamara et al. ,"Incorporating rules for responding into evolutionary games," Nature, 1999.
-Alger, "Evolutionarily stable preferences," PTRSB, 2023.

3) The connection of the present update mechanism (aspiration/philantropy) with previous works on aspiration dynamics should be clarified. In previous works (Ref [37-39]) aspiration has a very precise meaning: it is a payoff threshold below which individuals always change strategies, either to cooperation or defection. This is different than a setting where, by being aspirationalist, individuals will always choose to cooperate if their average payoff is low. In lines 416 and 417, we read, "Our model can recover the prior studies as a specific case." Could you explain how exactly? By the above definition of aspiration dynamics, it seems that no fixed motivation profile (of the present manuscript) could capture that update rule. Another potentially interesting question is whether these sorts of motivations would be robust against the traditional aspiration dynamics.

4) A deeper comparison to previous models might also help understanding the results on the network effects. Despite the differences pointed above, the two update mechanisms have natural similarities- they share that the decision to adopt a certain action depends on the difference between an aspiration level and the player's current payoff. in Figure 3, we can see that the kind of network has no effect on the frequency of cooperation. Wouldn't that be expected from Ref [38], where they study the traditional aspiration dynamics on networks? Furthermore, in lines 374-385 of the discussion, it is mentioned that "In our study, when individuals' behavioral motivations remain fixed, the global cooperation frequency remains robust against population structure, regardless of the density and distribution of connections. However we also note that once individuals' behavioral motivations start to evolve, population structure again plays a critical role in determining the evolutionary direction of behavioral motivation."This seems to imply that it's the motivations that create this network effect. However, the mechanisms by which individuals update actions (C and D) and update motivations are fundamentally different. The first is aspiration-like, and the second is imitation-like (death-birth).

The death-birth process is known to favor cooperation on networks (see your Ref. [8]). It is thereby not surprising that undemanding philanthropists and demanding aspirationalists can evolve after a critical b/c ratio, for a death-birth process on networks, given that these motivations directly translate into higher levels of cooperation. Will the same motivation evolve under, say, a birth-death process? This, instead, might be what makes the difference under the networks (not motivations/heuristics per se). It would be better to rewrite this paragraph for more clarity.

5) In Figure 6, it is not clear what each dot (and trajectory) means. We presume it refers to one individual, but that should be stated clearly. Also, what are the initial conditions exactly? A population composed of individuals with the four behavioral motivations indicated by open black dots? Or do you consider four different monomorphic populations as initial conditions in independent simulations? Why, in the figure caption, do you first refer to 300 simulations and then later to 200? In line 314, you probably meant Fig.6b.

6) While in the SI the authors show general conditions for any 2-by-2 game, they only run simulations for the donation game. It would be interesting to also have simulations for other games. What motivations would evolve for the snowdrift and stag-hunt?

7) Supplementary Figure 2. Is there any reason/intuition as to why the frequency of cooperation is now non-linear in the need threshold for a population of philanthropists but still (somewhat) linear for a population of aspirationalists? Also, at least for panel a) you could also show the analytical results from Eq.32. Would they match the simulations?

8) We do not really understand how weak \lambda comes in in Figure 6. How is it different from Supplementary Figure 3? Perhaps it's the intensity of selection \delta values that differ between the two figures? If so, what are the values? I understand that perhaps in the main text, the authors do not mention \delta in separate terms from \lambda to avoid confusion, but at least in the SI figure, the reader has already been introduced to it.

9) Supplementary Figure 5. It would be nice to see the same kind of plot for c = 0.2 as well, to capture the lower panels' regime from Figure 6. Do the results also agree? Also, please consider changing the axes' labels "slope" and "threshold" to intensity and need, respectively, because that's how you referred to them most of the times in the main text.

Minor comments:

-The manuscript should motivate earlier the pertinence of considering weighted networks and the interpretation of weights in this model.
-Lines 34-42 suggest that this study is about understanding the behavioral motivations behind each cooperation mechanisms, which is not the case.
-Line 36: cheats? Cheaters?
-Line 54: trasncendance -> transcendance
-Line 61: seen -> be seen
-Line 72: met -> meet
-Line 74: hard to -> hardly (or for whom it is hard to)
-Figure 1b: Please mention in the caption that the illustrative payoffs assume all weights between neighbours to be 1.
-Equation 3: e_ji -> e_ij
-Line 163: Evolutionary of cooperation -> Evolutionary dynamics of cooperation
-Line 169: perhaps you meant Figure 3e.
-Line 204: should the condition be \alpha < (b – c)/8? Or else, how did you derive it?
-Line 248: more likely to fix than a neutral mutant
-Figure 5: please explain meaning of inset values. One can infer this corresponds to \alpha and \lambda, but the authors can explicitly point this out
-Figure 5 caption: Dots->Squares
-Figure 5 caption: explain vertical dashed lines
-Line 266: appropriate?
-Figure 6: typo in the legend – should be x_C > ½ for blue and x_C < ½ for red. In the caption, regression instead of progression?
-Line 310: motivation range instead of "parameter range"
-Line 314: Fig. 5b -> Fig.6b
-Line 340: please consider citing some foundational papers on the framework of "adaptive dynamics" so that the reader knows what you mean by that
-Line 462: name -> game
-Line 489: formalization -> formalizations
-SI Eq. 10: sums in m should be in the set L (not N)?
-SI Eq. 11: keep the bar for average as in Eq. (9)
-If SI Eq. 18 is derived for linear games, how can it be used to obtain Equations 29 and 30 (for non-linear games)? What did I miss?
-Supplementary Figure 6. Are the parameters and other details the same as in Figure 3e of the main text? If so, please mention that in the caption.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Thank you for the opportunity to review this revised manuscript. I appreciate the authors' careful response to the issues I raised. I also found the comments by the other reviewers very valuable for my own understanding of this work.

As the authors emphasise in their response letter, a key aim of this manuscript is to integrate evolutionary game theory and human decision making by modelling the evolution of motivations (or heuristics) which can differ between individuals and across contexts. I must admit that even with the revised literature review, I remain unconvinced by this connection.

If the aim is to model the evolution of actual motivational heuristics, then the shape of the motivations in the model should at least be plausible abstractions of behavioural or motivational patterns documented in the psychological literature. Here, motivations are described by two dimensions: (a) the need threshold (i.e. how demanding or undemanding an individual is) and (b) the intensity (i.e. whether an individual cooperates until meeting their need threshold and defects above it or vice-versa).

They intensity dimension strikes me as the conceptually more important one, since it captures the central idea that cooperation is dependent on an internal, absolute reference point. The question is whether this is a plausible model of human decision making. The evidence presented refers to correlations between SES and prosocial behaviour, but such crosssectional correlations are at best suggestive and provide no direct evidence of within-person reference dependence. While such reference dependence has been discussed in the literature (for example, this assumption underlies the experimental practice of paying only some rounds of iterated games to avoid account-keeping), I am not aware of a formal model or empirical test of this idea. I would've liked to see a stronger case in the literature review that these are actually plausible shapes for human utility functions.

I want to emphasise that I find the manuscript interesting. This work clearly makes a contribution to the literature on aspiration dynamics. However, I am not convinced that it models the evolution of motivational heuristics which are plausible abstractions of actual motivational heuristics, which does limit the insight it offers for human decision making or the understanding of the evolution of cooperation.

I always sign my reviews.

Simon Columbus | MIT
simon@simoncolumbus.com

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
We would like to thank the authors for the comprehensive and clear answers. Overall, we find the answers satisfactory, and we believe this paper is a suitable contribution to Nature Communications.

We have two final remarks, however:

Regarding our first major concern, the authors provide a satisfactory answer that could be better explored in the manuscript. We believe the first result discussed by the authors — undemanding philanthropic and demanding aspirational behavioral motivations promote cooperation — still sounds rather trivial when reading the manuscript. The analysis described by the authors in their response, which consists in keeping the average threshold \alpha constant and varying the position of nodes with low and high \alpha on the network, is a very interesting one. However, this is not explored in the paper. Figure 3 assumes that \alpha values are assigned uniformly at random and no type of network correlation on \alpha is studied. Our suggestion is to either 1) implement the network correlation on \alpha analysis that the answer suggests; or 2) clarify in the manuscript why this result presented defies intuition; or 3) to be upfront and state it as a simple model consistency check.

Regarding Supplementary Figure 5, we commend the authors for the newly added panel b. We see how the “results agree qualitatively with the main text results” for high c (panel a), but not for low c (new panel b). As the authors point out in their reply to the reviewers, this figure differs from main Figure 6 in that it refers to a well-mixed population analysis, and that’s perhaps the reason for the disagreement. Therefore, it would be best to either provide some intuition why the results disagree for low c, or at least avoid saying that the results agree qualitatively.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Thank you for the opportunity to review this revised manuscript.

In my earlier review, I emphasised the need to more clearly link the proposed model to the existing literature on human motivation and highlighted discrepancies between features of the model and what the empirical literature claims. In their rebuttal, the authors argue that some of the features I highlighted -- in particular, the presence of internal reference points -- are outcomes of the model, rather than assumptions. I find this argument only somewhat convincing. There are two reasons. First, while it is true that internal reference points emerge from the model, allowing motivations to vary along two dimensions -- need threshold and intensity -- does make some assumptions about what features motivational heuristics could have. I am not fully convinced that the choice of these dimensions has a strong grounding in the existing literature, theoretical or empirical, on social decision making.

Second, if the aim is to model the evolution of actual human motivational heuristics -- and this is what the authors claim -- then the outcomes should, in my view, show at least some convergence with empirical observed motivational heuristics. That said, I recognise that this may be more of a matter of taste about the aims of evolutionary models, and as in my previous reviews, I happily accept that the paper makes a meaningful contribution to the study of a class of models which include such motivations.

I always sign my reviews.

Simon Columbus | simon@simoncolumbus.com

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
We thank the authors for the new revisions and clarifications. In our previous review we expressed our positive assessment regarding this manuscript and fit for Nature Communications. We believe it presents relevant results on the evolution of motivations that underlie cooperative behaviour. We just raised two minor remarks:

1) It struck us as trivial the observation that “undemanding philanthropic and demanding aspirational behavioral motivations promote cooperation”. The authors have complemented their analysis with a network study where threshold \alpha now depends on network properties. Essentially, the authors conclude that cooperation does not depend on thresholds being placed according to nodes’ degree. What ultimately drives the observed results is the average \alpha considered, and no network effects are observed. Our suggestion here remains: the authors could simply state that the effect of changing average \alpha leads to an average outcome that conforms with the expected results given the model definition.

2) The authors’ answer is clear, and the point clarified.

Overall, and despite minor point 1), we believe this paper is a suitable contribution to the journal.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
Review for "Evolutionary Dynamics of Behavioral Motivations"

SUMMARY.
In this paper, the authors propose a novel approach to study the evolution of cooperation on social networks. Traditionally, in this literature what evolves is whether or not people wish to cooperate. Instead, this study acknowledges that such cooperation decisions are often based on heuristics. Here, the authors consider a particular form of heuristics. They assume that cooperation decisions depend on the person's accumulated wealth in the previous round. This setup is inspired by previous experiments asking whether high-income individuals would behave more or less socially than low-income individuals.

To explore this question, the authors study a dynamics on two timescales. In the short run, individuals' heuristics are fixed. Here, people use their heuristics (and their previous experiences) to decide whether or not to cooperate in the next round. In the long run, also the individuals' heuristics might evolve (here, according to a death-birth process, in proportion to the payoffs that the heuristics generates on average).
The outcome depends on the benefit-to-cost ratio and on the network structure. In particular, when cooperation is particularly valuable (for high benefit-to-cost ratios), selection favors "undemanding philantropists" or "demanding aspirationalists" (i.e., heuristics that make it particularly likely the actor will cooperate).

GENERAL EVALUATION.
In my view, the authors study an interesting and underexplored topic. Traditionally, evolutionary game theory takes an ultimate perspective when explaining human cooperation: What factors might make cooperation favorable on an ultimate level? Instead, here the authors combine this ultimate perspective with a proximate perspective (what kind of heuristics would evolve to help people make their cooperative decisions)?

The methods used in the paper are state of the art. In particular, the paper presents several analytical results in the limit of weak selection (or, here, rather: limit of weak motivation intensity). In addition, I appreciate that the authors show the robustness of their analytical findings with simulations (also in the case that the motivation intensity is substantial).

I understand that I was added to the list of reviewers after there was already a first round of revisions. Out of respect for a streamlined publication process, I will thus not raise any new major topics that would require substantial revisions (I do not see any substantial gaps in the authors' analysis anyways).

Instead, let me comment on how I think the authors have responded to the previous comments by other reviewers.
First (regarding the comments of referee 1), I see the referee's point that the authors already seem to presume that individual heuristics take the form of a threshold function. However, I do feel the authors give a satisfactory answer: Although their functions are formally threshold functions, parameters can be adjusted in ways such that these functions can also implement alternative response paradigms (e.g., approximately linear responses). So if there is some evolution towards functions with a clear threshold, that is better interpreted as a result of the model, rather than a fixed assumption.
Second, I feel that also the requests of referee 2 have been taken into account in a satisfactory manner.

Overall, I think the paper makes a good contribution to the evolutionary game theory literature, and that it merits publication in Nature Communications.

MINOR COMMENTS.

(-) Figure 3: In the figure caption, the authors write that "undemanding philantropic and demanding aspirational motivations lead to abundant cooperation".
Given that all cooperation rates here are between 0.49 and 0.51, the term "abundant" is somewhat misleading. Perhaps the authors can say that "cooperation is favored" under these conditions.

(-) In line 373, the authors say they check the robustness of their results when the motivational intensity increases up to 2. If I understand the figure caption of the corresponding extended data figure 2 correctly, the intensity tested there is 1 (which of course is just as fine).

(-) Not for this paper, but perhaps some suggestion for future research: I saw that in the end of the SI document, the authors derive some adaptive dynamics equation for their system. I understand that this adaptive dynamics is not the focus of this paper, but at least I would be interested in a more formal analysis of the trajectories of this dynamics.

(-) Related to the previous point -- perhaps instead of representing a player's cooperation probability with a function with two free parameters, one might even be able to let the entire function evolve (e.g., the paper "The adaptive dynamics of function-valued traits", JTB 2006). However, that would be an entire project on its own, and probably a rather challenging one.

(Remarks on code availability)

---

**Replies to comments and suggestions from referee 1**

---

**Comments:**
The paper provides a model of dynamics of cooperative behavior among agents with specific utility function that can switch between "cooperative" and "selfish" depending on the parameter.

The results presented by the authors are interesting in terms of explaining what dynamics we should expect from the utility function and information structure that they have chosen. However, it is not clear what we should learn from this model. I describe the details in points:

We thank the referee for their careful reading of our manuscript. We have revised the main text to better articulate the contribution of our work, which we have also laid out in our responses to the comments raised below.

1) while the authors state in the introduction that there are too many models of cooperative mechanisms in the literature, they themselves invent yet another one. There are truly a lot of papers in the last 30 years that consider some dynamics of cooperation taking some utility function and showing some results. It is unclear to me what we can learn from yet another exercise in this vein with some new assumptions on the utility function and the structure of agents' interaction that seem mostly arbitrary. Why should we think that this specification of cooperative dynamics is "more true" than any other dynamics with different assumptions?

We agree with the observation that there are many theoretical studies focused on mechanisms that facilitate the evolution of cooperation. Indeed we make this point in our introduction, as the referee notes. Our objective is not to invent yet another model for the evolution of cooperation, but to better connect the theoretical study of cooperative behavior to the real world behavior of humans by studying the motivations that underpin individual acts of cooperation. And so the contribution of our work is not to construct a new utility function that facilitates the evolution of cooperation. What we study is the evolution of the shape of the utility function used by individuals when deciding to cooperate. Our results are then framed in terms of the relationship between evolved motivations (i.e., the motivation intensity and the need threshold) and propensity to cooperate – something which connects to the existing psychology literature, and can in principle be tested empirically in future work. To reiterate, this is quite different from a study that points out a new mechanism for the evolution of cooperation.

In order to make this point clear to the reader, we have updated the Introduction and Discussion sections of the main text. In our revised Introduction we have emphasized that "Unlike prior studies that focus on the evolution of behaviors, where the behavior yielding a higher payoff is more likely to be adopted, this paper considers the evolution of behavioral motivations that guide the choices of behaviors.". In the revised Discussion we point out that "In reality human beings face many opportunities to engage in minor acts of cooperation every day, and the decision whether to help another person is typically guided by a few simple heuristics, rather than a carefully optimized behavioral strategy. And so we have asked the natural question: what happens if natural selection acts at the level of the heuristics people use when deciding to cooperate?".

We have also taken care to explicitly place our analysis within the context of the psychological literature, which shows that willingness to engage in prosocial behavior can vary with socioeconomic status (Nature Communications, 12:4266, 2021; PNAS, 109:4086–4091, 2012). We assume that all individuals are willing to cooperate in the right circumstances, and we investigate evolution at the level of individual motivations (i.e., what those circumstances are) and its corresponding impact on cooperation frequency at the level of the population. This aspect of our work is summarized nicely by referee 3, who notes that "this work focuses on the indirect evolution of strategies (actions); that is, instead of a direct evolution of actions (e.g., whether to cooperate or defect), ..., this kind of model remains underexplored."

Because our utility function pertains to behavioral motivation, we also note that our work opens up the potential to study the evolution of motivations for other types of behavior than cooperation. For instance, in the evolution of risk preferences, once the payoff meets a threshold, players may switch to riskier, more profitable behaviors. We have updated the discussion to make this point clear. Finally, we note that few studies have analyzed the impact of population structure on the types of questions addressed in this work, due to its mathematical challenges. We show, remarkably, that population structure has little effect on behavior under fixed motivations, but plays a pivotal role in shaping the evolution of behavioral motivations themselves. Taken together we believe that these changes to the manuscript clearly set out how our work is distinct from models that simply propose new mechanisms for the evolution of cooperation.

2) The authors strangely do not mention social norms in their argument about what drives cooperation. In all social sciences there is a consensus by now that social norms are the main drivers of coopration behavior. How should we think about norms in the framework provided by the authors?

We agree that social norms play a central role in maintaining cooperative behavior. We would also argue that the theoretical study of the drivers of cooperation is best understood as the study of the emergence and maintenance of social norms. As noted above however, the goal of our work is not to propose a mechanism for the evolution of cooperation as such. We take as our starting point the assumption that all individuals are willing to participate in cooperation under some circumstances, and we study how that willingness evolves. We can thus think of social norms in the context of our framework as providing the baseline willingness of all individuals to cooperate in some circumstances. We address this point in our revised Discussion, where we also note the centrality of social norms to maintain cooperation, and point out that behavioral motivations of the type studied here both contribute to social norms, and likely constrain the ability of behavioral motivations to evolve (e.g., in a population where aspirations motivations are typical, individuals using philanthropic motivations may be perceived as norm violators, preventing such attitudes from evolving.) We suggest the co-evolution of motivations and reputations as a productive avenue for future study.

---

**Replies to comments and suggestions from reviewer 2**

---

**Comments:**

I have read the manuscript "Evolutionary dynamics of behavioral motivations" with great interest. The manuscript describes a framework for studying the evolution of self-evaluative heuristics, which have agents compare their past outcomes against their (subjective) needs before deciding whether to cooperate or defect. The manuscript distinguishes 'philanthropic' types (which cooperate only once their needs are met) and 'aspirational' types (which cooperate as long as their needs are not met), and varies the need threshold (more or less 'demanding'). The manuscript presents a wide range of results including on the evolution of cooperation given fixed motivational types and the evolutionary dynamics of motivations conditional on incentives and social network structure. I found many aspects of the manuscript intriguing and appreciate the careful effort to establish the robustness of the results. At the same time, I struggle somewhat to connect the model to what we know about human motivation.

We are grateful to the reviewer for the excellent summary of the manuscript and constructive comments for the improvement. We have followed all of the reviewer's suggestions to revise our manuscript.

First, I want to acknowledge that I am not an expert in evolutionary game theory. I read this manuscript as an experimentalist interested in (human) cooperation, albeit one with an interest in evolutionary theory. That said, the modelling work in this manuscript strikes me as very robust: the authors present converging analytic and computational results and explore a wide range of specifications (e.g. six different classes of networks). I hope that other reviewers can comment on technical matters in more detail.

We thank the referee for this point. Our theory has been based on the setup of small motivation intensity, which theoretically agrees with the simulation quite well. We have also performed simulations for larger motivation intensity, as shown in Supplementary Fig. 2. While the theoretical results deviate from the simulations in this case, the theoretical predictions still qualitatively agree with the simulation quite well. We have emphasized this point in the revised manuscript.

I am, however, less convinced by the framing of the model. The authors explicitly motivate the shape of the behavioural motivations with reference to Maslow's hierarchy of needs. Maslow's model, however, was originally proposed on purely theoretical grounds and has since received only limited empirical support (and much criticism). More specifically on the matter of cooperative behaviour, the manuscripts cites a swath of literature purporting that richer individuals are less cooperative, but the highest-quality evidence does not support this claim (e.g. Andreoni et al. 2021). In the specific case of Piff et al. (2012), a recent study failed to replicate the original findings under highly controlled conditions (Jung et al., 2023). Even if these findings were robust, however, they would be at best tangentially related to the model studied here, which suggests within-person changes in cooperation conditional on past earnings meeting an internal need threshold.

We are grateful to the referee for pushing us on this. We certainly concede the point that Maslow should not be taken as the current standard or the correct framework for understanding human motivations. The extent to which we draw on the hierarchy of needs in constructing our model is simply to assert that 1) willingness to engage in acts of cooperation may increase or decrease with an individual's current well-being (measured in terms of payoff or utility) and 2) when cooperation is "complex" – i.e. involves multiple steps each consisting of an individual act of cooperation – individuals may have different thresholds for different steps. Maslow provides a convenient and widely understood framing for talking about this basic idea, but we accept the point that this leaves the reader with the impression that we are specifically modelling the hierarchy of needs. In order to address this, we have updated the main text to remove references to the hierarchy of needs and Maslow, however we have retained the analogy of "meeting needs" in the description of the model since we still believe it aids understanding.

With respect to the literature cited, we are grateful to the referee for providing the references above, which we have cited in the revised manuscript. However we wish to clarify that our intention is not to claim that the literature supports either philanthropic or aspirational motivations for cooperation, only that there is (highly inconclusive) evidence supporting both. Indeed as one of the references provided by the reviewer states, "no empirical study can provide a definitive answer to a question as general as to whether individuals of high SES tend to behave more (or less) prosocially than low SES individuals by itself" [Andreoni et al. 2021]. This is precisely the type of situation in which modelling efforts can be of value, and this is why we have constructed a model with the flexibility to include both types of motivation, or neither since our framework also accommodates motivations with no association between income and prosocial behavior (when the motivation intensity is zero). In the revised introduction we have made it clear that we do not prefer one association over the other in describing human decision-making.

Perhaps the literature on scarcity and cooperative behaviour provides a better context for the models studied here (e.g. Elbaek et al., 2021). The ethnographic literature on this topic is also quite fitting (e.g. Townsend, 2020). However, this literature quite consistently evidences 'philanthropic' behaviour: deprivation (i.e. not meeting one's need threshold) reduces cooperative behaviour. One intriguing result of the present model is that it suggests a flipside to this: conditions under which 'aspirational' motivations prevail. Unfortunately, right now, the empirical literature only makes an appearance in the introduction, and rather superficially so, which made it difficult for this reader at least to connect the models back to what we know about human motivation and cooperative behaviour.

Thank you for these excellent suggestions. We have revised the discussion to better connect the model back to this literature and to highlight the predictions of our model and how they may be tested. In the revised discussion we note that the key predictions of our model relate the correlation between motivation intensity and need threshold to propensity to cooperate (Figure 6). This suggests that simply determining whether individuals are aspirationally or philanthropically motivated is not enough to easily predict their behavior, and instead we must look at whether people are demanding or undemanding (in the language of our model). We suggest that this question is best addressed by looking at how behavior changes in response to environmental shifts (e.g., periods of scarcity or abundance).

This is related to my other concern: outside of the introduction, the manuscript is very technical, including in the discussion. Part of the reason is surely that it is so full of exciting material. However, this does make it difficult to translate the modelling results back into the broader context. I especially struggled with the implications of the social network structure for the evolutionary dynamics. This is addressed across two paragraphs in the discussion (pp. 16–17), but these remain both technical *and* unspecific ("By considering the population structure..." – but exact what structures are to be considered is not mentioned here). I think a more high-level discussion which connects the model back to substantive application areas would make the manuscript more accessible and more valuable for many readers.

Thanks a lot for this great suggestion. For any population structure, we can calculate the critical benefit-to-cost $(b/c)^* = \eta_2/(\eta_3 - \eta_1)$, where $\eta_n$ is obtained from Eq. 10 in Methods. We have performed the theoretical analysis into six types of networks, we analyze four representative classes of networks: random regular networks (RR), Erdös-Rényi networks (ER), Watts-Strogatz small-world networks (SW) with rewiring probability 0.1, Barabási-Albert scale-free networks (BA). For each of these classes, we investigate 1000 networks of size $N = 80$ and of average node degree $\bar{d}$ chosen uniformly from $[4, N-4]$. Figure R1 shows the critical benefit-to-cost ratio $(b/c)^*$ for the evolution of prosocial behavioral motivation. Generally, the sparse network (i.e., node degree $\bar{d}$ is small) is beneficial to prosocial behavioral motivation and accordingly to cooperation. The real-world network is often much sparser than the completely connected networks, which might explain the prevailing cooperation.

We have added these discussions in the revised manuscript.

# Minor issues p. 9: Is it accurate to speak of 'cooperation frequency' when $L > 2$?

We thank the reviewer for pointing out this confusion. We have replaced 'cooperation frequency' with 'cooperation abundance' and define the cooperation abundance for multi-action donation games, given by $x_C = (1/N)\sum_{i\in\mathcal{N}}(s_i - 1)/(L-1)$, where $s_i \in \{1, 2, \cdots, L\}$ is individual $i$'s action.

p. 12: Would be good to plot Figures 5a and 5b on the same coordinates

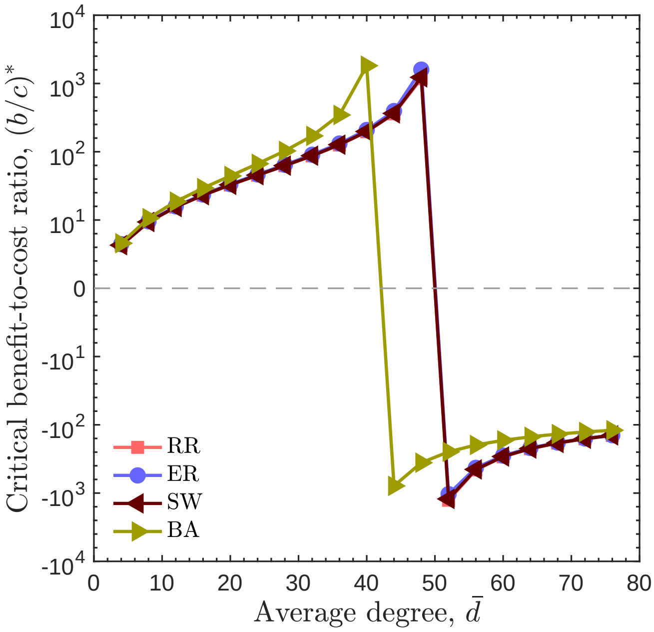



Figure R1: **Low connection density facilitates the evolution of prosocial behavioral motivation**. Each dot shown is the average of critical benefit-to-cost ratio over 1000 networks.

We plotted a figure with Figure 5a and 5b on the same coordinates (see Fig. R2). Since the dots overlap a lot, we take the original Figure in the main text.

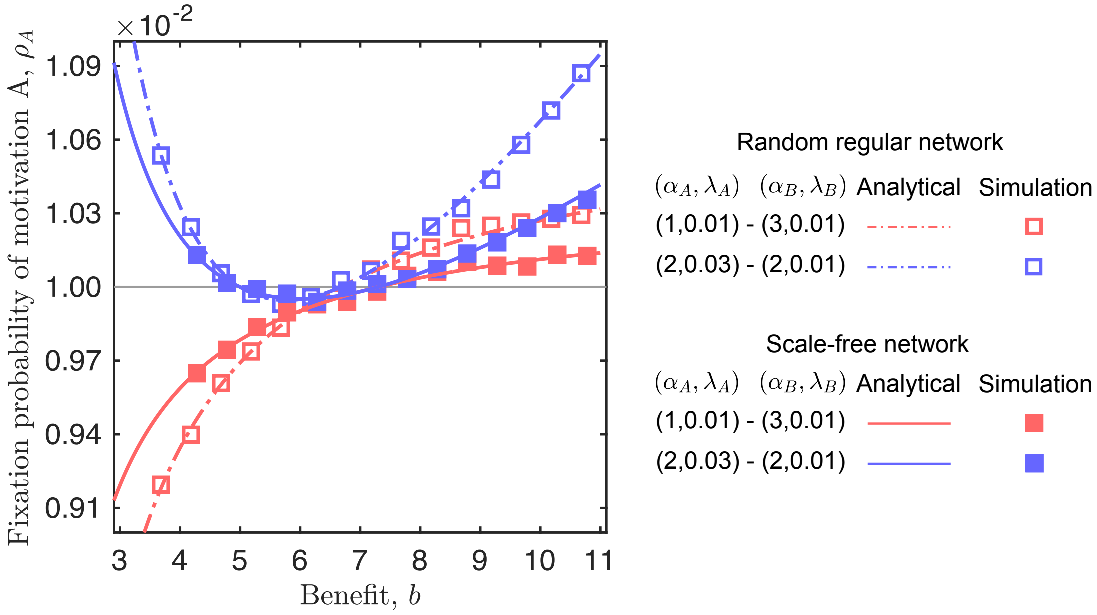



Figure R2: **Selection favors undemanding philanthropists over demanding philanthropists when the benefit-to-cost ratio exceeds a critical value**.

---

**Replies to comments and suggestions from reviewer 3**

---

**Comments:**

This manuscript studies the evolution of behavioral motivations in the context of prosocial decision-making. The model considers two dimensions of behavioural motivations: the need threshold (whether an individual is more or less needy) and the motivation intensity (whether an individual tends to be more cooperative when personally satisfied or unsatisfied). This study aims at understanding i) the effect of each motivation profile on cooperative behaviour and ii) which of these profiles evolves.

The authors employ several sorts of analysis. They start by deriving the cooperation frequency for populations with a fixed motivation (for weak intensity of selection). They consider several network structures and games (more than two actions and non-linear donation games). Then, the authors analyze the fixation probability of a given motivation profile in a population. Finally, the analytical results are complemented by simulations. The results are robust across the various methods: when the benefit-to-cost ratio of cooperation is above a certain threshold, undemanding philanthropic and demanding aspirational motivations can evolve, and thus, cooperation increases. This study offers 4 key conclusions:

-Undemanding philanthropists and demanding aspirationalists improve cooperation.

-Extending the action space beyond binary options (Cooperate / Defect) will either improve or impede cooperation, depending on the individuals need threshold.

-There is a critical benefit-to-cost ratio for cooperation, above which behavioral motivations evolve towards either undemanding philanthropists or demanding aspirationalists.

-This critical benefit-to-cost ratio depends on network topology.

We appreciate the reviewer's excellent summary of our manuscript and their constructive comments for improvement. We have incorporated all of the reviewer's suggestions in our revisions.

We found this study creative and relevant in synthetizing literate on the seemingly contradictory motives for individuals to cooperate. The authors point to previous psychology literature arguing that prosociality is high in both the lower and the higher social classes (which some papers refer to as the U-Shaped charitable giving profile). This paper provides an initial attempt to explain such phenomena through the lens of evolutionary game theory. We appreciate that this work focuses on the indirect evolution of strategies (actions); that is, instead of a direct evolution of actions (e.g., whether to cooperate or defect), the proposed model allows for the evolution of the mechanisms/heuristics/learning/motivations that govern the decision-making process. We agree with the authors that in evolutionary game theory literature, this kind of model remains underexplored. For that reason, this study advances our knowledge about what sort of motivations can evolve and can also inspire more research in this direction. We also value that different levels of needs are explored (which allows for a natural extension of aspiration dynamics to multi-action games).

We appreciate the reviewer's positive feedback on our manuscript and their accurate recognition of its significance.

We do, however, have some concerns and questions. Let's start with the main observations and then refer to minor comments, such as typos.

1) The conclusion relating cooperation with undemanding philanthropists and demanding aspirationalists directly follows from the model assumptions. As the need threshold (alpha) increases for aspirationalists the probability of cooperation directly increases; the same happens by decreasing alpha for philanthropists. Why is this surprising given the model definition? We believe the relevance of this conclusion should be stressed, otherwise the relation between cooperation and undemanding philanthropists / demanding aspirationalists should not be presented as a key conclusion.

This is an excellent point. We agree with the reviewer that for an aspirationalist, increasing the need threshold at the individual level indeed promotes more cooperation. However, in structured populations, an individual's actions depend on the comparison between their need threshold and the payoff obtained from all interactions with neighbors. The population structure often leads to a complex coupling of individuals' behavioral motivation and cooperative actions, making the global cooperation level worth investigating.

This effect can be even more unpredictable when individuals have different need thresholds. For example, consider two cases within the same heterogeneous structured population. In the first case, individuals occupying the largest 50% of highly connected nodes have a higher threshold, $\alpha + \epsilon$, while those in the least 50% have a lower threshold, $\alpha - \epsilon$. In the second case, the thresholds are reversed: individuals in the largest 50% of highly connected nodes have a lower threshold, $\alpha - \epsilon$, and those in the least 50% have a higher threshold, $\alpha + \epsilon$. Although the average threshold is identical in both cases, it is not intuitive which case will lead to higher cooperation levels. The variation in individuals' motivation intensity further complicates the outcome.

In the revised manuscript, Figure 3 now features a heterogeneous setup for the need threshold (where individuals differ in their thresholds) instead of a homogeneous setup (where all individuals have the same threshold, as shown in the prior manuscript). We emphasize that increasing the average need threshold fosters more cooperation, a result that defies intuition.

2) There is already some work that incorporates the idea of indirect evolution, which could be good to reference. To name a few,

-Akçay et al. "A theory for the evolution of other-regard integrating proximate and ultimate perspectives." PNAS, 2009.

-McNamara et al. ,"Incorporating rules for responding into evolutionary games," Nature, 1999.

-Alger, "Evolutionarily stable preferences," PTRSB, 2023.

We thank the referee for providing such valuable pertinent to out work. We have cited and discussed these references in the revised Discussion section.

3) The connection of the present update mechanism (aspiration/philantropy) with previous works on aspiration dynamics should be clarified. In previous works (Ref [37-39]) aspiration has a very precise meaning: it is a payoff threshold below which individuals always change strategies, either to cooperation or defection. This is different than a setting where, by being aspirationalist, individuals will always choose to cooperate if their average payoff is low. In lines 416 and 417, we read, "Our model can recover the prior studies as a specific case." Could you explain how exactly? By the above definition of aspiration dynamics, it seems that no fixed motivation profile (of the present manuscript) could capture that update rule. Another potentially interesting question is whether these sorts of motivations would be robust against the traditional aspiration dynamics.

We thank the reviewer for this excellent question. We describe the general threshold-based updating model in the Model section of the supplementary information. Briefly, individual $i$ forgoes action $k$ and switches to action $\ell$ with probability $g_i^{k\ell}\left(\delta\left(u-\alpha_i^{k\ell}\right)\right)$, where $u-\alpha_i^{k\ell}$ indicates how much $i$'s payoff exceeds his need threshold $\alpha_i^{k\ell}$ for transitioning from action $k$ to action $\ell$. The parameter $\delta$ ($>0$) measures the extent to which the payoff difference influences the action updating process. This action updating function can vary for individual $i$ and the two actions $k$ and $\ell$. The simplification of the general model can recover the two behavioral motivations in the main text and the traditional aspirational dynamics.

We introduce $g(y)=\frac{1}{1+\exp[-y]}$, which gives $1-g(y)=g(-y)$. To recover the philanthropic motivation in the main text (taking the two-action game for example, i.e., $k,\ell\in\{C,D\}$), we take

$$\begin{aligned} g_i^{CC}\left(\delta\left(u-\alpha^{CC}\right)\right)&=g\left(\delta(u-\alpha)\right),\\ g_i^{DC}\left(\delta\left(u-\alpha^{DC}\right)\right)&=g\left(\delta(u-\alpha)\right),\end{aligned}\tag{1}$$

which gives $g_i^{CD}\left(\delta\left(u-\alpha^{CD}\right)\right)=g\left(\delta(u-\alpha)\right)$ and $g_i^{DD}\left(-\delta\left(u-\alpha^{DD}\right)\right)=g\left(-\delta(u-\alpha)\right)$. As long as the obtained payoff $u$ is greater than $\alpha$, individuals are more likely to take cooperation than defection. If taking

$$\begin{aligned} g_i^{CC}\left(\delta\left(u-\alpha^{CC}\right)\right)&=g\left(-\delta(u-\alpha)\right),\\ g_i^{DC}\left(\delta\left(u-\alpha^{DC}\right)\right)&=g\left(-\delta(u-\alpha)\right),\end{aligned}\tag{2}$$

which gives $g_i^{CD}\left(\delta\left(u-\alpha^{CD}\right)\right)=g\left(\delta(u-\alpha)\right)$ and $g_i^{DD}\left(\delta\left(u-\alpha^{DD}\right)\right)=g\left(\delta(u-\alpha)\right)$, as long as the obtained payoff $u$ is greater than $\alpha$, individuals are more likely to take defection than cooperation, corresponding to the aspirational motivation. In above updating functions, the probability of transitioning to an action is independent of the original action, i.e., $g_i^{CC}(y)=g_i^{DC}(y)$ and $g_i^{CD}(y)=g_i^{DD}(y)$. However, if the probability depends on the original action, such as

$$\begin{aligned} g_i^{CC}\left(\delta\left(u-\alpha^{CC}\right)\right) &= g\left(\delta(u-\alpha)\right) \\ g_i^{DC}\left(\delta\left(u-\alpha^{DC}\right)\right) &= g\left(-\delta(u-\alpha)\right), \end{aligned} \tag{3}$$

which gives $g_i^{CD}\left(\delta\left(u-\alpha^{CD}\right)\right) = g\left(-\delta(u-\alpha)\right)$ and $g_i^{DD}\left(\delta\left(u-\alpha^{DD}\right)\right) = g\left(\delta(u-\alpha)\right)$, as long as the obtained payoff $u$ is greater than $\alpha$, the individual keeps the current strategy. However, if the obtained payoff $u$ is less than $\alpha$, the individual forgoes the current strategy, which aligns with prior studies on aspirational dynamics. All these updating functions are simplifications of the general model.

We further consider the evolution of our behavioral motivations in a population of individuals using the aspirational dynamics. Substituting Eq. (3) into Eq. (20) in SI, we have the cooperation abundance in the aspirational dynamics, given by

$$x_C = \frac{1}{2} - \frac{\delta}{8}c, \tag{4}$$

which indicates that the cooperation abundance in the traditional aspirational dynamics is independent of the threshold $\alpha$, and is always below 0.5 in the donation games. The analysis into Eq. (23) gives the condition for selection favoring the traditional aspirational dynamics over the philanthropic behavioral motivation if

$$(2\alpha - b)\left(-c\eta_2 + b(\eta_3 - \eta_1)\right) > 0, \tag{5}$$

and favoring the traditional aspirational dynamics over the aspirational behavioral motivation if

$$(b - 2c - 2\alpha)\left(-c\eta_2 + b(\eta_3 - \eta_1)\right) > 0. \tag{6}$$

Overall, in the prosocial environment, i.e., $-c\eta_2 + b(\eta_3 - \eta_1) > 0$, undemanding philanthropists (a small $\alpha$ in Eq. (5)) and demanding aspirationalists (a large $\alpha$), by fostering a high level of cooperation, can well resist the invasion of traditional aspirational individuals (see Fig. R3). We have added these discussions in the Discussion section and the revised supplementary information.

4) A deeper comparison to previous models might also help understanding the results on the network effects. Despite the differences pointed above, the two update mechanisms have natural similarities- they share that the decision to adopt a certain action depends on the difference between an aspiration level and the player's current payoff. in Figure 3, we can see that the kind of network has no effect on the frequency of cooperation. Wouldn't that be expected from Ref [38], where they study the traditional aspiration dynamics on networks?

We thank the referee for these insights. Reference [38] studies the classic aspiration dynamics of two actions on random regular graphs. They find that the evolutionary outcome of these dynamics on random regular networks is identical to that in well-mixed populations, with a structural coefficient $\sigma = 1$, indicating no cooperation cluster. Thus, the network has no effect on

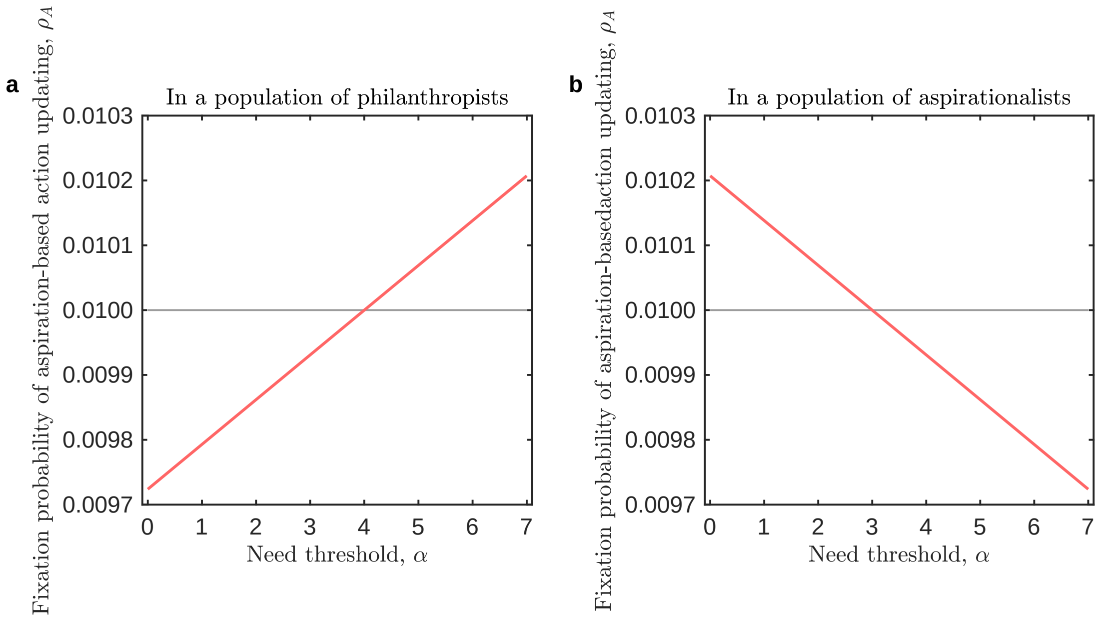



Figure R3: **Undemanding philanthropists and demanding aspirationalists can resist the invasion of mutants using aspiration-based action updating**. Presented are the fixation probability of a mutant using aspiration-based action updating in a population of philanthropists (**a**) and aspirationalists (**b**), for a broad range of need threshold $\alpha$. We consider the random regular networks with size $N = 100$ and node degree $d = 6$. Results show that mutants using aspiration-based action updating are disfavored against philanthropists with low need thresholds and aspirationalists with large need thresholds. Parameters: benefit $b = 8$, cost $c = 1$, $\delta = 0.01$ (corresponding to motivation intensity 0.01 for philanthropists and $-0.01$ for aspirationalists, the selection intensity 0.01 for aspiration-based action updating).

cooperation. We offer an alternative intuition applicable to any threshold-based action updating. In any heterogeneous network and under any threshold-based action updating, the frequency of any individual $i$ taking any action $\ell$ is given by:

$$x_i^\ell = \frac{1}{L} + \frac{\delta}{L} \sum_{k\in\mathcal{L}, k\neq\ell} \left[ \lambda_i^{k\ell} \left( \frac{1}{L} \sum_{m\in\mathcal{L}} M_{km} - \alpha_i^{k\ell} \right) - \lambda_i^{\ell k} \left( \frac{1}{L} \sum_{m\in\mathcal{L}} M_{\ell m} - \alpha_i^{\ell k} \right) \right]. \tag{7}$$

This shows that the frequency of an individual taking any action is fully determined by the payoff structure, their motivation intensity, and their need threshold, but is independent of their connections with other individuals.

We have emphasized both the intuition from prior studies and our understanding in the Discussion section.

Furthermore, in lines 374-385 of the discussion, it is mentioned that "In our study, when individuals' behavioral motivations remain fixed, the global cooperation frequency remains robust against population structure, regardless of the density and distribution of connections. However we also note that once individuals' behavioral motivations start to evolve, population structure again plays a critical role in determining the evolutionary direction of behavioral motivation."This seems to imply that it's the motivations that create this network effect. However, the mechanisms by which individuals update actions (C and D) and update motivations are fundamentally different. The first is aspiration-like, and the second is imitation-like (death-birth). The death-birth process is known to favor cooperation on networks (see your Ref. [8]). It is thereby not surprising that undemanding philanthropists and demanding aspirationalists can evolve after a critical b/c ratio, for a death-birth process on networks, given that these motivations directly translate into higher levels of cooperation. Will the same motivation evolve under, say, a birth-death process? This, instead, might be what makes the difference under the networks (not motivations/heuristics per se). It would be better to rewrite this paragraph for more clarity.

Excellent point. We agree with the reviewer that the updating process plays a critical role in the evolution of behavioral motivations. The evolution of undemanding philanthropists and demanding aspirationalists within a population of demanding philanthropists and undemanding aspirationalists arises from two key factors: (1) undemanding philanthropists and demanding aspirationalists can lead to a higher level of cooperation; (2) the death-birth process is more beneficial to cooperation for larger $b/c$ ratios. Thus, it aligns with the intuition that undemanding philanthropists and demanding aspirationalists can evolve under certain $b/c$ ratios in a death-birth updating scenario. The simulation results under birth-death updating further confirm these conclusions (see Fig. R4). However, this intuition assumes identical motivation intensity $\alpha$. When motivation intensity differs, the evolutionary direction could be reversed, which goes beyond the initial intuition and can be addressed within our analytical framework. We have rewritten this paragraph to clarify the mechanism.

5) In Figure 6, it is not clear what each dot (and trajectory) means. We presume it refers to one individual, but that should be stated clearly. Also, what are the initial conditions exactly? A population composed of individuals with the four behavioral motivations indicated by open black dots? Or do you consider four different monomorphic populations as initial conditions in independent simulations? Why, in the figure caption, do you first refer to 300 simulations and then later to 200? In line 314, you probably meant Fig.6b.

We apologize for the confusion. We have revised the caption and the description in the main text. Additionally, We have reviewed all figure captions to ensure clarity and corrected all typos. Each simulation starts from a monomorphic population with behavioral motivations $(\alpha, \lambda)$ in one of the following pairs: (0.75,0.05), (0.25,0.05), (0.25,-0.05), and (0.75,-0.05), represented by open black dots. The population undergoes $5 \times 10^7$ motivation updating steps. Each thin red line represents a resulting trajectory (i.e., the average behavioral motivation of the population)

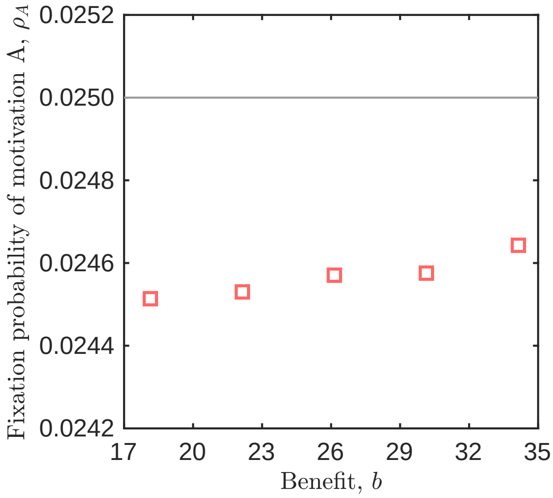



Figure R4: **Selection disfavors undemanding philanthropists over demanding philanthropists in the birth-death updating process**. We consider random regular networks of size $N = 40$ and node degree $d = 6$. Presented is the fixation probability of an undemanding philanthropist with $\lambda = 0.01$ and $\alpha = 1.0$ (denoted by motivation A) in a population of demanding philanthropists with $\lambda = 0.01$ and $\alpha = 4.0$ as a function of benefit $b$, in the birth-death updating. Squares represent the fixation probability by simulations. The fixation probability is determined by the fractions of simulations where the beneficial motivation A reached fixation out of $2 \times 10^7$ generations.

and each open red dot represents an ending behavioral motivation.

6) While in the SI the authors show general conditions for any 2-by-2 game, they only run simulations for the donation game. It would be interesting to also have simulations for other games. What motivations would evolve for the snowdrift and stag-hunt?

Excellent suggestion. We analyze the representative snowdrift game with payoff structure

$$\begin{array}{c c} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} b-\frac{1}{2} & b-1 \\ b & 0 \end{pmatrix} \end{array}, \tag{8}$$

where $b > 1$. Substituting the payoff structure into Eq. 23 in the Supplementary Information, we have the condition for selection favoring the motivation $A$ over $B$, given by

$$(x_C(\mathbf{A}) - x_C(\mathbf{B}))\left[b\left(\eta_2 + \eta_3 - \eta_1\right) - \frac{1}{2}\left(3\eta_2 - \eta_3 + \eta_1\right)\right] > 0. \tag{9}$$

Particularly, in the random regular networks with size $N$ and node degree $d$, selection favors the undemanding philanthropists or demanding aspirationalists over demanding philanthropists or undemanding aspirationalists if

$$b > \frac{(3d-1)\,N - 4d}{2\,(d+1)\,N - 8d}. \tag{10}$$

Analogously, we analyze the representative stag-hunt games with payoff structure

$$\begin{array}{c c} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{pmatrix} b & 0 \\ 1 & 1 \end{pmatrix} \end{array}, \tag{11}$$

where $b > 1$. We have the condition for selection favoring the motivation $A$ over $B$, given by

$$(x_C(\mathbf{A}) - x_C(\mathbf{B}))\,[b\,(\eta_2 + \eta_3 - \eta_1) - 2\eta_2] > 0. \tag{12}$$

In the random regular networks with size $N$ and node degree $d$, selection favors the undemanding philanthropists or demanding aspirationalists over demanding philanthropists or undemanding aspirationalists if

$$b > \frac{(2N-4)\,d}{(d+1)\,N - 4d}. \tag{13}$$

The analytical results well predict the simulations (see Fig. R5). We have added the investigation into the two games in the Supplementary Information.

7) Supplementary Figure 2. Is there any reason/intuition as to why the frequency of cooperation is now non-linear in the need threshold for a population of philanthropists but still (somewhat) linear for a population of aspirationalists? Also, at least for panel a) you could also show the analytical results from Eq.32. Would they match the simulations?

We thank the reviewer for this interesting point. We have shown the analytical results from Eq. 32 and they match the simulations very well (see Fig. R6). Here we provide a few intuitions about the non-linearity of the cooperation abundance $x_C$ in relation to the need threshold $\alpha$ for philanthropists, and the linearity for aspirationalists. Although both of them are threshold-based behavioral motivations, they result in different evolutionary processes, especially under high motivation intensity, such as for $|\lambda|$=1.

In a population of undemanding philanthropists (small $\alpha$), individuals are easily satisfied and thus choose to cooperate. This brings high payoffs to individuals taking defection, driving them to switch to cooperation. This switching from defection to cooperation further increases the payoffs for others, leading to more transitions to cooperation, resembling positive feedback. This mechanism explains the nearly full cooperation observed in a population of undemanding philanthropists. However, when philanthropists are demanding (large $\alpha$) and not easily satisfied, the choice of defection induces more individuals to switch to defection, until all individuals choose defection. Overall, the positive feedback leads to global cooperation in a population of undemanding philanthropists and to global defection in a population of demanding philanthropists, corresponding to the sharp decrease and non-linearity observed in Fig. R6.

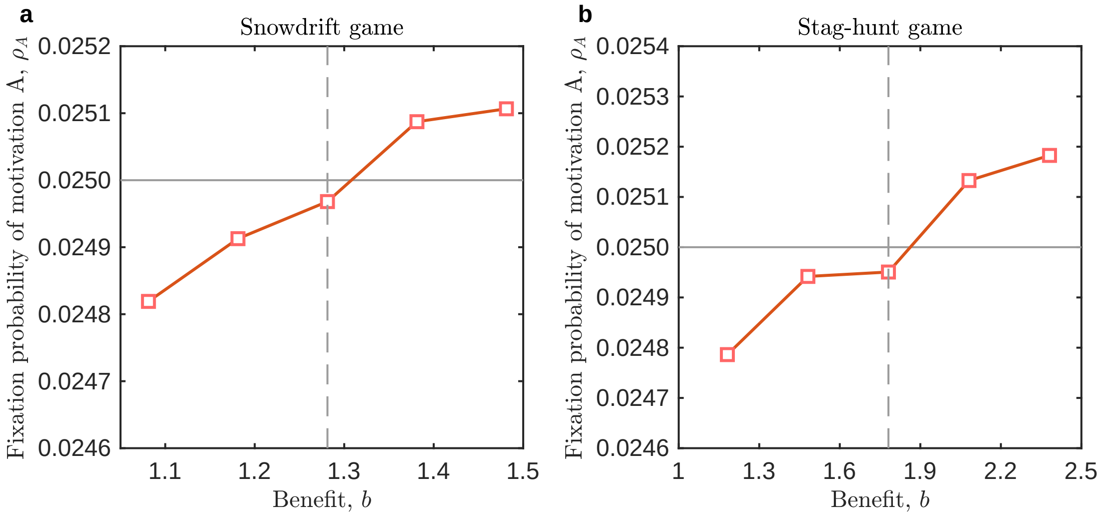



Figure R5: **Selection favors undemanding philanthropists over demanding philanthropists when the benefit of cooperation exceeds a critical value**. We consider random regular networks of size $N = 40$ and node degree $d = 6$. Presented is the fixation probability of an undemanding philanthropist with $\lambda = 0.01$ and $\alpha = 1.0$ (denoted by motivation A) in a population of demanding philanthropists with $\lambda = 0.01$ and $\alpha = 4.0$ as a function of benefit $b$, under snowdrift games (**a**) and stag-hunt games (**b**). Squares represent the fixation probability by simulations and vertical dashed lines mark the critical benefit for selection favoring undemanding philanthropists over demanding philanthropists. The fixation probability is determined by the fractions of simulations where the beneficial motivation A reached fixation out of $2 \times 10^7$ generations.

Conversely, aspirational motivation results in negative feedback. When the aspirationalists are undemanding, they easily switch to defection, but this decreases their payoffs, promoting a transition back to cooperation. Therefore, compared to philanthropists, it is harder for aspirationalists to reach a state of full cooperation or full defection, corresponding to the observed linearity.

We have updated Supplementary Figure 2 and added these details.

8) We do not really understand how weak $\lambda$ comes in in Figure 6. How is it different from Supplementary Figure 3? Perhaps it's the intensity of selection $\delta$ values that differ between the two figures? If so, what are the values? I understand that perhaps in the main text, the authors do not mention $\delta$ in separate terms from $\lambda$ to avoid confusion, but at least in the SI figure, the reader has already been introduced to it.

Sorry for the confusion. In Figure 6, weak $\lambda$ refers to the initial value of $\lambda$, which is small

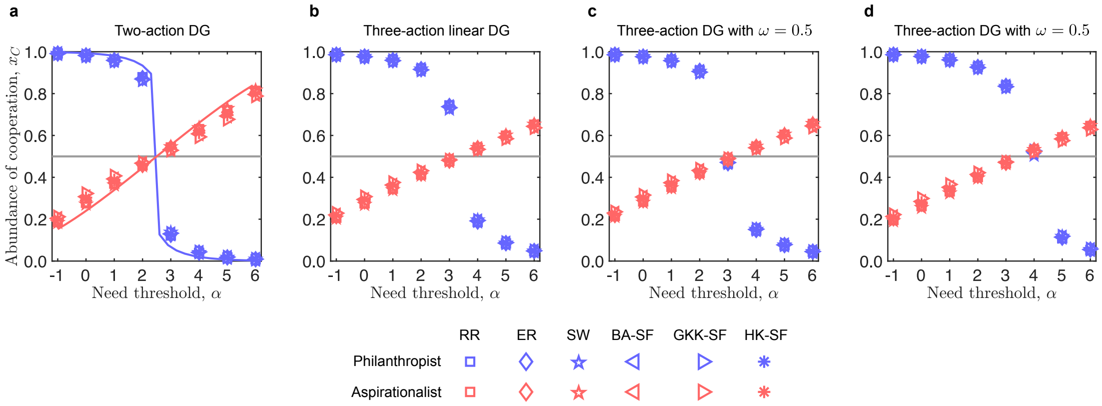



Figure R6: **Undemanding philanthropic and demanding aspirational behavioral motivations promote cooperation beyond weak motivation intensity**. Solid lines are analytical results by Eq. (32) in the Supplementary Information (blue for philanthropists and red for aspirationalists).

($|\lambda| = 0.05$). As the population evolves, individuals' motivation intensity increases, reaching $|\lambda| = 0.5$ and even $|\lambda| = 1.5$. In Supplementary Figure 3, the initial value of $\lambda$ is $|\lambda| = 0.5$, which is much larger than the initial value of 0.05 in Figure 6. We have revised the captions to clarify the variable setup.

9) Supplementary Figure 5. It would be nice to see the same kind of plot for c = 0.2 as well, to capture the lower panels' regime from Figure 6. Do the results also agree? Also, please consider changing the axes' labels "slope" and "threshold" to intensity and need, respectively, because that's how you referred to them most of the times in the main text.

Thank you for catching this. We have updated the figure to include a second panel with $c = 0.2$ and updated the axis labelling to be consistent with the rest of the manuscript. However we note this figure, which is based on evolutionary invasion analysis, applies to the well-mixed case (whereas Figure 6 shows results for a structured population). In the well-mixed scenario undemanding philanthropists are always disfavored, and the revised figure reflects that. However the revised figure does show that decreasing costs can change the dynamics of the well-mixed case in important ways by removing the intermediate stable state. We have updated the Figure legend to highlight this result.

Minor comments:
-The manuscript should motivate earlier the pertinence of considering weighted networks and the interpretation of weights in this model.

Sure. In human society, some interactions occur more frequently or last longer than others, resulting in different payoffs even when the actions of the participants are fixed. Here, we used weight to represent the interaction frequency or intensity, capturing the heterogeneity in interactions beyond just the number of connections. We have introduced this concept in the Model section.

-Lines 34-42 suggest that this study is about understanding the behavioral motivations behind each cooperation mechanisms, which is not the case.

We agree. We have edited the paragraph to remove this impression.

-Line 36: cheats? Cheaters?
-Line 54: trasncendance $->$ transcendance
-Line 61: seen $->$ be seen
-Line 72: met $->$ meet
-Line 74: hard to $->$ hardly (or for whom it is hard to)
-Figure 1b: Please mention in the caption that the illustrative payoffs assume all weights between neighbours to be 1.
-Equation 3: $e_{ji} -> e_{ij}$
-Line 163: Evolutionary of cooperation $->$ Evolutionary dynamics of cooperation
-Line 169: perhaps you meant Figure 3e.
-Line 204: should the condition be $\alpha < (b–c)/8$? Or else, how did you derive it?
-Line 248: more likely to fix than a neutral mutant

Exactly correct. The condition is $\alpha < (b–c)/8$.

-Figure 5: please explain meaning of inset values. One can infer this corresponds to $\alpha$ and $\lambda$, but the authors can explicitly point this out

We have revised the inset to make the parameters clear.

-Figure 5 caption: Dots $->$ Squares
-Figure 5 caption: explain vertical dashed lines

The vertical dashed lines in the original manuscript indicate the critical benefit-to-cost ratio $(b/c)^*$ for $\rho_A > 1/N$ by theoretical analysis. Since the crossing point of the analytical fixation probability and the horizontal line $y = 0.01$ also correspond to the critical benefit-to-cost ratio $(b/c)^*$, we have removed the vertical dashed lines to make the figure more concise.

-Line 266: appropriate?
-Figure 6: typo in the legend – should be $x_C > 1/2$ for blue and $x_C < 1/2$ for red. In the caption, regression instead of progression?

Exactly right. We have corrected typos and revised figures.

-Line 310: motivation range instead of "parameter range"
-Line 314: Fig. 5b − > Fig.6b
-Line 340: please consider citing some foundational papers on the framework of "adaptive dynamics" so that the reader knows what you mean by that

We have cited a few foundational papers about "adaptive dynamics".

-Line 462: name − > game
-Line 489: formalization − > formalizations
-SI Eq. 10: sums in m should be in the set L (not N)?
-SI Eq. 11: keep the bar for average as in Eq. (9)

We have corrected all typos and supplemented details to make the figures clear.

-If SI Eq. 18 is derived for linear games, how can it be used to obtain Equations 29 and 30 (for non-linear games)? What did I miss?

Equation 18 is derived for general games, applicable to both linear and nonlinear games. Specifically, by substituting

$$M_{k\ell} = -\frac{k-1}{L-1}c + \frac{\ell-1}{L-1}b \tag{14}$$

into Eq. 18, we obtain Eq. 27, which provides the condition for motivation A being favored in multi-action linear games. Similarly, by substituting

$$M_{k\ell} = -\frac{k-1}{L-1}c + \frac{1-\omega^{\ell-1}}{1-\omega^{L-1}}b \tag{15}$$

into Eq. 18, we derive Eq. 29, which outlines the condition for motivation A being favored in multi-action nonlinear games.

-Supplementary Figure 6. Are the parameters and other details the same as in Figure 3e of the main text? If so, please mention that in the caption.

Exactly right. We have supplemented parameter details.

---

**Replies to comments and suggestions from reviewer 4**

---

**Comments:**

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

We thank the referee for reviewing our paper and providing valuable suggestions for its improvement.

---

**Replies to comments and suggestions from referee 1**

---

**Comments:**

Thank you for the opportunity to review this revised manuscript. I appreciate the authors' careful response to the issues I raised. I also found the comments by the other reviewers very valuable for my own understanding of this work.

As the authors emphasise in their response letter, a key aim of this manuscript is to integrate evolutionary game theory and human decision making by modelling the evolution of motivations (or heuristics) which can differ between individuals and across contexts. I must admit that even with the revised literature review, I remain unconvinced by this connection.

If the aim is to model the evolution of actual motivational heuristics, then the shape of the motivations in the model should at least be plausible abstractions of behavioural or motivational patterns documented in the psychological literature. Here, motivations are described by two dimensions: (a) the need threshold (i.e. how demanding or undemanding an individual is) and (b) the intensity (i.e. whether an individual cooperates until meeting their need threshold and defects above it or vice-versa).

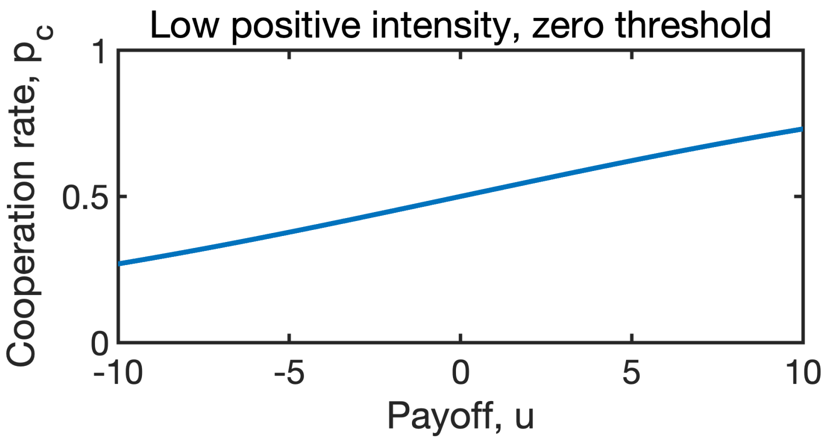
They intensity dimension strikes me as the conceptually more important one, since it captures the central idea that cooperation is dependent on an internal, absolute reference point. The question is whether this is a plausible model of human decision making. The evidence presented refers to correlations between SES and prosocial behaviour, but such crosssectional correlations are at best suggestive and provide no direct evidence of within-person reference dependence. While such reference dependence has been discussed in the literature (for example, this assumption underlies the experimental practice of paying only some rounds of iterated games to avoid account-keeping), I am not aware of a formal model or empirical test of this idea. I would've liked to see a stronger case in the literature review that these are actually plausible shapes for human utility functions.

I want to emphasise that I find the manuscript interesting. This work clearly makes a contribution to the literature on aspiration dynamics. However, I am not convinced that it models the evolution of motivational heuristics which are plausible abstractions of actual motivational heuristics, which does limit the insight it offers for human decision making or the understanding of the evolution of cooperation.

We really appreciate the thoughtful and constructive reviews that the referee has provided throughout this process. We agree with much of what you say and we certainly accept the importance of engaging properly with the Psychology literature when building a model of human decision-making and motivation. However, it's important to understand that the emergence of an internal "reference point" is an *outcome* of our analysis of the evolutionary dynamics of motivations, not an a priori assumption.

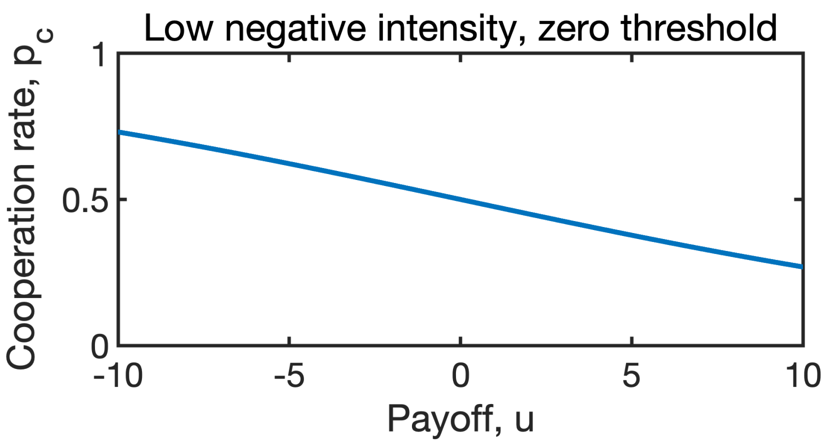
To explain why this is so, we must consider the space of possible motivation response curves that evolution has access to in our model. The referee is correct that the motivation response curves are characterized by a need threshold and need intensity. There is also a maximum and minimum payoff that can be derived from social interactions, for a given population size and a given benefit $B$ and cost $C$ of cooperation.

To see the different types of motivation response function that can evolve, let us first suppose that the need intensity is small and positive. In this case the function is close to linear and increasing:

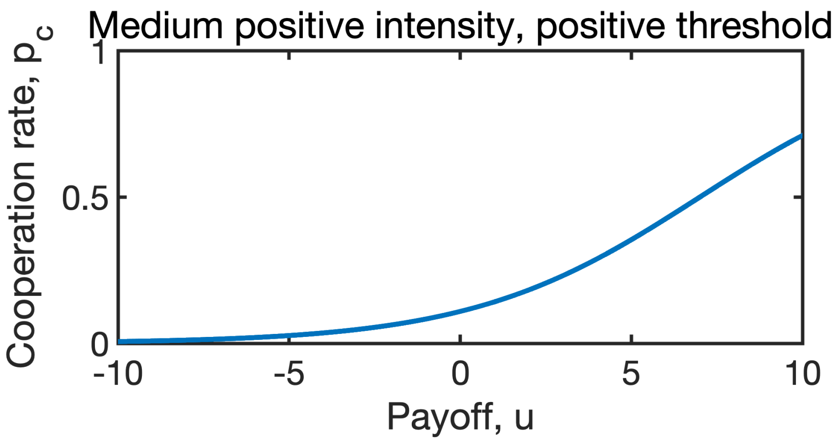



(where we have fixed the max and min payoffs to $\pm 10$ for the purposes of illustration)

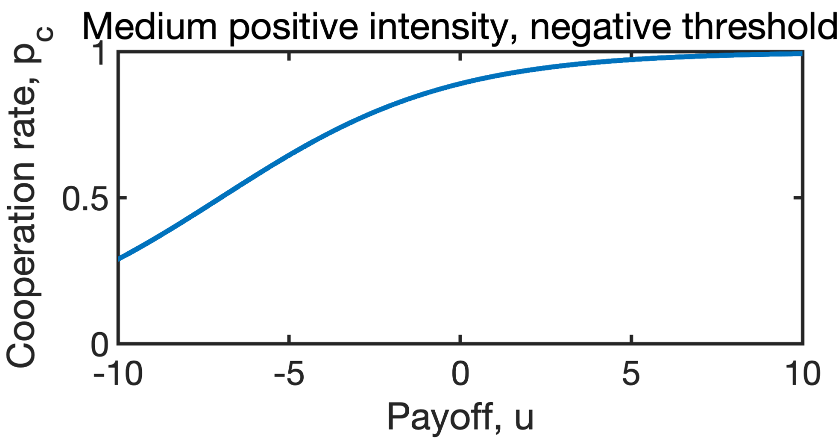
If we instead make the need intensity small and negative the response function is linear and decreasing:



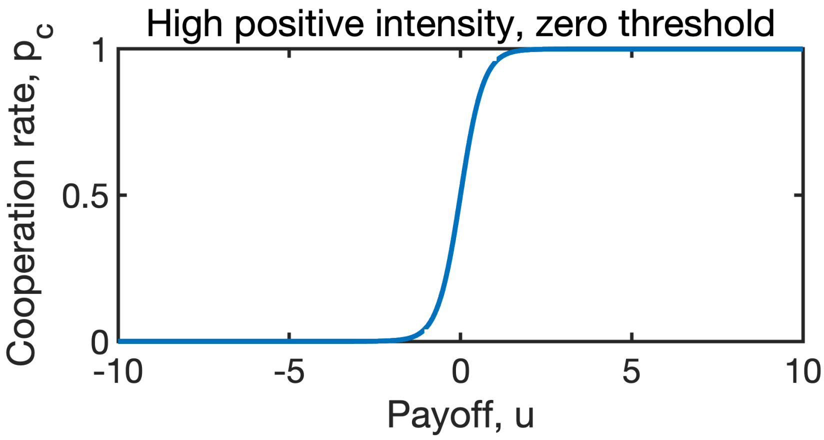
If we make the need intensity intermediate and positive, and the need threshold positive, we get a convex (accelerating) response function:




If we make the need intensity intermediate and positive, and the need threshold negative, we get a concave (decelerating) response function:



And of course if we make the need intensity high and positive we get the familiar S-shaped function




(where of course the final three curves have analogues for negative intensity).

When thinking about how our results connect to the literature then, it is not accurate to say our model assumes an internal reference point, any more than an empiricist who inferred a linear motivation response function would take this as evidence of an internal reference point. Empirically, an internal reference point would correspond to a sharp threshold, in which cooperation rate "switches" from low to high (or high to low) at some specific level of payoff, as seen in the final figure above. We do indeed find that evolution leads to such sharp thresholds, but this is an *outcome* of our model. This is (we believe) a novel evolutionary prediction that is to be tested.

We also think it's important to emphasize that the flexibility of our framework is even greater than described above. In the Supplementary Information we describe how to generalize our framework to even more complex motivation response functions. Our framework accommodates an arbitrary number of internal reference points and can incorporate arbitrarily complex action-updating functions. For example, our framework can model the U-shaped charitable giving profile highlighted by referee #2, where some studies find that individuals in both lower and higher income brackets donate a higher percentage of their income to charity. This giving profile is more intricate than the one explored in the main text, and our more general framework allows the introduction two internal reference points and two updating functions to capture this complexity. This is just one way in which we believe our theoretical framework can serve as a foundation for future research, including experimental studies designed to test these predictions.

We have revised the manuscript to better distinguish between the assumptions of our model and the predictions that arise from it.

---

**Replies to comments and suggestions from referee 2**

---

**Comments:**

We would like to thank the authors for the comprehensive and clear answers. Overall, we find the answers satisfactory, and we believe this paper is a suitable contribution to Nature Communications.

We appreciate the reviewer's satisfaction of our revisions last round and their constructive comments for further improvement. We have incorporated all of the reviewer's suggestions in our revisions.

We have two final remarks, however:

Regarding our first major concern, the authors provide a satisfactory answer that could be better explored in the manuscript. We believe the first result discussed by the authors — undemanding philanthropic and demanding aspirational behavioral motivations promote cooperation — still sounds rather trivial when reading the manuscript. The analysis described by the authors in their response, which consists in keeping the average threshold $\alpha$ constant and varying the position of nodes with low and high $\alpha$ on the network, is a very interesting one. However, this is not explored in the paper. Figure 3 assumes that $\alpha$ values are assigned uniformly at random and no type of network correlation on $\alpha$ is studied. Our suggestion is to either 1) implement the network correlation on $\alpha$ analysis that the answer suggests; or 2) clarify in the manuscript why this result presented defies intuition; or 3) to be upfront and state it as a simple model consistency check.

This is an excellent point. In the revised manuscript, we investigated the two cases within the same heterogeneous structured populations, In the first case, individuals occupying the largest 50% of highly connected nodes have a higher threshold, $\alpha + \epsilon$, while those in the least 50% have a lower threshold, $\alpha - \epsilon$. In the second case, the thresholds are reversed: individuals in the largest 50% of highly connected nodes have a lower threshold, $\alpha - \epsilon$, and those in the least 50% have a higher threshold, $\alpha + \epsilon$. Additionally, we have included a new paragraph in the main text to discuss how the results challenge intuition.

Regarding Supplementary Figure 5, we commend the authors for the newly added panel b. We see how the "results agree qualitatively with the main text results" for high c (panel a), but not for low c (new panel b). As the authors point out in their reply to the reviewers, this figure differs from main Figure 6 in that it refers to a well-mixed population analysis, and that's perhaps the reason for the disagreement. Therefore, it would be best to either provide some intuition why the results disagree for low c, or at least avoid saying that the results agree qualitatively.

We apologize for any confusion caused. To clarify, we first examine a well-mixed population of size $N$ based on Eq. (6). We have the key values

$$\begin{aligned}\eta_1 &= \frac{N-1}{2},\\ \eta_2 &= \frac{N-2}{2},\\ \eta_3 &= \frac{N^2-3N+3}{2(N-1)}.\end{aligned} \tag{1}$$

This leads to a negative critical value $(b/c)^* = -(N-1)$. According to Eq. (6) in the main text, for any positive values of $b$ and $c$, we always have $-c\eta_2+b(\eta_3-\eta_1) < 0$. Thus selection consistently favors the behavioral motivations that result in lower cooperation levels, such as $\mathbf{A}$ over $\mathbf{B}$ when $x_C(\mathbf{A}) < x_C(\mathbf{B})$. And the population evolve towards the demanding philanthropists (region I in Figure 6a) and undemanding aspirationalists (region III in Figure 6a). The case shown in Figure 6b is impossible in the well-mixed setup. The results presented in Supplementary Figure 5 are based on the evolutionary invasion analysis. Regardless of whether the cost $c$ is small (but positive, Supplementary Figure 5a) or large (Supplementary Figure 5b), we find that for the low threshold $\alpha$, selection favors the aspirationalists (with negative $\lambda$); for a high threshold $\alpha$, selection favors the philanthropists (with positive $\lambda$). Therefore, the results from the adaptive dynamics analysis in Supplementary Figure 5 qualitatively (i.e., stronger motivation intensity) agree with those based on Eq. (6) in the main text (i.e., weak motivation intensity). In the revised manuscript, we have added more details to avoid any further confusion.

---

**Replies to comments and suggestions from referee 2**

---

**Comments:**

Thank you for the opportunity to review this revised manuscript.

In my earlier review, I emphasised the need to more clearly link the proposed model to the existing literature on human motivation and highlighted discrepancies between features of the model and what the empirical literature claims. In their rebuttal, the authors argue that some of the features I highlighted – in particular, the presence of internal reference points – are outcomes of the model, rather than assumptions. I find this argument only somewhat convincing. There are two reasons. First, while it is true that internal reference points emerge from the model, allowing motivations to vary along two dimensions – need threshold and intensity – does make some assumptions about what features motivational heuristics could have. I am not fully convinced that the choice of these dimensions has a strong grounding in the existing literature, theoretical or empirical, on social decision making.

Second, if the aim is to model the evolution of actual human motivational heuristics – and this is what the authors claim – then the outcomes should, in my view, show at least some convergence with empirical observed motivational heuristics. That said, I recognise that this may be more of a matter of taste about the aims of evolutionary models, and as in my previous reviews, I happily accept that the paper makes a meaningful contribution to the study of a class of models which include such motivations.

We have highlighted the limitations in Discussion as follows "While our theoretical approach provides new insights into the evolution of motivations for cooperation, it has certain limitations. First, human motivational heuristics are far more complex and diverse than the two factors — need threshold and motivation intensity — considered in our model. Second, although our findings build on classical evolutionary dynamics and offer novel theoretical perspectives, empirical validation remains essential. Future experimental studies could examine how environmental fluctuations shape cooperative behavior and how cooperation adapts to periods of scarcity and abundance. Incorporating additional factors could further enhance our understanding of human behavior and the motivations that drive it."

---

**Replies to comments and suggestions from referee 3**

---

**Comments:**

We thank the authors for the new revisions and clarifications. In our previous review we expressed our positive assessment regarding this manuscript and fit for Nature Communications. We believe it presents relevant results on the evolution of motivations that underlie cooperative behaviour. We just raised two minor remarks: 1) It struck us as trivial the observation that "undemanding philanthropic and demanding aspirational behavioral motivations promote cooperation". The authors have complemented their analysis with a network study where threshold $\alpha$ now depends on network properties. Essentially, the authors conclude that cooperation does not depend on thresholds being placed according to nodes' degree. What ultimately drives the observed results is the average $\alpha$ considered, and no network effects are observed. Our suggestion here remains: the authors could simply state that the effect of changing average $\alpha$ leads to an average outcome that conforms with the expected results given the model definition. 2) The authors' answer is clear, and the point clarified. Overall, and despite minor point 1), we believe this paper is a suitable contribution to the journal.

We appreciate the reviewer's satisfaction of our revisions last round. We have summarized the main conclusion in a single sentence in the revised manuscript,"Put simply, altering the average need threshold ($\alpha$) produces an overall outcome that aligns with the expected results based on the model's definition."

---

**Replies to comments and suggestions from referee 5**

---

**Comments:**

SUMMARY. In this paper, the authors propose a novel approach to study the evolution of cooperation on social networks. Traditionally, in this literature what evolves is whether or not people wish to cooperate. Instead, this study acknowledges that such cooperation decisions are often based on heuristics. Here, the authors consider a particular form of heuristics. They assume that cooperation decisions depend on the person's accumulated wealth in the previous round. This setup is inspired by previous experiments asking whether high-income individuals would behave more or less socially than low-income individuals.

To explore this question, the authors study a dynamics on two timescales. In the short run, individuals' heuristics are fixed. Here, people use their heuristics (and their previous experiences) to decide whether or not to cooperate in the next round. In the long run, also the individuals' heuristics might evolve (here, according to a death-birth process, in proportion to the payoffs that the heuristics generates on average). The outcome depends on the benefit-to-cost ratio and on the network structure. In particular, when cooperation is particularly valuable (for high benefit-to-cost ratios), selection favors "undemanding philantropists" or "demanding aspirationalists" (i.e., heuristics that make it particularly likely the actor will cooperate).

GENERAL EVALUATION. In my view, the authors study an interesting and underexplored topic. Traditionally, evolutionary game theory takes an ultimate perspective when explaining human cooperation: What factors might make cooperation favorable on an ultimate level? Instead, here the authors combine this ultimate perspective with a proximate perspective (what kind of heuristics would evolve to help people make their cooperative decisions)? The methods used in the paper are state of the art. In particular, the paper presents several analytical results in the limit of weak selection (or, here, rather: limit of weak motivation intensity). In addition, I appreciate that the authors show the robustness of their analytical findings with simulations (also in the case that the motivation intensity is substantial). I understand that I was added to the list of reviewers after there was already a first round of revisions. Out of respect for a streamlined publication process, I will thus not raise any new major topics that would require substantial revisions (I do not see any substantial gaps in the authors' analysis anyways). Instead, let me comment on how I think the authors have responded to the previous comments by other reviewers. First (regarding the comments of referee 1), I see the referee's point that the authors already seem to presume that individual heuristics take the form of a threshold function. However, I do feel the authors give a satisfactory answer: Although their functions are formally threshold functions, parameters can be adjusted in ways such that these functions can also implement alternative response paradigms (e.g., approximately linear responses). So if there is some evolution towards functions with a clear threshold, that is better interpreted as a result of the model, rather than a fixed assumption. Second, I feel that also the requests of referee 2 have been taken into account in a satisfactory manner. Overall, I think the paper makes a good contribution to the evolutionary game theory literature, and that it merits publication in Nature

Communications.

We sincerely appreciate the reviewer's insightful summary of our manuscript and their assistance in addressing concerns raised by other reviewers. We have carefully implemented all of the reviewer's suggestions in our revisions.

MINOR COMMENTS. (1) Figure 3: In the figure caption, the authors write that "undemanding philantropic and demanding aspirational motivations lead to abundant cooperation". Given that all cooperation rates here are between 0.49 and 0.51, the term "abundant" is somewhat misleading. Perhaps the authors can say that "cooperation is favored" under these conditions.

We thank the referee for this suggestion. We have revised the caption of Figure 3.

(2) In line 373, the authors say they check the robustness of their results when the motivational intensity increases up to 2. If I understand the figure caption of the corresponding extended data figure 2 correctly, the intensity tested there is 1 (which of course is just as fine).

We thank the reviewer for this point. We have corrected this typo.

(3) Not for this paper, but perhaps some suggestion for future research: I saw that in the end of the SI document, the authors derive some adaptive dynamics equation for their system. I understand that this adaptive dynamics is not the focus of this paper, but at least I would be interested in a more formal analysis of the trajectories of this dynamics.

We agree, and we had planned to do more in this direction. However the points of zero selection gradient can only be calculated numerically for this model, which makes a formal analysis a complicated task. And so we decided to defer this to future research where the adaptive dynamics can be explored more fully. We have however edited the SI to point out the potential for multi-stability in the selection intensity $\lambda$, as illustrated in Supplementary Figure 5.

(4) Related to the previous point – perhaps instead of representing a player's cooperation probability with a function with two free parameters, one might even be able to let the entire function evolve (e.g., the paper "The adaptive dynamics of function-valued traits", JTB 2006). However, that would be an entire project on its own, and probably a rather challenging one.

We thank the reviewer for this excellent point and providing a valuable reference. We have incorporated this promising direction for future research in the Discussion section.