On the origins of human sociality

Akdeniz, A.

Publication date
2023
Document Version
Final published version

tinbergen institute
GRADUATE PROGRAM

On the Origins of Human Sociality

Aslıhan Akdeniz

Universiteit van Amsterdam

On the Origins of Human Sociality

Aslıhan Akdeniz

ISBN: 978 90 361 0557 6

Cover photo: Generated by DALL-E 2

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. 810 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

On the Origins of Human Sociality

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 9 maart 2023, te 16.00 uur

door Aslihan Akdeniz
geboren te Üsküdar

Promotiecommissie

Promotor:	prof. dr. C.M. van Veelen	Universiteit van Amsterdam
Copromotor:	prof. dr. S. Shalvi	Universiteit van Amsterdam
Overige leden:	prof. dr. A.J.H.C. Schram	Universiteit van Amsterdam
	prof. dr. K.R.L. Janmaat	Universiteit van Amsterdam
	dr. A. Ule	Universiteit van Amsterdam
	prof. dr. S. Suetens	Tilburg University
	dr. S. Mathew	Arizona State University
	prof. dr. C. Efferson	University of Lausanne

Faculteit Economie en Bedrijfskunde

Acknowledgements

‘Bummer!’ Matthijs would exclaim whenever we got upset over something, and that was one of the things I learned (or maybe, copied, to be more precise) from him over the years we worked together. And now what a bummer it is that we reached the end of our time together as trainer and trainee. The first instance that I reached out to Matthijs was in the second year of my master’s. I was extremely enthusiastic to work on evolution in my PhD. So, one day, I knocked on his office door and asked if he had time to talk; he responded “sure, about life?”. Now looking back, I realize how indicating this instance was of him as a mentor and a collaborator. Even though talking about science constituted the bulk of our meetings, our conversations were never restricted to research. We talked about many things regarding ‘life’ over the years of my PhD, and he inspired me to think about things in ways that I wouldn’t have imagined without his perspective. I would like to thank you, Matthijs, for everything you have taught me, about being a good scientist and otherwise. I hope to continue having nice lunches at Bloem and discussing all sorts of ‘life’ matters.

There is one more person whose mentoring was extremely valuable for me. Sarah, even though I was not one of your students, you always guided me and became a role model. My visit at your department in Arizona was among the best experiences of my PhD. The couple of months I spent in the desert had an immense impact on me, scientifically and personally. I am more than grateful to work with you. Our collaboration was not only fun, but it also moved my research in a direction that I was extremely interested in. It moreover enabled me to conduct a study at my alma mater in Istanbul, which had great emotional significance for me. I would like to thank you for all these things.

I was also lucky to have wonderful friends who helped me in various ways throughout the PhD. Some provided feedback on my research, some listened to me complain, some offered mental support, some provided fun to take the weight of other things, and some helped me think about who I want to become as a researcher. Many friends did more than one

of these things. I would like to thank one person specifically, who supported me along all these dimensions –and many more. Thank you, Ayşe Gül, for being a friend, a colleague, and a mentor. I would also like to thank Johan and Konstantinos for their fellowship starting at the Tinbergen Institute and going towards our PhDs at UvA (and hopefully much further). Thank you, Chris, for being a great co-author and a fun office mate. Thank you, Kathi, Andreas, Davide, Sneha, and Margarita, for all the fun conversations and foosball games. There are many more to list, but with the fear of missing anyone out, I would like to thank all my friends and colleagues at CREED for making the past few years enjoyable and bearable. I also greatly appreciate the support I received from the UvA and TI staff, especially from Wilma, Arianne, and Robert, over the years. Thanks to them, the PhD journey was a much less bumpy ride than it would have been otherwise.

Most of all, I would like to thank my parents, Hatice and Nadir, and my partner, Rik, for all their love and support. It's without a doubt that someone's love and support in tough times are more indicative than in good times. And being someone's kid or partner, you happen to go through tough times for sure. They are the ones that see us at our most vulnerable, furious, and desperate –which, I can say, happens regularly during a PhD. Therefore, I really mean it when I say that I would not have made it without them. Thank you, mom and dad, for being there for me since the beginning of my journey in Istanbul and for providing me with every opportunity you can so that I can be happy, healthy, and free. And thank you, Rik, for joining me in the second act of my journey in Amsterdam and for making everything in my life infinitely better from there on, I look forward to spending the upcoming acts together. And finally, I would like to thank the youngest (and furriest) member of our household and definitely the best economist in the house, Pigou.

Aslı

Utrecht, January 2023

Contents

Chapter 1

Introduction

Human nature contains both good and evil. We observe displays of extreme prosociality and substantial cruelty in human societies around the world. To answer the question of how these two extremes can co-exist, and why we observe any prosociality at all, given its costs, I study the origins of human sociality. If we understand where our social preferences and behaviours come from, we can then understand why people do or do not care about one another, and why there exist high levels of heterogeneity in sociality across societies, across individuals within a society, and across time and space within an individual.

My dissertation, therefore, centres around exploring the origins of and the variation in human sociality. In my projects, I study how humans evolved to be ultra-social while the extent of sociality in other animals, and especially in non-human primates, remained relatively restricted. In doing so, I focus on the mechanisms that can explain the co-existence of good and evil in us, and look for unique elements in human social interactions, or a combination thereof, to pinpoint the cause of divergence in human sociality.

Chapter 2 examines a critical assumption from the group selection literature. Group selection models combine selection pressure at the individual level with selection pressure at the group level. Cooperation can be costly for individuals, but beneficial for the group, and therefore, if individuals are sufficiently much assorted, and cooperators find themselves in groups with disproportionately many other cooperators, cooperation can evolve. The existing literature on group selection generally assumes that competition between groups takes place in a well-mixed population of groups, where any group competes with any other group equally intensely. Competition between groups however might very well occur locally; groups may compete more intensely with nearby than with far-away groups. We show that if competition between groups is indeed local, then the evolution of cooperation

can be hindered significantly by the fact that groups with many cooperators will mostly compete against neighbouring groups that are also highly cooperative, and therefore harder to outcompete. The existing empirical method for determining how conducive a group structured population is to the evolution of cooperation also implicitly assumes global between-group competition, and therefore gives (possibly very) biased estimates.

Chapter 3 experimentally tests the role of partner choice in the evolution of honesty. There are many situations in which the ability to lie can help one get ahead. Most people however hesitate to lie and regularly prefer to tell the truth, even if lying could work to their advantage. Why honesty exists therefore is an open question. We explore the possibility that it has evolved because of partner choice. In a lab experiment, we find that in a situation in which their partner will know more than they do, subjects do prefer to be matched with honest partners. We also find that they are right in doing so because honest partners will behave consistently, and on average more, prosocially. We do not find support for the possibility that honest partners behave more prosocially in order to avoid the choice between lying and revealing having been selfish. Instead, our findings are consistent with an explanation in which honest people behave more prosocially because they also have a harder time justifying selfish behaviour to themselves.

Chapter 4 presents an extensive literature review, where we compare the explanations for human cooperation from the theoretical literature with the empirical observations from the experimental literature, and argue that there is a mismatch between the two. A considerable share of the literature on the evolution of human cooperation considers the question of why we have not evolved to play the Nash equilibrium in prisoners' dilemmas or public goods games. In order to understand human morality and pro-social behaviour, we suggest it would actually be more informative to investigate why we have not evolved to play the subgame perfect Nash equilibrium in sequential games, such as the ultimatum game and the trust game. The "rationally irrational" behaviour that can evolve in such games gives a much better match with actual human behaviour, including elements of morality such as honesty, responsibility, and sincerity, as well as the more hostile aspects of human nature, such as anger and vengefulness. The mechanism at work here is commitment, which does not need population structure, nor does it need interactions to be repeated. We argue that this shift in focus can not only help explain why humans have evolved to know wrong from right but also why other animals, with similar population structures and similar rates of repetition, have not evolved similar moral sentiments. The suggestion that the evolutionary function of morality is to help us commit to otherwise irrational behaviour stems from the work of Robert Frank (1987; 1988), which has played

a surprisingly modest role in the scientific debate to date.

Chapter 5 examines whether deviations from selfishness in the ultimatum game can be explained by noise or mistakes in individuals' choices. In this chapter, we review, upgrade, and synthesize existing models from evolutionary game theory on the ultimatum game, and we compare their predictions with the existing experimental evidence. We find that the results in Gale et al. (1995) and Rand et al. (2013) are primarily driven by bias in the mutations. We make versions with local instead of global mutations for both. This minimizes the bias and changes the results. We also consider Quantal Response Equilibria in combination with the assumption that individuals are selfish after all. The Quantal Response Equilibrium is the noisy twin of the Nash equilibrium, and looking at this combination we explore an alternative explanation for what we observe in the lab, namely noise instead of deviations from selfishness. Finally, we provide a refurbished version of the model of commitment in Nowak et al. (2000). The de-biased version of the model in Rand et al. (2013) becomes a special case of this more general model (with the possibility for commitment muted). We find that the experimental evidence does not align with the models in Gale et al. (1995), Rand et al. (2013), or our de-biased versions of them, and that it also rejects the combination of selfishness and the Quantal Response Equilibrium. All of these models predict that the distribution of minimal acceptable offers should start with high frequencies at 0, end with low frequencies at 1, and have decreasing frequencies in between, which is not what is found in lab experiments. Instead, the experimental evidence is in line with a commitment-based explanation, where the ability to commit to rejecting unfair offers, while being *ex-post* suboptimal, can be *ex-ante* beneficial.

Chapter 2

The cancellation effect at the group level[1]

2.1 Introduction

There is a wide variety of positions on the role of group selection in human evolution. One end of the spectrum considers group selection to be a key ingredient of human evolution (Haidt, 2012; Richerson et al., 2016; Sober and Wilson, 1998; Wilson and Wilson, 2007). The other side suggests that "group selection has no useful role to play in psychology or social science" (Pinker, 2015); see also (Wade, 1978; Williams, 1966). In this paper we will not resolve this controversy, nor take a position in this debate, but what we will do is consider a crucial element that has been missing, both from the current group selection models, and from the current empirical approach to establishing how conducive group structure is to the evolution of cooperation.

The defining characteristic of a group selection model is that it captures the opposing effects of selection at the individual level, where defectors do better than cooperators within groups, and selection at the group level, where groups with more cooperators do better than groups with fewer cooperators (Wilson and Wilson, 2007). The existing models within the group selection literature all share the property that competition between groups happens globally; all groups compete with all other groups equally intensely (Boyd and Richerson, 2009; Luo, 2014; Luo and Mattingly, 2017; Simon, 2010; Simon et al., 2013; Traulsen and Nowak, 2006; van Veelen et al., 2014). This is a useful simplification if the aim is to illustrate the possibility of a tug of war between the different levels of selection.

[1]This chapter is based on Akdeniz and van Veelen (2020).

It may however not always be particularly realistic. Groups themselves typically live in a structured population of groups, where neighbouring groups compete with each other more than they do with groups that are further away. Local dispersal would then imply that groups with many cooperators are typically surrounded by groups that also contain many cooperators, compared to the groups that surround groups with many defectors. More cooperative groups therefore might also be subject to more intense competition at the group level. This can significantly dampen the benefits of being a cooperative group, which, in turn, affects the balance between selection at the individual and at the group level. In models without group structure a similar phenomenon, but then at the individual level, is called the cancellation effect (Taylor, 1992a;b; Wilson et al., 1992). We show that the cancellation effect also exists at the group level, where it plays out in a more complex way, and that it can make a sizable difference for the conditions under which cooperation can evolve by group selection. This also has empirical implications. The current standard approach to determining how large the benefit to the group should be, compared to the cost to the individual, for cooperation to evolve by group selection implicitly assumes global between-group competition (Aoki and Nozawa, 1984; Bell et al., 2009; Bowles, 2006; Crow and Aoki, 1984; Langergraber et al., 2011; Walker, 2014; Weir and Cockerham, 1984). If competition between groups is not global, but at least to some extent local, then this procedure paints too positive a picture of how favourable conditions are for the evolution of cooperation by group selection.

2.2 Model

In order to study the difference between global and local group competition, we consider a stylized model, in which m groups consisting of n individuals live on a cycle (Figure 2.2.1). Individuals can either be a cooperator (C) or a defector (D). In every time period, one of three types of events will happen: individual reproduction, group reproduction, or migration. These events happen with probabilities p, q and r, respectively, where $p + q + r = 1$. We compare two different processes for group reproduction, one with local and one with global between-group competition.

If an individual reproduction event occurs, first a random group is selected, where all groups have equal probability of being chosen. Then an individual from the selected group is chosen to reproduce. Within the group, defectors get a payoff of 1 and cooperators get a payoff of $1 - c$. The intensity of selection w is then used to transform these payoffs to

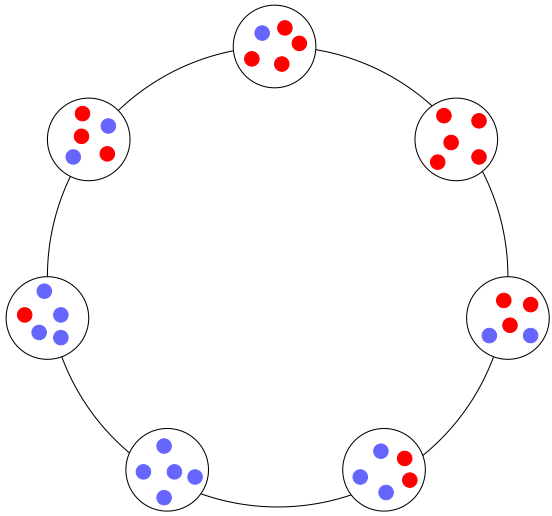


Figure 2.2.1: An example of a population state on a cycle with $m = 7$ groups of $n = 5$ individuals each. The blue dots indicate cooperators and the red dots indicate defectors.

values f_C and f_D:

$$f_C = 1 - wc \qquad \text{and} \qquad f_D = 1$$

The probabilities with which individuals are chosen for reproduction within the group are proportional to these values. Whenever an individual reproduces, an individual from the same group is chosen to die, where each individual, including the parent, but excluding the offspring, is chosen with probability $\frac{1}{n}$ (Figure 2.2.2a).

If a group reproduction event occurs, then one group is chosen to reproduce, and one group is chosen to die. The groups are numbered $i = 1, ..., m$, and k_i is the number of cooperators in group i. These groups live on a cycle, so i and $i + 1$ are neighbouring groups, and so are groups 1 and m. The group payoff of group i is 1 plus b times the share of cooperators in the group. The intensity of selection w is then used to transform these payoffs to values

$$g(k_i) = 1 + w\frac{k_i}{n}b$$

We consider two update processes for group reproduction; Birth-Death and Shift. In both of them, first a group is chosen for reproduction, where each group's probability of

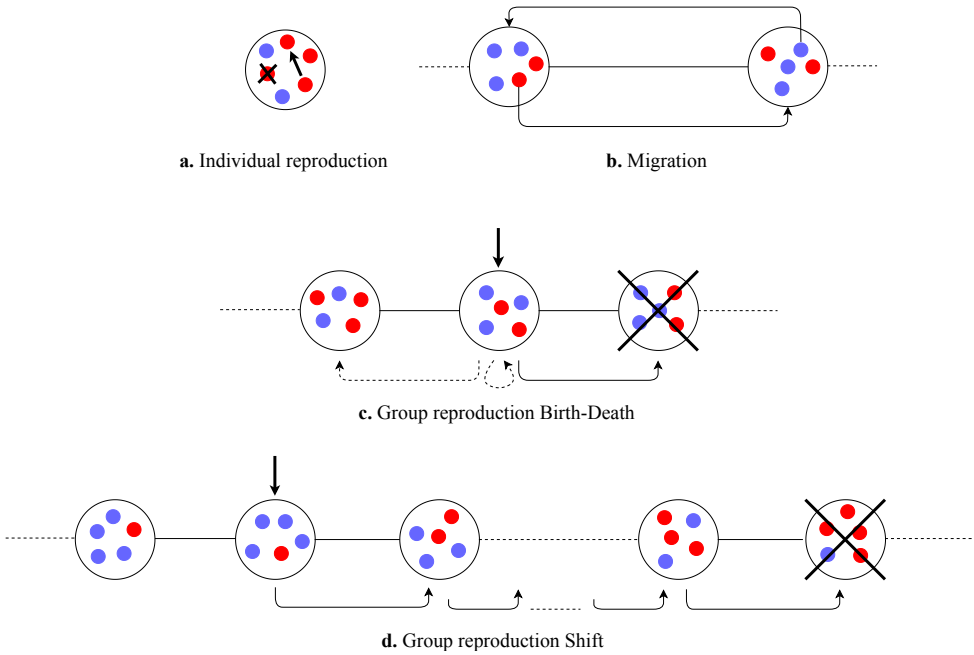


Figure 2.2.2: **(a)** At an individual reproduction event, one individual reproduces, and one individual within the same group dies. In this example, one defector is chosen for reproduction, and another defector is chosen to die, so the overall group composition has not changed. Defectors have a higher chance of being chosen for individual reproduction than cooperators do. **(b)** At a migration event, two individuals from neigbouring groups trade places. **(c)** In the Birth-Death process, the group that is chosen to reproduce produces an identical offspring group. This offspring group then replaces one of the neighbouring groups, or, with a small probability, it replaces the parent group itself. **(d)** In the Shift process, the group that is chosen to reproduce also produces an identical offspring group, but here any group can be chosen to die, including the parent group. If the parent group and the dying group are more than 1 position apart, all groups between them move over one position. In both processes, groups with many cooperators have a higher chance of being chosen for group reproduction than groups with many defectors.

being chosen is proportional to their value $g(k_i)$. With Birth-Death, the offspring group then replaces the left or the right neighbour of the parent group, both with probability $\frac{m-1}{2m}$, and it replaces its own parent group with probability $\frac{1}{m}$ (Figure 2.2.2c). This makes competition at the group level local. With Shift, each group, including the parent group, but excluding the offspring group, is chosen to die with probability $\frac{1}{m}$. Unless the offspring group replaces the parent group, the new group is placed either to the right or to the left of the parent group, with equal probability, and every other group in between the parent group and the dying group moves over one spot (Figure 2.2.2d). With Shift, every

group is equally likely to die, irrespective of the composition of their neighbouring groups. Competition between groups is therefore global, as it is in the standard group selection models that have a well-mixed population of groups.

Finally, if a migration event happens, then a random pair of individuals from neighbouring groups trade places (Figure 2.2.2b).

2.3 Results

We first analyse this model in the limit of weak selection using inclusive fitness. We can do this, because the effects that being a cooperator instead of a defector has on individual reproduction rates, and on individuals death rates, as well as the effects it has on reproduction and death rates of groups, satisfy generalized equal gains from switching in the limit of weak selection (van Veelen, 2018; van Veelen et al., 2017a).

The fitness effects of the focal individual being a cooperator instead of a defector are given in Figure 2.3.1. Conditional on an individual event happening, and it happening in the group of the focal individual, the probability that any given individual is chosen to reproduce is proportional to its payoffs, scaled by the individual intensity of selection. That implies that these probabilities are the individual's own value, which is either f_C or f_D, over the sum of these values for everyone within the same group, including the individual itself. In the limit of weak selection, that amounts to a decrease proportional to $\frac{1}{n}c$ for the focal individual due to the decrease in the numerator of this probability, and an increase proportional to $\frac{1}{n^2}c$ for everyone, including the focal individual, due to the decrease in the denominators. Those add up to the changes in individual reproduction rates given in Figure 2.3.1. Individual death rates are unaffected.

Conditional on a group event happening, the probability that group i is chosen to reproduce is proportional to $g(k_i)$, which is the average level of cooperation within the group, scaled by the group intensity of selection. The effect of being a cooperator instead of a defector on the group average k_i is $\frac{1}{n}b$, and in the limit of weak selection, the effect on the group reproduction probability is proportional to $\frac{1}{m}\frac{1}{n}b$ through an increase in the numerator for the group of the focal individual, and $-\frac{1}{m^2}\frac{1}{n}b$ through an increase in the denominator for every group, including the group of the focal individual. Those amount to the changes in group reproduction rates given in Figure 2.3.1.

For Birth-Death, an increase in the reproduction rate of the group that the focal individual is in increases the death rates of the two neighbouring groups, and reduces the death rates

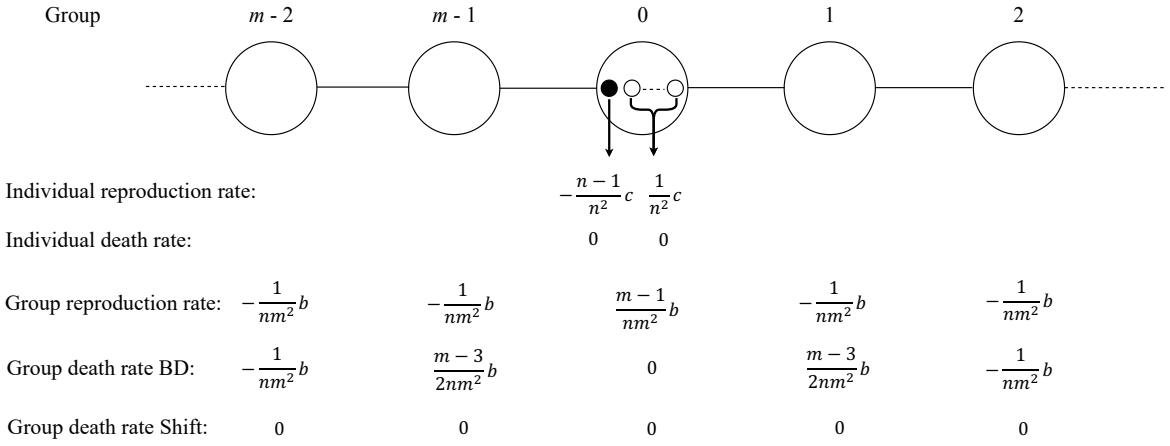


Figure 2.3.1: An overview of all the fitness effects in the limit of weak selection, conditional on an individual event happening in the group of the focal individual, or a group event happening, respectively. The black dot represents the focal individual.

of all other groups. For Shift, all groups have a probability $\frac{1}{m}$ of dying, so changes in group reproduction rates do not affect any group's death rate.

In Section 3 of the Supporting Information we derive and discuss these effects in detail. In the limit of $w \downarrow 0$, we also add them up, weighted by the relatedness of the individuals affected. The relatedness between two individuals whose groups are i steps apart is defined as the low mutation limit of

$$r_i = \frac{q_i - \bar{q}}{1 - \bar{q}}$$

where q_i denotes the stationary identical-by-descent probability for the two individuals, and $\bar{q}$ denotes the average identical-by-descent probability of a focal individual to all the individuals in the population, including the focal individual itself (see Section 4 in the Supporting Information for further details). This implies that these relatednesses are relative measures, that are positive for individuals in close by groups and negative for individuals in far away groups, and that they sum up to 0.

For the Birth-Death process we then find that cooperators are selected for if

$$-p\frac{1}{m}(1 - r_0)\frac{1}{n}c + qn\left(r_0 - \left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)\right)\frac{1}{nm}b > 0 \qquad (2.1)$$

The first term reflects all changes in individual reproduction rates. The probability that

an individual event happens is p. The probability that if it does, it happens in the group of the focal individual is $\frac{1}{m}$. If we write the effect on the individual reproduction rate of the focal individual as $-\frac{1}{n}c + \frac{1}{n^2}c$, then we can also see the individual effects as a combination of a reduction in individual reproduction rate of the focal individual by $\frac{1}{n}c$, and an increase in individual reproduction rate of $\frac{1}{n^2}c$ for everyone in the group, including the focal individual. With n individuals per group, the latter is equivalent to an effect of $\frac{1}{n}c$ on a randomly chosen individual from the same group, including the focal individual. This randomly chosen individual is related r_0 to the focal individual.

The term $qnr_0\frac{1}{nm}b$ reflects the effects through changes in group reproduction rates. The probability that a group event happens is q, and if it does, all n individuals in the group reproduce. If we write the effect on the group reproduction rate of the focal individual as $\frac{1}{nm}b - \frac{1}{nm^2}b$, then we can also see the group effects as a combination of an increase in reproduction rate of the group the focal individual is in by $\frac{1}{nm}b$, and a decrease in reproduction rate of $\frac{1}{nm^2}b$ for all groups, including the group the focal individual is in. The latter is equivalent to an effect of $\frac{1}{nm}b$ on a randomly chosen group, including the group the focal individual is in. A randomly chosen individual from a randomly chosen group is related $\sum_{i=0}^{m-1} r_i = 0$ to the focal individual.

The term $-qn\left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)\frac{1}{nm}b$ reflects the effects through changes in group death rates. This matches the group replacement rule for Birth-Death, where a reproducing group replaces itself with probability $\frac{1}{m}$, and one of its neighbouring groups with probability $\frac{m-1}{m}$. A randomly chosen individual from the neighbouring groups is related r_1 to the focal individual.

For Shift, almost everything is the same, and the only thing that is different is that all group death rates are unaffected. That makes the counterpart of Condition (2.6) simpler.

$$-p\frac{1}{m}(1-r_0)\frac{1}{n}c + qnr_0\frac{1}{nm}b > 0 \tag{2.2}$$

There are two differences between these two conditions. The first is that r_0 will not the same between the two processes, even if everything else (that is: p, q, r, n and m) is equal. In the Supporting Information we calculate how r_0 depends on those five parameters for both processes, and it turns out that r_0 tends to be higher for Birth-Death than for Shift. Therefore, if this was the only difference, it would actually be easier to evolve cooperation in Birth-Death than it would be for Shift. The second difference is that Condition (2.6) has a $-\left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)$ term that is absent in Condition (2.8). This term reflects the cancellation effect, and it makes the evolution of cooperation harder.

We can rewrite both inequalities as conditions on the b/c-ratio. Condition (2.6) for Birth-Death then becomes

$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{n\left(r_0 - \left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)\right)} \tag{2.3}$$

Condition (2.8) for Shift becomes

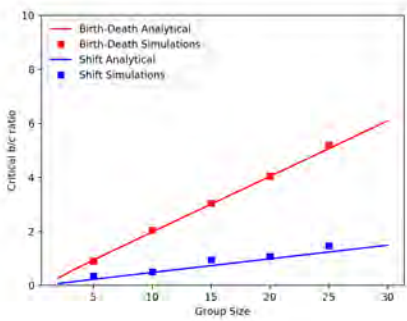
$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{nr_0} \tag{2.4}$$

One can also consider a more general class of processes that are the same as Birth-Death and Shift with respect to their individual reproduction, but that vary in how local between-group competition is. For simplicity, we can assume for all processes that if a group is chosen to reproduce, then the parent group itself is chosen to die with probability $\phi_0 = \frac{1}{m}$. For the remainder of the probabilities $\phi_i, i = 1, ..., m-1$ we only assume symmetry ($\phi_j = \phi_{m-j}$) and, since they are probabilities, $\sum_{i=0}^{m-1}\phi_i = 1$. Groups between the reproducing group and the dying group then move over in the same way as they do in Shift. If we do, we find a more general condition that encompasses Conditions (2.10) and (2.4).

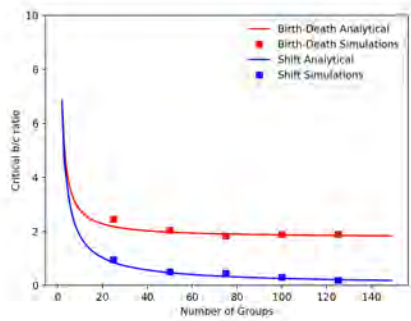
$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{n\left(r_0 - \sum_{i=0}^{m-1}\phi_i r_i\right)} \tag{2.5}$$

Birth-Death is a special case of this larger collection of models with $\phi_0 = \frac{1}{m}$, $\phi_1 = \phi_{m-1} = \frac{m-1}{2m}$ and $\phi_i = 0$, for $i = 2, ..., m-2$. To get Condition (2.10) for Birth-Death, we use $r_1 = r_{m-1}$. Shift is a special case with $\phi_i = \frac{1}{m}$ for all i, and to get Condition (2.4), we use $\sum_{i=0}^{m-1} r_i = 0$. It should be noted that the relatednesses in Condition (5) are still endogenous; they depend on the process we choose.

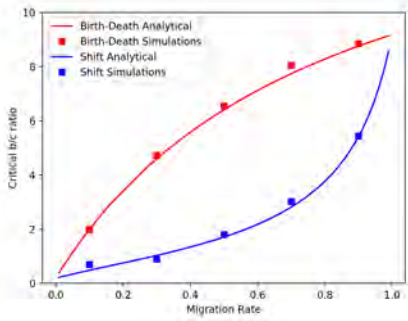
In Section 4 of the Supporting Information, we derive analytical expressions for relatednesses for Birth-Death and Shift by extending the method from Grafen (2007) to group structured populations. That gives us relatednesses at neutrality, which is appropriate in the case of weak selection. Figure 2.3.2 shows how the critical b/c-ratios in Conditions (2.10) and (2.4) depend on the group size, the number of groups, and the migration rate, if we fill in those relatednesses.



(a) $m = 50, r = 0.1$



(b) $n = 10, r = 0.1$



(c) $m = 50, n = 10$

Figure 2.3.2: Critical b/c ratios in the limit of weak selection for Birth-Death (red lines) and Shift (blue lines), as well as simulation results at an intensity of selection $w = 0.1$, for Birth-Death (red squares) and Shift (blue squares). In panels **(a)** and **(b)**, one in every ten events is a migration event ($r = 0.1$). In panels **(a)** and **(c)**, the number of groups is set to $m = 50$. In panels **(b)** and **(c)**, the group size is set to $n = 10$. Probabilities p and q are chosen so that the average individual is as likely to die as a result of an individual reproduction event as it is to die from a group reproduction event under neutral selection: $p = (1 - r)\frac{n}{n+1}$ and $q = (1 - r)\frac{1}{n+1}$. Similar to a model without population structure at the group level (Traulsen and Nowak, 2006), larger group sizes **(a)** and larger migration rates **(c)** increase the critical b/c ratio, and larger numbers of groups **(b)** decrease it. For $m \to \infty$, the threshold for Shift goes to 0, because r_0 then converges to 1. The gap between Birth-Death and Shift is there for a range of group sizes, numbers of groups, and migration rates. The gap between the two processes disappears when the migration rate vanishes, in which case the dynamics are such that all groups are at within-group fixation almost all of the time (see Sections 2 and 8 of the Supporting Information for why that makes the gap disappear). The gap also disappears when the migration rate is close to 1, and the whole population is shaken and stirred between any two reproduction events. Section 7 of the Supporting Information gives the complete argument why relatednesses r_0 and r_1 being close to 0 not only means that the right hand sides of Conditions (2.10) and (2.4) should be similar, but that they actually are exactly the same in the limit of $r \to 1$.

Once we move away from the limit of weak selection, the model quickly becomes intractable. To study the model not in the limit of weak selection, we ran simulations. The critical b/c-ratios we find for an intensity of selection of $w = 0.1$ are also shown in Figure 2.3.2. For intensity of selection $w = 0.5$ they can be found in the Supporting Information. These simulations show a similar gap between Birth-Death and Shift, and they suggest that the analytical results in the limit of weak selection are quite informative here (Wu et al., 2013). Section 5 in the Supporting Information also contains a mathematical proof that the threshold for Birth-Death is always higher than the one for Shift as soon as the number of groups exceeds 3. For $m = 2$ or $m = 3$, the two different update processes imply the same dynamic.

2.4 Relation to other models

We chose our model in order to make it as simple as possible to illustrate the difference between global and local between-group competition. We expect that the cancellation effect at the group level will show up in all models with local between-group competition. Other models that allow the scale of between-group competition to vary may however not allow for such relatively straightforward comparisons. We will go over a few other models that one could also combine with ways to model local between-group competition.

In Luo (2014) and van Veelen et al. (2014), the rate at which individuals reproduce is an individual characteristic, which is independent of the population state. It can either be high or low, depending on whether an individual is a defector or a cooperator. Also the reproduction rate of a group only depends on the number of cooperators in the group itself; it will be high if there are many, and low if there are few. This implies that the ratio of group events to individual events depends on the population state; if everyone is a cooperator, this ratio will be lower than if everyone is a defector. In our model, the probabilities p and q are fixed, and being a cooperator only has an effect on the individual reproduction rate, conditional on the group it is in being chosen to host an individual event. This is less realistic, and perhaps also less elegant, but it does make it easier to capture the effect of local between-group competition in relatively concise formulas.

In Traulsen and Nowak (2006), individual reproduction events make groups grow bigger, and when they reach maximum capacity, occasionally an individual reproduction event does not lead to another individual within the group dying, but to the group splitting into two daughter groups. Cooperators reduce their own reproduction rate, but increase the reproduction rate of others in their group. Individuals in groups with many cooper-

ators reproduce more often, and therefore they also make their groups split more often. In our model, group reproduction events are not triggered by individual reproduction events.

There are also differences in the methods used to derive analytical solutions. Traulsen and Nowak (2006) assume a separation of timescales, by considering the case where the probability that a group splits as a result of an individual reproduction event is vanishingly small. That results in a nested Moran process, for which they compute fixation probabilities in the limit of weak selection. That is different from our analysis. What is somewhat similar, is that the group reproduction stage ends up being condensed in both. In their case, it is the result of the separation of timescales. We simply assume that a group as a whole reproduces in one go, thereby bundling a sequence of individual reproduction events and a splitting event together. That is, again, not particularly realistic or elegant, but it does help avoid having to make other, perhaps more consequential unrealistic assumptions in order to be able to derive analytical solutions. Also, in Section 8.2 of the Supporting Information, we do take a somewhat similar approach by considering the limit of $p \to 1$, but without assuming selection to be weak. There, we find that the difference between Birth-Death and Shift disappears with the separation of timescales. The reason why it does is similar to the reason why it dissipates without migration, when the dynamics also make groups be at within-group fixation almost all of the time.

In order to find analytical solutions, Luo (2014); Simon (2010); Simon et al. (2013); van Veelen et al. (2014) all assume a dynamic equilibrium, where every individual group will keep changing composition, but in the equilibrium distribution of group types in the population as a whole, these changes balance. They moreover consider a limit where of both the number of groups and the group size approach infinity. It may be possible to create a version of their model, where groups are situated on the cycle as well, but their approach to deriving analytical solutions would not generalize in a straightforward way.

Our model, where groups replace other groups, would fall under Multi-Level Selection 2 in the classification of Okasha (2006), under "old group selection" in terms of West et al. (2007), or "replacement group selection" in terms of Molleman et al. (2013). This is not necessarily an unrealistic possibility; see Soltis et al. (1995). If, instead of replacing other groups more often, successful groups produce more offspring, which then migrate to other groups, then that would classify as Multi-Level Selection 1, "new group selection", or "contagion group selection". Such a model would fit Rousset and Billiard (2000), who present a model with localized dispersal on a cycle, but without group level events. They

do not interpret theirs as a group selection model, nor do they discuss the cancellation effect, but in their analysis, relatednesses with individuals in neighbouring demes do play a similar role. Both their and our model can also be seen as examples of metapopulation models (Hanski, 1998; 1999).

2.5 Discussion and implications for empirical studies

Our results show that in models of group selection, the evolution of cooperation can be quite a bit harder if between-group competition is local instead of global. The difference in critical b/c-ratios can be more than substantial between Birth-Death, which has completely local between-group competition, and Shift, for which between-group competition is completely global. The particular structure we considered - the cycle - is obviously very simple, and not particularly realistic. It may represent some populations, if they are constrained by geographic characteristics, such as rivers, or chains of mountains. For example, Howell (1952) has noted that the Shilluk, a Nilotic tribe, are organized into divisions of settlements situated along the west bank of the Nile in a linear fashion. For this population, a 1-dimensional model is a good approximation. For most populations, however, the cycle is not a good model. We do nonetheless think that its simplicity allows us to demonstrate a more general effect, which we expect will also occur with more realistic and complex ways in which groups can be located in a higher-dimensional spatial structure.

It is probably less unrealistic to assume that between-group competition is at least to some extent local. Straightforward examples of local between-group competition are warfare in the Enga society, which happens within the same ethnic group (Wiessner, 2019; Wiessner et al., 2010; Wiessner and Pupu, 2012), endemic warfare in the Asabano society in the pre-contact era (Lohmann, 2014), or feuds among the Shilluk settlements mentioned above, which usually take place among direct neighbours (Howell, 1952). Examples for which one can reasonably assume that between-group competition is global, on the other hand, will be much harder to find.

This also has empirical implications. The current, well-established approach in empirical studies concerning group selection is to measure F_{ST}'s - the empirical equivalent of r_0 in our model - in order to determine how large the benefit to the group should be, compared to the cost to the individual, for cooperation to evolve. The condition that Bell et al. (2009) uses for when cooperation will be selected for by group selection (see also Aoki and

Nozawa (1984); Bowles (2006; 2009); Crow and Aoki (1984); Langergraber et al. (2011); Rusch (2018); Walker (2014); Weir and Cockerham (1984)) is

$$\frac{\beta(w_g, p_g)}{\beta(w_{ig}, p_{ig})} > \frac{1 - F_{ST}}{F_{ST}}$$

Here $\beta(w_g, p_g)$ is the increase in mean fitness of the group as a result of an increase in the frequency of cooperators, or altruists, and $\beta(w_{ig}, p_{ig})$ is the decrease in fitness of an individual as a result of switching from defection to cooperation. The idea is that this criterion separates the fitness effects, on the left hand side of the inequality, from a measure that characterizes the population structure, on the right hand side of the inequality. In a setting where the fitness effects constitute a linear public goods game, played within groups that compete with each other globally, such a separation can indeed be made in this way (see Section 9 of the Supporting Information, and van Veelen (2020). A small, collateral finding here is that in such a setting, one should compute the F_{ST} without, and not with replacement, as is usually done.)

If we were to measure $\beta(w_g, p_g)$ in a setting in which competition between groups is not actually global, but to some degree local, then the resulting value for $\beta(w_g, p_g)$ would not only reflect the effect of cooperators on the average fitness within the group, but a mixture of these fitness effects and the cancellation effect. A moderate value for $\beta(w_g, p_g)$ can both be the result of a moderate group benefit and the absence of the cancellation effect, and a high group benefit combined with the cancellation effect at the group level. In the latter case, the negative effect of having neighbouring groups with many cooperators, combined with the positive correlation between being a cooperative group and having neighbouring groups with many cooperators, would bias the estimated effect of – all else equal – the number of cooperators on average fitness within the group downwards. In other words, this term would end up absorbing the cancellation effect. In order to disentangle all fitness effects and the cancellation effect, one would have to estimate a more complex statistical model, which would not only use the composition of the own group as an explanatory variable of the average fitness within the group, but also include the composition of neighbouring groups as an explanatory variable. This would then have to be combined, not just with the relatedness within groups, but also with the relatedness with individuals in neighbouring groups.

What most empirical papers do, however, is only estimate the F_{ST}, which is then taken as an indication of how conducive the population structure is to cooperation. The implicit assumption in that approach is that competition between groups is global. We have seen

that the absence or presence of the cancellation effect – which is part of the population structure – can make a huge difference for how much the group needs to benefit from cooperators in it, relative to the individual costs, in order for cooperation to spread in the population. If competition between groups is not global, but at least to some extent local, then this procedure therefore paints a too positive picture of how favourable conditions are for the evolution of cooperation by group selection.

Appendix

2A.1 The cancellation effect at the individual level

The cancellation effect at the individual level was discovered by Wilson et al. (1992) and Taylor (1992a;b). Before then, it was more or less generally thought that as soon as interacting individuals are related, there is scope for cooperation to evolve (see Hamilton, 1971; Boyd, 1982; Grafen, 1983; and other references in Wilson et al., 1992; Taylor, 1992a; and Taylor, 1992b, for exceptions). The idea was that positive relatedness means that cooperators are more likely to interact with cooperators than defectors are, which implies that, while cooperators pay the cost of cooperating, they will also be on the receiving end of cooperation more often. If they are indeed sufficiently much more often the recipients of cooperation than defectors are, and if the benefits are sufficiently large, then the cost of cooperation can be offset by the increase in benefits received.

What Wilson et al. (1992) and Taylor (1992a;b) discovered is that being around cooperators is not necessarily unambiguously good news. While being around more cooperators means receiving more cooperation – which is good – it can also mean being around individuals that cooperate more with one another, and therefore constitute more fierce competition – which is not good. For cases in which relatedness is caused by local dispersal, receiving more cooperation and facing more intense competition can go hand in hand, and therefore, if individuals have opportunities for cooperating that are as local as their competition is, no benefit is large enough to get costly cooperation to evolve. The main insight provided by Wilson et al. (1992) and Taylor (1992a;b) is therefore that it is not enough to be related; what is needed is a discrepancy between the relatednesses to the individuals with whom one has the opportunity to cooperate, and those with whom one has to compete. While the idea that positive relatedness alone would be enough for the evolution of cooperation was inspired by Hamilton's rule (Hamilton, 1964a;b), it was pointed out that, by defining the fitness effects of cooperation versus defection appropriately, the cancellation effect can actually be identified within the framework of Hamilton's rule; see Allen et al. (2012); Grafen (2007); Taylor (1992b), and Section 7 in van Veelen et al. (2017b), all in settings where the game between individuals satisfies equal gains from switching.[2]

[2]Ohtsuki (2012) analyzes a model that does not satisfy equal gains from switching. Other papers on the scale of cooperation versus the scale of competition are Queller (1992; 1994) and West et al. (2002), but they take a different approach. While Allen et al. (2012); Grafen (2007); Taylor (1992b) and van Veelen et al. (2017b) include the cancellation effect by making sure to account for all fitness effects, and keeping the definition of relatedness the same, Queller (1992; 1994) and West et al. (2002) get Hamilton's

2A.2 The cancellation effect at the group level

Group selection models aim at capturing the opposing effects of selection at the individual level, where defectors do better than cooperators within groups, and selection at the group level, where groups with more cooperators do better than groups with fewer cooperators (Richerson et al., 2016; Sober and Wilson, 1998; Wilson and Wilson, 2007). The existing models within the group selection literature share the property that competition between individuals happens within groups, and that competition between groups happens in a setting where all groups compete with all other groups equally intensely (Luo, 2014; Luo and Mattingly, 2017; Simon, 2010; Simon et al., 2013; Traulsen and Nowak, 2006; van Veelen et al., 2014). This last property is a useful simplification if the aim is to illustrate the possibility of a tug of war between the different levels of selection. It is, however, not particularly realistic. Groups themselves typically live in a structured population of groups, where they compete with their neighbouring groups more often than they do with groups that are farther apart. Local dispersal would then imply that groups with many cooperators are typically surrounded by groups that also contain many cooperators, compared to the groups that surround groups with many defectors, and therefore are also subject to more intense competition. This can significantly dampen the benefits of being a cooperative group, which in turn affects the balance between selection at the individual and at the group level.

In order to study the cancellation effect at the group level, and how it affects the balance between selection at different levels, we consider a stylized model, in which groups live on a cycle. We look at two replacement rules for groups. The first replacement rule is Birth-Death, where groups replace their direct neighbours. The second replacement rule is Shift, where a group at one position can reproduce, and a group anywhere else can die, and all groups in between the two just move over. With Birth-Death, groups compete with their direct neighbours, and the cancellation effect at the group level is the largest it can be. With Shift, there is no cancellation at the group level at all.

The rest of the Supporting Information is organized as follows. In Section 2A.3, we describe the model, including both replacement rules. Analytical thresholds for the benefit-to-cost ratios in the limit of weak selection, both for Birth-Death and Shift, are derived in Section 2A.4. These thresholds are expressed in terms of the parameters of the model and relatednesses. The relatednesses are endogenous themselves and also depend on the parameters of the model, so we compute these in Section 2A.5. In Section 2A.6, the

rule to hold by changing the relatedness into *effective relatedness*. When we will do computations to include the cancellation effect at the group level, we will follow the first approach.

benefit-to-cost ratios and the relatednesses are combined and it is shown that the critical benefit-to-cost ratio for the Birth-Death process is always higher than the critical benefit-to-cost ratio for the Shift process, making group selection models that assume away the cancellation effect more optimistic about the conditions for the evolution of cooperation. Section 2A.7 describes the simulations. The simulation results, which are not in the limit of weak selection, are compared to the thresholds in the limit of weak selection in Section 2A.8. In Section 2A.9, we derive analytical results for some limits other than the limit of weak selection, and in Section 2A.10 we discuss the empirical implications.

2A.3 The model

2A.3.1 Two update rules on the cycle

We consider a simple model, in which groups are situated on a cycle. Both the group size and the number of groups are fixed; at each point in time, there are m groups consisting of n individuals. Each individual can be either a cooperator (C) or a defector (D). In every time period, one of three types of events will happen: an individual can replace another individual within a group, a group can replace another group, or two individuals from two neighbouring groups can change places. These individual, group, and migration events happen with probabilities p, q and r, respectively, and, without loss of generality, we assume that $p + q + r = 1$.

In order to demonstrate the difference between global and local between-group competition, we compare two different replacement rules for group reproduction: Birth-Death (BD) and Shift. Competition between groups is local in BD, and global in Shift. Individual and migration events happen in the same way in both processes.

2A.3.2 Individual events

If an individual event occurs – which happens with probability p – then a random group is selected, and within that group, an individual is chosen to produce an identical offspring. All groups have equal probability of being chosen to host an individual level event. Within the group, defectors get an individual payoff of 1 and cooperators get an individual payoff of $1 - c$. The intensity of selection w is then used to transform these payoffs to values f_C and f_D:

$$f_C = 1 - w + w\,(1 - c) = 1 - wc \qquad \text{and} \qquad f_D = 1 - w + w\,(1) = 1$$

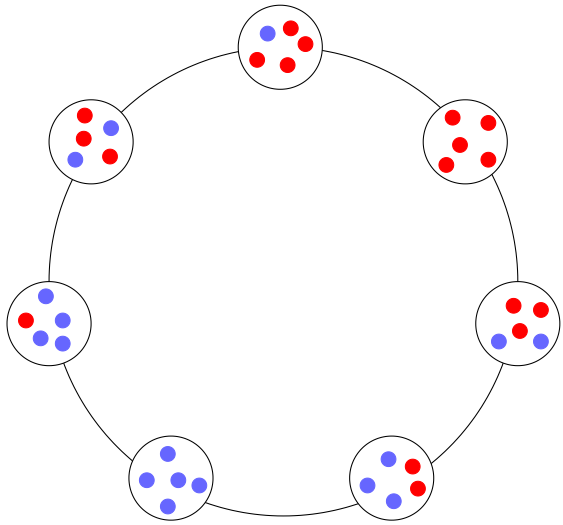


Figure 2A.1: An example of a population state on a cycle with $m = 7$ groups of $n = 5$ individuals each. The blue dots indicate cooperators and the red dots indicate defectors. Each group has one right and one left neighbour.

The probabilities with which individuals are chosen for reproduction within the group are proportional to these values.[3] The probability $p_C(k_i)$ that a cooperator is chosen, and the probability $p_D(k_i)$ that a defector is chosen, in a group with k_i cooperators, then become:

$$p_C(k_i) = \frac{k_i f_C}{k_i f_C + (n - k_i) f_D} = \frac{k_i(1 - wc)}{n - k_i wc}$$
$$p_D(k_i) = \frac{(n - k_i) f_D}{k_i f_C + (n - k_i) f_D} = \frac{n - k_i}{n - k_i wc} = 1 - p_C(k_i)$$

Whenever an individual reproduces, someone from the same group is chosen to die, where each individual, including the parent, but excluding the offspring, is chosen with probability $\frac{1}{n}$.

If an individual event happens, the number of cooperators in that group can go up or down by one, or remain constant, depending on who reproduces and who dies within the group. Once a group reaches a state where all individuals are cooperators or all individuals are defectors, individual selection can not change the state of that group. Individual selection on its own, within a given group, therefore constitutes a Markov process with two absorbing

[3] One could call these values "fitnesses", in line with Nowak (2006). We chose to not use that term here, and reserve the word fitness, and fitness effects, for (effects on) expected numbers of offspring.

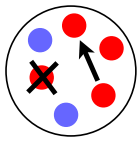


Figure 2A.2: An example of the individual reproduction events. A defector is chosen for reproduction and produces an identical offspring. Another defector is chosen to die. In this case, the overall group composition has not changed.

states.

2A.3.3 Group events

If a group event occurs – which happens with probability q – then one group is chosen to reproduce, and one group is chosen to die. Which group reproduces depends on the distribution of cooperators among groups. If the groups are numbered from 1 to m, then a population state is a vector k, in which $k_i \in \{0, 1, ..., n\}$ is the number of cooperators in group i, for $i = 1, ..., m$. Groups i and $i + 1$ are neighbouring groups, for $i = 1, ..., m - 1$, as well as groups 1 and m, which makes this a cycle. The group payoff of group i is $1 + \frac{k_i}{n}b$. The intensity of selection w is then used to transform these payoffs to values

$$g(k_i) = 1 - w + w\left(1 + \frac{k_i}{n}b\right) = 1 + w\frac{k_i}{n}b$$

Both in BD and in Shift, first a group is chosen for reproduction, where each group's probability of being chosen is proportional to their value $g(k_i)$ as given below.

$$q(k_i, K) = \frac{g(k_i)}{\sum_{j=1}^{m} g(k_j)} = \frac{1 + w\frac{k_i}{n}b}{n + w\frac{K}{n}b}$$

where $K = \sum_{j=1}^{m} k_j$ is the total number of cooperators in the population. If a group is chosen for reproduction, it produces an identical offspring group. Which group is being replaced depends on which replacement rule is used at the group level.

Birth-Death (BD) The offspring group replaces its own parent group with probability $\frac{1}{m}$, and it replaces either the left or the right neighbour of the parent group, both with probability $\frac{m-1}{2m}$. If the offspring group replaces the parent group, the population state does not change. The possibility to replace the parent group is included in order to make the analytical comparison between the two processes more straightforward. This does not have any profound consequences; any process with that possibility is equivalent to a process

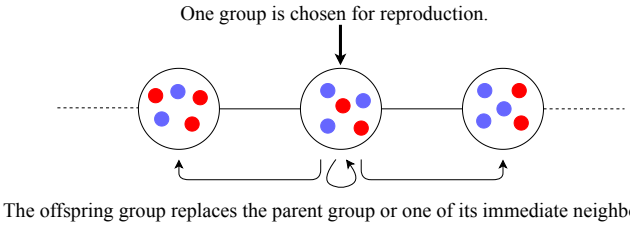


Figure 2A.3: The BD process. One of the groups is chosen to reproduce, proportional to group values, and produces an identical offspring group. The offspring group replaces one of the neighbouring groups, or, with a small probability, it replaces the parent group itself.

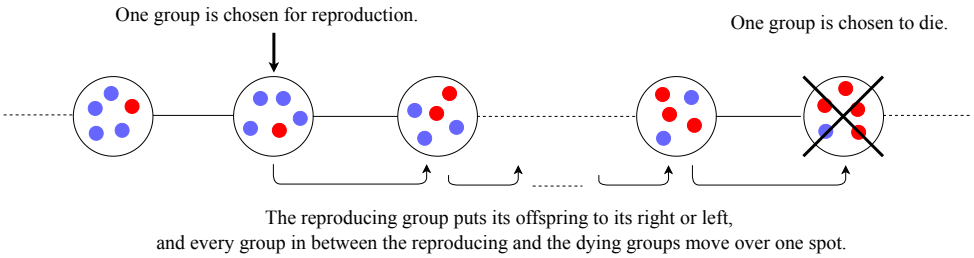


Figure 2A.4: The Shift process. One of the groups is chosen to reproduce, proportional to group values, and produces an identical offspring group. The offspring replaces one of the groups chosen randomly from the whole pool of groups.

without that possibility, and with a lower probability of having a group reproduction event at all.

Shift Each group, including the parent group, but excluding the offspring group, is chosen to die with probability $\frac{1}{m}$. Hence, with Shift, every group is equally likely to die and competition between groups is therefore global, as it is in the standard group selection models with a well-mixed population of groups. Once a group dies, the offspring group is either placed to the right or to the left of the parent group, with equal probability, and every other group between the parent group and the dying group moves over one spot.

Group selection by itself also constitutes a Markov process with multiple absorbing states. Once all groups have the same fraction of cooperators, the population state can no longer change by group level events. Therefore, the states in which all groups have the same composition would be the absorbing states of a Markov process with group level events

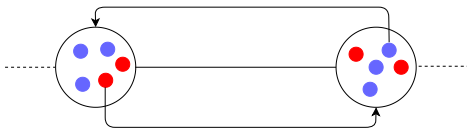


Figure 2A.5: An example of a migration event, where a defector from the group on the left and a cooperator from the group on the right change places

only.

2A.3.4 Migration events

If a migration event occurs – which happens with probability r – then a random group is selected, and within that group, a random individual is chosen to migrate. All groups have equal probability of being chosen to host a migration event, and within the group, all individuals are equally likely to be chosen to become the migrant. Then a coin toss determines whether the individual moves to the group on the left or to the group on the right. Within the receiving group, a randomly chosen individual trades places with the first individual.

Migration makes a difference for which sets of states are absorbing. Without migration, the set that consists of all population states in which some groups consist of cooperators only, other groups consist of defectors only, but no group is mixed, is absorbing. Group events can still make a population transition from one population state within this set to another population state within this set, but no group or individual level event can make the population transition from a state within this set to one outside it. With migration, on the other hand, this is not an absorbing set of population states. When there is migration, the only sets of states that are absorbing are the set that only contains the state where everyone is a cooperator, and the set that only contains the state where everyone is a defector. Besides these two singleton sets, no sets of states are absorbing if we include migration.

This observation is important for understanding why, as we will see later, very low migration rates will reduce the gap between the critical b/c ratios for the two processes. Once absorbed within the set where all groups are homogeneous, the dynamics for the different replacement rules are only different in speed, while the differences in fixation probabilities disappear.

2A.3.5 Alternative modeling choices

There are places in the model where, with the literature on group selection in mind, one can easily think of other ways to determine who reproduces and who is replaced. The reason for the choices we made is typically that they allow us to derive relatively tractable analytical solutions, which help illustrate the cancellation effect. Alternative choices would produce the same gap between full cancellation at the group level (BD) and no cancellation at the group level (Shift), but some would come with (much) more complicated analytical solutions.

2A.3.6 Luo (2014)

In Luo (2014) and van Veelen et al. (2014), the reproduction rate of an individual is an individual characteristic, where defectors have a higher individual reproduction rate than cooperators do. That implies that a group with more defectors is more likely to host an individual reproduction event than a group with fewer defectors. In our model, every group is equally likely to host an individual reproduction event. Another implication of the choices in Luo (2014) and van Veelen et al. (2014) is that there, the ratio of group level reproduction events to individual level reproduction events depends on how many cooperators there are in the population as a whole; with many cooperators, group level reproduction events happen more frequently, relative to individual level reproduction events, compared to a population with fewer cooperators. In our model, this ratio is $\frac{p}{q}$, which is constant.

The modeling choices in Luo (2014), and similar ones in Simon (2010) and Simon et al. (2013), produce a model that is in some respects more elegant than ours. When combined with a structured population of groups like the cycle, and with different replacement rules for group reproduction, the cancellation effect would be present and absent in their model in the same way as it is in our version. While our model may be a bit less elegant, it does allow for a more straightforward derivation of the formulas for the critical b/c ratios. (Just to be sure: groups in Luo (2014) and van Veelen et al. (2014) compete globally, making this otherwise equivalent to the Shift model without migration).

A final difference, not in the model itself, but in the approach to deriving analytical solutions, is that Luo (2014); Simon (2010); Simon et al. (2013), and van Veelen et al. (2014) all take limits of group size n and number of groups m going to infinity, while we derive analytical solutions in the limit of weak selection.

2A.3.7 Traulsen & Nowak (2006)

In Traulsen and Nowak (2006), individuals within groups affect their own and each other's individual reproduction rates. Cooperators lower their own individual reproduction rate, and increase the individual reproduction rates of their fellow group members. When a group is at maximimum capacity, there is a small probability that at an individual reproduction event, the offspring does not replace another group member, but makes the group split. Because individuals in all-cooperator groups reproduce more frequently than individuals in all-defector groups, all-cooperator groups also split more often, and therefore produce more offspring groups. This is a bit further removed from our model, but it would of course be possible to make versions of this model where groups are situated on a cycle, and, in case a group splits, it either replaces a neighbouring group (BD), or a randomly chosen group (Shift).

There are also differences in the methods used to derive analytical solutions. One reason not to use an extended version of their model is that their method for finding analytical solutions makes assumptions that make it hard to identify the cancellation effect at the group level by comparing Birth-Death and Shift. Traulsen and Nowak (2006) assume a separation of timescales, by considering the case where the probability that a group splits as a result of an individual reproduction event is vanishingly small. This implies that almost all of the time, groups will be at maximum capacity, and will consist of one type of individual only. The separation of timescales results in a nested Moran process, for which they compute fixation probabilities in the limit of weak selection. In Section 2A.9.5 we take a similar approach by considering the limit of $p \to 1$, but without assuming selection to be weak. There, we find that the difference between Birth-Death and Shift disappears with the separation of timescales. The reason why it does is similar to the reason why it does without migration, when the dynamics also make groups be at within-group fixation almost all of the time. This is described in Section 2A.8.

Traulsen and Nowak (2006) also have a version with (global) migration. In the limit they consider, migration events happen with a frequency that is of the same order of magnitude as splitting events. This leaves the separation of timescales intact, which implies that groups will still either be all-cooperator or all-defector groups most of the time. For identifying a difference between Birth-Death and Shift, we would need migration to occur sufficiently frequently to keep at least some groups away from within group fixation. Although their model does not allow for a straightforward extension to a structured population of groups, in which we can easily identify the cancellation effect at the group level, we do find a few consistent patterns. The number of groups m, group size n, and migration

rate r all affect the critical b/c ratio in our model in ways that are similar to how they affect that ratio in their model.

2A.3.8 Public goods game

Although the game in our model does have the key properties of a public goods game, where paying individual costs come with collective benefits, one can also define payoffs so that the game looks like a public goods game already at the individual level. Such a version of the model would leave all model assumptions unchanged, except for the individual and group payoff functions. The values for a cooperator and a defector in group i, in which there are k_i cooperators, then become:

$$f_C^{PGG}(k_i) = 1 - w + w\left(1 + \frac{k_i}{n}b - c\right) = 1 + w\left(\frac{k_i}{n}b - c\right)$$

and

$$f_D^{PGG}(k_i) = 1 - w + w\left(1 + \frac{k_i}{n}b\right) = 1 + w\left(\frac{k_i}{n}b\right)$$

where w is the intensity of selection. The value for the group is defined as:

$$g(k_i)^{PGG} = 1 - w + w\left(1 + \frac{k_i}{n}(b - c)\right) = 1 + w\left(\frac{k_i}{n}(b - c)\right)$$

where w is the intensity of selection. We assume that $b > c$.

In this formulation, the payoffs seem to make this a public goods game, already at the individual level, as all individuals get a higher payoff when all individuals play C (when they all get $b - c$), compared to what all individuals get when all play D (when they all get 0). In our setting, this is however not really an improvement for all at the individual level. The probability of a group being chosen to host an individual event is $\frac{1}{m}$, and this probability is therefore independent of the population state. When all individuals have the same payoff, whether it is all large or all small, their probability of reproducing, conditional on their group being chosen to host the individual reproduction event, is $\frac{1}{n}$. These payoffs therefore only seemingly introduce the public goods nature at the individual level. In alternative settings, where groups with many cooperators also host more individual reproduction events than groups with many defectors, the reproduction rates of everyone will be higher if everyone is a cooperator rather than everyone being a defector, but now this is compensated by an also elevated death rate, again neutralizing all "gains from cooperation" at the individual reproduction level. The only difference between the formulations is how the change in individual reproduction rate depends on the total num-

ber of cooperators in the group. In the formulation we chose, being a cooperator and not a defector decreases ones individual reproduction rate by $\frac{1}{n-wk_ic} - \frac{1-wc}{n-w(k_i+1)c}$, if k_i is the number of cooperators among the other members of the group. In this "PGG" formulation, that difference is $\frac{1+w\frac{k_i}{n}b}{n+wk_i(b-c)} - \frac{1+w\left(\frac{k_i+1}{n}b-c\right)}{n+w(k_i+1)(b-c)}$. In other words, in our formulation, the reduction in individual reproduction rate gets a bit larger when more others cooperate; while in this alternative PGG formulation, the reduction in individual reproduction rate gets a bit smaller.

This alternative formulation does furthermore link individual and group values, by making the latter the average of the former. The gains in group reproduction rate relate in a more straightforward way; the $b - c$ in the PGG version replaces the b in our version.

Given that the effective differences between these two versions of the model are only minor details, it is not surprising that simulations also give very similar results. For obtaining analytic results, however, our version is easier to work with.

2A.4 Analytical results in the limit of weak selection

We would like to derive critical b/c ratios, above which cooperation is selected for, in the limit of weak selection. In order to do that, we will go over the effects of being a cooperator instead of a defector on reproduction rates and death rates. The b and the c are just model parameters, so they are not the fitness benefits and fitness costs of cooperation. The effects that we compute below do amount to those fitness benefits and costs. Because these effects satisfy equal gains from switching locally (they are additive in the limit of weak selection), we can follow an inclusive fitness approach, where these effects are weighted with the relatednesses to the individual that the effects are on, in order to determine the direction of selection (van Veelen et al., 2017a;b). We will use this approach to derive critical b/c ratios, which are therefore formulated in terms of model parameters. In the following subsections, we consider a case where one individual switches from being a defector to being a cooperator, and we calculate the effects of this change by the focal individual on every individual in the population, weighted by the corresponding relatednesses.

Since the probabilities concerning reproduction events, both at the individual level and at the group level, are the same in BD and in Shift, the effects of being a cooperator instead of a defector on reproduction rates will be the same for both. Moreover, in both processes,

the individual death rate is the same. The difference between the two processes, therefore, will only start showing up when we compute the effects on group death rates.

2A.4.1 Changes in reproduction rates

Changes in individual reproduction rates

Changes in my reproduction rate If I am a C instead of a D, with i other C players in my group, and my group is chosen for an individual update, I change my probability of being chosen for reproduction from $\frac{1}{n-iwc}$ to $\frac{1-wc}{n-(i+1)wc}$, which is a difference of $\frac{1-wc}{n-(i+1)wc} - \frac{1}{n-iwc}$. If we take the derivative with respect to w for both terms, we get:

$$\frac{d\left(\frac{1-wc}{n-(i+1)wc}\right)}{dw} = -\frac{n-(i+1)}{(n-(i+1)wc)^2}\cdot c \qquad \text{and} \qquad \frac{d\left(\frac{1}{n-iwc}\right)}{dw} = \frac{i}{(n-iwc)^2}\cdot c$$

Evaluated at $w = 0$, we get:

$$\left.\frac{d\left(\frac{1-wc}{n-(i+1)wc}\right)}{dw}\right|_{w=0} = -\frac{n-(i+1)}{n^2}\cdot c \qquad \text{and} \qquad \left.\frac{d\left(\frac{1}{n-iwc}\right)}{dw}\right|_{w=0} = \frac{i}{n^2}\cdot c$$

which means that the change in probability of being chosen for reproduction, for w close to 0 (weak selection), can be approximated by

$$\frac{1-wc}{n-(i+1)wc} - \frac{1}{n-iwc} \approx -\frac{n-1}{n^2}c\cdot w$$

Changes in the reproduction rate of a C in my group If I am a C instead of a D, with i other C players in my group, and my group is chosen for an individual update, I change other C's individual probability of being chosen for reproduction from $\frac{1-wc}{n-iwc}$ to $\frac{1-wc}{n-(i+1)wc}$, which is a difference of $\frac{1-wc}{n-(i+1)wc} - \frac{1-wc}{n-iwc}$. If we take the derivative with respect to w, we get:

$$\frac{d\left(\frac{1-wc}{n-(i+1)wc}\right)}{dw} = -\frac{n-(i+1)}{(n-(i+1)wc)^2}\cdot c \qquad \text{and} \qquad \frac{d\left(\frac{1-wc}{n-iwc}\right)}{dw} = -\frac{n-i}{(n-iwc)^2}\cdot c$$

Evaluated at $w = 0$, we get:

$$\left.\frac{d\left(\frac{1-wc}{n-(i+1)wc}\right)}{dw}\right|_{w=0} = -\frac{n-(i+1)}{n^2}\cdot c \qquad \text{and} \qquad \left.\frac{d\left(\frac{1-wc}{n-iwc}\right)}{dw}\right|_{w=0} = -\frac{n-i}{n^2}\cdot c$$

which means that the change in their probability of being chosen for reproduction, for w close to 0 (weak selection), can be approximated by

$$\frac{1-wc}{n-(i+1)wc} - \frac{1-wc}{n-iwc} \approx \frac{1}{n^2}c \cdot w$$

Changes in the reproduction rate of a D in my group If I am a C instead of a D, with i other C players in my group, and my group is chosen for an individual update, I change other D's individual chance of being chosen for reproduction from $\frac{1}{n-iwc}$ to $\frac{1}{n-(i+1)wc}$, which is a difference of $\frac{1}{n-(i+1)wc} - \frac{1}{n-iwc}$. If we take the derivative with respect to w, we get:

$$\frac{d\left(\frac{1}{n-(i+1)wc}\right)}{dw} = \frac{i+1}{(n-(i+1)wc)^2} \cdot c \qquad \text{and} \qquad \frac{d\left(\frac{1}{n-iwc}\right)}{dw} = \frac{i}{(n-iwc)^2} \cdot c$$

Evaluated at $w = 0$, we get:

$$\left.\frac{d\left(\frac{1}{n-(i+1)wc}\right)}{dw}\right|_{w=0} = \frac{i+1}{n^2} \cdot c \qquad \text{and} \qquad \left.\frac{d\left(\frac{1}{n-iwc}\right)}{dw}\right|_{w=0} = \frac{i}{n^2} \cdot c$$

which means that the change in their probability of being chosen for reproduction, for w close to 0 (weak selection), can be approximated by

$$\frac{1}{n-(i+1)wc} - \frac{1}{n-iwc} \approx \frac{1}{n^2}c \cdot w$$

The effects on individual reproduction rates should be multiplied by $p\frac{1}{m}$ in order to account for the probability with which an individual event happens, and that my group is chosen to host it. The changes in individual reproduction rates within the group add up to 0, as they should with a fixed group size. There are no effects on individual reproduction rates in other groups.

Changes in group reproduction rates

Changes in my group's reproduction rate If I am a C instead of a D, with i other C players in my group, $i = 0, 1, ..., n-1$, and j other C's in the population as a whole, $j = i, i+1, ..., i+n(m-1)$, then if a group level event happens, I change my group's probability of being chosen for reproduction from $\frac{1+w\frac{i}{n}b}{m+w\frac{j}{n}b}$ to $\frac{1+w\frac{i+1}{n}b}{m+w\frac{j+1}{n}b}$, which is a difference

of $\frac{1+w\frac{i+1}{n}b}{m+w\frac{j+1}{n}b} - \frac{1+w\frac{i}{n}b}{m+w\frac{j}{n}b}$. If we take the derivative of both terms with respect to w, we get:

$$\frac{d\left(\frac{1+w\frac{i+1}{n}b}{m+w\frac{j+1}{n}b}\right)}{dw} = \frac{\frac{m(i+1)-(j+1)}{n}\cdot b}{\left(m+w\frac{j+1}{n}b\right)^2} \qquad \text{and} \qquad \frac{d\left(\frac{1+w\frac{i}{n}b}{m+w\frac{j}{n}b}\right)}{dw} = \frac{\frac{mi-j}{n}\cdot b}{\left(m+w\frac{j}{n}b\right)^2}$$

Evaluated at $w = 0$, we get:

$$\left.\frac{d\left(\frac{1+w\frac{i+1}{n}b}{m+w\frac{j+1}{n}b}\right)}{dw}\right|_{w=0} = \frac{m(i+1)-(j+1)}{nm^2}\cdot b \qquad \text{and} \qquad \left.\frac{d\left(\frac{1+w\frac{i}{n}b}{m+w\frac{j}{n}b}\right)}{dw}\right|_{w=0} = \frac{mi-j}{nm^2}\cdot b$$

which means that the change in my group's probability of being chosen for reproduction, for w close to 0 (weak selection), can be approximated by

$$\frac{1+w\frac{i+1}{n}b}{m+w\frac{j+1}{n}b} - \frac{1+w\frac{i}{n}b}{m+w\frac{j}{n}b} \approx \frac{m-1}{nm^2}b\cdot w$$

Changes in other groups' reproduction rates If I am a C instead of a D, with i other C players in my group, $i = 0, 1, ..., n-1$, and j other C's in the population as a whole, $j = i, i+1, ..., i+n(m-1)$, then if a group level event happens, I change the probability of being chosen for reproduction of a random other group with k cooperators from $\frac{1+w\frac{k}{n}b}{m+w\frac{j}{n}b}$ to $\frac{1+w\frac{k}{n}b}{m+w\frac{j+1}{n}b}$, which is a difference of $\frac{1+w\frac{k}{n}b}{m+w\frac{j+1}{n}b} - \frac{1+w\frac{k}{n}b}{m+w\frac{j}{n}b}$. If we take the derivative with respect to w, we get:

$$\frac{d\left(\frac{1+w\frac{k}{n}b}{m+w\frac{j+1}{n}b}\right)}{dw} = \frac{\frac{mk-(j+1)}{n}\cdot b}{\left(m+w\frac{j+1}{n}b\right)^2} \qquad \text{and} \qquad \frac{d\left(\frac{1+w\frac{k}{n}b}{m+w\frac{j}{n}b}\right)}{dw} = \frac{\frac{mk-j}{n}\cdot b}{\left(m+w\frac{j}{n}b\right)^2}$$

Evaluated at $w = 0$, we get:

$$\left.\frac{d\left(\frac{1+w\frac{k}{n}b}{m+w\frac{j+1}{n}b}\right)}{dw}\right|_{w=0} = \frac{mk-(j+1)}{nm^2}\cdot b \qquad \text{and} \qquad \left.\frac{d\left(\frac{1+w\frac{k}{n}b}{m+w\frac{j}{n}b}\right)}{dw}\right|_{w=0} = \frac{mk-j}{nm^2}\cdot b$$

which means that the change in that other group's probability of being chosen for reproduction, for w close to 0 (weak selection), can be approximated by

$$\frac{1+w\frac{k}{n}b}{m+w\frac{j+1}{n}b} - \frac{1+w\frac{k}{n}b}{m+w\frac{j}{n}b} \approx -\frac{1}{nm^2}b\cdot w$$

The effects on group reproduction rates should be multiplied by q in order to account for the probability with which a group event happens. The changes in group reproduction rates add up to 0, as they should with a fixed number of groups.

Overall effect through changes in reproduction rates The combined effects, close to neutrality, all weighted with the corresponding relatednesses, are:

$$\frac{p}{m}\left(-\frac{n-1}{n^2}+(n-1)r_s\frac{1}{n^2}\right)c\cdot w+q\left(\frac{m-1}{nm^2}+(n-1)r_s\frac{m-1}{nm^2}-n(m-1)r_o\frac{1}{nm^2}\right)b\cdot w$$

where r_s is the relatedness between an individual and a randomly chosen other individual from the same group, and r_o is the relatedness between an individual and a randomly chosen individual from a randomly chosen other group.

In Section 2A.5, we derive identities concerning different relatednesses. We can use one of those to rewrite the combined effects. Equation (2.44) states that $r_o=-\frac{1}{n(m-1)}-\frac{n-1}{n(m-1)}r_s$, and using that, we can see rewrite the combined effects as

$$-\frac{p}{m}\frac{n-1}{n^2}(1-r_s)c\cdot w+q\frac{1}{nm}\left(1+(n-1)r_s\right)b\cdot w$$

Equation (2.13) moreover states that $r_0=\frac{1}{n}+\frac{n-1}{n}r_s$, and therefore that $\frac{n-1}{n}(1-r_s)=1-r_0$ and $1+(n-1)q_s=nq_0$, where r_0 is the relatedness between an individual and a randomly chosen individual from the same group – which is labeled group 0 – including the individual itself. The only difference with r_s is that there the individual itself is excluded, hence the simple relation between r_0 and r_s. We can use this to rewrite the combined effects as

$$-\frac{p}{m}\frac{1}{n}(1-r_0)c\cdot w+q\frac{1}{nm}\left(nr_0\right)b\cdot w$$

2A.4.2 Changes in death rates

Changes in individual death rates Individual death rates do not change when a player switches from being D to C.

Changes in group death rates Changes in group death rates differ between the two update processes at the group level.

Birth-Death

Changes in my group's death rate If I am a C instead of a D, with i other C players in my group, $i=0,1,...,n-1$, and j other C's in the population as a whole, $j=i,i+1,...,i+n(m-1)$, then, if a group level event happens, we have seen that I

change the reproduction rate of my own group by approximately $\frac{m-1}{nm^2}b \cdot w$. Since with probability $\frac{1}{m}$ the offspring group replaces the parent group, that comes with an increase in death rate equal to

$$\frac{1}{m}\frac{m-1}{nm^2}b \cdot w$$

We have also seen that I change the reproduction rate of all other groups by approximately $-\frac{1}{nm^2}b \cdot w$, including my two neighbouring groups. If one of those is chosen for reproduction, it replaces my group with probability $\frac{m-1}{2m}$, reducing my death rate by

$$2 \cdot \frac{1}{2}\frac{m-1}{m}\frac{1}{nm^2}b \cdot w$$

These two cancel out exactly, so the overall effect on the death rate of my group is 0, as my neighbours are now less likely to be chosen for reproduction and replace my group, but my group is more likely to reproduce and replace itself.

Changes in the death rate of the two neighbouring groups Following a similar argument, we find that the probability that one of my next-door neighbour groups is replaced changes by:

$$\frac{1}{2}\frac{m-1}{m}\frac{m-1}{nm^2}b \cdot w - \frac{1}{2}\frac{m-1}{m}\frac{1}{nm^2}b \cdot w - \frac{1}{m}\frac{1}{nm^2}b \cdot w$$

The first term is the product of the probability that, if my group is chosen for reproduction, it replaces a given neighbour, and the increase in my groups probability to reproduce. The second term is the product of the probability of the neighbouring group of the neighbouring group to replace the neighbouring group, if chosen for reproduction, and the decrease in their probability of being chosen for reproduction. The third term is the product of the probability of the neighbouring group to replace itself, if chosen for reproduction, and the decrease in their probability of being chosen for reproduction. This sum can be rewritten as:

$$\left(\frac{1}{2}(m-1)(m-2) - 1\right)\frac{1}{nm^3}b \cdot w = \frac{1}{2}(m-3)\frac{1}{nm^2}b \cdot w$$

This effect is the same for both right and left next-door neighbours.

Changes in the death rates of other groups The probability of being replaced changes by:

$$2 \cdot -\frac{1}{2}\frac{m-1}{m}\frac{1}{nm^2}b \cdot w - \frac{1}{m}\frac{1}{nm^2}b \cdot w = -\frac{1}{nm^2}b \cdot w$$

for any of the $m-3$ other groups. The first term on the left hand side is twice the product

of the probability that a given neighbouring group replaces a given group, times the change in reproduction probability of those neighbouring groups. The second term is the product of the probability a group replaces its parent group, when chosen for reproduction, and the change in reproduction probability of such a group.

The effects on group reproduction rates should be multiplied by q in order to account for the probability with which a group event happens.

Shift

Changes in group death rates For the Shift process, group death rates do not change when a player switches from being D to C.

Overall effect through changes in death rates

Birth-Death The overall effect through death rates again combines the effects on individual death rates (which are zero) and the effects on group death rates affecting self, group members, and members of other groups, each weighted with the corresponding relatedness measure:

$$\begin{aligned}
&\frac{p}{m}\cdot 0 + q\left(0-(n-1)r_s\cdot 0 + nr_1\frac{1}{2}\frac{m-3}{nm^2}b\cdot w - n\left(\sum_{i=2}^{m-2} r_i\right)\frac{1}{nm^2}b\cdot w + nr_{m-1}\frac{1}{2}\frac{m-3}{nm^2}b\cdot w\right)\\
&\quad = q\left(nr_1\frac{1}{2}\frac{m-3}{nm^2} + nr_1\frac{1}{nm^2} - n\left(\sum_{i=1}^{m-1} r_i\right)\frac{1}{nm^2} + nr_{m-1}\frac{1}{nm^2} + nr_{m-1}\frac{1}{2}\frac{m-3}{nm^2}\right)b\cdot w\\
&\quad = q\left(nr_1\frac{m-3}{nm^2} + 2nr_1\frac{1}{nm^2} - n\left(\sum_{i=1}^{m-1} r_i\right)\frac{1}{nm^2}\right)b\cdot w\\
&\quad = q\left(nr_1\frac{m-1}{nm^2} + (1+(n-1)r_s)\frac{1}{nm^2}\right)b\cdot w\\
&\quad = q\frac{1}{nm^2}\left(n(m-1)r_1 + 1 + (n-1)r_s\right)b\cdot w\\
&\quad = q\frac{1}{nm^2}\left(n(m-1)r_1 + nr_0\right)b\cdot w
\end{aligned}$$

where we used the identity $r_1 = r_{m-1}$, and Equations (2.42), and (2.13) from Section 2A.5.

Shift For Shift, the overall effect through changes in death rates is zero, since the effects on death rates of both individuals and groups are zero.

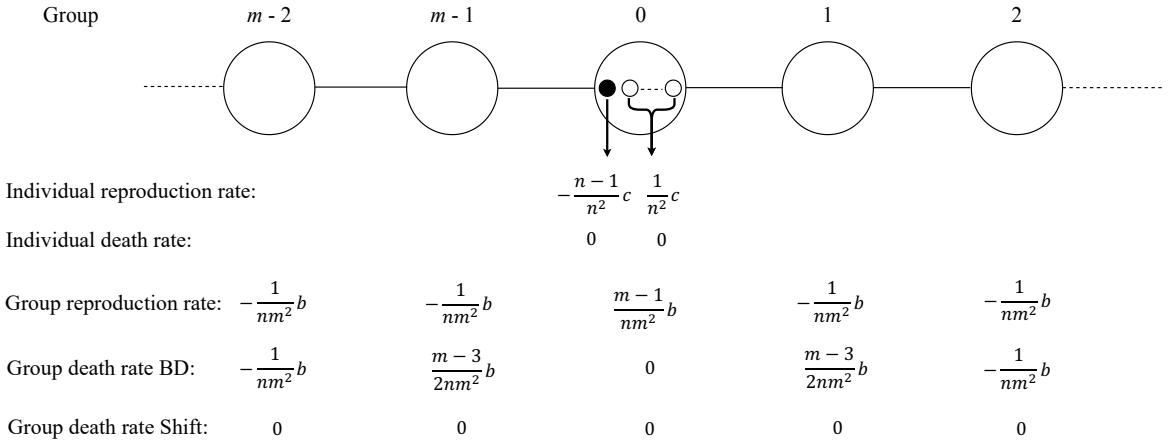


Figure 2A.1: An overview of all the fitness effects in the limit of weak selection, conditional on an individual event happening in the group of the focal individual, or a group event happening, respectively. The black dot represents the focal individual.

2A.4.3 Overall effect of switching from D to C

The overall effect of a player switching from playing D to C would be the effect on reproduction rates minus the effect on death rates.

Birth-Death The overall effect for the BD process is

$$\begin{aligned}
&-p\frac{1}{mn}(1-r_0)c\cdot w + q\frac{m}{nm^2}\left(nr_0\right)b\cdot w - q\frac{1}{nm^2}\left(n(m-1)r_1 + nr_0\right)b\cdot w \\
&\quad = -p\frac{1}{mn}(1-r_0)c\cdot w + q\frac{1}{nm^2}((m-1)(nr_0) - (m-1)nr_1)b\cdot w \\
&\quad = -p\frac{1}{mn}(1-r_0)c\cdot w + q\frac{m-1}{nm^2}(nr_0 - nr_1)b\cdot w
\end{aligned}$$

This effect is positive if

$$-p\frac{1}{m}(1-r_0)\frac{1}{n}c + q\frac{m-1}{m}(nr_0 - nr_1)\frac{1}{nm}b > 0$$

This gathers the effects as summarized in Figure 2A.1. The probability that an individual event happens is p. The probability that if it does, it happens in the group of the focal individual is $\frac{1}{m}$. If we write the effect on the individual reproduction rate of the focal individual as $-\frac{1}{n}c + \frac{1}{n^2}c$, then we can also see the individual effects as a combination of a

reduction in individual reproduction rate of the focal individual by $\frac{1}{n}c$, and an increase in individual reproduction rate of $\frac{1}{n^2}c$ for everyone in the group, including the focal individual. With n individuals per group, the latter is equivalent to an effect of $\frac{1}{n}c$ on a randomly chosen individual from the same group, including the focal individual. This randomly chosen individual is related r_0 to the focal individual.

The probability that a group event happens is q. For all groups other than the focal group and its direct neighbours, the group reproduction rate and the group death rate go down by the same amount. The difference between the change in group reproduction rate and the change in the group death rate is $\frac{m-1}{nm^2}b$ for the focal group, and $-\frac{1}{2}\frac{m-1}{nm^2}b$ for both neighbouring groups, which adds up to $-\frac{m-1}{nm^2}b$. In the group of the focal individual, there are n individuals, who on average are related r_0 to the focal individual, and also in the neighbouring group there are n individuals, and those are related r_1 to the focal individual.

One way to rewrite this inequality would be

$$-p\frac{1}{m}\left(1-r_0\right)\frac{1}{n}c+qn\left(r_0-\left(\frac{1}{m}r_0+\frac{m-1}{m}r_1\right)\right)\frac{1}{nm}b>0 \tag{2.6}$$

The term $qnr_0\frac{1}{nm}b$ now reflects the effects through changes in group reproduction rates, while the term $-qn\left(\frac{1}{m}r_0+\frac{m-1}{m}r_1\right)\frac{1}{nm}b$ reflects the effects through changes in group death rates. This matches the group replacement rule for Birth-Death, where a reproducing group replaces itself with probability $\frac{1}{m}$, and one of its neighbouring groups with probability $\frac{m-1}{m}$.

Another way to rewrite this condition would be

$$q\frac{m-1}{m}n(r_0-r_1)b>p(1-r_0)c$$

which implies that, in the BD process, cooperation is selected for if

$$\frac{b}{c}>\frac{p}{q}\frac{m}{m-1}\frac{1-r_0}{n(r_0-r_1)} \tag{2.7}$$

Shift Since the effect on death rates is zero, the overall effect would be equal to the effect on reproduction rates:

$$-p\frac{1}{mn}(1-r_0)c\cdot w+q\frac{1}{nm}\left(nr_0\right)b\cdot w$$

This effect is positive if

$$-p\frac{1}{m}(1-r_0)\frac{1}{n}c + qnr_0\frac{1}{nm}b > 0 \tag{2.8}$$

This gathers the effects for Shift, which are also summarized in Figure 2A.1. The effects on individual reproduction rates are the same as for Birth-Death. On the group level, there are now only changes in reproduction rates and no changes in death rates. The changes in group reproduction rates can be seen as a combination of an increase in group reproduction rate of the group of the focal individual by $\frac{1}{nm}c$, and a decrease in group reproduction rate of $\frac{1}{nm^2}c$ for every group, including the group of the focal individual. The latter is equivalent to an effect of $\frac{1}{nm}c$ on a randomly chosen group, including the group of the focal individual. A randomly chosen individual from the same group, including the focal individual itself, is related r_0 to the focal individual, and a randomly chosen individual from a randomly chosen group, including the group of he focal individual, is related $\sum_{i=0}^{m-1}\frac{1}{n}r_i = 0$ to the focal individual.

If we simplify this condition, we get

$$qnr_0b > p(1-r_0)c$$

which implies that, in the Shift process, cooperation is selected for if

$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{nr_0} \tag{2.9}$$

2A.4.4 Comparing thresholds

There are two differences between these two thresholds. To pinpoint the first, we can write Condition (2.7), which gives the threshold for Birth-Death, as

$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{n\left(r_0 - \left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)\right)}$$

This condition has has a $-\left(\frac{1}{m}r_0 + \frac{m-1}{m}r_1\right)$ term that is absent in Condition (2.9). This term reflects the cancellation effect, and it pushes the threshold up. The second difference is that r_0 will not the same across the two processes, even if everything else (that is: p, q, r, n and m) is equal. In Section 2A.5, we calculate how r_s, and thereby also r_0, depends on those five parameters for both processes, and it turns out that r_0 is higher for BD than for Shift. Therefore, if it was not for the first difference, the critical b/c ratio would

actually be *lower* for BD than for Shift.

2A.4.5 Hamilton's rule

We derived these thresholds using inclusive fitness. Therefore, it is worth pointing out that the b and the c in those thresholds are model parameters, and that they are *not* the fitness benefits and fitness costs of cooperation. It would therefore not be correct to read these formulas themselves as versions of Hamilton's rule, where the b and the c would represent the fitness benefits and costs, and the right hand side would replace $\frac{1}{r}$ (Nowak et al., 2010; Ohtsuki et al., 2006). These parameters do however determine the size of the fitness effects, as we have seen in the derivations that preceded Conditions (2.7) and (2.9).

2A.4.6 More general version and interpretation

The processes we consider are Birth-Death, with completely local between-group competition, and Shift, with completely global between-group competition. One can however also consider a more general class of processes that are the same as Birth-Death and Shift with respect to their individual reproduction, but that vary in how local between-group competition is. To keep them comparable, we can assume for all processes that if a group is chosen to reproduce, then the parent group itself is chosen to die with probability $\phi_0 = \frac{1}{m}$. The remainder of the probabilities can be chosen freely, and are given by $\phi_i, i = 1, ..., m-1$, where we do assume symmetry ($\phi_j = \phi_{m-j}$) and, since they are probabilities, $\sum_{i=0}^{m-1} \phi_i = 1$. Groups between the reproducing group and the dying group then move over in the same way as they do in Shift. **Changes in group death rates** For such a process, we can write the effect through the change in death rate of group i as a sum of the change in its death rate as a result of the focal group reproducing, and the changes as a result of all other groups reproducing:

$$\frac{m-1}{nm^2} b \cdot \phi_i \cdot w + \sum_{j=1}^{m-1} -\frac{1}{nm^2} b \cdot \phi_{|i-j|} \cdot w$$

This can be rewritten as

$$\left(\frac{1}{nm}-\frac{1}{nm^2}\right)b\cdot\phi_i\cdot w+\sum_{j=1}^{m-1}-\frac{1}{nm^2}b\cdot\phi_{|i-j|}\cdot w$$
$$=\frac{1}{nm}b\cdot\phi_i\cdot w+\sum_{j=0}^{m-1}-\frac{1}{nm^2}b\cdot\phi_{|i-j|}\cdot w$$
$$=\frac{1}{nm}b\cdot\phi_i\cdot w-\frac{1}{nm^2}b\cdot w$$

Here, we use $\sum_{j=0}^{m-1}\phi_{|i-j|}=\sum_{j=0}^{m-1}\phi_j=1$.

Overall effect of switching from D to C The overall effect then becomes

$$-p\frac{1}{mn}(1-r_0)c\cdot w+q\frac{m}{nm^2}(nr_0)\,b\cdot w-q\cdot n\sum_{i=0}^{m-1}\left(\frac{1}{nm}b\cdot\phi_i\cdot w-\frac{1}{nm^2}b\cdot w\right)r_i$$
$$=-p\frac{1}{mn}(1-r_0)c\cdot w+q\frac{1}{nm}\left(nr_0-n\sum_{i=0}^{m-1}\phi_i r_i\right)b\cdot w$$

Here we use $\sum_{i=0}^{m-1}r_i=0$. This is positive i

$$-p\frac{1}{m}(1-r_0)\frac{1}{n}c+q\left(nr_0-n\sum_{i=0}^{m-1}\phi_i r_i\right)\frac{1}{nm}b>0 \tag{2.10}$$

Birth-Death is a special case of this larger collection with $\phi_0=\frac{1}{m}$, $\phi_1=\phi_{m-1}=\frac{m-1}{2m}$ and $\phi_i=0$, for $i=2,...,m-2$. To get Condition (2.6) for Birth-Death, we use $r_1=r_{m-1}$. Shift is a special case with $\phi_i=\frac{1}{m}$ for all i, and to get Condition (2.8), we use $\sum_{i=0}^{m-1}r_i=0$.

In the descriptions of how Conditions (2.6) and (2.8) summarize the fitness effects, we have seen that the first term in both of them, which is also the first term here, summarizes the effects of changes in individual reproduction rates. We have also seen how the first half of the second term, $qnr_0\frac{1}{nm}b$, summarizes the effects of changes in group reproduction rates. The second half of the second term, $-qn\sum_{i=0}^{m-1}\phi_i r_i\frac{1}{nm}b$, finally, summarizes the effect of changes in the group death rates. This term therefore captures the cancellation effect at the group level for these models.

We can also write the condition as

$$\frac{b}{c}>\frac{p}{q}\frac{1-r_0}{n\left(r_0-\sum_{i=0}^{m-1}\phi_i r_i\right)} \tag{2.11}$$

It should be noted though that the relatednesses are endogenous; they also depend on the process. Since we focus on Birth-Death and Shift, we only compute relatednesses for those.

The whole of Section 2A.4.6 follows a very helpful suggestion by one of the reviewers. Also the use of Equation (2.13) earlier on, and the focus on and interpretation of Conditions (2.6) and (2.8) follow suggestions of this reviewer.

2A.5 Relatedness

In this section, we will calculate relatednesses r_i between individuals that are i groups apart, both for Birth-Death and for Shift, in the limit of weak selection, using identical-by-descent (IBD) probabilities. Two individuals are considered IBD if they descend from a common ancestor, and no mutations have occurred along their lineage. We will first derive the relatedness measures by assuming a mutation probability u per individual reproduction (and, at a group reproduction event, all individuals in the group reproduce with the same mutation probability), and then take the limit of $u \downarrow 0$ to find the no-mutation limit of relatedness measures. To do so, we will first derive the recurrence relations for IBD probabilities by assuming a stationary distribution $\{q_i\}$ and then derive the no-mutation limit for relatedness using the following identity (Durrett, 2008; Grafen, 2007; Malécot, 1948; Rousset, 2004; Taylor et al., 2007a;b)

$$r_i = \frac{q_i - \bar{q}}{1 - \bar{q}} \tag{2.12}$$

where q_i denotes the stationary IBD probability for two individuals whose groups are i steps apart and $\bar{q}$ denotes the average IBD probability of a focal individual to all the individuals in the population, including self.

Two observations will be useful in our later derivations. The first is that, by symmetry, $q_i = q_{m-i}$ for $0 \leq i \leq m$. The second is that, for $i = 0$, we can relate q_0 – the IBD probability for two members from the same group, drawn with replacement – and q_s – the same probability without replacement – in a straightforward way:

$$q_0 = \frac{1 + (n-1)q_s}{n}. \tag{2.13}$$

since the individual's relatedness to herself is 1.

2A.5.1 Birth-Death

In the BD process, an individual can be replaced if:

- An individual event happens, the group the individual is in is chosen to host it, and within the group, she is chosen to be replaced.
- A group event happens, one of the neighbouring groups of the group the individual is in is chosen to reproduce, and that group replaces her group.
- A group event happens, the group that the individual belongs to is chosen, and replaces itself.
- A migration event occurs, the individual's group is chosen to be one of the groups in which the migration takes place, and the individual is chosen to be swapped.

Combining these events, we can derive the recurrence relations for the stationary IBD probabilities.

$i = 1$ For $i = 1$, we have the following recurrence relation:

$$q_1 = \underbrace{p\frac{2}{m}\frac{1}{n}(1-u)q_1}_{\substack{\text{replaced in an}\\ \text{individual event}}} + \underbrace{q\left(\frac{1}{m}\frac{m-1}{m}(1-u)(q_0+q_2)+\frac{2}{m^2}(1-u)q_1\right)}_{\text{replaced in a group event}} + \underbrace{r\frac{1}{m}\frac{1}{n}\left(2q_2+\frac{2(n-1)}{n}q_s+\frac{1}{n}q_1\right)}_{\text{replaced in a migration event}}$$
$$+\left(\underbrace{p\left(\frac{m-2}{m}+\frac{2}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through}\\ \text{individual events}}} + \underbrace{q\left(\frac{m-4}{m}+\frac{m+1}{m^2}+\frac{m-1}{m^2}\right)}_{\text{no change through group events}} + \underbrace{r\left(\frac{m-3}{m}+\frac{2}{m}\frac{n-1}{n}+\frac{1}{m}\frac{(n-1)^2}{n^2}\right)}_{\text{no change through migration events}}\right)q_1$$

Here we use the probabilities with which different replacement events happen. We will sometimes refer to the two individuals in two neighbouring groups as the two focal individuals, and to their groups as the two focal groups.

If an individual event happens, the probability that any given group is chosen to host it is $\frac{1}{m}$. The probability that any individual within that group is chosen to be replaced is $\frac{1}{n}$. Therefore, for two given individuals in neighbouring groups, that adds up to $\frac{2}{m}\frac{1}{n}$. There are a few ways in which they can both not be affected. One of the other $m-2$ groups can be chosen, which happens with probability $\frac{m-2}{m}$; or one of the two neighbouring groups can be chosen, while some other individual is replaced, which happens with probability $\frac{2}{m}\frac{n-1}{n}$.

If a group event happens, then one of the two neighbouring groups can replace the other –

which happens with probability $\frac{1}{m}\frac{m-1}{m}\frac{1}{2}$ – or the other can replace the one, which happens with the same probability. In both cases, q_0 is the relevant IBD probability. Also, a neighbour outside the focal pair of groups can replace one of the two in the focal pair, which again happens with a probability that is twice $\frac{1}{m}\frac{m-1}{m}\frac{1}{2}$. In this case, q_2 is the relevant IBD probability. Finally, with a probability of twice $\frac{1}{m^2}$, one of the focal groups replaces itself, in which case, q_1 is the relevant IBD probability. Nothing happens to the pair if a group is chosen to reproduce that is neither of the two groups within the focal pair, nor one of their direct neighbours, which happens with probability $\frac{m-4}{m}$. Also nothing happens if one of the neighbours of the focal pair is chosen, and they replace themselves ($\frac{1}{m^2}$) or their other neighbour ($\frac{1}{m}\frac{m-1}{m}\frac{1}{2}$). These probabilities together times 2 is $\frac{1}{m}\frac{m+1}{m^2}$. Finally, nothing happens if one of the two focal groups is chosen, and replaces a group outside the focal pair, which happens with probability $\frac{1}{m}\frac{m-1}{m^2}$.

With a migration event, every neighbouring pair is doing an exchange with probability $\frac{1}{m}$. This pair consists of both focal groups with probability $\frac{1}{m}$, and it is a pair that consists of one of the two focal groups and its neighbour on the other side with probability $\frac{2}{m}$. In the last case, the focal individual is chosen with probability $\frac{1}{n}$, and q_2 is the relevant IBD probability. In case the exchange is between the focal pair of groups, the two focal individuals themselves are chosen to switch with probability $\frac{1}{n^2}$, one of them is swapped with an individual that is not the other with probability $\frac{n-1}{n^2}$, and the other is swapped with an individual that is not the one with the same probability. In the first case, nothing changes with regard to the IBD probabilities, and in the latter two cases, q_s is the relevant IBD probability. Nothing happens at migration if one of the other $m-3$ pairs is chosen ($\frac{m-3}{m}$), a focal group and a neighbouring group outside the focal pair is chosen, but the focal individual is not chosen ($\frac{2}{m}\frac{n-1}{n}$), or the focal pair itself is chosen, but two individuals other than the focal ones trade places ($\frac{1}{m}\frac{(n-1)^2}{n^2}$).

If we gather the terms with q_1 on the left-hand side, we get

$$\begin{aligned}\left(1-p\left(\frac{2}{m}\frac{1}{n}(1-u)+\frac{m-2}{m}+\frac{2}{m}\frac{n-1}{n}\right)-q\left(\frac{2}{m^2}(1-u)+\frac{m-2}{m}\right)\right.\\ \left.-r\left(\frac{1}{m}\frac{1}{n^2}+\frac{m-3}{m}+\frac{2}{m}\frac{n-1}{n}+\frac{1}{m}\frac{(n-1)^2}{n^2}\right)\right)q_1\\ =q\frac{m-1}{m^2}(1-u)(q_0+q_2)+r\frac{2}{mn}\left(q_2+\frac{n-1}{n}q_s\right)\end{aligned}$$

If we use the identity $q_0=\frac{1}{n}+\frac{n-1}{n}q_s$, and multiply left and right by $\frac{m}{2}$, this can be rewritten as

$$\left(\frac{m}{2} - \frac{p}{n}(1-u) - p\left(\frac{m-2}{2} + \frac{n-1}{n}\right) - \frac{q}{m}(1-u) - q\left(\frac{m-2}{2}\right)\right.$$
$$\left. -r\left(\frac{1}{2}\frac{1}{n^2} + \frac{m-3}{2} + \frac{n-1}{n} + \frac{1}{2}\frac{(n-1)^2}{n^2}\right)\right) q_1$$
$$= \left(\frac{q}{2}\frac{m-1}{m}(1-u) + \frac{r}{n}\right)(q_0 + q_2) - r\frac{1}{n^2}$$

If we furthermore use that $p + q + r = 1$, this can be further simplified to

$$\left(1 - (1-u)\left(\frac{p}{n} + \frac{q}{m}\right) - p\frac{n-1}{n} - r\left(\frac{n-1}{n}\right)^2\right) q_1 = \left((1-u)\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(q_0 + q_2) - \frac{r}{n^2} \tag{2.14}$$

$1 < i < m-1$ The recurrence relations for $1 < i < m-1$ are derived in a similar way. The differences with the recurrence relation for $i = 1$ arise because the two groups no longer are each other's neighbours, which means that the groups can no longer replace each other, nor can they exchange individuals.

$$q_i = \underbrace{p\frac{2}{m}\frac{1}{n}(1-u)q_i}_{\substack{\text{replaced in an}\\ \text{individual event}}} + \underbrace{q\left(\frac{1}{m}\frac{m-1}{m}(1-u)(q_{i-1}+q_{i+1}) + \frac{2}{m^2}(1-u)q_i\right)}_{\text{replaced in a group event}} + \underbrace{r\frac{2}{m}\frac{1}{n}(q_{i-1}+q_{i+1})}_{\substack{\text{replaced in a}\\ \text{migration event}}}$$
$$+ \left(\underbrace{p\left(\frac{m-2}{m} + \frac{2}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through}\\ \text{individual events}}} + \underbrace{q\left(\frac{m-6}{m} + 2\frac{m+1}{m^2} + 2\frac{m-1}{m^2}\right)}_{\text{no change through group events}} + \underbrace{r\left(\frac{m-4}{m} + \frac{4}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through}\\ \text{migration events}}}\right) q_i$$

If we gather the terms with q_i on the left-hand side, we get:

$$\left(1 - p\left(\frac{2}{m}\frac{1}{n}(1-u) + \frac{m-2}{m} + \frac{2}{m}\frac{n-1}{n}\right) - q\left(\frac{2}{m^2}(1-u) + \frac{m-2}{m}\right) - r\left(\frac{m-4}{m} + \frac{4}{m}\frac{n-1}{n}\right)\right) q_i$$
$$= \left(q\frac{m-1}{m^2}(1-u) + r\frac{2}{mn}\right)(q_{i-1} + q_{i+1})$$

If we multiply left and right by $\frac{m}{2}$, and use $p + q + r = 1$, this can be rewritten as

$$\left(\frac{p}{n} + q + \frac{2r}{n} - (1-u)\left(\frac{p}{n} + \frac{q}{m}\right)\right) q_i = \left((1-u)\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(q_{i-1} + q_{i+1}) \tag{2.15}$$

$i = 0$ To derive q_s – and therewith q_0 – we will use the recurrence relation concerning the

IBD probabilities for two individuals within the same group.

$$q_s = p\frac{1}{m}\frac{2}{n^2}(1-u) + p\left(\frac{1}{m}\frac{2(n-1)}{n^2}(1-u) + \frac{1}{m}\frac{n-2}{n} + \frac{m-1}{m}\right)q_s + q\left(\frac{1}{m^2} + \frac{m-1}{m^2}\right)(1-u)^2 q_s$$
$$+q\left(\frac{m-3}{m} + \frac{m+1}{m^2} + \frac{m-1}{m^2}\right)q_s + r\frac{4}{mn}q_1 + r\left(\frac{m-2}{m} + \frac{2(n-2)}{mn}\right)q_s$$

The relevant probabilities in case of an individual event are $\frac{1}{m}\frac{2}{n^2}$ for the focal group being chosen and one focal individual replacing the other, or vice versa; $\frac{1}{m}\frac{2(n-1)}{n^2}$ for the focal group being chosen, and any individual other than the other focal one replacing one of the focal ones (including herself), or vice versa; $\frac{1}{m}\frac{n-2}{n}$ for the focal group being chosen, and one of the other individuals being replaced; and $\frac{m-1}{m}$ for a different group being chosen.

In case of a group event, the relevant probabilities are $\frac{1}{m^2}$ for the focal group replacing itself; $\frac{m-1}{m^2}$ for the focal group being replaced by a neighbouring group; $\frac{m-3}{m}$ for a group being chosen to reproduce that is not the focal group nor a neighbour; $\frac{m+1}{m^2}$ for a neighbouring group being chosen to reproduce, and replacing itself or it's other neighbour; and $\frac{m-1}{m^2}$ for the focal group being chosen to reproduce, and replacing a neighbouring group and not itself.

In case of a migration event, the relevant probabilities are $\frac{4}{mn}$ for the focal group to exchange a member with the neighbouring group on the right or on the left, and one of the two focal individuals being chosen; $\frac{2(n-2)}{mn}$ for the exchange happening between the focal group and any of the two neighbouring groups, and neither of the two focal individuals being replaced; and $\frac{m-2}{m}$ for the exchange happening in any of the other $m-2$ pairs.

The equation can be rewritten as

$$q_s = p\frac{1}{m}\frac{2}{n^2}(1-u) + p\left(1 - \frac{2}{mn^2} + \frac{2}{mn}\left(1-\frac{1}{n}\right)(-u)\right)q_s + q\left(\frac{1}{m}\right)(1-u)^2 q_s$$
$$+q\left(\frac{m-1}{m}\right)q_s + r\frac{4}{mn}q_1 + r\left(1 - \frac{4}{mn}\right)q_s$$

Expressing q_1 as a function of q_s, this becomes

$$q_1 = \frac{mn}{4r}\left(1 - p + p\frac{2}{mn}\left(u + \frac{1-u}{n}\right) - q\left(\frac{m-1}{m} + \frac{(1-u)^2}{m}\right) - r + r\frac{4}{mn}\right)q_s - \frac{p}{r}\frac{1-u}{2n}$$

Finally, we use $p + q + r = 1$ to write

$$q_1 = \frac{mn}{4r}\left(p\frac{2}{mn}\left(u + \frac{1-u}{n}\right) + q\frac{1}{m} - q\frac{(1-u)^2}{m} + r\frac{4}{mn}\right)q_s - \frac{p}{r}\frac{1-u}{2n} \tag{2.16}$$

Solving the system Suppose, for $1 \leq i \leq m-1$, q_i has the following form:

$$q_i = s_i q_1 \tag{2.17}$$

where $s_1 = s_{m-1} = 1$, and $\lim_{u\to 0} s_i = 1$ for $1 < i < m-1$, since all IBD probabilities approach 1 in the limit of no mutation. If we rewrite Equation (2.15) with the assumption in Equation (2.17), we get the following equality:

$$\left(\frac{p}{n} + q + \frac{2r}{n} - (1-u)\left(\frac{p}{n} + \frac{q}{m}\right)\right) s_i q_1 = \left((1-u)\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(s_{i-1} + s_{i+1}) q_1$$

Assuming that $q_1 \neq 0$, this is also:

$$\left(\frac{p}{n} + q + \frac{2r}{n} - (1-u)\left(\frac{p}{n} + \frac{q}{m}\right)\right) s_i = \left((1-u)\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(s_{i-1} + s_{i+1}) \tag{2.18}$$

Since $q_i \to 1$ for all i as $u \downarrow 0$, we cannot directly use Equation (2.12) to calculate relatedness, as both the numerator and the denominator approach zero. Therefore, we will apply L'Hôpital's rule and calculate relatednesses as:

$$r_i = \frac{q_i' - \bar{q}'}{-\bar{q}'} \tag{2.19}$$

Here the derivatives are taken with respect to u, and evaluated in the limit of $u \downarrow 0$. In order to determine q_i' and $\bar{q}'$, we want to find s_i' by taking derivatives on both sides of Equation (2.18).

$$\left(\frac{p}{n} + q + \frac{2r}{n} - (1-u)\left(\frac{p}{n} + \frac{q}{m}\right)\right) s_i' + \left(\frac{p}{n} + \frac{q}{m}\right) s_i$$
$$= \left((1-u)\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(s_{i-1}' + s_{i+1}') - \frac{q}{2}\frac{m-1}{m}(s_{i-1} + s_{i+1})$$

If we evaluate this in the limit of $u \downarrow 0$, then we can also use that $s_i = 1$ for all i in that limit.

$$\left(q\frac{m-1}{m} + \frac{2r}{n}\right) s_i' + \frac{p}{n} + \frac{q}{m} = \left(\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(s_{i-1}' + s_{i+1}') - q\frac{m-1}{m}$$

This can be reorganized as follows:

$$s_i' - \frac{s_{i-1}' + s_{i+1}'}{2} = -\frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}} \tag{2.20}$$

Summing both sides of the above equation for $2 \leq i \leq m-2$ yields

$$\sum_{i=2}^{m-2} s'_i - \frac{1}{2}\sum_{i=2}^{m-2}(s'_{i-1} + s'_{i+1}) = -\sum_{i=2}^{m-2} \frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}}$$

$$\sum_{i=2}^{m-2} s'_i - \frac{1}{2}(s'_1 + s'_{m-1} + s'_2 + s'_{m-2}) - \frac{1}{2}\sum_{i=3}^{m-3} 2s'_i = -(m-3)\frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}}$$

$$s'_2 + s'_{m-2} - \frac{1}{2}(s'_2 + s'_{m-2}) = -(m-3)\frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}}$$

$$s'_2 = -(m-3)\frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}} \tag{2.21}$$

where we used the identity $s'_1 = s'_{m-1} = 0$ – since s_1 and s_{m-1} are constant – and $s'_2 = s'_{m-2}$ – in the third and the fourth lines, respectively. Using the final equation above and the identity $s'_1 = 0$, we can derive the limit values for all s'_i as given below:

$$s'_i = -(i-1)(m-i-1)\frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}} \tag{2.22}$$

for $2 \leq i \leq m-2$.

Before we can plug these into the relatedness formula, it will be helpful to write $\bar{q}'$ differently. Since $\bar{q} = \frac{1}{m}\sum_{i=0}^{m-1} q_i$, we can take the derivatives of all the terms separately.

$$\bar{q}' = \frac{1}{m}\left(q'_0 + q'_1 + q'_{m-1} + \sum_{i=2}^{m-2} q_i\right) = \frac{1}{m}\left(q'_0 + q'_1 + q'_{m-1} + \sum_{i=2}^{m-2}(s'_i q_1 + s_i q'_1)\right)$$

Now we can use that $q_1 = 1$ and $s_i = 1$ in the limit of $u \downarrow 0$ for all i, that $q'_1 = q'_{m-1}$, and that $s'_1 = s'_{m-1} = 0$ (because $s_1 = s_{m-1} = 1$ regardless of u) when rewriting $\bar{q}'$.

$$\begin{aligned}\bar{q}' &= \frac{1}{m}\left(q'_0 + q'_1 + q'_{m-1} + \sum_{i=2}^{m-2}(s'_i + q'_1)\right) = \frac{1}{m}\left(q'_0 + (m-1)q'_1 + \sum_{i=2}^{m-2} s'_i\right) \\ &= \frac{1}{m}\left(q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i\right)\end{aligned}$$

We can now plug this in Equation (2.12) to get

$$r_0 = \frac{\frac{m-1}{m}(q'_0 - q'_1) - \frac{1}{m}\sum_{i=1}^{m-1} s'_i}{-\frac{1}{m}\left(q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i\right)} = \frac{(m-1)(q'_1 - q'_0) + \sum_{i=1}^{m-1} s'_i}{q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i} \tag{2.23}$$

$$r_1 = \frac{-\frac{1}{m}(q'_0 - q'_1) - \frac{1}{m}\sum_{i=1}^{m-1} s'_i}{-\frac{1}{m}\left(q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i\right)} = \frac{-(q'_1 - q'_0) + \sum_{i=1}^{m-1} s'_i}{q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i} \tag{2.24}$$

$$r_i = \frac{s'_i - \frac{1}{m}(q'_0 - q'_1) - \frac{1}{m}\sum_{i=1}^{m-1} s'_i}{-\frac{1}{m}\left(q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i\right)} = r_1 - \frac{m s'_i}{q'_0 + (m-1)q'_1 + \sum_{i=1}^{m-1} s'_i} \tag{2.25}$$

where

$$\sum_{i=1}^{m-1} s'_i = -\frac{(m-1)(m-2)(m-3)}{6} \frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}} \tag{2.26}$$

In order to have a formula for relatedness that depends only on the parameters of the model (p, q, r, m and n), we still need to express q'_0 and q'_1 in terms of these parameters. In order to be able to do that, we will first express q'_s in terms of those parameters.

Step 1 is to take the first derivative with respect to u on both sides of Equation (2.14), and evaluate them at $u = 0$.

$$\left(1 - p - \frac{q}{m} - r\left(\frac{n-1}{n}\right)^2\right) q'_1 + \left(\frac{p}{n} + \frac{q}{m}\right) q_1 = \left(\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(q'_0 + q'_2) - \left(\frac{q}{2}\frac{m-1}{m}\right)(q_0 + q_2)$$

At $u = 0$, $q_0 = q_1 = q_2 = 1$, and hence

$$\left(1 - p - \frac{q}{m} - r\left(\frac{n-1}{n}\right)^2\right) q'_1 + \frac{p}{n} = \left(\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)(q'_0 + q'_2) - q$$

Also $q'_0 = \frac{n-1}{n} q'_s$ and $q'_2 = s'_2 q_1 + s_2 q'_1$, and hence, with $q_2 = q_1 = 1$ and $s_2 = 1$, at $u = 0$, the expresion above becomes

$$\left(1 - p - \frac{q}{m} - r\left(\frac{n-1}{n}\right)^2\right) q'_1 + \frac{p}{n} = \left(\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)\left(\frac{n-1}{n} q'_s + s'_2 + q'_1\right) - q$$

or

$$\left(1 - p - \frac{q}{2}\frac{m+1}{m} - r\frac{(n-1)^2 + n}{n^2}\right) q'_1 + \frac{p}{n} = \left(\frac{q}{2}\frac{m-1}{m} + \frac{r}{n}\right)\left(\frac{n-1}{n} q'_s + s'_2\right) - q$$

Step 2 is to take the first derivative with respect to u in Equation (2.16)

$$\begin{aligned} q_1' =& \frac{mn}{4r}\left(p\frac{2}{mn}\left(u+\frac{1-u}{n}\right)+q\frac{1}{m}-q\frac{(1-u)^2}{m}+r\frac{4}{mn}\right)q_s' \\ &+\frac{mn}{4r}\left(p\frac{2}{mn}\left(1-\frac{1}{n}\right)+2q\frac{(1-u)}{m}\right)q_s+\frac{p}{r}\frac{1}{2n} \end{aligned}$$

Evaluated at $u=0$, where also $q_s=1$, this is

$$\begin{aligned} q_1' &= \left(\frac{p}{r}\frac{1}{2n}+1\right)q_s'+\frac{mn}{4r}\left(p\frac{2}{mn}\left(1-\frac{1}{n}\right)+2q\frac{1}{m}\right)+\frac{p}{r}\frac{1}{2n} \\ &= \left(\frac{p}{r}\frac{1}{2n}+1\right)q_s'+\frac{p}{2r}\left(1-\frac{1}{n}\right)+\frac{qn}{2r}+\frac{p}{r}\frac{1}{2n} \\ &= \left(\frac{p}{r}\frac{1}{2n}+1\right)q_s'+\frac{p}{2r}+\frac{qn}{2r} \end{aligned}$$

If we combine these two steps, we get

$$\begin{aligned} \left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\left(\frac{p}{r}\frac{1}{2n}+1\right)q_s'+\frac{p}{2r}+\frac{qn}{2r}\right)+\frac{p}{n} \\ =\left(\frac{q}{2}\frac{m-1}{m}+\frac{r}{n}\right)\left(\frac{n-1}{n}q_s'+s_2'\right)-q \end{aligned}$$

or

$$\begin{aligned} &\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{2n}+1\right)q_s'-\left(\frac{q}{2}\frac{m-1}{m}+\frac{r}{n}\right)\frac{n-1}{n}q_s' \\ &=\left(\frac{q}{2}\frac{m-1}{m}+\frac{r}{n}\right)s_2'-\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{2r}+\frac{qn}{2r}\right)-q-\frac{p}{n} \end{aligned}$$

or

$$\begin{aligned} &\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)q_s'-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}q_s' \\ &=\left(q\frac{m-1}{m}+\frac{2r}{n}\right)s_2'-\left(\frac{n}{r}-\frac{pn}{r}-\frac{qn}{2r}\frac{m+1}{m}-\frac{(n-1)^2+n}{n}\right)\left(\frac{p}{n}+q\right)-2\left(\frac{p}{n}+q\right) \end{aligned}$$

which gives

$$
\begin{aligned}
q'_s &= \frac{\left(q\frac{m-1}{m}+\frac{2r}{n}\right)s'_2 - \left(\frac{p}{n}+q\right)\left(\frac{n}{r}-\frac{pn}{r}-\frac{qn}{2r}\frac{m+1}{m}-\frac{(n-1)^2}{n}+1\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{\left(q\frac{m-1}{m}+\frac{2r}{n}\right)s'_2 - \left(\frac{p}{n}+q\right)\left(3-\frac{1}{n}-\frac{qn}{2r}\frac{m+1}{m}+\frac{n}{r}-\frac{pn}{r}-n\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}}
\end{aligned}
$$

We can simplify this using the formula for s'_2 from Equation (2.21), repeated below.

$$
s'_2 = -(m-3)\frac{\frac{p}{n}+q}{q\frac{m-1}{m}+r\frac{2}{n}}
$$

If we look at the first term in the numerator in Equation (2.27), we see that the coefficient of s'_2 is equal to the denominator of s'_2, so we can rewrite the formula for q'_s as follows:

$$
\begin{aligned}
q'_s &= \frac{-(m-3)\left(\frac{p}{n}+q\right)-\left(\frac{p}{n}+q\right)\left(3-\frac{1}{n}-\frac{qn}{2r}\frac{m+1}{m}+\frac{n}{r}-\frac{pn}{r}-n\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}-\frac{qn}{2r}\frac{m+1}{m}+\frac{n}{r}-\frac{pn}{r}-n\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}}
\end{aligned}
$$

We can further rewrite the numerator,

$$
\begin{aligned}
q'_s &= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}-\frac{qn}{2r}\frac{m+1}{m}+n\left(\frac{1-p-r}{r}\right)\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}-\frac{qn}{2r}\frac{m+1}{m}+n\frac{q}{r}\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{(n-1)^2+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}}
\end{aligned}
$$

where we used $p+q+r=1$ in the second line.

Then, we can rewrite the denominator,

$$\begin{aligned}
q'_s &= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r\frac{n^2-2n+1+n}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\left(1-p-\frac{q}{2}\frac{m+1}{m}-r+r\frac{n-1}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\left(q-\frac{q}{2}\frac{m+1}{m}+r\frac{n-1}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\left(\frac{q}{2}\frac{m-1}{m}+r\frac{n-1}{n^2}\right)\left(\frac{p}{r}\frac{1}{n}+2\right)-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}} \\
&= \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{2rn}\frac{m-1}{m}+q\frac{m-1}{m}+p\frac{n-1}{n^3}+r\frac{2(n-1)}{n^2}-\left(q\frac{m-1}{m}+\frac{2r}{n}\right)\frac{n-1}{n}}
\end{aligned}$$

after which we arrive at

$$q'_s = \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}} \tag{2.27}$$

Equation (2.13) moreover implies

$$q'_0 = \frac{n-1}{n}q'_s \tag{2.28}$$

which links q'_s to q'_0, and Step 2 above gave us

$$q'_1 = \left(\frac{p}{r}\frac{1}{2n}+1\right)q'_s + \frac{p+nq}{2r} \tag{2.29}$$

which links q'_s to q'_1. This implies that we have everything we need to complete Equations (2.23), (2.24) and (2.25) for the BD process.

2A.5.2 Shift

In the Shift process, an individual can be replaced if:

- An individual event happens, the group the individual is in is chosen to host it, and then within the group, she is chosen to be replaced.
- A group event happens, and a neighbouring group, or its offspring group, pushes the individual's group one position away from where it was, or replaces it. Probabilities for those events are also derived in Allen and Nowak (2012) for the Shift process, where all positions are occupied by individuals instead of groups.

- A migration event occurs, the individual's group is chosen to be one of the groups in which the migration takes place, and the individual is chosen to be swapped.

Combining these events, we can derive the recurrence relations for the stationary IBD probabilities.

$i = 1$ For $i = 1$, we have the following recurrence relation:

$$q_1 = \underbrace{p\frac{2}{m}\frac{1}{n}(1-u)q_1}_{\substack{\text{replaced in an} \\ \text{individual event}}} + \underbrace{q\left(\frac{(m-1)(1-u)}{m^2}q_0 + \frac{1-u}{m}q_1 + \frac{(m-2+(1-u))}{m^2}q_2\right)}_{\text{replaced in a group event}}$$
$$+ \underbrace{r\frac{1}{m}\frac{1}{n}\left(2q_2 + \frac{2(n-1)}{n}q_s + \frac{1}{n}q_1\right)}_{\text{replaced in a migration event}}$$
$$+ \left(\underbrace{p\left(\frac{m-2}{m} + \frac{2}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through} \\ \text{individual events}}} + \underbrace{q\left(\frac{(m-1)(m-2)}{m^2}\right)}_{\text{no change through group events}} + \underbrace{r\left(\frac{m-3}{m} + \frac{2}{m}\frac{n-1}{n} + \frac{1}{m}\frac{(n-1)^2}{n^2}\right)}_{\text{no change through migration events}}\right) q_1$$

The probabilities for individual and migration events are the same as in BD. The probabilities for group events are different. To get the probabilities for group events right, it is important, in the face of equivalent ways to define this update rule, to have an unambiguous rule for who ends up at which location after a group reproduction event. The reproducing group always stays put. If it is also chosen to die, its offspring group takes its place and no group moves. If not, then with probability one half, the offspring group occupies the position to the left of the parent group, and every group in between the reproducing and the dying group moves in the same direction. Also with probability one half, the offspring group occupies the position on the right, and every group in between the reproducing and the dying group moves in that direction.

Now, for a given pair of neighbouring positions, the group on the left replaces the group on the right, if the left one is chosen to reproduce, not chosen to also die, and reproduces on the right. The right one replaces the left one if the right one is chosen to reproduce, not chosen to also die, and reproduces to the left. Both events happen with probability $\frac{1}{2}\frac{m-1}{m^2}$, and after this, two randomly chosen members of the neighbouring groups are IBD with probability $(1-u)q_0$.

The left one replaces itself with probability $\frac{1}{m^2}$. The right one too. After this, two randomly chosen members of the neighbouring groups are IBD with probability $(1 -$

$u)q_1$.

The group to the left of the left one reproduces to the right and pushes the left group to the right position if the group to the left of the left one is chosen to reproduce, the left group and the group to the left of the left one are both not chosen to die, and the reproducing group reproduces to the right. This happens with probability $\frac{1}{2}\frac{m-2}{m^2}$. The mirror image of that happens with the same probability. After this, two randomly chosen members of the neighbouring groups are IBD with probability $(1-u)q_1$.

These probabilities and the probabilities with which they replace themselves add up to $\frac{1}{m}$, and we have seen that, for both, the IBD probability is $(1-u)q_1$.

The left one is replaced by the group to the left of it, if the left one is chosen to die, any group other than the left one and its left neighbour is chosen to reproduce, and the reproducing group reproduces to the right. The right one is replaced by the group to the right of it, if the right one is chosen to die, any group other than the right one and its right neighbour is chosen to reproduce, and the reproducing group reproduces to the left. Both events happen with probability $\frac{1}{2}\frac{m-2}{m^2}$, and after this, two randomly chosen members of the neighbouring groups are IBD with probability q_2.

The left one is replaced by the offspring of the group to the left of it, if the left one is chosen to die, its left neighbour is chosen to reproduce, and it reproduces to the right. The right one is replaced by the offspring of the group to the right of it, if the right one is chosen to die, its right neighbour is chosen to reproduce, and it reproduces to the left. Both events happen with probability $\frac{1}{2}\frac{1}{m^2}$, and after this, two randomly chosen members of the neighbouring groups are IBD with probability $(1-u)q_2$.

After all other events, the groups at the two given neighbouring locations are both not the offspring group, and were neighbouring groups in the period before the group event, too.

All of those group event probabilities can also be found in Allen and Nowak (2012). The only differences are that we derive them in a forward looking way, while they do it in a backward looking way, and, since we may have more than one individual at any site, we do not have $q_0 = 1$.

If we gather all terms with q_1 on the left hand side, we get

$$\left(1-p\left(\frac{2}{m}\frac{1}{n}(1-u)+\frac{m-2}{m}+\frac{2}{m}\frac{n-1}{n}\right)-q\left(\frac{1-u}{m}+\frac{(m-1)(m-2)}{m^2}\right)\right.$$
$$\left.-r\left(\frac{1}{m}\frac{1}{n^2}+\frac{m-3}{m}+\frac{2}{m}\frac{n-1}{n}+\frac{1}{m}\frac{(n-1)^2}{n^2}\right)\right)q_1=$$
$$q\left(\frac{(m-1)(1-u)}{m^2}q_0+\frac{(m-2+(1-u))}{m^2}q_2\right)+r\frac{1}{m}\frac{1}{n}\left(2q_2+\frac{2(n-1)}{n}q_s\right)$$

If we use the identity $q_0=\frac{1+(n-1)q_s}{n}$, and multiply left and right by m, this can be rewritten as

$$\left(m-p\left(m-\frac{2}{n}u\right)-q\left(m-\frac{2m-2}{m}-u\right)-r\left(m-\frac{4n-2}{n^2}\right)\right)q_1=$$
$$\left(q\frac{m-1}{m}+r\frac{2}{n}\right)(q_0+q_2)-q\frac{u}{m}((m-1)q_0+q_2)-r\frac{2}{n^2}$$

If we furthermore use that $p+q+r=1$, this can be further simplified to

$$\left(q\frac{2(m-1)}{m}+r\frac{2(2n-1)}{n^2}+u\left(p\frac{2}{n}+q\right)\right)q_1=\left(q\frac{m-1}{m}+r\frac{2}{n}\right)(q_0+q_2)-q\frac{u}{m}((m-1)q_0+q_2)-r\frac{2}{n^2} \tag{2.30}$$

$1<i<m-1$ The recurrence relations for $1<i<m-1$ are derived in a similar way. The differences with the recurrence relation for $i=1$ arise because the two groups no longer are each other's neighbours, which means that the groups can no longer replace each other, nor can they swap individuals.

$$q_i=\underbrace{p\frac{2}{m}\frac{1}{n}(1-u)q_i}_{\substack{\text{replaced in an}\\ \text{individual event}}}+\underbrace{q\left(\frac{(m-i)(i-1+(1-u))}{m^2}q_{i-1}+\frac{1-u}{m}q_i+\frac{i(m-i-1+(1-u))}{m^2}q_{i+1}\right)}_{\text{replaced in a group event}}$$
$$+\underbrace{r\frac{2}{m}\frac{1}{n}(q_{i-1}+q_{i+1})}_{\substack{\text{replaced in a}\\ \text{migration event}}}$$
$$+\left(\underbrace{p\left(\frac{m-2}{m}+\frac{2}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through}\\ \text{individual events}}}+\underbrace{q\left(\frac{(m-i)(m-i-1)+i(i-1)}{m^2}\right)}_{\text{no change through group events}}+\underbrace{r\left(\frac{m-4}{m}+\frac{4}{m}\frac{n-1}{n}\right)}_{\substack{\text{no change through}\\ \text{migration events}}}\right)q_i$$

If we gather the terms with q_i on the left-hand side, we get

$$\left(1 - p\left(\frac{2}{m}\frac{1}{n}(1-u) + \frac{m-2}{m} + \frac{2}{m}\frac{n-1}{n}\right) - q\frac{(m-i)(m-i-u)+i(i-u)}{m^2} - r\left(\frac{m-4}{m} + \frac{4}{m}\frac{n-1}{n}\right)\right) q_i$$

$$= \left(q\frac{(m-i)(i-u)}{m^2} + r\frac{2}{m}\frac{1}{n}\right) q_{i-1} + \left(q\frac{i(m-i-u)}{m^2} + r\frac{2}{m}\frac{1}{n}\right) q_{i+1}$$

If we multiply both sides with m, and use $p+q+r=1$ again, we can rewrite this as

$$\left(q\frac{2i(m-i)}{m} + r\frac{4}{n} + u\left(p\frac{2}{n} + q\right)\right) q_i = \left(q\frac{i(m-i)}{m} + r\frac{2}{n}\right)(q_{i-1} + q_{i+1}) - q\frac{u}{m}((m-i)q_{i-1} + iq_{i+1}) \tag{2.31}$$

$i = 0$ To derive q_s – and therewith q_0 – we use the recurrence relation concerning the IBD probabilities for two individuals from one and the same group.

$$q_s = p\frac{1}{m}\frac{2}{n^2}(1-u) + p\left(\frac{1}{m}\frac{2(n-1)}{n^2}(1-u) + \frac{1}{m}\frac{n-2}{n} + \frac{m-1}{m}\right) q_s + q\left(\frac{(1-u)^2}{m} + \frac{m-1}{m}\right) q_s$$
$$+ r\frac{4}{mn}q_1 + r\left(\frac{m-2}{m} + \frac{2(n-2)}{mn}\right) q_s$$

Again, the probabilities for individual and migration events are the same as in BD. If a group event happens, then any given group reproduces and replaces itself with probability $\frac{1}{m^2}$, is replaced by its left neighbour with probability $\frac{1}{2}\frac{m-1}{m^2}$, and by its right neighbour with the same probability. That adds up to $\frac{1}{m}$, and the IBD probability dilutes to $(1-u)^2 q_s$ in all of these cases. In all other cases, it remains q_s.

Expressing q_1 as a function of q_s, this becomes

$$q_1 = \frac{mn}{4r}\left(1 - p + p\frac{2}{mn}\left(u + \frac{1-u}{n}\right) - q\left(\frac{(1-u)^2}{m} + \frac{m-1}{m}\right) - r + r\frac{4}{mn}\right) q_s - \frac{p}{r}\frac{1-u}{2n}$$

Finally, we use $p+q+r=1$ to write

$$q_1 = \frac{mn}{4r}\left(p\frac{2}{mn}\left(u + \frac{1-u}{n}\right) + q\frac{1}{m} - q\frac{(1-u)^2}{m} + r\frac{4}{mn}\right) q_s - \frac{p}{r}\frac{1-u}{2n} \tag{2.32}$$

This turns out to be the same equation for Shift as for BD.

Solving the system Suppose again that q_i has the following form for $1 \leq i \leq m-1$:

$$q_i = s_i q_1$$

where $s_1 = s_{m-1} = 1$, and $\lim_{u \to 0} s_i = 1$ for $1 < i < m-1$. Then, we can rewrite Equation (2.31) as follows:

$$\left(q\frac{2i(m-i)}{m} + r\frac{4}{n} + u\left(p\frac{2}{n} + q\right)\right)s_i q_1 \\ = \left(q\frac{i(m-i)}{m} + r\frac{2}{n}\right)(s_{i-1} + s_{i+1})q_1 - q\frac{u}{m}((m-i)s_{i-1} + is_{i+1})q_1$$

Assuming that $q_1 \neq 0$, this is also:

$$\left(q\frac{2i(m-i)}{m} + r\frac{4}{n} + u\left(p\frac{2}{n} + q\right)\right)s_i = \left(q\frac{i(m-i)}{m} + r\frac{2}{n}\right)(s_{i-1} + s_{i+1}) - q\frac{u}{m}((m-i)s_{i-1} + is_{i+1}) \tag{2.33}$$

Since $q_i \to 1$ for all i as $u \downarrow 0$, we cannot directly use Equation (2.12) to calculate relatedness, as both the numerator and the denominator approach zero. Therefore, we will apply L'Hôpital's rule and calculate relatednesses as:

$$r_i = \frac{q_i' - \bar{q}'}{-\bar{q}'}$$

Here the derivatives are taken with respect to u, and evaluated in the limit of $u \downarrow 0$. In order to determine q_i' and $\bar{q}'$, we want to find s_i' by taking derivatives with respect to u on both sides of Equation (2.33).

$$\left(q\frac{2i(m-i)}{m} + r\frac{4}{n} + u\left(p\frac{2}{n} + q\right)\right)s_i' + \left(p\frac{2}{n} + q\right)s_i = \\ \left(q\frac{i(m-i)}{m} + r\frac{2}{n}\right)(s_{i-1}' + s_{i+1}') - \frac{q}{m}((m-i)s_{i-1} + is_{i+1}) - q\frac{u}{m}\left((m-i)s_{i-1}' + is_{i+1}'\right)$$

If we evaluate this in the limit of $u \downarrow 0$, then we can also use that $s_i = 1$ for all i in that limit.

$$\left(q\frac{2i(m-i)}{m} + r\frac{4}{n}\right)s_i' + \left(p\frac{2}{n} + q\right) = \left(q\frac{i(m-i)}{m} + r\frac{2}{n}\right)(s_{i-1}' + s_{i+1}') - q$$

This can be reorganized as follows

$$\left(q\frac{i(m-i)}{m}+r\frac{2}{n}\right)(2s'_i-(s'_{i-1}+s'_{i+1}))=-\left(p\frac{2}{n}+2q\right)$$

If we divide everything by 2 and by the first term, we get the following equation.

$$s'_i-\frac{s'_{i-1}+s'_{i+1}}{2}=-\frac{\frac{p}{n}+q}{q\frac{i(m-i)}{m}+r\frac{2}{n}} \tag{2.34}$$

We will call the right hand side of this equation $-\eta_i$. If we do so, and we sum both sides of the equation over $2 \leq i \leq m-2$, we get

$$\sum_{i=2}^{m-2} s'_i-\frac{1}{2}\sum_{i=2}^{m-2}(s'_{i-1}+s'_{i+1})=-\sum_{i=2}^{m-2}\eta_i$$

$$\sum_{i=2}^{m-2} s'_i-\frac{1}{2}(s'_1+s'_{m-1}+s'_2+s'_{m-2})-\frac{1}{2}\sum_{i=3}^{m-3}2s'_i=-\sum_{i=2}^{m-2}\eta_i$$

$$s'_2+s'_{m-2}-\frac{1}{2}(s'_2+s'_{m-2})=-\sum_{i=2}^{m-2}\eta_i$$

$$s'_2=-\sum_{i=2}^{m-2}\eta_i \tag{2.35}$$

which, together with Equation (2.34), implies that

$$s'_j=-(j-1)\sum_{i=2}^{m-2}\eta_i+2\sum_{i=2}^{j-1}(j-i)\eta_i$$

The derivations of Equations (2.23), (2.24) and (2.25) in the previous subsection do not depend on the update process; they only depend on the assumption that $q_i = s_i q_1$ for $1 \leq i \leq m-1$. Here we make the same assumption for Shift. Hence, we can use the same equations for Shift as we did for BD. Because Equations (2.23), (2.24) and (2.25) also feature the sum of s'_j values, it will help to compute that sum as well. Since $s'_1 = s'_{m-1} = 0$, summing from $j = 2$ to $m-2$ gives the same answer as summing from $j = 1$ to $m-1$.

$$\sum_{j=1}^{m-1} s'_j=\sum_{j=2}^{m-2} s'_j=-\sum_{j=2}^{m-2}(j-1)\sum_{i=2}^{m-2}\eta_i+2\sum_{j=2}^{m-2}\sum_{i=2}^{j-1}(j-i)\eta_i$$

The first part of the above summation can be calculated as follows

$$\sum_{j=2}^{m-2}(j-1)\sum_{i=2}^{m-2}\eta_i = (1+2+...+m-3)\sum_{i=2}^{m-2}\eta_i = \frac{(m-3)(m-2)}{2}\sum_{i=2}^{m-2}\eta_i \tag{2.36}$$

The second part of the summation can be calculated as follows

$$\begin{aligned}\sum_{j=2}^{m-2}\sum_{i=2}^{j-1}(j-i)\eta_i = &\underbrace{\eta_2}_{j=3} + \underbrace{2\eta_2+\eta_3}_{j=4} + \underbrace{3\eta_2+2\eta_3+\eta_4}_{j=5} + ... + \\ &+ ... + \underbrace{(m-4)\eta_2+(m-5)\eta_3+...+3\eta_{m-5}+2\eta_{m-4}+\eta_{m-3}}_{j=m-2}\end{aligned}$$

When we extend the summation in this way, we see that for every η_i, the first time it appears, it is multiplied by 1; the next time it is multiplied by 2; and so on; and the last point the same term appears, it is multiplied by $(m-2-i)$. Therefore, we can rewrite the above summation as follows

$$\sum_{j=2}^{m-2}\sum_{i=2}^{j-1}(j-i)\eta_i = \sum_{i=2}^{m-3}(1+2+...+m-2-i)\eta_i = \sum_{i=2}^{m-3}\frac{(m-2-i)(m-1-i)}{2}\eta_i \tag{2.37}$$

Now, if we combine the two parts found in (2.36) and (2.37), we see that

$$\begin{aligned}\sum_{j=2}^{m-2}s'_j &= -\sum_{j=2}^{m-2}(j-1)\sum_{i=2}^{m-2}\eta_i + 2\sum_{j=2}^{m-2}\sum_{i=2}^{j-1}(j-i)\eta_i \\ &= -\frac{(m-3)(m-2)}{2}\sum_{i=2}^{m-2}\eta_i + 2\sum_{i=2}^{m-3}\frac{(m-2-i)(m-1-i)}{2}\eta_i\end{aligned}$$

We can write the terms with η_2 and η_{m-2} separately, and use that $\eta_2 = \eta_{m-2}$:

$$\begin{aligned}
\sum_{j=2}^{m-2} s'_j =& -\frac{(m-3)(m-2)}{2}(\eta_2 + \eta_{m-2}) - \frac{(m-3)(m-2)}{2}\sum_{i=3}^{m-3}\eta_i \\
&+ (m-4)(m-3)\eta_2 + 2\sum_{i=3}^{m-3}\frac{(m-2-i)(m-1-i)}{2}\eta_i \\
=& -(m-3)(m-2)\eta_2 - \frac{(m-3)(m-2)}{2}\sum_{i=3}^{m-3}\eta_i \\
&+ (m-4)(m-3)\eta_2 + 2\sum_{i=3}^{m-3}\frac{(m-2-i)(m-1-i)}{2}\eta_i \\
=& -2(m-3)\eta_2 + \sum_{i=3}^{m-3}\left(2\frac{(m-2-i)(m-1-i)}{2} - \frac{(m-3)(m-2)}{2}\right)\eta_i
\end{aligned}$$

Rearranging gives the formula below.

$$\sum_{j=1}^{m-1} s'_j = \sum_{j=2}^{m-2} s'_j = -2(m-3)\eta_2 + \sum_{i=3}^{m-3}\left(\frac{m^2 - m - 2 - 2i(2m-3) + 2i^2}{2}\right)\eta_i \qquad (2.38)$$

Unfortunately, there is no closed-form solution for the values s'_i and their sum. However, it is possible to find numerical solutions once we fix the population size.

In order to have a formula for relatedness that only depends on the parameters of the model (p, q, r, m and n), we still need to express q'_0 and q'_1 in terms of these parameters. In order to be able to do that, we will first express q'_s in terms of those parameters.

Step 1 is to take the first derivative with respect to u on both sides of Equation (2.30), and evaluate them at $u = 0$.

$$\left(q\frac{2(m-1)}{m} + r\frac{2(2n-1)}{n^2}\right)q'_1 + \left(p\frac{2}{n} + q\right)q_1 = \left(q\frac{m-1}{m} + r\frac{2}{n}\right)(q'_0 + q'_2) - \frac{q}{m}((m-1)q_0 + q_2)$$

At $u = 0$, also $q_0 = q_1 = q_2 = 1$, and hence

$$\left(q\frac{2(m-1)}{m} + r\frac{2(2n-1)}{n^2}\right)q'_1 + \left(p\frac{2}{n} + 2q\right) = \left(q\frac{m-1}{m} + r\frac{2}{n}\right)(q'_0 + q'_2)$$

Also $q_0' = \frac{n-1}{n} q_s'$ and $q_2' = s_2' q_1 + s_2 q_1'$, and hence, with $q_2 = q_1 = 1$ and $s_2 = 1$ at $u = 0$, this is also

$$\left(q\frac{2(m-1)}{m} + r\frac{2(2n-1)}{n^2}\right) q_1' + \left(p\frac{2}{n} + 2q\right) = \left(q\frac{m-1}{m} + r\frac{2}{n}\right)\left(\frac{n-1}{n} q_s' + s_2' + q_1'\right)$$

or

$$\left(q\frac{m-1}{m} + r\frac{2(n-1)}{n^2}\right) q_1' + \left(p\frac{2}{n} + 2q\right) = \left(q\frac{m-1}{m} + r\frac{2}{n}\right)\left(\frac{n-1}{n} q_s' + s_2'\right)$$

Step 2 is to take the first derivative with respect to u in Equation (2.32)

$$\begin{aligned} q_1' =& \frac{mn}{4r}\left(p\frac{2}{mn}\left(u + \frac{1-u}{n}\right) + q\frac{1}{m} - q\frac{(1-u)^2}{m} + r\frac{4}{mn}\right) q_s' \\ &+ \frac{mn}{4r}\left(p\frac{2}{mn}\left(1 - \frac{1}{n}\right) + 2q\frac{(1-u)}{m}\right) q_s + \frac{p}{r}\frac{1}{2n} \end{aligned}$$

Evaluated at $u = 0$, where also $q_s = 1$, this is

$$\begin{aligned} q_1' &= \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{mn}{4r}\left(p\frac{2}{mn}\left(1 - \frac{1}{n}\right) + 2q\frac{1}{m}\right) + \frac{p}{r}\frac{1}{2n} \\ &= \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p}{2r}\left(1 - \frac{1}{n}\right) + \frac{qn}{2r} + \frac{p}{r}\frac{1}{2n} \\ &= \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p}{2r} + \frac{qn}{2r} \end{aligned}$$

If we combine these two steps, we get

$$\begin{aligned} \left(q\frac{m-1}{m} + r\frac{2(n-1)}{n^2}\right)\left(\left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p}{2r} + \frac{qn}{2r}\right) + \left(p\frac{2}{n} + 2q\right) \\ = \left(q\frac{m-1}{m} + r\frac{2}{n}\right)\left(\frac{n-1}{n} q_s' + s_2'\right) \end{aligned}$$

or

$$\begin{aligned} \left(q\frac{m-1}{m} + r\frac{2(n-1)}{n^2}\right)\left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' - \left(q\frac{m-1}{m} + r\frac{2}{n}\right)\frac{n-1}{n} q_s' = \\ \left(q\frac{m-1}{m} + r\frac{2}{n}\right) s_2' - \left(q\frac{m-1}{m} + r\frac{2(n-1)}{n^2}\right)\left(\frac{p}{2r} + \frac{qn}{2r}\right) - \left(p\frac{2}{n} + 2q\right) \end{aligned}$$

or

$$\left(\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{m-1}{m}\frac{1}{n}+p\frac{n-1}{n^3}\right)q'_s=$$
$$\left(q\frac{m-1}{m}+r\frac{2}{n}\right)s'_2-\left(\frac{q}{r}\frac{n}{2}\frac{m-1}{m}+\frac{(n-1)}{n}\right)\left(\frac{p}{n}+q\right)-\left(p\frac{2}{n}+2q\right)$$

which gives

$$q'_s=\frac{\left(q\frac{m-1}{m}+r\frac{2}{n}\right)s'_2-\left(\frac{p}{n}+q\right)\left(3-\frac{1}{n}+\frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}$$

If we plug in the formula for s'_2 from (2.35), we get the following expression for q'_s:

$$\begin{aligned}q'_s&=\frac{-\left(q\frac{m-1}{m}+r\frac{2}{n}\right)\sum_{i=2}^{m-2}\frac{\frac{p}{n}+q}{q\frac{i(m-i)}{m}+r\frac{2}{n}}-\left(\frac{p}{n}+q\right)\left(3-\frac{1}{n}+\frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}\\&=\frac{-\left(\frac{p}{n}+q\right)\left(\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}+3-\frac{1}{n}+\frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}\end{aligned}\tag{2.39}$$

As in BD, Equation (2.13) moreover implies

$$q'_0=\frac{n-1}{n}q'_s\tag{2.40}$$

which links q'_s to q'_0, and Step 2 above gave us

$$q'_1=\left(\frac{p}{r}\frac{1}{2n}+1\right)q'_s+\frac{p+nq}{2r}\tag{2.41}$$

which links q'_s to q'_1. This implies that we have everything we need to complete Equations (2.23), (2.24) and (2.25) for the Shift process.

2A.5.3 Three useful identities

The definition of relatedness (Equation 2.12) implies that relatednesses have to add up to 0.

$$\sum_{i=0}^{m-1}r_i=\sum_{i=0}^{m-1}\frac{q_i-\bar{q}}{1-\bar{q}}=\frac{m\bar{q}-m\bar{q}}{1-\bar{q}}=0$$

Equation (2.13) relates the relatedness within the group including self and excluding self in an obvious way, which we repeat here:

$$r_0 = \frac{1 + (n-1)r_s}{n}$$

Together, these imply that

$$\sum_{i=1}^{m-1} r_i = -r_0 = -\frac{1}{n} - \frac{(n-1)r_s}{n} \tag{2.42}$$

If we define r_o as the relatedness found through IBD probabilities, as in the previous subsection, to a randomly drawn individual from another group, where all other groups are equally likely to be drawn, then this is

$$r_o = \frac{1}{m-1}r_1 + \frac{1}{m-1}r_2 + \frac{1}{m-1}r_3 + \cdots + \frac{1}{m-1}r_{m-1} = \frac{1}{m-1}\sum_{i=1}^{m-1} r_i$$

or

$$\sum_{i=1}^{m-1} r_i = (m-1)r_o \tag{2.43}$$

Combining Equations (2.42) and (2.43), we get

$$r_o = -\frac{1}{n(m-1)} - \frac{(n-1)r_s}{n(m-1)} \tag{2.44}$$

None of the three identities depends on the update process, so they apply to BD as well as Shift.

2A.5.4 Alternative derivation of the three identities

In the limit of $u \downarrow 0$, if two individuals are identical, they are identical by descent. Therefore, if we derive these identities more generally, using conditional probabilities, then they will coincide in this limit. Consider the dynamical system at hand, which is a Markov chain, in which every state is a vector k, where $k_i \in \{0, 1, ..., n\}$ is the number of cooperators in group i, for $i = 1, ..., m$. For every such population state, one can imagine hypothetical chance experiments, and define differences in conditional probabilities as we will below. These can then be aggregated, with weights attached to the population states. The weights could represent how often these states are visited relative to each other (this would be the rare-mutation dimorphic distribution from Allen and Tarnita, 2014, or the rare-mutation conditional distribution from Allen and McAvoy, 2019), but for now, all that matters is that p_k is the weight of population state k, and that $\sum_k p_k = 1$.

Within-group relatedness Consider the following hypothetical chance experiment for a given population state k. Draw a random individual from the population, with every individual equally likely to be drawn. After this, go back to the same group, and randomly draw another individual from it. Then, for this state, one could define the proto-relatedness as

$$r_{s,k} = P_{s,k}(C|C) - P_{s,k}(C|D)$$

The subscript s for same group indicates that this measure is about the relatedness between two different individuals within the same group. The subscript k indicates which population state it pertains to.

Within group relatedness can now be defined as

$$r_s = \sum_k p_k \cdot r_{s,k} = \sum_k p_k \left(P_{s,k}(C|C) - P_{s,k}(C|D)\right)$$

Relatedness with an individual from a random other group Now think of another chance experiment for a given population state k. Draw a random individual from the population, again with every individual equally likely to be drawn. After this, go to a different group, with all other groups equally likely to be chosen, and randomly draw another individual from that group. Then, for this state, proto-relatedness is

$$r_{o,k} = P_{o,k}(C|C) - P_{o,k}(C|D)$$

The subscript o for other group indicates that this measure is about the relatedness between two individuals in different groups.

Relatedness between two individuals from randomly chosen different groups can now be defined as

$$r_o = \sum_k p_k \cdot r_{o,k} = \sum_k p_k \left(P_{o,k}(C|C) - P_{o,k}(C|D) \right)$$

Relatedness between individuals that are i groups apart With groups situated on the cycle, we can also define other chance experiments for a given population state k. First draw a random individual from the population, as before. After this, with probability $\frac{1}{2}$ go to the group that is i steps to the left of the first group, and with probability $\frac{1}{2}$ go to the group that is i steps to the right of the first group, $i = 1, ..., m-1$. Randomly draw another individual from that group. Then, for this state, the i-step proto-relatedness, can now be defined as

$$r_{i,k} = P_{i,k}(C|C) - P_{i,k}(C|D)$$

for $i = 1, ..., m-1$, where $r_i = r_j$ if $i = m - j$ for all $1 \leq i \leq m-1$. The subscript i indicates that this measure is about the relatedness between two individuals in groups that are i steps away from each other.

Relatedness between two individuals that are i groups apart can now be defined as

$$r_i = \sum_k p_k \cdot r_{i,k} = \sum_k p_k \left(P_{i,k}(C|C) - P_{i,k}(C|D) \right)$$

Identity I In order to relate these relatednesses to each other, assume that we consider a state k, for which $K = \sum_{j=0}^{m} k_j$ is the total number of cooperators in the population as a whole. Now imagine a chance experiment where we first draw a random individual from the population, with all individuals equally likely to be chosen, and then – without replacement – another individual from the population as a whole, with all remaining individuals equally likely to be drawn. Conditional on the first being a cooperator, the chance that the second is a cooperator can be written in two different ways, which must be equal to each other:

$$\frac{n-1}{nm-1} P_{s,k}(C|C) + \frac{n(m-1)}{nm-1} P_{o,k}(C|C) = \frac{K-1}{nm-1}$$

Conditional on the first being a defector, one can also express the chance that the second is a cooperator in two equivalent ways

$$\frac{n-1}{nm-1}P_{s,k}(C|D) + \frac{n(m-1)}{nm-1}P_{o,k}(C|D) = \frac{K}{nm-1}$$

These two identities together imply that

$$(n-1)\left(P_{s,k}(C|C) - P_{s,k}(C|D)\right) + n(m-1)(P_{o,k}(C|C) - P_{o,k}(C|D)) = -1$$

which implies that

$$r_{o,k} = -\frac{1}{n(m-1)} - \frac{n-1}{n(m-1)}r_{s,k}$$

Because the p_k add up to 1, this state-wise identity implies that if we aggregate over states accordingly, the following holds:

$$r_o = -\frac{1}{n(m-1)} - \frac{n-1}{n(m-1)}r_s$$

This is Equation (2.44).

Identity II One can also go over the groups, according to their distance to the group from which the first individual was drawn. Then the equalities become:

$$\frac{n-1}{nm-1}P_{s,k}(C|C) + \frac{n}{nm-1}\sum_{i=1}^{m-1} P_{i,k}(C|C) = \frac{K-1}{nm-1}$$

and

$$\frac{n-1}{nm-1}P_{s,k}(C|D) + \frac{n}{nm-1}\sum_{i=1}^{m-1} P_{i,k}(C|D) = \frac{K}{nm-1}$$

These two identities together imply that

$$(n-1)(P_{s,k}(C|C) - P_{s,k}(C|D)) + n\left(\sum_{i=1}^{m-1} P_{i,k}(C|C) - \sum_{i=1}^{m-1} P_{i,k}(C|D)\right) = -1$$

which can be rewritten as

$$\sum_{i=1}^{m-1} r_{i,k} = -\frac{1}{n} - \frac{n-1}{n}r_{s,k}$$

Because the p_k add up to 1, this state-wise identity implies that if we aggregate over states accordingly, the following holds:

$$\sum_{i=1}^{m-1} r_i = -\frac{1}{n} - \frac{n-1}{n} r_s$$

This is Equation (2.42).

Identity III Given Identity I, Identity II is also equivalent to

$$\sum_{i=1}^{m-1} r_i = (m-1) r_o$$

This is Equation (2.43).

2A.6 Birth-Death versus Shift

In Section 2A.4, we have found the critical b/c ratios for the the Birth-Death and the Shift process. For the BD process, the critical ratio was given in Condition (2.7). It is repeated below.

$$\frac{b}{c} > \frac{p}{q} \frac{m}{m-1} \frac{1-r_0}{n(r_0 - r_1)}$$

For the Shift process, the critical ratio was given in Condition (2.9). This is also repeated below.

$$\frac{b}{c} > \frac{p}{q} \frac{1-r_0}{n r_0}$$

These thresholds are expressed as functions of p, q, r, m, and n, as well as relatednesses r_0 and r_1 – or, equivalently, r_s and r_1. These relatednesses will typically be different for different update processes.

In Section 2A.5, we have computed r_0 and r_1, both for the Birth-Death process and for the Shift process, expressing them as functions of p, q, r, m and n as well. For both processes,

Equations (2.23) and (2.24), reproduced below, apply.

$$r_0 = \frac{(m-1)(q_1' - q_0') + \sum_{i=1}^{m-1} s_i'}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'}$$
$$r_1 = \frac{-(q_1' - q_0') + \sum_{i=1}^{m-1} s_i'}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'}$$

In these formulas, we still need to fill in q_1', q_0' and $\sum_{i=1}^{m-1} s_i'$, and these will differ between the two processes. For BD, we have Equations (2.26), (2.27), (2.28) and (2.29), reproduced below.

$$\sum_{i=1}^{m-1} s_i' = -\frac{(m-1)(m-2)(m-3)}{6} \frac{\frac{p}{n} + q}{q\frac{m-1}{m} + r\frac{2}{n}}$$
$$q_s' = \frac{-\left(\frac{p}{n} + q\right)\left(m - \frac{1}{n} + \frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m} + q\frac{1}{n}\frac{m-1}{m} + p\frac{n-1}{n^3}}$$
$$q_0' = \frac{n-1}{n} q_s'$$
$$q_1' = \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p+nq}{2r}$$

For Shift, we have Equations (2.38), (2.39), (2.40) and (2.41), reproduced below, where $\eta_i = \frac{\frac{p}{n}+q}{q\frac{i(m-i)}{m}+r\frac{2}{n}}$ is given by Equation (2.34).

$$\sum_{j=1}^{m-1} s_j' = \sum_{j=2}^{m-2} s_j' = -2(m-3)\eta_2 + \sum_{i=3}^{m-3} \left(\frac{m^2 - m - 2 - 2i(2m-3) + 2i^2}{2}\right)\eta_i$$
$$q_s' = \frac{-\left(\frac{p}{n} + q\right)\left(\sum_{i=2}^{m-2} \frac{q\frac{m-1}{m} + r\frac{2}{n}}{q\frac{i(m-i)}{m} + r\frac{2}{n}} + 3 - \frac{1}{n} + \frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m} + q\frac{1}{n}\frac{m-1}{m} + p\frac{n-1}{n^3}}$$
$$q_0' = \frac{n-1}{n} q_s'$$
$$q_1' = \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p+nq}{2r}$$

For both processes, we have $q_0' = \frac{n-1}{n} q_s'$ and $q_1' = \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p+nq}{2r}$, even though the value of q_s' will differ between the two processes. Therefore, we can use these formulas to express difference $q_1' - q_0'$ in the relatedness formulas in terms of model parameters and q_s'

only.

$$q_1' - q_0' = \left(\frac{p}{r}\frac{1}{2n} + 1\right) q_s' + \frac{p+nq}{2r} - \frac{n-1}{n} q_s' = \left(\frac{p}{r}\frac{1}{2n} + 1 - \frac{n-1}{n}\right) q_s' + \frac{p+nq}{2r}$$
$$= \left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' + \frac{p+nq}{2r}$$

Now, if we plug this back into the relatedness formulas, we get the following expressions:

$$r_0 = \frac{(m-1)\left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' + (m-1)\frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}{\left(\frac{n-1}{n} + (m-1)\left(\frac{p}{r}\frac{1}{2n} + 1\right)\right) q_s' + (m-1)\frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}$$
$$r_1 = \frac{-\left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' - \frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}{\left(\frac{n-1}{n} + (m-1)\left(\frac{p}{r}\frac{1}{2n} + 1\right)\right) q_s' + (m-1)\frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}$$

Using the relatedness formulas and the relationships between IBD probabilities q_s', q_0' and q_1', we can rewrite $1 - r_0$ and $r_0 - r_1$ as follows

$$1 - r_0 = \frac{mq_0'}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'} = \frac{m\frac{n-1}{n} q_s'}{\left((m-1)\left(\frac{p}{r}\frac{1}{2n} + 1\right) + \frac{n-1}{n}\right) q_s' + (m-1)\frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}$$
$$r_0 - r_1 = \frac{m(q_1' - q_0')}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'} = \frac{m\left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' + m\frac{p+nq}{2r}}{\left((m-1)\left(\frac{p}{r}\frac{1}{2n} + 1\right) + \frac{n-1}{n}\right) q_s' + (m-1)\frac{p+nq}{2r} + \sum_{i=1}^{m-1} s_i'}$$

If we plug these into the formulas for the critical ratios, we get, for BD,

$$\begin{aligned}
\frac{b}{c} &> \frac{p}{q}\frac{1}{n}\frac{m}{m-1}\frac{\frac{mq_0'}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'}}{\frac{m(q_1' - q_0')}{q_0' + (m-1)q_1' + \sum_{i=1}^{m-1} s_i'}} \\
&= \frac{p}{q}\frac{1}{n}\frac{m}{m-1}\frac{q_0'}{q_1' - q_0'} \\
&= \frac{p}{q}\frac{1}{n}\frac{m}{m-1}\frac{\frac{n-1}{n} q_s'}{\left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' + \frac{p+nq}{2r}} \\
&= \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{q_s'}{\left(\frac{p}{r}\frac{1}{2n} + \frac{1}{n}\right) q_s' + \frac{p+nq}{2r}}
\end{aligned} \tag{2.45}$$

And for Shift, we get

$$\begin{aligned}
\frac{b}{c} &> \frac{p}{q}\frac{1}{n}\frac{\frac{mq_0'}{q_0'+(m-1)q_1'+\sum_{i=1}^{m-1}s_i'}}{\frac{(m-1)(q_1'-q_0')+\sum_{i=1}^{m-1}s_i'}{q_0'+(m-1)q_1'+\sum_{i=1}^{m-1}s_i'}} \\
&= \frac{p}{q}\frac{1}{n}\frac{mq_0'}{(m-1)(q_1'-q_0')+\sum_{i=1}^{m-1}s_i'} \\
&= \frac{p}{q}\frac{1}{n}\frac{m\frac{n-1}{n}q_s'}{(m-1)\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)q_s'+(m-1)\frac{p+nq}{2r}+\sum_{i=1}^{m-1}s_i'} \\
&= \frac{p}{q}\frac{n-1}{n^2}\frac{mq_s'}{(m-1)\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)q_s'+(m-1)\frac{p+nq}{2r}+\sum_{i=1}^{m-1}s_i'} \\
&= \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{q_s'}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)q_s'+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}s_i'}
\end{aligned} \tag{2.46}$$

What we will do in this section, is to compare these two critical b/c ratios, and show that the one for BD is always higher than the one for Shift.

We start by comparing the IBD probabilities q_s', which now get a superscript, depending on the update process.

$$\begin{aligned}
(q_s')^{BD} &= -\frac{\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}} \\
(q_s')^{Shift} &= -\frac{\left(\frac{p}{n}+q\right)\left[\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}+3-\frac{1}{n}+\frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right]}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}} \\
(q_s')^{BD}-(q_s')^{Shift} &= -\frac{\left(\frac{p}{n}+q\right)\left(m-\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}-3\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}
\end{aligned}$$

$$(q_s')^{BD}=(q_s')^{Shift}-\frac{\left(\frac{p}{n}+q\right)\left(m-3-\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}:=(q_s')^{Shift}-A \tag{2.47}$$

For the terms $\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}$ within the sum above, we have the following calculations.

$$\begin{aligned}
\frac{i(m-i)}{m}-\frac{m-1}{m} &= \frac{i(m-i)-m+1}{m}=\frac{im-i^2-m+1}{m}=\frac{m(i-1)-(i^2-1)}{m} \\
&= \frac{m(i-1)-(i-1)(i+1)}{m}=\frac{(i-1)(m-i-1)}{m}
\end{aligned}$$

For any $2 \leq i \leq m-2$, we have $i-1>0$ and $m-i-1>0$, which implies

$$\begin{aligned}
0 &< \frac{i(m-i)}{m} - \frac{m-1}{m} \\
\frac{m-1}{m} &< \frac{i(m-i)}{m} \\
q\frac{m-1}{m} &< q\frac{i(m-i)}{m} \\
q\frac{m-1}{m} + r\frac{2}{n} &< q\frac{i(m-i)}{m} + r\frac{2}{n} \\
\frac{q\frac{m-1}{m} + r\frac{2}{n}}{q\frac{i(m-i)}{m} + r\frac{2}{n}} &< 1
\end{aligned}$$

Therefore, every term within the summation in the numerator of Equation (2.47) is less than one. Hence, the sum is less than $m-3$, and the numerator is positive, which implies that A is positive. Therefore, $(q_s')^{BD}$ is less than $(q_s')^{Shift}$, and since we know that both q_s''es are negative, this implies that $(q_s')^{BD}$ is "more negative" than $(q_s')^{Shift}$.

To compare the critical ratios for the two processes – given below – we need to compare their last terms, as their first three terms are identical.

$$\frac{b}{c} > \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q_s')^{BD}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q_s')^{BD}+\frac{p+nq}{2r}} \tag{2.48}$$

for BD, and

$$\frac{b}{c} > \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q_s')^{Shift}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q_s')^{Shift}+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift}} \tag{2.49}$$

for Shift.

2A.6.1 When m is odd

In the calculations in this and the next subsection, we assume that $m>3$ and $n>1$ since these are the interesting cases. If $m \leq 3$, the two update processes become identical, and so do their thresholds. And if $n=1$, in every individual event, the offspring replaces the parent and nothing changes in the population state.

First assume that m is odd – the case where m is even is treated below. If m is odd, we

can rewrite $\sum_{i=2}^{m-2}(s_i')^{Shift}$ as follows:

$$\begin{aligned}\sum_{i=2}^{m-2}(s_i')^{Shift} &= -2(m-3)\frac{\frac{p}{n}+q}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} + \sum_{i=3}^{m-3}\left(\frac{m^2-m-2-2i(2m-3)+2i^2}{2}\right)\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\ &= -2(m-3)\frac{\frac{p}{n}+q}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} + \sum_{i=3}^{(m-1)/2} 2m\left(\frac{m-1}{m}-\frac{i(m-i)}{m}\right)\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\ &= -2(m-3)\frac{\frac{p}{n}+q}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} - 2m\sum_{i=3}^{(m-1)/2}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\end{aligned} \tag{2.50}$$

Going from the first to the second line, we gather the coefficients of $\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}$ for i and $m-i$ within the sum since that is when these terms are the same. These coefficients then add up to $\frac{m^2-m-2-2i(2m-3)+2i^2}{2}+\frac{m^2-m-2-2(m-i)(2m-3)+2(m-i)^2}{2}=2(m-1)-2i(m-i)$.

We need to find out whether the critical ratio for the BD process is always larger than that of the Shift process; hence, we need to compare the right-hand sides of the above two inequalities that give the critical b/c ratios, given in Conditions (2.48) and (2.49). To do so, we start with comparing a few terms to zero; and step-by-step, we will make our way to the equations above. We start by showing that the four terms below are positive:

- Term 1:

$$\frac{2m}{m-1}-\frac{2m}{2m\frac{r}{qn}\left(m-\frac{1}{n}\right)+m-1}>0$$

  since $2m\frac{r}{qn}\left(m-\frac{1}{n}\right)>0$. Hence,

$$\frac{2m}{m-1}-\frac{\frac{qn}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}>0$$

- Term 2: For $1<i<m-1$,

$$\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}>0$$

- Term 3: As seen before,

$$\frac{i(m-i)}{m}-\frac{m-1}{m}=\frac{(i-1)(m-i-1)}{m}$$

For i ranging from 2 to $m-2$, we have $i-1>0$ and $m-i-1>0$. Therefore,

$$\frac{i(m-i)}{m}-\frac{m-1}{m}>0$$

- Term 4:

$$\frac{2(m-3)}{m-1}-\frac{2(m-3)}{2m\frac{r}{qn}\left(m-\frac{1}{n}\right)+m-1}>0$$

since $2m\frac{r}{qn}\left(m-\frac{1}{n}\right)>0$, and also

$$\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}>0$$

Hence,

$$\left(\frac{2(m-3)}{m-1}-\frac{\frac{qn}{r}\frac{m-3}{m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}>0$$

Since each of the terms above are individually positive for $1<i<m-1$, their products and sums will be positive as well. If we multiply the first three terms for a given i, the resulting product will be positive. If we sum these products over i, where $3\leq i\leq\frac{m-1}{2}$, the resulting sum will be positive as well, since each term in the summation is positive. And finally, we add the fourth term above to the summation to reach the expression below, which is again positive:

$$\begin{aligned}0<&\left(\frac{2(m-3)}{m-1}-\frac{\frac{qn}{r}\frac{m-3}{m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}\\&+\sum_{i=3}^{(m-1)/2}\left(\frac{2m}{m-1}-\frac{\frac{qn}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\end{aligned}$$

Now, we split this expression into two parts, depending on the sign of the terms, and put the terms with a negative coefficient on the left hand side:

$$
\begin{aligned}
0 < {} & \frac{2(m-3)}{m-1}\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} - \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\frac{q^{\frac{m-3}{m}}}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} \\
& + \sum_{i=3}^{(m-1)/2}\left(\frac{2m}{m-1}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right) - \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(q^{\frac{i(m-i)}{m}}-q^{\frac{m-1}{m}}\right)\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\
0 < {} & \frac{2(m-3)}{m-1}\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} - \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\frac{q^{\frac{2(m-2)}{m}}-q^{\frac{m-1}{m}}}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} \\
& + \sum_{i=3}^{(m-1)/2}\left(\frac{2m}{m-1}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right) - \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(q^{\frac{i(m-i)}{m}}-q^{\frac{m-1}{m}}\right)\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\
0 < {} & \frac{2(m-3)}{m-1}\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} - \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\frac{q^{\frac{2(m-2)}{m}}-q^{\frac{m-1}{m}}}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} \\
& + \sum_{i=3}^{(m-1)/2}\frac{2m}{m-1}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\
& - \sum_{i=3}^{(m-1)/2}\frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(q^{\frac{i(m-i)}{m}}-q^{\frac{m-1}{m}}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}
\end{aligned}
$$

This leads to

$$
\begin{aligned}
& \frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\sum_{i=2}^{(m-1)/2}\frac{q^{\frac{i(m-i)}{m}}-q^{\frac{m-1}{m}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) \\
& \qquad < \frac{1}{m-1}\left(2(m-3)\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}} + 2m\sum_{i=3}^{(m-1)/2}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)
\end{aligned}
$$

Notice that the whole term inside the parentheses on the right hand side is the summation from Equation (2.50) divided by $-\left(\frac{p}{n}+q\right)$, so we can rewrite it as $-\frac{1}{\frac{p}{n}+q}\sum_{i=2}^{m-2}(s_i')^{Shift}$:

$$\frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\sum_{i=2}^{(m-1)/2}\frac{q^{\frac{i(m-i)}{m}}-q^{\frac{m-1}{m}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) \qquad < \frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

$$\frac{\frac{n}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\sum_{i=2}^{(m-1)/2}\frac{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}-\left(q^{\frac{m-1}{m}}+r^{\frac{2}{n}}\right)}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) < \frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

$$\frac{2\frac{n}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\sum_{i=2}^{(m-1)/2}\left(1-\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)\right) < \frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

$$\frac{\frac{n}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\sum_{i=2}^{m-2}\left(1-\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)\right) < \frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

$$\frac{\frac{n}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) < \frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

Going from the third to the fourth line above, we use that the terms for i and $m-i$ in the latter summation are always the same. Also we rewrite $\sum_{i=2}^{m-2}1$ as $m-3$ in the last line. Now, multiply every term with $\frac{p}{n}+q$ to get;

$$\frac{\frac{p+nq}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) \qquad < \qquad \frac{1}{m-1}\left(-\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

Now, if we multiply both sides with -1, the sign also changes:

$$-\frac{\frac{p+nq}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right) \qquad > \qquad \frac{1}{m-1}\left(\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

Another step of re-arranging the terms above gives us

$$-\frac{p+nq}{2r}\left(\frac{m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right) \qquad > \qquad \frac{1}{m-1}\left(\sum_{i=2}^{m-2}(s_i')^{Shift}\right)$$

Multiplying and dividing the left hand side by the term $\left(\frac{p}{n}+q\right) / \left(\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}\right)$:

$$-\frac{p+nq}{2r}\left(\frac{\frac{\left(\frac{p}{n}+q\right)\left(m-3-\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}}{\frac{\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}}\right) \quad > \quad \frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift} \tag{2.51}$$

Notice that the term in the numerator within the parentheses on the left hand side is equal to $A = (q_s')^{Shift} - (q_s')^{BD}$ from Equation (2.47), and that the denominator is equal to $-(q_s')^{BD}$ from Equation (2.27).

$$-\frac{p+nq}{2r}\left(\frac{(q_s')^{Shift}-(q_s')^{BD}}{-(q_s')^{BD}}\right) \quad > \quad \frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

$$\frac{p+nq}{2r}\left(\frac{(q_s')^{Shift}-(q_s')^{BD}}{(q_s')^{BD}}\right) \quad > \quad \frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

If we divide both sides by $(q_s')^{Shift}$,

$$\frac{p+nq}{2r}\left(\frac{(q_s')^{Shift}-(q_s')^{BD}}{(q_s')^{Shift}(q_s')^{BD}}\right) \quad < \quad \frac{1}{m-1}\frac{1}{(q_s')^{Shift}}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

where the sign of the inequality changes since we multiply both sides with a negative term. Using $\frac{(q_s')^{Shift}-(q_s')^{BD}}{(q_s')^{Shift}(q_s')^{BD}} = \frac{1}{(q_s')^{BD}} - \frac{1}{(q_s')^{Shift}}$, we can rewrite the above inequality as follows,

$$\frac{p+nq}{2r}\left(\frac{1}{(q_s')^{BD}}-\frac{1}{(q_s')^{Shift}}\right) \quad < \quad \frac{1}{m-1}\frac{1}{(q_s')^{Shift}}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

$$\frac{p+nq}{2r}\frac{1}{(q_s')^{BD}} \quad < \quad \frac{p+nq}{2r}\frac{1}{(q_s')^{Shift}}+\frac{1}{m-1}\frac{1}{(q_s')^{Shift}}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

Add $\frac{p}{r}\frac{1}{2n}+\frac{1}{n}$ to both sides

$$\frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{(q_s')^{BD}} \quad < \quad \frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{(q_s')^{Shift}}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift}\frac{1}{(q_s')^{Shift}}$$

Reverse the numerator and the denominator on both sides of the inequality,

$$\begin{aligned}
\frac{1}{\frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{(q'_s)^{BD}}} &> \frac{1}{\frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{(q'_s)^{Shift}}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s'_i)^{Shift}\frac{1}{(q'_s)^{Shift}}} \\
\frac{(q'_s)^{BD}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{BD}+\frac{p+nq}{2r}} &> \frac{(q'_s)^{Shift}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{Shift}+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s'_i)^{Shift}}
\end{aligned}$$

where the sign of the inequality changes again on the first line above since we are reversing the fractions that we are comparing. Now, if we multiply both sides by $\frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}$, we arrive at the inequality:

$$\frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q'_s)^{BD}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{BD}+\frac{p+nq}{2r}} > \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q'_s)^{Shift}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{Shift}+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s'_i)^{Shift}}$$

In this last inequality, the left hand side is the critical b/c ratio for the BD process, and the right hand side is the critical b/c ratio for the Shift process, given in Conditions (2.48) and (2.49), respectively.

2A.6.2 When m is even

If m is even, we can rewrite $\sum_{i=2}^{m-2}(s'_i)^{Shift}$ as follows:

$$\begin{aligned}
\sum_{i=2}^{m-2}(s'_i)^{Shift} &= -2(m-3)\frac{\frac{p}{n}+q}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}+\sum_{i=3}^{m-3}\left(\frac{m^2-m-2-2i(2m-3)+2i^2}{2}\right)\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\
&= -2(m-3)\frac{\frac{p}{n}+q}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}-2m\sum_{i=3}^{m/2-1}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{\frac{p}{n}+q}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}} \\
&\quad -\frac{(m-2)^2}{4}\frac{\frac{p}{n}+q}{q^{\frac{\frac{m}{2}\frac{m}{2}}{m}}+r^{\frac{2}{n}}}
\end{aligned} \tag{2.52}$$

This we can do for reasons similar to the ones when m is odd.

To compare the critical ratios for the two update processes from Equations (2.48) and (2.49), we are going to follow a very similar path to the case above in Section 2A.6.1. We start with adding a fifth term to the terms given in the previous subsection; and step-by-step, we are going to reach the equations for the critical ratios of the two update processes. We have shown previously that the Terms 1 through 4 are positive. Here, we add another

term and show that it is also positive:

- Term 5:

$$\frac{(m-2)^2}{4(m-1)} - \frac{(m-2)^2}{8m\frac{r}{qn}\left(m-\frac{1}{n}\right)+4(m-1)} > 0$$

since $8m\frac{r}{qn}\left(m-\frac{1}{n}\right) > 0$, therefore,

$$\frac{(m-2)^2}{4(m-1)} - \frac{\frac{qn}{r}\frac{(m-2)^2}{8m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}} > 0$$

and

$$\frac{1}{q\frac{m^2}{4m}+r\frac{2}{n}} > 0$$

Hence,

$$\left(\frac{(m-2)^2}{4(m-1)} - \frac{\frac{qn}{r}\frac{(m-2)^2}{8m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\frac{1}{q\frac{m^2}{4m}+r\frac{2}{n}} > 0$$

Now, if we multiply Term 1, 2 and 3 and sum these products over $3 \le i \le \frac{m-2}{2} = \frac{m}{2} - 1$, add Term 4 to the summation from the previous subsection, we know that the resulting term will be positive. Now, we add Term 5 above, which is also positive, to get

$$\begin{aligned}
0 < &\left(\frac{2(m-3)}{m-1} - \frac{\frac{qn}{r}\frac{m-3}{m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\frac{1}{q\frac{2(m-2)}{m}+r\frac{2}{n}} \\
&+ \sum_{i=3}^{m/2-1}\left(\frac{2m}{m-1} - \frac{\frac{qn}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\left(\frac{i(m-i)}{m} - \frac{m-1}{m}\right)\frac{1}{q\frac{i(m-i)}{m}+r\frac{2}{n}} \\
&+ \left(\frac{(m-2)^2}{4(m-1)} - \frac{\frac{qn}{r}\frac{(m-2)^2}{8m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\right)\frac{1}{q\frac{m^2}{4m}+r\frac{2}{n}}
\end{aligned}$$

We are re-arranging the expression above such that every term with a negative (positive) sign is on the left hand side (right hand side),

$$\frac{\frac{qn}{r}\left(\frac{2(m-2)}{m}-\frac{m-1}{m}\right)}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}+\sum_{i=3}^{m/2-1}\frac{\frac{qn}{r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}$$
$$+\frac{\frac{qn}{r}\frac{(m-2)^2}{8m}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\frac{1}{q^{\frac{m^2}{4m}}+r^{\frac{2}{n}}}$$
$$<\frac{2(m-3)}{m-1}\frac{1}{q^{\frac{2(m-2)}{m}}+r^{\frac{2}{n}}}+\sum_{i=3}^{m/2-1}\frac{2m}{m-1}\left(\frac{i(m-i)}{m}-\frac{m-1}{m}\right)\frac{1}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}$$
$$+\frac{(m-2)^2}{4(m-1)}\frac{1}{q^{\frac{m^2}{4m}}+r^{\frac{2}{n}}}$$

Here, we used $\frac{2(m-2)}{m}-\frac{m-1}{m}=\frac{m-3}{m}$ on the first line. Realize that the right hand side in the last line above is equal to $-\frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\sum_{i=2}^{m-2}(s_i')^{Shift}$ where $\sum_{i=2}^{m-2}(s_i')^{Shift}$ is defined in Equation (2.52).

Using similar steps as in the previous subsection and the fact that

$$\frac{\frac{m}{2}\left(m-\frac{m}{2}\right)}{m}-\frac{m-1}{m}=\frac{m^2}{4m}-\frac{m-1}{m}=\frac{(m-2)^2}{4m}$$

and

$$\frac{q^{\frac{m^2}{4m}}-q^{\frac{m-1}{m}}}{q^{\frac{m^2}{4m}}+r^{\frac{2}{n}}}=1-\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{m^2}{4m}}+r^{\frac{2}{n}}}$$

we get

$$\frac{\frac{n}{2r}}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+\frac{2r}{n}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)<-\frac{1}{m-1}\frac{1}{\frac{p}{n}+q}\sum_{i=2}^{m-2}(s_i')^{Shift}$$

Multiplying both sides by $-\left(\frac{p}{n}+q\right)$, which changes the sign of the inequality, and multiplying and dividing the left hand side by $\left(\frac{p}{n}+q\right)/\left(\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}\right)$ gives

$$-\frac{p+nq}{2r}\frac{1}{m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}}\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+\frac{2r}{n}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)>\frac{1}{m-1}\sum_{i=2}^{m-2}(s_i')^{Shift}$$

$$\frac{p+nq}{2r}\left(\frac{\left(\frac{p}{n}+q\right)\left(m-3-\sum_{i=2}^{m-2}\frac{q^{\frac{m-1}{m}}+r^{\frac{2}{n}}}{q^{\frac{i(m-i)}{m}}+r^{\frac{2}{n}}}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}\right)\Bigg/\left(-\frac{\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}\right)\quad>\quad\frac{1}{m-1}\sum_{i=1}^{m-1}(s_i')^{Shift}$$

which is the same inequality as Inequality (2.51). Following the same steps after Inequality (2.51) in the previous subsection, we arrive at the result that

$$\frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q'_s)^{BD}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{BD}+\frac{p+nq}{2r}} > \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{(q'_s)^{Shift}}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)(q'_s)^{Shift}+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}(s'_i)^{Shift}}$$

In this last inequality, the left hand side is the critical b/c ratio for the BD process, and the right hand side is the critical b/c ratio for the Shift process, given in Conditions (2.48) and (2.49), respectively.

2A.6.3 Limit results for the number of groups approaching infinity

In this section, we explore the results in the limit where the number of groups m approaches infinity.

BD For the BD process, we first repeat and rewrite Condition (2.45)

$$\begin{aligned}\frac{b}{c} &> \frac{p}{q}\frac{1}{n}\frac{m}{m-1}\frac{\frac{n-1}{n}q'_s}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)q'_s+\frac{p+nq}{2r}} \\ &= \frac{p}{q}\frac{1}{n}\frac{m}{m-1}\frac{\frac{n-1}{n}}{\frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{q'_s}}\end{aligned}$$

as well as Equation (2.27)

$$q'_s = \frac{-\left(\frac{p}{n}+q\right)\left(m-\frac{1}{n}+\frac{qn}{2r}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}$$

Since $\lim_{m\to\infty} q'_s = \infty$, the critical b/c ratio for the BD process converges to

$$\begin{aligned}\frac{b}{c} &> \frac{p}{q}\frac{1}{n}\frac{\frac{n-1}{n}}{\frac{p}{r}\frac{1}{2n}+\frac{1}{n}} \\ &= \frac{p}{q}\frac{n-1}{n}\frac{1}{\frac{p}{2r}+1}\end{aligned} \tag{2.53}$$

The limit result of the critical ratio we found in Condition (2.53) is in line with numerical solutions for large m.

Shift For the Shift process, we first repeat and rewrite Condition (2.46)

$$\begin{aligned}\frac{b}{c} &> \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{q'_s}{\left(\frac{p}{r}\frac{1}{2n}+\frac{1}{n}\right)q'_s+\frac{p+nq}{2r}+\frac{1}{m-1}\sum_{i=1}^{m-1}s'_i}\\ &= \frac{p}{q}\frac{n-1}{n^2}\frac{m}{m-1}\frac{1}{\frac{p}{r}\frac{1}{2n}+\frac{1}{n}+\frac{p+nq}{2r}\frac{1}{q'_s}+\frac{1}{m-1}\frac{1}{q'_s}\sum_{i=1}^{m-1}s'_i}\end{aligned}$$

We also have, from Equations (2.39) and (2.38), reproduced below, where $\eta_i = \frac{\frac{p}{n}+q}{q\frac{i(m-i)}{m}+r\frac{2}{n}}$ is given by Equation (2.34).

$$\begin{aligned}q'_s &= \frac{-\left(\frac{p}{n}+q\right)\left(\sum_{i=2}^{m-2}\frac{q\frac{m-1}{m}+r\frac{2}{n}}{q\frac{i(m-i)}{m}+r\frac{2}{n}}+3-\frac{1}{n}+\frac{q}{r}\frac{n}{2}\frac{m-1}{m}\right)}{\frac{pq}{r}\frac{1}{2n}\frac{m-1}{m}+q\frac{1}{n}\frac{m-1}{m}+p\frac{n-1}{n^3}}\\ \sum_{j=1}^{m-1}s'_j &= -2(m-3)\eta_2+\sum_{i=3}^{m-3}\left(\frac{m^2-m-2-2i(2m-3)+2i^2}{2}\right)\eta_i\end{aligned}$$

Now consider the sum in the numerator in q'_s. Since each term in this summation is less than 1, we have

$$\sum_{i=2}^{m-2}\frac{m-1+m\frac{r}{q}\frac{2}{n}}{i(m-i)+m\frac{r}{q}\frac{2}{n}} < m-3$$

and hence q'_s decreases at most proportional to m. On the other hand, $\sum_{j=1}^{m-1}s_j$ decreases proportional to m^2. Therefore, the term with $\sum_{j=1}^{m-1}s_j$ in the denominator of the critical b/c ratio for the Shift process dominates in the limit $m\to\infty$, and hence, the critical b/c ratio for Shift converges to

$$\frac{b}{c} > 0 \tag{2.54}$$

The limit result of the critical ratio we found in Condition (2.54) is in line with numerical solutions for large m.

Birth-Death versus Shift in the limit where number of groups approach infinity In the previous subsection, we solved for the critical b/c ratios in the limit where the number of groups m approaches infinity. The limit results are repeated below for convenience.

$$\frac{b}{c} > \frac{p}{q}\frac{n-1}{n}\frac{1}{\frac{p}{2r}+1}$$

for BD, and

$$\frac{b}{c} > 0$$

for Shift. From these formulas, it is immediately clear that the critical b/c ratio for Shift is lower than the one for BD in the limit $m \to \infty$.

2A.7 Simulations

Since our model quickly becomes intractable once we move away from the limit of weak selection, we also ran numerical simulations with different intensities of selection. In this section, we describe the details of those simulations.

We programmed the simulation version of model, presented in Section 2A.3, in Matlab, where the population state is represented by a vector at each time step. In the remainder of the text, we will call this "the population vector". Each entry in the population vector represents the number of cooperators in a group at a given location on the circle. The group size is fixed in our model, and therefore the number of defectors is implied by the number of cooperators. This vector represents a circle, and therefore the first and the last entry in the vector are treated like neighbours in any relevant group competition or migration event. At the end of each time step, we update the population vector depending on the changes that happened in that time step. Each simulation run starts with a population vector of zeros, except for the first entry, which is a 1. This represents a mutant cooperator in a population of defectors. Since the cycle we use in our model is a transitive graph (Taylor et al., 2007a), the location of the initial mutant does not matter.[4] As mentioned in Section 2A.3, there are only two absorbing states in our model: one in which there are only cooperators and one with only defectors. We made sure that each individual simulation run presented in our results was long enough such that it reached either one of the two absorbing states.

We used random draws to decide what type of event – individual, group, or migration – happens in a given time step, and to determine the details of the corresponding event.[5]

[4] We ran robustness checks where we placed the initial mutant at different locations on the population vector, and the results did not change significantly. Therefore, we decided to keep the location of the mutant fixed in the first group.

[5] In all of the random draws in our simulations, we used the *rand* function in Matlab. For the cases where we needed integers, we used the appropriate combination of our population size parameters and the *ceil* function in Matlab. For example, when we needed to choose a group randomly, we used the combination *ceil(rand*m)*, where *rand* draws a number from the uniform distribution on the interval

At each time step, we draw a random number from a uniform distribution between 0 and 1,[6] and,

- **if the random draw is lower than p, an individual event happens.** In this case, we draw a random integer between 1 and m to decide in which group the individual event will occur. Individual payoffs within the group are defined as presented in Section 2A.3; cooperators within the group get a payoff of $1 - wc$, and defectors get a payoff of 1. Using these payoffs and the number of cooperators in the group that is chosen for the individual event, we define the probabilities with which the population state changes in this particular event as follows. Suppose that there are k_i cooperators in the group, then with probability p^+, the number of cooperators in the group increases by one – a cooperator is chosen to reproduce and a defector is chosen to die – and, with probability p^-, the number of cooperators decreases by one – a defector is chosen to reproduce and a cooperator is chosen to die – where

$$p^+ = \frac{k_i(1-wc)}{n-k_iwc}\frac{n-k_i}{n}$$
$$p^- = \frac{n-k_i}{n-k_iwc}\frac{k_i}{n}$$

  With the remaining probability $1 - p^+ - p^-$, the number of cooperators does not change, which would represent the case in which the reproducing and the dying individuals are both cooperators, or the case in which because they are both defectors. After defining the probabilities p^+ and p^- within the group, we draw another uniform random variable between 0 and 1 to decide which of the above events happens within the group. If the random number is lower than p^+, the number of cooperators in the group increases by one; if it is higher than or equal to p^+ but lower than $p^+ + p^-$, the number of cooperators in the group decreases by one; and, if it is higher than $p^+ + p^-$, the number of cooperators in the group does not change.

- **if the random draw is higher than or equal to p but lower than p + q, a group event happens.** In this case, we define the group payoff for group i as $g(k_i) = 1 + w\frac{k_i}{n}b$, where k_i is the number of cooperators in group i at the beginning of the current time step. Then, based on our model, group i reproduces with probability

$[0, 1]$, multiplying this number by m turns it into a random draw from the uniform distribution on the interval $[0, m]$, and *ceil* rounds the numbers up such that we end up with only integers between 1 and m, where each integer is equally likely to be drawn.

[6]The default precision of the *rand* function in Matlab is 4 decimal points, therefore, the probability of drawing a specific number is approximately zero. For our decisions on what type of events are happening in the simulations, it does not matter substantially where we place the cases of the random draws being equal to certain numbers, e.g. that the random draw is exactly equal to p.

$q(k_i, K)$, where

$$q(k_i, K) = \frac{g(k_i)}{\sum_{j=1}^{m} g(k_j)} = \frac{1 + w\frac{k_i}{n}b}{n + w\frac{K}{n}b}$$

and $K = \sum_{j=1}^{m} k_j$ is the total number of cooperators in the population. Together, these give n thresholds; threshold j is $\sum_{i=1}^{j} q(k_i, K)$. Then we draw a uniform random number between 0 and 1, and if it is between threshold $j-1$ and j, group j is chosen to reproduce. Then, two additional uniform random numbers are drawn to choose the dying group and the location of the offspring group. In the BD process, one of the random draws is compared to $1/m$ to determine if the reproducing group is chosen to die or not. The other random draw is compared to $1/2$ to decide whether the group above or below in the population vector is chosen to die, in case the reproducing group is not chosen to die. The population vector is updated such that the number of cooperators in the place of the dying group becomes equal to the number of cooperators in its reproducing neighbour. In the Shift process, one of the random draws is used to determine which of the groups from the whole population of groups is chosen to die, and the other random draw is compared to $1/2$ to determine the direction in which the offspring group is placed in the population vector. Every entry between the reproducing group and the dying group in the vector is shifted by one spot in the chosen direction.

- **if the random draw is higher than or equal to $\mathbf{p+q}$, a migration event happens.** By drawing a random integer between 1 and m, we choose the first group to take part in the migration event. In our model, migration events occur between immediate neighbours. With the use of another random draw, we decide which neighbour of the first group is going to take part in the migration event as well. Within each group, a random individual is picked to migrate in our model, and, only in the case that the individuals migrating are of different types, i.e. one is a cooperator and the other is a defector, the population state changes. Making use of the number of cooperators given by the population vector, we calculate the probabilities that the first group sends a cooperator or a defector, and the probabilities that the second group sends a cooperator or a defector. Then, we draw two random numbers to decide what type of player is sent by each group (this is done in a similar fashion to other type of choices mentioned above).

Depending on the type of the event, and what happens in a given event, the population vector at the end of each time step is updated as described. This procedure is repeated

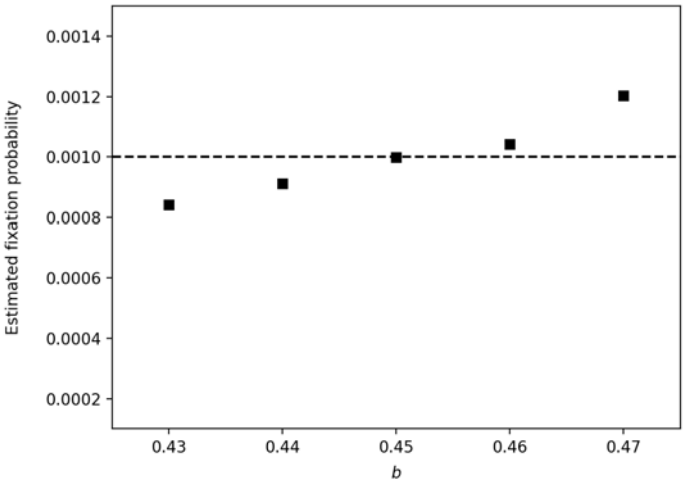


Figure 2A.1: An example of the last few steps in the iterative process of finding the critical b/c-ratio. Here, $m = 50$ and $n = 20$, which makes the fixation probability in the neutral process $\frac{1}{1000}$. Furthermore, this is the Birth-Death process, $c = 0.1$, $w = 0.5$, $p = \frac{18}{21}$, $q = \frac{9}{210}$, and $r = \frac{1}{10}$. In this particular instance, the number of fixations for $b = 0.45$ was almost spot on (999 out of 1,000,000), so we settled on a b/c-ratio of 4.5. In other instances, we rounded to the nearest value for b with 2 digits precision. This procedure results in one critical b/c-ratio, and therefore one point in Figure 2A.2a.

until we reach one of the absorbing states, or once the maximum time steps set is reached (in the latter case, we extend the maximum time steps and re-run the same setup until the run ends due to reaching fixation). At the end of each individual simulation run, we record whether cooperators or defectors fixated in that run.

This simulation program is used for finding critical b/c-ratio's for different choices of n, m, p, q, and r. For any given combination of n, m, p, q, and r, we do this by choosing $c = 0.1$, choosing a b, and running the simulation program a 1,000,000 times. This gives an estimate of the fixation probability, which we compare to $\frac{1}{nm}$, which is the fixation probability under neutral selection. The b is then adjusted accordingly, until we find a fixation probability that is indistinguishable from $\frac{1}{nm}$.

2A.8 Theoretical results and simulations

In this section, we take the critical b/c ratios in the limit of weak selection, calculated in Section 2A.4, combine them, as we did in Section 2A.6, with the relatednesses in the limit of weak selection, calculated in Section 2A.5, and plot them for a variety of parameter combinations. We combine those analytical results in the limit of weak selection with simulation results not in the limit of weak selection. As we vary group size n, the number of groups m, and migration rate r, we want to choose the probabilities of group versus individual events such that the ratio of probabilities for an individual to die in an individual event and in a group event remains constant under neutral selection. In this case, we choose them so that these probabilities are always equally large. At neutrality, the probability that an individual dies in an individual event is $p\frac{1}{m}\frac{1}{n}$ and the probability that an individual dies in a group event is $q\frac{1}{m}$. Keeping them equal therefore requires

$$p\frac{1}{m}\frac{1}{n} = q\frac{1}{m}$$

$$p = nq$$

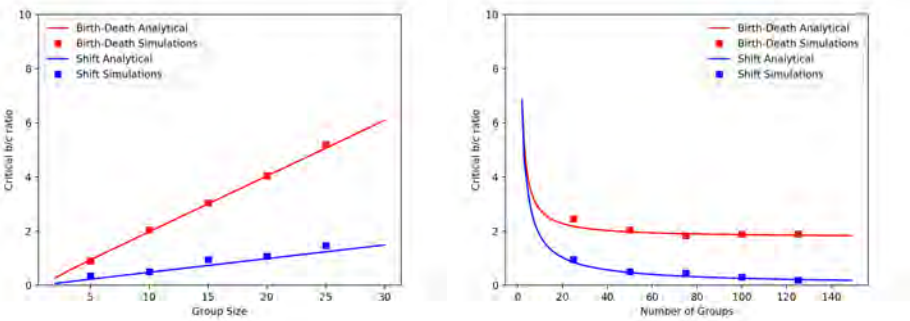
We also have the condition $p + q + r = 1$. Together with the equality above, this implies that

$$p + q = 1 - r$$
$$(n+1)q = 1 - r$$
$$q = (1-r)\frac{1}{n+1} \rightarrow p = (1-r)\frac{n}{n+1}$$

In Figure 2A.1 and Figure 2A.2, we choose $r = 0.1$, and in Figure 2A.3, r varies.

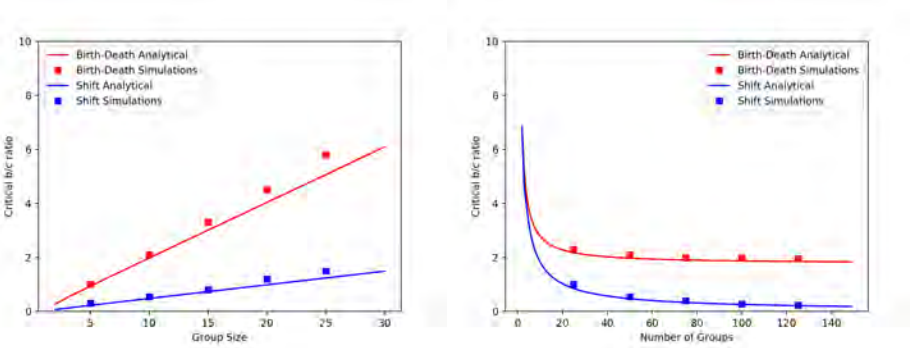
In Figure 2A.1a and Figure 2A.2a, we see that the thresholds in both processes increase with group size, and that the gap between the critical b/c ratios for the two processes is there for a range of group sizes. In Figure 2A.1b and Figure 2A.2b, we see that the thresholds decrease with the number of groups, and, again, that the gap between the critical b/c ratios for the two processes is there for a range of numbers of groups. Both the increase of the thresholds with group size and the decrease with the number of groups are in line with the results of Traulsen and Nowak (2006). The gap between the thresholds for the two processes in relative terms becomes (very) large for (very) large numbers of groups.

In Figure 2A.3, we see that both thresholds increase with the migration rate. This is



(a) $m = 50$ **(b)** $n = 10$

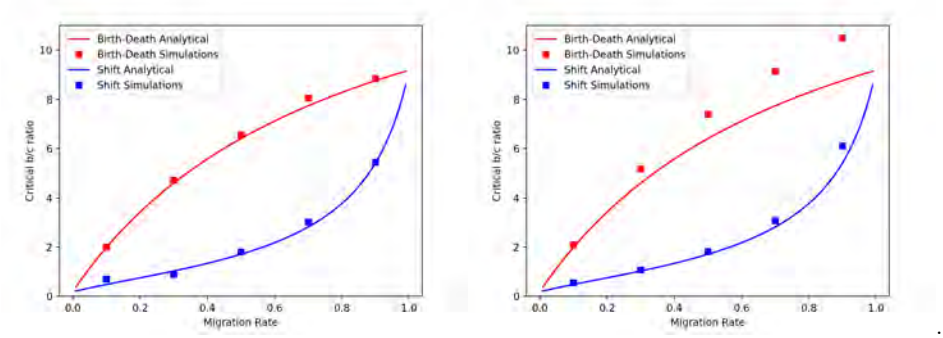
Figure 2A.1: Results for the critical b/c ratios, combined with simulations with $1,000,000$ independent runs, where $w = 0.1$, $r = 0.1$, $p = 0.9\frac{n}{n+1}$ and $q = 0.9\frac{1}{n+1}$. In **(a)**, the effect of increasing the group size is shown for a fixed number of groups. In **(b)**, the effect of increasing the number of groups is shown for a fixed group size.



(a) $m = 50$ **(b)** $n = 10$

Figure 2A.2: Same as Figure 2A.1, but with $w = 0.5$.

understandable, since a higher migration rate will result in lower relatedness. We also see that the gap disappears for migration rates close to 0 and migration rates close to 1. For migration rates close to 0, and not too few groups, there is a fair chance that the process will spend a lot of time in population states where all groups are at within-group fixation, or, in other words, where all groups consist of only cooperators or only defectors. Once this has happened, there is a fair chance that the population will get to a state where these all-cooperator groups form one sequence of groups on the circle. Without mutation or migration, this will remain true from then onwards. In Section 2A.9, we will see that if we imagine a mutant group, where all individuals are cooperators, the fixation probability



(a) $w = 0.1$ **(b)** $w = 0.5$

Figure 2A.3: Results for the critical b/c ratios for different migration rates r, for $n = 10, m = 50$, again combined with simulations with $1,000,000$ independent runs.

of such a mutant group is the same in either process. If cooperators on average gain ground on (or lose ground to) defectors, they do so faster with Shift than with BD, but the condition under which they gain or loose is the same for both processes. The reason for the difference in speed is that with Shift, in all intermediate population states with a string of all-cooperator groups and a string of all-defector groups, any all-cooperator group has a real chance of replacing any all-defector group and vice versa. On the other hand, in BD, all the action is at the two boundaries between the strings of all-cooperator groups and all-defector groups, while group reproduction events not on the boundary are inconsequential for the population state. This implies that the boundary moves much more with Shift than it does with BD, and if the boundary moves in expectation in favour of the all-cooperator groups, it will move faster with Shift than with BD. However, the speed turns out not to matter for the fixation probability at the group level of a mutant all-cooperator group. The speed does matter if there are mixed groups, because if there are mixed groups, the ground gained at the group level will balance against the ground lost at the individual level. With cooperative groups winning faster in Shift, they can therefore overcome a larger within-group decay of cooperators. Absent mixed groups, however, speed does not matter anymore, and the difference in fixation probabilities disappears.

At migration rates r close to 1, it is not hard to understand why the difference between BD and Shift disappears as well. When almost all events are migration events, the population is shaken and stirred between any two reproduction events, be it at the individual or at the group level, and all that matters is what being a cooperator does to an individual's own birth rate and what it does to its group's birth rate. Those effects are the same in

both processes, and therefore the difference between the two processes should disappear at very high migration rates.

This is also reflected in the formulas for the thresholds. In the limit of $r \uparrow 1$, only finite population effects remain, and relatednesses become $r_s = r_1 = -\frac{1}{nm-1}$ and $r_0 = \frac{1}{n} + \frac{n-1}{n} r_s = \frac{1}{n} - \frac{n-1}{n}\frac{1}{nm-1} = \frac{m-1}{nm-1}$. This implies that $1 - r_0 = \frac{(n-1)m}{nm-1}$ and $r_0 - r_1 = \frac{m}{nm-1}$.

Condition (2.7), the critical b/c-ratio for BD, then becomes

$$\frac{b}{c} > \frac{p}{q}\frac{m}{m-1}\frac{1-r_0}{n(r_0-r_1)} = \frac{p}{q}\frac{m}{m-1}\frac{n-1}{n}$$

Condition (2.9), the critical b/c-ratio for Shift, then becomes

$$\frac{b}{c} > \frac{p}{q}\frac{1-r_0}{nr_0} = \frac{p}{q}\frac{m}{m-1}\frac{n-1}{n}$$

These thresholds are the same.

2A.9 Analytical solutions, without migration, and for limit cases that are not the limit of weak selection

There are some limit cases, other than the limit of weak selection, for which it also becomes feasible to derive analytical solutions for fixation probabilities. These results will help understand why the difference between the critical b/c ratios for BD and for Shift disappears when migration rates get close to 0. In both limits we consider here, we assume that there is no migration ($r = 0$), we assume strong selection ($w = 1$), and we assume that there is a separation of timescales. In the first limit, where $p \to 1$, group events are rare, and in almost all events, an individual reproduces. This limit is similar to the limit for which Traulsen and Nowak (2006) derive their formula [**1**]. They do consider the limit of weak selection, while we consider strong selection – which means that by scaling the b and c, we can really consider any intensity of selection. In the second limit, where $p \to 0$, individual events are rare, and in almost all events, a group reproduces.

2A.9.1 Fixation probabilities of mutant groups

To compute fixation probabilities in both limits, it will be useful first to compute the fixation probability of a "mutant group", with l cooperators in it, in a population of

groups, all of which contain k cooperators. This is the process at the group selection time scale[7]. It will only visit states in which there is one string of subsequent groups with k cooperators, and one string of subsequent groups with l cooperators – besides of course the absorbing states, where all groups have k cooperators or all groups have l cooperators. Therefore, the population state can be denoted by a single number i, $0 \leq i \leq m$, representing the number of groups with l cooperators.

2A.9.2 Birth-Death

For the Birth-Death process, the only way in which i can change is if one of the groups on the two boundaries between the strings of different group types is chosen for reproduction; one of the two k-groups on the boundary if i is to go down, and one of the two l-groups on the boundary if i is to go up. Moreover, it needs to replace the group on the other side of the boundary, and not its other neighbour or itself, which happens with probability $\frac{1}{2}\frac{m-1}{m}$. The probability of a transition from i to $i-1$ in the Birth-Death process then becomes $T^-_{i,BD} = 2 \cdot \frac{1}{2}\frac{m-1}{m} \cdot \frac{1+\frac{k}{n}b}{i(1+\frac{l}{n}b)+(m-i)(1+\frac{k}{n}b)}$, while the probability of a transition from i to $i+1$ is $T^+_{i,BD} = 2 \cdot \frac{1}{2}\frac{m-1}{m} \cdot \frac{1+\frac{l}{n}b}{i(1+\frac{l}{n}b)+(m-i)(1+\frac{k}{n}b)}$. That makes the down-up ratio for the Birth-Death process equal to

$$\frac{T^-_{i,BD}}{T^+_{i,BD}} = \frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}$$

2A.9.3 Shift

For the Shift process, change is much more likely. State i changes any time the group that is chosen for reproduction is of a different type than the group that is chosen for death. The probability of a transition from i to $i-1$ is $T^-_{i,Shift} = \frac{(m-i)(1+\frac{k}{n}b)}{i(1+\frac{l}{n}b)+(m-i)(1+\frac{k}{n}b)}\frac{i}{m-1}$, while the probability of a transition from i to $i+1$ is $T^+_{i,Shift} = \frac{i(1+\frac{l}{n}b)}{i(1+\frac{l}{n}b)+(m-i)(1+\frac{k}{n}b)}\frac{m-i}{m-1}$. That makes the down-up ratio for the Shift process also equal to

$$\frac{T^-_{i,Shift}}{T^+_{i,Shift}} = \frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}$$

[7]The fixation probabilities follow more or less directly from the fixation probabilities in the standard Birth-Death process on the cycle (Nowak, 2006; Ohtsuki and Nowak, 2006), or the standard Shift process (Allen and Nowak, 2012). The only difference is that in the standard versions, there are individuals on a cycle, and here we have groups on a cycle instead.

2A.9.4 Birth-Death and Shift

Since the down-up ratios are the same for both, the fixation probability of a group with l cooperators, while all other groups have k of them, is also the same for Birth-Death and Shift:

$$\tau_{k\to l}=\frac{1}{1+\sum_{j=1}^{m-1}\prod_{i=1}^{j}\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}}=\frac{1}{1+\sum_{j=1}^{m-1}\left(\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}\right)^{j}}=\frac{1}{1+\frac{1-\left(\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}\right)^{m}}{1-\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}}-1}=\frac{1-\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}}{1-\left(\frac{1+\frac{k}{n}b}{1+\frac{l}{n}b}\right)^{m}}$$

This can be rewritten as

$$\tau_{k\to l}=\frac{\left(1+\frac{l}{n}b\right)-\left(1+\frac{k}{n}b\right)}{\left(1+\frac{l}{n}b\right)^{m}-\left(1+\frac{k}{n}b\right)^{m}}\frac{\left(1+\frac{l}{n}b\right)^{m}}{1+\frac{l}{n}b}=\frac{\frac{l-k}{n}b\left(1+\frac{l}{n}b\right)^{m-1}}{\left(1+\frac{l}{n}b\right)^{m}-\left(1+\frac{k}{n}b\right)^{m}}$$

In the limit of $p\to 1$, we will use $\tau_{0\to n}$ and $\tau_{n\to 0}$, and in the limit of $p\to 0$, we will use $\tau_{k\to k+1}$ and $\tau_{k\to k-1}$ for $1\leq k\leq m-1$.

2A.9.5 The limit $p\to 1$

In this limit, group events are rare, while there are individual events almost all of the time. This implies that within her group, an initial mutant will either have gone extinct, or gone to fixation, before a group event happens. With the separation of timescales, we can first concentrate on the probabilities of fixation within the group happening. With j being the number of cooperators in the group, a defector replaces a cooperator within the group, and a cooperator replaces a defector, with probabilities

$$T_j^-=\frac{n-j}{n-jc}\frac{j}{n}\qquad\text{and}\qquad T_j^+=\frac{j(1-c)}{n-jc}\frac{n-j}{n}$$

The down-up ratio therefore is $\frac{T_j^-}{T_j^+}=\frac{1}{1-c}$, making the fixation probabilities within the group

$$\sigma_C=\frac{1-\frac{1}{1-c}}{1-\left(\frac{1}{1-c}\right)^{n}}=\frac{c(1-c)^{n-1}}{1-(1-c)^{n}}\qquad\text{and}\qquad\sigma_D=\left(\frac{1}{1-c}\right)^{n-1}\frac{1-\frac{1}{1-c}}{1-\left(\frac{1}{1-c}\right)^{n}}=\frac{c}{1-(1-c)^{n}}$$

After fixation within the group, all groups will be homogeneous. As soon as the population consists of homogeneous groups, the population state can only change with a group event. Therefore, the fixation probability of a mutant will be the product of the fixation probability of the mutant within the group – which is the same for both processes, because

this does not involve group events – and the fixation probability of the all-mutant group in the population – which is also the same for both processes, as we have seen above.

$$\rho_C = \sigma_C \tau_{0\to n} = \frac{1-\frac{1}{1-c}}{1-\left(\frac{1}{1-c}\right)^n} \frac{1-\frac{1}{1+b}}{1-\left(\frac{1}{1+b}\right)^m} = \frac{c(1-c)^{n-1}}{1-(1-c)^n} \frac{b(1+b)^{m-1}}{(1+b)^m-1}$$

and

$$\rho_D = \sigma_D \tau_{n\to 0} = \frac{1-(1-c)}{1-(1-c)^n} \frac{1-(1+b)}{1-(1+b)^m} = \frac{c}{1-(1-c)^n} \frac{b}{(1+b)^m-1}$$

If we do consider the limit of weak selection, then we can approximate σ_C with $\frac{1}{n}\left(1-\frac{n-1}{2}c\right)$ for small c, and $\tau_{0\to n}$ with $\frac{1}{m}\left(1+\frac{m-1}{2}b\right)$ for small b. That implies that in the limit of weak selection, $\rho_C > \frac{1}{nm}$ if

$$\frac{b}{c} > \frac{n-1}{m-1}.$$

Condition [**1**] in Traulsen and Nowak (2006) is

$$\frac{b}{c} > 1 + \frac{n}{m}.$$

This condition applies to their model without migration, and also with almost all events being individual replacements. In their model, the individual reproduction rate in all-cooperators groups (and therefore also their group reproduction rate) is $b-c$ higher than the individual reproduction rate in all-defector groups. In our model, the group reproduction rate is increased by b, which explains why there is a 1 on the right hand side in their inequality, which one can also write as $\frac{b-c}{c} > \frac{n}{m}$, and not in ours.

Similarly, we can approximate σ_D with $\frac{1}{n}\left(1+\frac{n-1}{2}c\right)$ for small c, and $\tau_{n\to 0}$ with $\frac{1}{m}\left(1-\frac{m-1}{2}b\right)$ for small b. In the limit of weak selection, that gives the same threshold for when $\rho_D < \frac{1}{nm}$.

Traulsen and Nowak (2006) also consider a version with migration, but because the migration rate is taken to be proportional to the probability that an individual reproduction event induces the group to split (which is vanishingly small), this implies that these events are also only happening occasionally, and if they do, the migrants almost always either take over their new group or go extinct there before anything else happens that is not an individual event.

2A.9.6 The limit $p \to 0$

In this limit, individual events are rare, while there are group events almost all of the time. The population will therefore regularly be in a state in which all groups have the same composition. When each group has the same number of cooperators, the population state can only change as the result of an individual event. If it does, then that event can make the number of cooperators within one group go up by one, or go down by one. After that, a sequence of group events either make the new "mutant group" go extinct, or go to fixation. Fixation of the mutant group happens with probability $\tau_{k\to k+1}$ if the individual event had a cooperator replace a defector, and with probability $\tau_{k\to k-1}$ if a defector replaced a cooperator. With rare individual events, the population, on the larger time scale, therefore moves between states where all groups have the same number of cooperators, and this number goes up or down by at most 1. The probability that it goes down by one is the probability with which a defector reproduces, which is $\frac{n-k}{n-kc}$, times the probability that a cooperator dies, which is $\frac{k}{n}$, times the fixation probability of the "mutant group", which is $\tau_{k\to k-1}$. The probability that it goes up by one is the probability with which a cooperator reproduces, which is $\frac{k(1-c)}{n-kc}$, times the probability that a defector dies, which is $\frac{n-k}{n}$, times the fixation probability of the "mutant group", which is $\tau_{k\to k+1}$. The down-up ratio at the larger timescale therefore is

$$\frac{\mathbf{T}_k^-}{\mathbf{T}_k^+} = \frac{1}{1-c}\frac{\tau_{k\to k-1}}{\tau_{k\to k+1}}$$

They can be smaller or larger than 1, depending on the parameters. With these down-up ratios at the larger timescale, we find the overall fixation probabilities of cooperators and defectors.

$$\lim_{p\to 0}\rho_C = \tau_{0\to 1}\frac{1}{1+\sum_{k=1}^{n-1}\prod_{l=1}^{k}\frac{\mathbf{T}_l^-}{\mathbf{T}_l^+}} = \tau_{0\to 1}\frac{1}{1+\sum_{k=1}^{n-1}\left(\frac{1}{1-c}\right)^k\prod_{l=1}^{k}\frac{\tau_{l\to l-1}}{\tau_{l\to l+1}}}$$

and

$$\lim_{p\to 0}\rho_D = \tau_{n\to n-1}\frac{\prod_{k=1}^{n-1}\frac{\mathbf{T}_k^-}{\mathbf{T}_k^+}}{1+\sum_{k=1}^{n-1}\prod_{l=1}^{k}\frac{\mathbf{T}_l^-}{\mathbf{T}_l^+}} = \tau_{n\to n-1}\frac{\left(\frac{1}{1-c}\right)^{n-1}\prod_{l=1}^{n-1}\frac{\tau_{l\to l-1}}{\tau_{l\to l+1}}}{1+\sum_{k=1}^{n-1}\left(\frac{1}{1-c}\right)^k\prod_{l=1}^{k}\frac{\tau_{l\to l-1}}{\tau_{l\to l+1}}}$$

The fixation probability ρ_C of a cooperator therefore is larger than the fixation probability ρ_D of a defector, if

$$\tau_{0\to 1} > \tau_{n\to n-1}\left(\frac{1}{1-c}\right)^{n-1}\prod_{l=1}^{n-1}\frac{\tau_{l\to l-1}}{\tau_{l\to l+1}},$$

or, in other words, if

$$1 > \left(\frac{1}{1-c}\right)^{n-1} \frac{\prod_{l=1}^{n} \tau_{l\to l-1}}{\prod_{l=0}^{n-1} \tau_{l\to l+1}}$$
$$1 > \frac{1}{(1-c)^{n-1}} \frac{1}{(1+b)^{m-1}}$$

In the limit of weak selection, this can be approximated with

$$1 > (1+(n-1)c)\,(1-(m-1)b)\,,$$

and, still in the limit of weak selection, this is true if

$$\frac{b}{c} > \frac{n-1}{m-1}.$$

The effect of the different separations of time scales therefore is the same in the limit of weak selection, and in both of them, the cancellation effect dissipates; in the first, because there are almost no mixed groups, in the second because the population almost never consists of differently composed groups. Note also that this formula makes perfect sense for $n = 1$, in which case any positive group benefit will make cooperators do better than defectors, and for $m = 1$, in which case no benefit is ever high enough to offset positive costs of cooperation.

2A.10 Empirical implications

2A.10.1 F_{ST}

Many empirical papers estimate F_{ST}'s. Our definition of r_s follows the definition of relatedness in Rousset (2004) and Durrett (2008), applied to members of the same group, and our definition of relatedness between members of different groups applies the same definition to individuals from different groups. Many empiricists however use other definitions. We assume that the definitions empiricists use are equivalent to the definition from Rousset (2004) and Durrett (2008), applied to members of the same group, but we have not found a reference where this is proven formally. What we will do below is to indicate how the two definitions used by most empiricists are the same as the probabilistic definition of relatedness within the group that we also used at the end of Section 2A.5. There, relatedness between two individuals in the same group, for a given population state k, is defined as $r_{s,k} = P_{s,k}\,(C|C) - P_{s,k}\,(C|D)$, where the probabilities refer to the outcomes

of two subsequent draws from the same group, without replacement. Relatedness r_s is then defined as the weighted average $r_s = \sum_k q_k r_{s,k}$ of those relatednesses over different states, where the weights q_k of states are given by the invariant distribution of the Markov chain in the limit of no mutation, normalized after excluding the absorbing states where all individuals are cooperators, or all are defectors (Allen and Tarnita, 2014).

Our model has groups of equal size, and let f_i be the fraction of groups with i cooperators out of n group members in population state k (we suppress the population state k in the notation). The overall frequency of cooperators and defectors follow from these fractions of group types: $\sum_{i=0}^{n} f_i \frac{i}{n} = p$ and $\sum_{i=0}^{n} f_i \frac{n-i}{n} = 1-p$. Also the fractions of group types must add up to 1; $\sum_{i=0}^{n} f_i = 1$. Then, we can rewrite $r_{s,k}$ as

$$
\begin{aligned}
r_{s,k} = P_{s,k}(C|C) - P_{s,k}(C|D) &= \frac{\sum_{i=0}^{n} f_i \frac{i}{n} \frac{i-1}{n-1}}{\sum_{i=0}^{n} f_i \frac{i}{n}} - \frac{\sum_{i=0}^{n} f_i \frac{n-i}{n} \frac{i}{n-1}}{\sum_{i=0}^{n} f_i \frac{n-i}{n}} = \frac{\sum_{i=0}^{n} f_i \frac{i}{n} \frac{i-1}{n-1}}{p} - \frac{\sum_{i=0}^{n} f_i \frac{n-i}{n} \frac{i}{n-1}}{1-p} \\
&= \frac{\sum_{i=0}^{n} f_i \frac{i}{n} \frac{i-1}{n-1}(1-p)}{p(1-p)} - \frac{\sum_{i=0}^{n} f_i \frac{n-i}{n} \frac{i}{n-1} p}{p(1-p)} \\
&= \frac{\sum_{i=0}^{n} f_i \frac{i}{n} \left(1 - \frac{n-i}{n-1}\right)(1-p)}{p(1-p)} - \frac{\sum_{i=0}^{n} f_i \frac{n-i}{n} \frac{i}{n-1} p}{p(1-p)} \\
&= \frac{\sum_{i=0}^{n} f_i \frac{i}{n}(1-p) - \sum_{i=0}^{n} f_i \frac{i}{n} \frac{n-i}{n-1} + \sum_{i=0}^{n} f_i \frac{i}{n} \frac{n-i}{n-1} p}{p(1-p)} - \frac{\sum_{i=0}^{n} f_i \frac{n-i}{n} \frac{i}{n-1} p}{p(1-p)} \\
&= \frac{p(1-p) - \sum_{i=0}^{n} f_i \frac{i}{n} \frac{n-i}{n-1}}{p(1-p)}
\end{aligned}
$$

The final expression matches the way F_{ST} is regularly defined (see Equation 2.55), except for one detail. With groups numbered from $j = 1, ..., m$, the normal definition would have $\sum_j^m p_j(1-p_j)$ as a second term in the numerator, where $p_j(1-p_j)$ is the variance within group j, and $p(1-p)$ is the variance in the population as a whole. In our case, with f_i counting the numbers of groups with i cooperators out of n group members, this average within group variance would be $\sum_{i=0}^{n} f_i \frac{i}{n} \frac{n-i}{n}$. In our formula, we have $\sum_{i=0}^{n} f_i \frac{i}{n} \frac{n-i}{n-1}$ instead.

Here it is worth noting that the variance is half the expected squared difference between two independent draws from the same distribution. If X_1 and X_2 are two independent draws for the same random variable (which therefore have the same first and second moment), then the expected value of the squared difference between them is

$$E[X_1 - X_2]^2 = E\left[X_1^2 - X_2^2 - 2X_1X_2\right] = 2E\left[X_1^2\right] - E^2[X_1] = 2Var(X_1)$$

In our case, X_1 and X_2 represent drawing an individual from one and the same group,

where drawing a cooperator makes the value of X to be 1, and drawing a defector makes it 0. If we draw two individuals without replacement, instead of with replacement, then they are no longer independent. We can however still compute the expected squared difference between the draws. This is

$$\frac{i}{n}\frac{i-1}{n-1}\cdot 0^2+\frac{i}{n}\frac{n-i}{n-1}\cdot 1^2+\frac{n-i}{n}\frac{i}{n-1}\cdot 1^2+\frac{n-i}{n}\frac{n-i-1}{n-1}\cdot 0^2=2\frac{i(n-i)}{n(n-1)}$$

Half of this is what is used instead of the within-group variance. For every state k, the "without-replacement" version of the F_{ST} is therefore equivalent to the definition as a difference in conditional probabilities $r_{s,k}=P_{s,k}\left(C|C\right)-P_{s,k}\left(C|D\right)$. For large group sizes n, the difference between the with and without replacement version of the within-group variance disappears.

$$F_{ST}=\frac{\text{population variance}-\text{avergage within group variance}}{\text{population variance}} \tag{2.55}$$

This definition of the F_{ST} features both the average within-group variance $\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}$ (with replacement) and the overall variance. There is yet another definition of F_{ST}, which expresses it in terms of within-group and between-group variance. We can take the following steps to get to that expression.

$$\begin{aligned}
r_{s,k} &= \frac{p(1-p)-\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}}{p(1-p)}\\
&= \frac{p-\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}-p^2}{p-\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}+\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2}\\
&= \frac{\sum_{i=0}^{n} f_i\frac{i}{n}-\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}-p^2}{\sum_{i=0}^{n} f_i\frac{i}{n}-\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}+\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2}\\
&= \frac{\sum_{i=0}^{n} f_i\frac{i}{n}\left(1-\frac{n-i}{n-1}\right)-p^2}{\sum_{i=0}^{n} f_i\frac{i}{n}\left(1-\frac{i-1}{n-1}\right)+\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2}\\
&= \frac{\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2}{\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}\quad+\quad\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2}
\end{aligned}$$

Here $\sum_{i=0}^{n} f_i\frac{i}{n}\frac{n-i}{n-1}$ is the without replacement version of within-group variance as before. If we call $\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-p^2=\sum_{i=0}^{n} f_i\frac{i}{n}\frac{i-1}{n-1}-\left(\sum_{i=0}^{n} f_i\frac{i}{n}\right)^2$ the without replacement version of between-group variance, we would read this as

$$F_{ST}=\frac{\text{between-group variance}}{\text{average within-group variance}+\text{between-group variance}} \tag{2.56}$$

Again, for large group sizes n, the with or without replacement versions are close.

2A.10.2 Empirical estimates

The condition for cooperation to be selected for by group selection used in, for instance, Bell et al. (2009) (see also Aoki and Nozawa 1984; Bowles 2006; 2009; Crow and Aoki 1984; Langergraber et al. 2011; Rusch 2018; Walker 2014; Weir and Cockerham 1984) is

$$\frac{\beta(w_g, p_g)}{\beta(w_{ig}, p_{ig})} > \frac{1 - F_{ST}}{F_{ST}}$$

Here $\beta(w_g, p_g)$ is the increase in the mean fitness of the group as a result of an increase in the frequency of cooperators, or altruists, and $\beta(w_{ig}, p_{ig})$ is the decrease in fitness of an individual as a result of switching from defection to cooperation. The idea is that this criterion separates the fitness effects, on the left hand side of the inequality, from a measure that characterizes the population structure, on the right hand side of the inequality. In a setting with a linear public goods game, played within groups that compete with each other in a well mixed population of groups, such a separation can indeed be made in this way.

Suppose the fitness of a cooperator and a defector in a group with i cooperators (for cooperators: including the individual itself) are $w_{C,i} = 1 + \frac{i-1}{n-1}b - c$ and $w_{D,i} = 1 + \frac{i}{n-1}b$. Being a cooperator instead of a defector would then give $n - 1$ others in the group a fitness benefit of $\frac{1}{n-1}b$, adding up to an aggregate fitness benefit of b, at a fitness cost to the individual of c. The average fitness within the group would go from $\frac{iw_{C,i}+(n-i)w_{D,i}}{n} = 1 + \frac{i}{n}(b - c)$ to $\frac{(i+1)w_{C,i+1}+(n-i-1)w_{D,i+1}}{n} = 1 + \frac{i+1}{n}(b - c)$, which amounts to an increase of $\frac{1}{n}(b - c)$ as a result of an increase in the frequency of cooperators within that group of $\frac{1}{n}$. This makes $\beta(w_g, p_g) = b - c$. The fitness effect measured by $\beta(w_{ig}, p_{ig})$ is similarly interpreted as c. Rewriting $\frac{b-c}{c} > \frac{1-F_{ST}}{F_{ST}}$ as $(b - c)F_{ST} > c(1 - F_{ST})$ and then as $F_{ST}b > c$ gives Hamilton's rule, with $r = F_{ST}$.

If we were to measure $\beta(w_g, p_g)$ in a setting in which competition between groups is not actually global, but to some degree local, then the resulting value for $\beta(w_g, p_g)$ would not only reflect the effect of cooperators on the average fitness within the group, but a mixture of these fitness effects and the cancellation effect. A moderate value for $\beta(w_g, p_g)$ can both be the result of a moderate group benefit and the absence of the cancellation effect, and a high group benefit combined with the cancellation effect at the group level. In the latter case, the negative effect of having neighbouring groups with many cooperators, combined with the positive correlation between being a cooperative group and having neighbouring

groups with many cooperators, would bias the estimated effect of – all else equal – the number of cooperators on average fitness within the group downwards. In other words, this term would end up absorbing the cancellation effect. In order to disentangle all fitness effects and the cancellation effect, one would have to estimate a more complex statistical model, which would not only use the composition of the own group as explanatory variable of the average fitness within the group, but also include the composition of neighbouring groups as an explanatory variable. This would be hard to estimate because it will require sufficiently high independent variation to overcome multicollinearity, but, if successful, it would separate the positive effect of the cooperation within the group from the negative effect of having a cooperative neighbouring group.

What most empirical papers do, however, is only estimate the F_{ST}, which is then taken as an indication of how conducive the population structure is to cooperation. We have seen that the absence or presence of the cancellation effect – which is part of the population structure – can however make a huge difference for how much the group needs to benefit from cooperators in it, relative to the individual costs, in order for cooperation to spread in the population.

Chapter 3

The evolution of honesty by partner choice[1]

3.1 Introduction

Why people lie is no mystery. Everyone acts on the basis of what they think is true, and if one can manipulate what others believe, then one can manipulate what others do. Why honesty exists, on the other hand, is not immediately clear. Revealing information because it is true, even if doing so would have negative consequences for oneself, requires some mechanism that compensates for the cost of honesty. Explanations that have been suggested for the evolution of honesty are similar to the ones that are proposed for the evolution of cooperation or altruism. Some authors suggest lying aversion can be explained by population structure (Krakauer and Pagel, 1995); others suggest that it can be explained by interactions being repeated (Rich and Zollman, 2016). We explore the possibility that, while telling the truth can be costly, being committed to telling the truth can be beneficial, because it makes one a more attractive partner.[2]

In situations in which one party knows or observes something that the other does not, honesty can make a difference since being honest can help commit to prosocial behaviour. In a partnership where only one of the partners for instance observes the result of a joint effort, the partner that is in the know would have an incentive to misinform the other

[1] This chapter is based on joint work with Stephan Jagau, Shaul Shalvi, and Matthijs van Veelen.

[2] Partner choice (Heintz et al., 2016) and commitment (Akdeniz and van Veelen, 2021; Frank, 1987; 1988) have been suggested as explanations for cooperative or pro-social behaviours in general, but here we focus on honesty (notice that in Heintz et al. (2016), honesty is an umbrella term for anything prosocial, and does not mean lying aversion). In the Discussion section we will come back to the relation between commitment and partner choice, and briefly discuss how it is also possible to have one without the other.

parties, and underreport the true gains in order to appropriate a larger portion of it for herself. When choosing a partner, one would therefore be happy to find someone who cannot help but tell the truth, and thereby opens herself up to scrutiny by the other parties. This preference for honest partners in mutually beneficial exchanges could then balance against the costs of honesty, and sustain a preference for truth-telling in human societies.

In this paper, we investigate the mechanism outlined above. We first of all test if, in a game with asymmetric information, honest others are indeed preferred as partners, and if honest partners indeed behave more prosocially. We find that the answer to both of these questions is yes. The findings from our lab experiment show that, in a trust game where only one of the partners observes the multiplier, honest individuals behave on average more prosocially than dishonest individuals, and that this is anticipated by others and creates a preference for honest partners. We also ask the question if the pathway for the connection between honesty and prosociality is conscious, in the following sense. Lying averse people by definition have a hard time lying. Therefore, when doing something not particularly prosocial is accompanied by a choice between, on the one hand, lying and, on the other, not lying and revealing they did something selfish, they could prefer to do the prosocial thing in order to avoid putting themselves in this difficult position. We find no evidence for this pathway. Instead, the outcome of our experiment is consistent with the possibility that lying averse people might have a harder time justifying selfish behaviour to themselves in the first place, or that they are just across the board more prosocial people.

3.2 Experimental design

We use a two-part design to examine the role of honesty in partner choice in a situation with asymmetric information. In the first part, we measure the relative lying aversion of subjects using a die-rolling task (Bicchieri et al., 2020; Conrads et al., 2013; Fischbacher and Föllmi-Heusi, 2013; Leib et al., 2021; Maréchal et al., 2017; Shalvi et al., 2011). Subjects observe a die-roll on their screen and are asked to report the outcome. Which die-roll a subject observes is chosen randomly, and each roll is equally likely. This happens for three rounds. Every time a subject reports a 5, she earns 75 points; and every time she reports another number, she earns 0 points. The potential earnings in the die-rolling task are therefore 0, 75, 150, or 225 points, depending on how many times a subject reports a 5. We define the lying aversion of a subject as the ratio of times she did not misreport a 5 out of the number of times she could have misreported a 5.

In the second part of the experiment, we assign subjects to groups of three. The group assignment within a session is done as follows. We first randomly assign one third of the subjects to the role of trustor. The remaining subjects are assigned the role of trustee. Each trustor is then matched with two trustees, one randomly drawn from the more lying averse half of the trustee population, and the other randomly drawn from the less lying averse half. In these groups, they play a version of the trust game for one round (Berg et al., 1995; Clots-Figueras et al., 2016; Güth et al., 2014), where trustors do not choose how much to send, but who to send it to, and where the multiplier is uncertain. Each group consists of one trustor, who is endowed with 50 points, and two candidate trustees, who are endowed with 0 points. The trustor observes the die-rolls that the candidate trustees saw, as well as what they reported, and chooses a partner to send her endowment to. The points sent by the trustor are then multiplied, either by 2 or by 4. Which one it is, is determined by chance, and both multipliers are equally likely. The trustees observe the multiplier, but the trustor does not. The trustees then choose how much they want to send back to the trustor, if they are chosen. While trustees make their decisions, we also elicit how much the trustor expects the trustees to send back, in case they are chosen, and depending on the multiplier.

We have two, between-subjects treatments: communication (C) and no communication (NC). In the communication treatment, there is cheap-talk messaging, where the candidate trustees also send a message about the multiplier to the trustor, together with the amount of points to send back. They choose between "the multiplier is 2" and "the multiplier is 4". Hence, if the multiplier is low, there is not much room for trustees to adjust their behaviour; however, if it is high, trustees face a choice between behaving prosocially, or increasing their payoff without damaging their image by pretenting that the multiplier is low. In the no communication treatment, there is no messaging.

The two treatments are there in order to investigate the mechanism by which honest subjects end up behaving more prosocially. Not all people are selfish, but even selfish people generally are not too keen on admitting they are selfish, and this is true also in experiments in which there is no way in which gaining a reputation for selfishness can have repercussions (Andreoni and Bernheim, 2009; Dana et al., 2007; Pillutla and Murnighan, 1995). Therefore, we assume that subjects in our experiment also generally care about not appearing selfish in their trust game behaviour. Assuming that they do, subjects that are not lying averse can have their cake and eat it –to some degree, if the multiplier is 4. If it is, they can report that the multiplier is 2, and return an amount that would be fair if the multiplier was 2, but not if the multiplier is 4. This way they seem prosocial,

in so far as the trustor believes the message, but keep more money for themselves. For lying averse subjects, this would be harder. When the multiplier is 4, the choice for lying averse subjects is likely to become one between, on the one hand, acting prosocially, and on the other hand, acting selfishly and admitting to it. In that case, they might just choose to always return a fair amount for the true multiplier, also if it is high, and tell the truth. If this is the mechanism by which lying averse individuals end up making more generous choices, then a difference between the more and less lying averse would occur in the communication treatment, and only if the multiplier is 4, that is, when there is something to lie about. In the no communication treatment, trustors may form a belief about the multiplier based on the amount that the trustee sends back, but in the absence of the possibility to make any explicit statements about the true multiplier, we would expect the impact of lying aversion on prosociality to be smaller or non-existent.

Once the candidate trustees decide how much they would like to send back if they are chosen (and their messages in the communication treatment), and the trustor submits her beliefs, the trust game ends. The trustor earns the points sent back by the trustee she chose as her partner, and she earns additional points from the belief elicitation. In the communication treatment, the trustor also observes the message sent by the trustee she chose, and responds by choosing between "I believe you" and "I do not believe you". The chosen trustee earns the amount she received (50 points sent by the trustor times the multiplier) minus the amount she chose to send back. In the communication treatment, she also observes the response of the trustor to her message about the multiplier. The non-chosen trustee earns the outcome of a random draw between 0 and 100, if the multiplier is 2; and between 0 and 200, if the multiplier is 4. The reason behind the random draw for the trustee that is not chosen, and the relatively large amounts that can be earned in the first part, is that we do not want subjects to be strategically honest in the die-rolling task. Being chosen in the second part is in expectation still better than not being chosen, but the difference is small enough to make sure that, if they are in the lab just to make money, the best way to do that is to lie in the first part. This may seem a counter-intuitive choice, but the purpose of this experiment is not to show that honest people can walk out of the lab with more money than dishonest people. We test hypotheses about how the benefits of being committed to telling the truth come about – and a defining characteristic of being committed to telling the truth, is that one does not immediately switch to lying, whenever that is more opportune.

After the trust game, we elicit participants' social value orientation (SVO), and collect some demographic information such as gender and age (see the Methods section in the

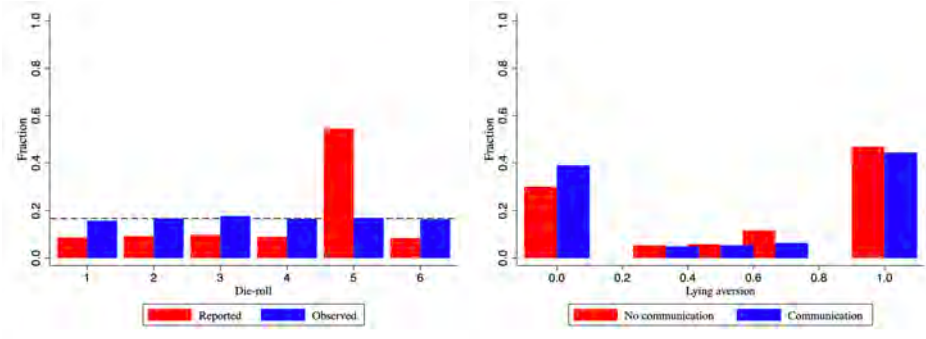
Supplementary Information for details).

If honesty evolved due to its hypothesized role in partner choice, we should first of all observe that honest individuals act more prosocially. Our first hypothesis is therefore that honest trustees would send back more points compared to dishonest trustees in the trust game. Moreover, if the reason why honest partners act more prosocially is the choice between, on the one hand, admitting to selfish behaviour and, on the other hand, behaving prosocially and telling the truth, we should observe that a stronger connection between honesty and prosociality in the presence of communication. Our second and third hypotheses are therefore that the average shares of points sent back and the difference between the shares of points sent back by honest and dishonest trustees would be larger in the communication treatment than in the no communication treatment, respectively. We expect both of these effects to be driven by the cases where the multiplier is high, since only then there is a reason to deceive one's partner. The choosing side should moreover anticipate the benefits of, and have a preference for, honest partners, for honesty to play a role in partner choice. Therefore, our fourth hypothesis is that trustors would expect honest trustees to send back more points, and choose them more often as their partners than dishonest trustees. Finally, we hypothesize that trustors would anticipate an impact of communication, especially on honest trustees (which are our fifth and sixth hypotheses; see the Methods section in the Supplementary Information for details on our hypotheses).

3.3 Results

Below we will first summarize the main findings from the die-rolling task and the trust game, and then give a detailed account of the results on trustee behaviour, trustor expectations and partner choice in the trust game. In our analyses, we normalize how much trustees send back and trustors expect to receive back by considering the proportion of the total amount available to the trustee that she returns and is expected to return.

Figure 3.2.1 presents the results of the die-rolling task. The distribution of reported rolls differ significantly from the distribution of actual rolls, with a spike at the reports of 5, and a corresponding downward shift on the reports of other numbers (Figure 3.2.1a). The majority of subjects either never lie, even though they could have benefited from lying, or they lie every time they can benefit from it, with no significant difference between the two treatments (Figure 3.2.1b). Since our experiment is designed in a way so that the groups in the trust game contain one trustee from the more lying averse half of the population

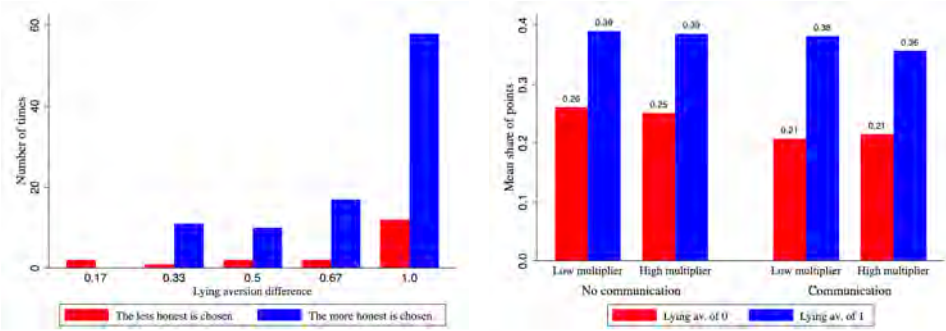


(a) Observed and reported die-rolls **(b) Distribution of lying aversion**

Figure 3.2.1: Lying aversion. (a) The frequencies of die-rolls observed by subjects (blue bars) is close to 1/6 for every number between 1 and 6, as expected. The frequencies of reported die-rolls (red bars) are consistent with significant over-reporting of 5s, but also considerable truthful reporting. **(b)** Our measure of lying aversion is the ratio of the number of instances in which the subject could have benefited from lying, but chose not to, and the number of instances they could have benefited from lying. For someone that observed zero fives, this number therefore can be 0, 1/3. 2/3 or 1 – where 0 means the subject always lied and 1 means the subject never lied. For someone that observed one five, this number can be 0, 1/2 or 1, for someone that observed 2 fives, 0 or 1, and for the 2 subjects that observed 3 fives, this is not defined (who are therefore excluded from our analyses.

of trustees and one trustee from the less lying averse half within a session, this typically generates relatively diverse pairs (see also Figure 3.3.1a). In the remainder of the paper, we will restrict our analyses on trustee behaviour and trustor expectations to trustees with a lying aversion measure of 1 or 0 to simplify the classification of individuals as honest and dishonest. Our results are however robust to the inclusion of all trustees (see the Supplementary Information for the analyses on the full sample).

If we look at the data for both treatments together, we first of all find that trustors mostly choose the more honest trustee as their partner (Figure 3.3.1a). They also expect honest trustees to send back more points (Figure 3.3.1b); and they are right in doing so, although they do overestimate the difference between honest and dishonest trustees (Figure 3.3.2). Next we look at the treatment effect to examine whether honest people act prosocially to avoid lying to their partners. We find no evidence for this. If we consider the behaviour when the multiplier is high in the communication treatment, we do not find a significant difference between subjects with a lying aversion measure of 0 and subjects with a lying aversion measure of 1. Instead, we find quite the opposite, as the overall difference between the behaviour of subjects with a lying aversion measure of 0 and subjects with a lying aversion measure of 1 seems to be driven mostly by the difference in



(a) Partner choice **(b) Trustor expectations**

Figure 3.3.1: Trustor behaviour. (a) Partner choice decisions of trustors. In red the number of times the less honest candidate is chosen, in blue the number of times the more honest partner is chosen, both depending on the difference in lying aversion of the two candidate trustees. **(b)** Averages of the shares of points trustors expected to be returned, based on the lying aversion measure of the trustees, the treatment and the multiplier in the trust game.

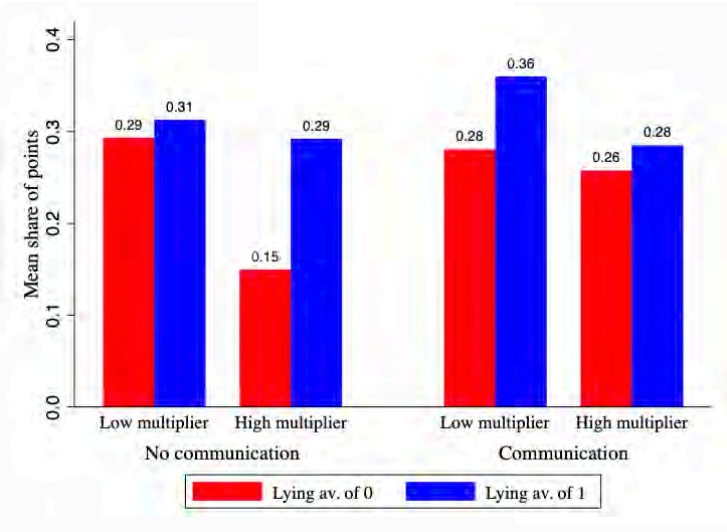


Figure 3.3.2: Trustee behaviour. Averages of the shares of points sent back by trustees, based on the lying aversion measure of the trustees, the treatment and the multiplier in the trust game.

the no communication treatment, when the multiplier is high (see Figure 3.3.2). Next we discuss trustee behaviour, trustor expectations and partner choice in more detail.

Trustee behaviour We can start with a simple statistical analysis, where we just compare

NC & Low multiplier	0.9706
NC & High multiplier	0.0018
C & Low multiplier	0.2391
C & High multiplier	0.6828
Low multiplier	0.4168
High multiplier	0.0169
NC	0.0913
C	0.2429
Overall	0.0413

Table 3.3.1: Mann Whitney test results. p-values of the Mann Whitney tests comparing the returns by trustees with a lying aversion measure of 0 versus 1, based on the treatment and/or the multiplier. The last row reports the result for the overall comparison where we aggregate observations for both treatment and multiplier. In the cases where there is a significant difference, those with a lying aversion of 1 send back more points than those with a lying aversion of 0.

the shares of points returned by those with a lying aversion measure of 0 with that of those with a lying aversion measure of 1 for different partitions of the sample (Table 3.3.1). If we do that for the combinations of treatment and multiplier depicted in Figure 3.3.2, then we only see a significant difference between the not lying averse and the lying averse for the high multiplier in the NC treatment (p-value: 0.0018, which rejects the null hypothesis of equality against the Bonferroni corrected critical value for multiple hypothesis testing). None of the other differences are anywhere close to significant (p-values: 0.9706 for the low multiplier in NC, 0.2391 for the low multiplier in C, and 0.6828 for the high multiplier in C). Here we should stress that we hypothesized that there would be a larger difference between the lying averse and the not lying averse for the combination of high multiplier and communication, than for the combination of high multiplier and absence of communication. What we observe however is that the difference is non-existent in the first case, whereas it is striking in the latter. A remarkable feature of the data is that the average shares that trustees return are not too far away from 30% in all cases, with one exception. That one exception is that when the multiplier is high and there is no communication, then those that are classified as not lying averse on average return 15% (Figure 3.3.2). This is remarkable, because, obviously, 15% of the amount for a multiplier of 4 equals 30% of the amount for a multiplier of 2.

If we aggregate observations for both multipliers, and compare the returns by honest and dishonest trustees within each treatment separately, then neither of the two is significant (p-values: 0.0913 for NC, and 0.2429 for C). If, on the other hand, we aggregate observations for both treatments, and compare the returns by honest and dishonest trustees

for each multiplier separately, again neither of the two is significant (p-value: 0.0169 for the high multiplier, p-value: 0.4168 for the low multiplier, compared to the Bonferroni corrected critical value for multiple hypothesis testing). If we aggregate observations for both multipliers and treatments, the difference between the shares of points sent back by the lying averse and the not lying averse is only marginally significant (p-value: 0.0413).[3] In the cases where there is a significant difference between the two groups, those with a lying aversion of 1 are the ones that act more prosocially.

Also a more elaborate statistical analysis confirms this. Table 3.3.2 presents ordinary least squares (OLS) regression results, where the dependent variable is the the share of points sent back by the trustee. The independent variables are the lying aversion of the trustee (for testing our first hypothesis), the level of the multiplier, the communication treatment, the interaction variable between the treatment and the multiplier (for testing our second hypothesis), and the interaction variable between the treatment, the multiplier and the lying aversion of the trustee (for testing our third hypothesis). Results show that honest trustees behave more prosocially than dishonest trustees in the no communication treatment (Column 1) and overall (Columns 3-5), but not in the communication treatment (Column 2). Altogether, these findings provide support for our first hypothesis that honest individuals act more prosocially than dishonest individuals in a situation with asymmetric information.

We had also hypothesized that a pathway from honesty to prosociality would be that lying averse individuals might prefer to send back a fair amount of points corresponding to the true multiplier, in order to avoid the choice between lying and revealing having been selfish. When the multiplier is low, this does not create a problem, because one can then only send back a decent amount for the low multiplier. When there is no communication, this also would not create a problem, because they are not asked to report the multiplier. Therefore, we expect this effect to be present when the multiplier is high, and where there is communication. In line with Figure 3.3.2 and the results of the non-parametric tests in Table 3.3.1, this is not what we find, as shown by the insignificance of the interaction variables in Columns 4-6 of Table 3.3.2. This provides evidence against our second and third hypotheses that there would be higher levels of prosociality in the presence of communication, due to honest individuals acting more prosocially to avoid admitting to selfish behaviour.

[3]We can run a robustness check on this result using the SVO measures collected at the end of the experiment. Mann-Whitney test comparing the SVO angles of trustees with a lying aversion measure of 0 versus 1 suggests a statistically significant difference between the two groups of trustees (p-value: 0.0037).

	(1) Ratio	(2) Ratio	(3) Ratio	(4) Ratio	(5) Ratio	(6) Ratio	(7) Ratio
Lying av.	0.0781**	0.0537	0.0649**	0.0657**	0.0786**	0.0422	
	(0.032)	(0.161)	(0.013)	(0.013)	(0.015)	(0.173)	
Multiplier	-0.0735**	-0.0512	-0.0613**	-0.0718**	-0.0736**	-0.0321	
	(0.040)	(0.173)	(0.017)	(0.041)	(0.038)	(0.430)	
C			0.0282	0.0179	0.0172	0.0192	
			(0.268)	(0.661)	(0.675)	(0.640)	
(C=1)#(M=1)				0.0209	0.0500	-0.0194	
				(0.684)	(0.425)	(0.729)	
(C=1)#(M=1)#(LA=1)					-0.0511		
					(0.335)		
(C=0)#(M=1)#(LA=0)						-0.100*	-0.149***
						(0.072)	(<0.001)
Constant	0.264***	0.295***	0.265***	0.270***	0.264***	0.282***	0.299***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
N	111	111	222	222	222	222	222
Treatment	NC	C	Both	Both	Both	Both	Both
adj. R^2	0.053	0.018	0.042	0.038	0.037	0.046	0.045

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3.2: OLS regression results on trustee behaviour. Share of points sent back by trustees, based on the lying aversion measure of the trustee; in the NC treatment (Column 1), in the C treatment (Column 2), in both treatments including the treatment variable (Column 3), the interaction variable between the treatment and the multiplier (Column 4), and the interaction variable between the treatment, the multiplier, and the lying aversion of the trustee (Column 5). Columns 6 and 7 report the results for our unexpected finding, where the interaction variable is included for the NC treatment when the multiplier is high for the not lying averse subjects. The data is restricted to the trustees with a lying aversion measure of 0 or 1. For the interaction variables, C=1 (0) means (no) communication treatment, M=1 means high multiplier, and LA=1 (0) means lying aversion measure of 1 (0).

Instead of the difference being driven by the lying averse sending back larger shares when the multiplier is high and there is communication, also with the OLS regression results, we observe that it is mainly driven by them sending back more when the multiplier is high *and there is no communication* (see Columns 6-7 in Table 3.3.2). Our communication treatment therefore appears to have a prosocializing effect on behaviour, however, in contrast to our expectations, it is on the not lying averse individuals. One interesting

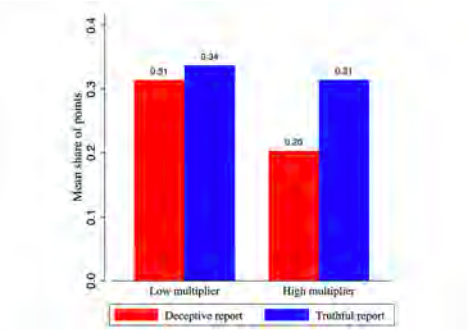


Figure 3.3.3: Trustee behaviour based on misreporting in the communication treatment. Averages of the shares of points returned by trustees based on whether the trustee misreported the multiplier in the communication treatment. When the multiplier is high, those who report truthfully send back on average 31% of the available points, and those who misreport send back on average 20%. The difference is statistically significant.

point here is that relatively few trustees make use of the communication opportunity to keep a prosocial image while being selfish, and that those who make use of it are not necessarily the ones classified as dishonest in the die-rolling task. Out of the 55 trustees facing the high multiplier in the communication treatment, 19 misreport the multiplier. Those who do so act less prosocially compared to those who report truthfully (Mann-Whitney test p-value: 0.008, see also Figure 3.3.3). There however does not seem to be a correlation between misreporting in the trust game and misreporting in the die-rolling task. If we regress the probability of misreporting in the trust game on the lying aversion of the trustee from the die-rolling task, the estimated coefficient on lying aversion is highly statistically insignificant (p-value: 0.992, see the Supplementary Information for details). The disconnect between misreporting in the two parts of the experiment could however be due to the relatively small number of observations of misreports in the second part of our experiment. Out of the 56 trustees facing the low multiplier in the communication treatment, 7 misreport the multiplier. Those who do so, however, still send back amounts that are consistent with the low multiplier (see the left-most bar in Figure 3.3.3). This, in combination with the low number of observations in this case, suggests that reporting a high multiplier, when it is in fact low, is likely to be a mistake.

Trustor expectations Table 3.3.3 presents the OLS regression results, where the dependent variable is the share of points expected back by the trustor. The independent variables are the lying aversion of the trustee (for testing our fourth hypothesis), the level of the multiplier, the communication treatment, the interaction variable between the

	(1) Ratio	(2) Ratio	(3) Ratio	(4) Ratio	(5) Ratio	(6) Ratio
Lying av.	0.131***	0.158***	0.145***	0.145***	0.146***	0.143***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Multiplier	-0.00676	-0.0100	-0.00842	-0.00676	-0.00676	-0.00676
	(0.377)	(0.352)	(0.201)	(0.373)	(0.374)	(0.375)
C			-0.0303	-0.0286	-0.0286	-0.0249
			(0.275)	(0.318)	(0.321)	(0.396)
(C=1)#(M=1)				-0.00329	-0.00101	0.000149
				(0.802)	(0.961)	(0.994)
(C=1)#(M=1)#(LA=1)					-0.00421	-0.00635
					(0.888)	(0.832)
Lying av. trustor						0.0427
						(0.181)
Constant	0.260***	0.216***	0.253***	0.252***	0.251***	0.225***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
N	216	222	438	438	438	438
Treatment	NC	C	Both	Both	Both	Both
adj. R^2	0.089	0.210	0.146	0.144	0.142	0.149

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3.3: OLS regression results on trustor expectations. Share of points expected back by trustors, based on the lying aversion measure of the trustee; in the NC treatment (Column 1), in the C treatment (Column 2), in both treatments including the treatment variable (Column 3), the interaction variable between the treatment and the (hypothetical) multiplier (Column 4), the interaction variable between the treatment, the (hypothetical) multiplier, and the lying aversion of the trustee (Column 5), and the lying aversion of the trustor (Column 6); with the errors clustered at the trustor level since each trustor submits multiple beliefs. The data is restricted to trustor expectations for the trustees with a lying aversion measure of 0 or 1. For the interaction variables, C=1 means communication treatment, M=1 means high multiplier, and LA=1 means lying aversion measure of 1 for the trustee.

treatment and the multiplier (for testing our fifth hypothesis), and the interaction variable between the treatment, the multiplier, and the lying aversion of the trustee (for testing our sixth hypothesis). Results show a very strong positive correlation between how lying averse one's partner is and how prosocial one expects them to be (Columns 1-6). Trustors expect to receive on average 37.9% of the available points from trustees with a lying aversion measure of 1, while they expect only 23.3% from trustees with a lying aversion

	(1) Chosen	(2) Chosen	(3) Chosen	(4) Chosen
The more lying averse trustee	1.576***	1.540***	1.620***	1.653***
	(<0.001)	(0.001)	(<0.001)	(<0.001)
C		0.0690		-0.0637
		(0.914)		(0.900)
N	70	70	115	115

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3.4: Logit regression results on partner choice. The logistic regression for the probability that the more lying averse candidate trustee is chosen as a partner by the trustor. Columns 1 and 2 restrict the data to the trustee pairs with a lying aversion difference of 1, including and excluding the treatment variable; and Columns 3 and 4 include the data on all trustee pairs with a non-zero lying aversion difference, including and excluding the treatment variable.

of 0 (Mann-Whitney test p-value < 0.001). They do not anticipate any impact of the communication treatment (Columns 3-5). These findings provide support for our fourth hypothesis on trustors anticipating honest individuals to be more prosocial, but against our fifth and sixth hypotheses on the anticipation of an impact of communication on the trustee population as a whole, or on the lying averse trustees.

The finding that trustors place great value on the information about the trustees' lying aversion, but do not anticipate any other effect, might be explained by the general pattern in trustee behaviour. As shown previously, honest and dishonest trustees behave similarly in the communication treatment, and in the no communication treatment when the multiplier is low; and honest trustees behave significantly more prosocially in the no communication treatment when the multiplier is high. Therefore, although it might not always benefit the trustor to pick an honest partner, it never hurts. We expect this to be a general pattern, also outside of the lab. Picking an honest partner is likely to be beneficial, but unlikely to be harmful in many partnerships. The selective pressure on the choosing side could therefore be focused mostly on the detection of honesty, and not so much on how the specifics of an interaction affect the behaviour of honest versus dishonest partners. While gathering information likely involves cognitive and other costs (Drugowitsch et al., 2012; Sullivan, 1994), it might not bring additional benefits, once we know how lying averse our partner is. This could have made the choosing side more attentive to information on lying aversion compared to other, seemingly less important, details of an interaction.

Partner choice Table 3.3.4 presents the logistic (logit) regression results, where the dependent variable is the probability that the more honest trustee in a trustee pair is chosen. Columns 1 and 2 restrict the data to trustors who were presented with a trustee pair with a lying aversion difference of 1, and Columns 3 and 4 include data for all trustors who were presented with a trustee pair with any non-zero lying aversion difference. Results show that honest trustees are chosen significantly more often as partner, with no significant difference between treatments (Columns 2 and 4). In line with the previous findings on the share of points trustors expect from honest versus dishonest trustees, honest trustees are therefore trusted significantly more often in a situation with asymmetric information. These findings provide evidence for our last hypothesis, which concerns the existence of a preference for honest partners.

One prerequisite for honesty to play a part in partner choice is obviously that others can tell. Trustors in our experiment make their partner choice decisions with full information on the lying aversion of the candidate trustees –as finely measured as possible by the die-rolling task. To what extent we are able to tell who is honest and who is not in real life, is an interesting open question. If possible, one would like to be perceived as honest, but lie when opportune, and an invasion of a type that seems honest, but is not, would undermine the selective advantage of the truly honest. There would therefore be selection pressure also on the ability to detect true commitment to honesty. Even though deception occurs regularly in real life, and sometimes with substantial consequences, there is considerable evidence suggesting that one can tell whether someone is (being) honest under certain conditions. For instance, people seem to be relatively good at detecting deception when they know their interaction partners (Anderson et al., 2002; DePaulo, 1994; DePaulo and Kashy, 1998; Ten Brinke et al., 2016; Von Hippel and Trivers, 2011; Vrij and Mann, 2005). Also, our brains are argued to have specialized modules to detect cheating (Cosmides and Tooby, 1992; Stone et al., 2002). The fact that honesty still is one of the most important virtues one looks for in prospective partners (Cottrell et al., 2007; Regan et al., 2000) moreover suggests that our detection capabilities are adequate enough to be utilized in judging others.[4]

3.4 Discussion

Why honesty exists is an open question. Truthfully revealing information, that could otherwise be used for one's benefit, requires a mechanism that compensates for the costs

[4]See also Heintz et al. (2016) for a discussion on how social traits and strategic vigilance, that is the ability to spot genuine displays of those traits, could have co-evolved.

of honesty. We investigate how partner choice can bring such benefits and enable the evolution of honesty. More specifically, we study whether honesty credibly signals commitment to prosociality, and create a demand for honest partners in mutually beneficial interactions.

Most partnerships involve reporting private information that individuals have. Since the interests of the involved parties do not usually align, the partner that is in the know would be tempted to deceive others for her own benefit. This temptation would likely destabilize partnerships unless partners can find a way of committing to treating each other fairly. This is where honesty can offer a solution. By increasing the (cognitive) costs of lying, it can make one prefer behaving prosocially to avoid lying to others, and enable trust between partners. The choice between behaving prosocially and lying about having been selfish is especially pertinent when partners are explicitly communicating with each other. When there is no communication between partners, the commitment power of honesty to prosocial behaviour would no longer be relevant.

Using a lab experiment, we test the mechanism outlined above. We examine whether honest people are in general more prosocial in a situation with asymmetric information, whether the connection between honesty and prosociality goes through the impact of communication, and whether these are anticipated by others and create a preference for honest partners. The results of our experiment show that honest people are indeed on average more prosocial in a trust game with an uncertain multiplier, that this is anticipated by others and results in a preference for honest partners in the trust game. However, as indicated by the null impact of our communication treatment on honest trustees, the connection between honesty and prosociality does not seem to go through communication. Honest individuals behave consistently prosocially independent of whether there is communication or not. Instead, dishonest individuals react to our communication treatment. They behave significantly less prosocially than honest individuals when this requires them to deceive their partners only implicitly (in the no communication treatment), but not when it requires an explicit lie to their partners (in the communication treatment). Below we discuss potential reasons for our findings on communication.

Deception in real life takes many different forms. People lie by both commission and omission, and by behaving in a way that would lead their partners to hold inaccurate beliefs. In our no communication treatment, trustees can lead their partner to wrongfully believe that the multiplier is low by sending back an amount that would be consistent with the low multiplier, when it is high. How much one sends back therefore has informational value by itself. When deception involves such behavioral misleading and not explicit

misreporting, we observe that individuals classified as dishonest, but not as honest, deceive their partners by sending back amounts that would be consistent with the low multiplier. Honest trustees however send back amounts consistent with the actual multiplier they observe. This suggests that our die-rolling task could be capturing an aversion to both implicit behavioural misleading and explicit verbal misleading, where those classified as dishonest in the die-rolling task still avoid the latter.

Both honest and dishonest subjects in our experiment generally refrain from lying in the trust game. The existence of a partner on the receiving end of a lie, without any strategic role of communication, therefore seems to push the majority of our subjects to report the multiplier truthfully (and behave accordingly prosocially). Moreover, in contrast to the findings on the ultimatum game (Boles et al., 2000; Kriss et al., 2013), introducing the possibility to send a deceptive message about the size of the pie increases prosocial behaviour in our experiment. The decision to *trust* in our trust game, as opposed to no such decision in the ultimatum game, might therefore have created an interaction environment that is less prone to deception. The timing of communication could have further reinforced this prosocializing impact of communication in our experiment. Existing literature suggests that the timing of communication affects one's propensity to lie and that people are most prosocial closest to the time of communication (Andersson and Wengström, 2012; Bhattacharya et al., 2020; Blume and Ortmann, 2007; Casella et al., 2018; Gneezy, 2005). Since communication happens at the same time as the choice of prosociality in our experiment, it might have nudged not only the honest but also the dishonest to report truthfully. It is therefore important in future research to investigate the impact of other forms of communication on prosociality of honest individuals, especially in settings where communication can be used to influence the recipient's behaviour (Blume et al., 1998; Cai and Wang, 2006; Crawford and Sobel, 1982; Sánchez-Pagés and Vorsatz, 2007).

Here we examined how honesty can be advantageous in partner choice. However, there is another, additional channel through which the ability to commit can help an individual. Honesty can bring benefits not only by increasing the chances that one is chosen as a partner, but also by changing how one's partner behaves during a partnership. Consider, for instance, the standard trust game (Berg et al., 1995; Clots-Figueras et al., 2016; Güth et al., 2014), where the trustor does not choose who to partner with, but chooses how much of her endowment to send to her existing partner. The partner choice channel is therefore absent in this version, but the trustee's honesty can still change the trustor's behaviour to her benefit. Given that honest individuals act on average more prosocially

than dishonest ones, and that this is anticipated by others, trustors are likely to be willing to send larger amounts to honest trustees. This would increase the size of the pie to be shared, and result in larger gains to be attained from the partnership. Partner choice is therefore one of the channels through which commitment power of honesty to prosociality can work, but it is not the only one; the ability to influence partner behaviour in a given interaction brings benefits as well (see Akdeniz and van Veelen (2021) for an overview on the role of commitment in a broad spectrum of human social behaviours).

Appendix

3A.1 Methods

General outline and procedure of the experiment[5] The experiment is conducted at the CREED lab of the University of Amsterdam. It is computerized using oTree (Chen et al., 2016). We invited 411 subjects to the lab, with 207 (204) subjects randomly assigned to the NC (C) treatment. We ran three pilot sessions with a total of 57 subjects before running our main sessions. After the pilot sessions, we updated our instructions and added comprehension questions. The data from the pilot sessions are excluded in our analyses. All participants received their earnings from the experiment plus a €4 show-up fee. Experimental payoffs are expressed in points during the experiment. At the end of the experiment the points are converted into euros at a conversion rate of €0.05 per point. The experiment consists of four parts and a survey. In the first part of the experiment, subjects are asked to report the outcome of a die roll they see on their screen for three rounds (Conrads et al., 2013; Fischbacher and Föllmi-Heusi, 2013; Leib et al., 2021; Maréchal et al., 2017; Shalvi et al., 2011). In the second part, subjects are assigned into groups of three to play a simplified version of the trust game (Berg et al., 1995) with partner choice and an uncertain multiplier. In the third and fourth parts, subjects make six resource allocations decisions (Crosetto et al., 2019; Murphy et al., 2011) and four investment decisions (Gneezy and Potters, 1997),[6] respectively. At the end of the experiment, subjects are asked to complete a survey. Subjects are paid their earnings from either the first part or the second part of the experiment, in order to eliminate potential income effects in the second part (Drouvelis and Sonnemans, 2017). Which stage is paid is determined randomly. Additionally, they are paid for two of the resource allocation tasks – once in the sender and once in the receiver role, and for one round of the investment task. Which of these tasks are paid is also determined randomly. The participation fee and any additional amount of money participants earned is paid to them privately in cash at the end of the experiment. Subjects were matched with different participants for the trust game, and the (different rounds of) the SVO task, in order to eliminate reciprocity concerns.

[5] Our pre-registration form can be found on:
https://osf.io/792p4/?view_only=ca1b5bb7acdc4f5e9ec4b48bfdf509ce.

[6] The investment task was included in our design since an earlier version of our design used the standard trust game where the trustor's choice of how much to send can be affected by her risk preferences. Since this channel was eliminated in the final version of our trust game, we exclude the investment task from our analysis.

Die-rolling task Subjects are presented with the video of a randomly chosen die-roll on the screen and are asked to report the outcome for three rounds (Kocher et al., 2018). In each round they earn 75 points if they report a 5, and 0 points if they report any other number. The minimum amount a subject can earn from the die-rolling task is therefore 0 points, and the maximum is 225 points. We picked the die-rolling task since it is a simple task and it has been shown to correlate with cheating in several real-life situations (Cohn and Maréchal, 2018; Dai et al., 2018; Hanna and Wang, 2017; Potters and Stoop, 2016).

We measure the lying aversion of subjects as the ratio of the number of times they did not misreport a die-roll to the number of times they could have misreported to increase their earnings over the three rounds of the die-rolling task. Our measure of lying aversion is therefore defined as:

$$\text{lying aversion}_i = \frac{\#\text{of times subject } i \text{ could have gained from misreporting, but did not}}{\#\text{of times subject } i \text{ could have gained from misreporting}}$$

If the result of each die roll is a 5, there is no room for the subject to lie in order to increase their earnings. In this case, the denominator in our measure of lying aversion takes the value of 0, which makes the lying aversion not defined. This case occurred twice in our experiment, and we omit those observations in our analysis. We assume that subjects would only have an incentive to lie up, i.e., they would lie only when they can increase their earnings by doing so. Our lying aversion measure therefore takes values between 0 and 1, decreases with the number of times subjects misreport a die-roll, and hence, increases with the lying aversion of the subject. A subject with a lying aversion measure of 0 and a subject with a lying aversion measure of 1 are the least and the most honest extremes within our measure, respectively. We exclude the one subject who reported down, i.e., who observed a 5 but reported another number, from our analysis.

Partner choice and trust game We begin by outlining the order of events in our trust game in the no communication treatment. Afterwards we will describe how our treatments differ in detail. Within a session, we first randomly assign one third of the subjects to the role of trustor. The remaining subjects are assigned the role of trustee. We then split the trustees within the session into two groups, where one group includes the upper half and the other includes the lower half of the lying aversion distribution. Each trustor is then matched with two trustees, one randomly drawn from the first group, and the other randomly drawn from the second group. This is done with the purpose of eliminating uninformative observations by excluding trustee pairs with the same lying aversion measures as much as possible. We can however not always avoid trustee pairs with

the same level of lying aversion due to limited session sizes. The results presented here exclude those cases. Since most of our subjects have lying aversion measures of either 0 or 1, the majority of the trustee pairs presented to trustors have a lying aversion difference of 1 (see Figure 1 in the main text).

In the trust game, the trustors are endowed with 50 points and have to pick one of the trustees they are matched with to send these 50 points to. Before picking their partner, they observe the die rolls and the reports of the two trustees from the die-rolling task. The trustors can therefore easily deduce which trustee is more lying averse. Upon sending, the 50 points are multiplied, either by 2 or by 4, where both multipliers are equally likely. The trustees, but not the trustors, observe the multiplier that is drawn. Both trustees choose how many points they want to send back to the trustor, if they are the chosen partner.

While trustees make their choices, we simultaneously elicit the trustors' beliefs on what they expect each trustee to return if the multiplier is 2, and if it is 4. Each trustor therefore submits four beliefs, two of which correspond to the true multiplier. We incentivize the trustors' beliefs for accuracy. For each trustee, trustors have a chance of winning 50 points, in addition to their earnings from the game. The probability of winning the prize is determined by how close their expectation is to how much the trustee returns, for the true multiplier. More specifically, if the trustor submits a belief of X points for a trustee and a multiplier level, and the trustee actually returns Y points, then the probability p of winning 50 points is:

$$p = 1 - \left(\frac{|X - Y|}{\text{Trustor endowment} * \text{Multiplier}} \right)^{1/2} = 1 - \left(\frac{|X - Y|}{50 * \text{Multiplier}} \right)^{1/2}$$

We then compare this probability to a random draw between 0 and 1, and if the probability p is higher than the random draw, the trustor wins the prize for that trustee; and if it is lower, the trustor does not win the prize. The trustors therefore has an incentive to submit their correct beliefs as it maximizes their chance of winning the prize.

After the trustees make their choices and the trustors submit their beliefs, the game ends. The trustees learn who is selected by the trustor, and the choice of that trustee is implemented: The trustor earns the amount sent back by the trustee she chose, and the chosen trustee earns the endowment she received times the multiplier minus how much she chose to send back. The non-chosen trustee earns the outcome of a random draw between 0 and 100, if the multiplier is 2; and between 0 and 200, if the multiplier is 4.

In the communication treatment, the trust game is played in exactly the same way, except that there is cheap-talk messaging. The trustees choose a message about the level of the multiplier to send to the trustor, together with the amount of points to send back. They choose between reporting that the multiplier was 2, or that it was 4. At the end of the trust game, the trustor observes the message sent by the trustee she chose, together with the amount sent back, and sends a reply. She chooses between saying that she does or does not believe the trustee. The reply for the trustor is included so that the trustor has *the last word*.

We take the share of points returned by the trustees and the share of points expected by the trustors as measures of actual and expected prosociality, respectively. Based on the existing literature and the mechanism outlined in the main text, we form the following hypotheses. Hypotheses 1-3 are about trustee behaviour. They test the relationship between prosociality, lying aversion and communication. Hypotheses 4-6 are about trustor beliefs. They test the relationship between expectations about partner's prosociality, partner's lying aversion and communication. And finally, Hypothesis 7 tests whether there is a preference for honest partners in the trust game.

- **Hypothesis 1 (H1):** There is a positive correlation between lying aversion and prosociality.
- **Hypothesis 2 (H2):** The existence of communication makes people behave more prosocially.
- **Hypothesis 3 (H3):** The impact of communication on prosocial tendencies is stronger for people with higher degrees of lying aversion.
- **Hypothesis 4 (H4):** People, in general, expect a positive correlation between lying aversion and prosociality.
- **Hypothesis 5 (H5):** People anticipate a positive impact of communication on prosocial tendencies.
- **Hypothesis 6 (H6):** People anticipate a stronger impact of communication on prosocial tendencies of the lying averse individuals.
- **Hypothesis 7 (H7):** The more lying averse people are chosen more often as a partner.

The impact of communication in H2-3 and H5-6 is anticipated in the cases where the multiplier is high. If the multiplier is 2, there is no (apparent) reason to misreport the

multiplier. However, if the multiplier is 4, there is a tension between being prosocial, and being selfish and lying about it.

Analyses We have two types of statistical tests on trustee behaviour and trustor expectations. We run ordinary least squares (OLS) regressions, and we report Mann-Whitney tests results as a non-parametric check. We use robust standard errors in all of our regressions. Moreover, we cluster errors at the trustor level in regressions for trustor expectations since each trustor submits four separate beliefs. We run logistic (logit) regressions for studying the partner choice behaviour of trustors.

Social value orientation (SVO) measure We elicit subjects' SVO using the slider task (Crosetto et al., 2019; Murphy et al., 2011). Subjects make six resource allocation decisions. The alignment between payoffs for themselves and the other participant they are matched with is varied across decisions. The SVO angles calculated from the allocation decisions are used as a control variable of subjects' prosocial attitudes outside of the trust game.

3A.2 Results

3A.2.1 Treatment assignment

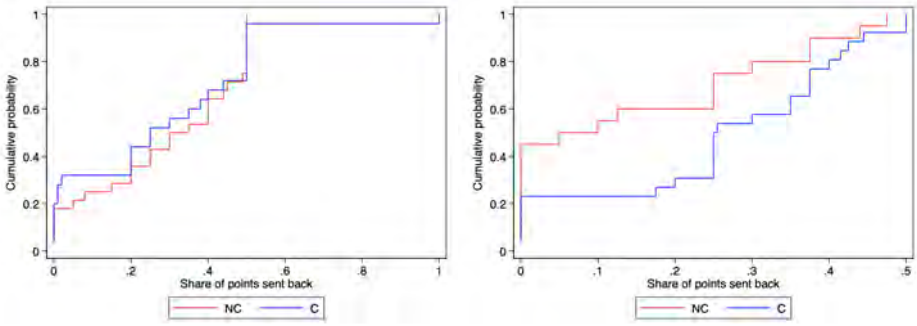
The distributions of subject attitudes across treatments, presented in Table 3A.1, suggest that our treatment assignment was random regarding subject characteristics.

Treatment	Lying aversion	SVO	Gender
C	0.535	16.492	0.522
	(0.459)	(12.982)	-
NC	0.565	16.149	0.565
	(0.450)	(13.245)	-

Table 3A.1: Summary statistics of subject characteristics across treatments: Mean values and standard deviations in parentheses. Mann-Whitney tests give a p-value of 0.6431 for the lying aversion comparison, and a p-value of 0.9354 for SVO angle comparison across treatments.

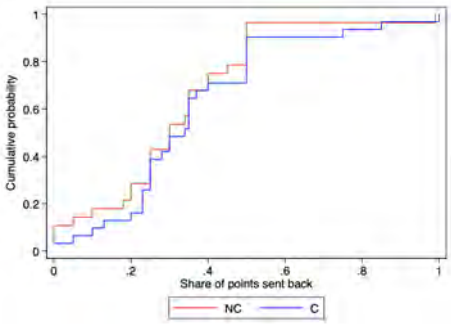
3A.2.2 Impact of communication

Distribution of trustee returns across treatments can be found in Figure 3A.1, and the distribution of trustor expectations across treatments can be found in Figure 3A.2.

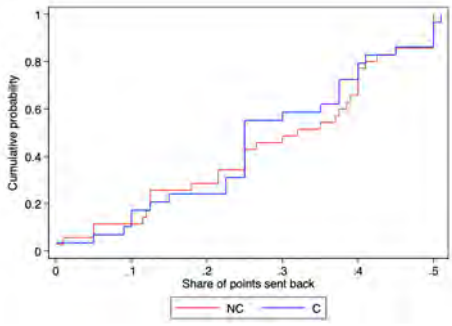


(a) Low multiplier & lying aversion = 0

(b) High multiplier & lying aversion = 0

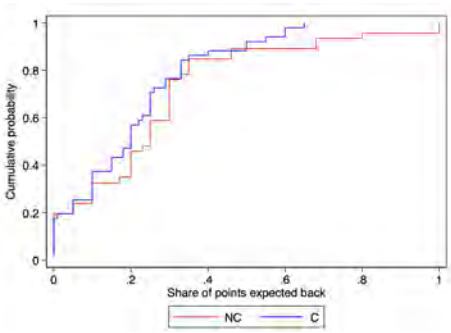


(c) Low multiplier & lying aversion = 1

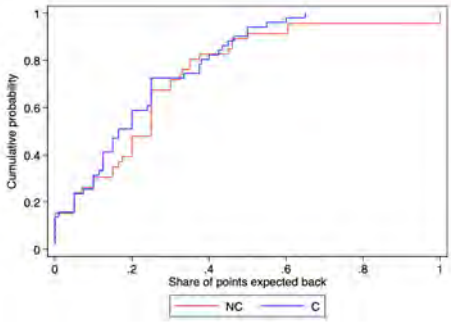


(d) High multiplier & lying aversion = 1

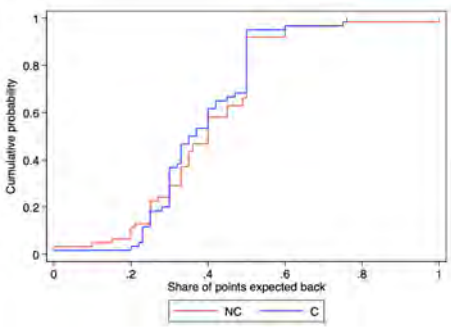
Figure 3A.1: Cumulative distribution functions of the share of points sent back by the trustees across treatments, for the cases with (a) low multiplier and trustees with a lying aversion measure of 0, (b) high multiplier and trustees with a lying aversion measure of 0, (c) low multiplier and trustees with a lying aversion measure of 1, and (d) high multiplier and trustees with a lying aversion measure of 1.



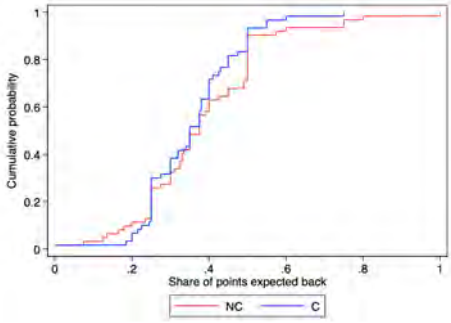
(a) Low multiplier & lying aversion of trustee = 0



(b) High multiplier & lying aversion of trustee = 0



(c) Low multiplier & lying aversion of trustee = 1



(d) High multiplier & lying aversion of trustee = 1

Figure 3A.2: Cumulative distribution functions of the share of points expected back by the trustors across treatments, for the cases with (a) low multiplier and trustees with a lying aversion measure of 0, (b) high multiplier and trustees with a lying aversion measure of 0, (c) low multiplier and trustees with a lying aversion measure of 1, and (d) high multiplier and trustees with a lying aversion measure of 1.

	(1) Misreport	(2) Misreport
Lying av.	-0.00587	-0.0347
	(0.992)	(0.950)
Constant	-0.636	-0.524
	(0.123)	(0.172)
N	55	68

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.2: Connection between misreporting the multiplier in the trust game and misreporting in the die-rolling task Logistic regression results on the probability that a trustee misreports the multiplier in the communication treatment when the multiplier is high, based on her lying aversion measure from the die-rolling task. Column I restricts data to trustees with a lying aversion of 0 or 1, Column II includes all trustees.

3A.2.3 Reports on the multiplier

Out of a total of 134 trustees the communication treatment, 102 trustees reported the multiplier truthfully, and 32 trustees misreported it. Out of 68 trustees who observed the high multiplier, 25 trustees misreported it. If we restrict observations to trustees with a lying aversion of 0 or 1, as done in the main text, we see that 19 out of 55 trustees misreported the multiplier. There does not seem to be a relation between misreporting the multiplier down in the trust game with misreporting the die-rolling outcomes in the die-rolling task (Table 3A.2). Out of 66 trustees who observed the low multiplier, 7 trustees misreported it upwards.

3A.3 Robustness checks on regression results

The tables below contain robustness checks we performed on our results. Tables 3A.1 and 3A.2 report results using alternative specifications for dividing trustee population as lying averse and not lying averse, for trustee returns and trustor expectations, respectively. Tables 3A.3 and 3A.4 report results including the full set of the control variables, for trustee returns and trustor expectations, respectively.

We measure the lying aversion of subjects with the die-rolling task. In order to eliminate the luck component in the die-rolling task, without it taking extremely long, we let subjects perform the task for three rounds. An individual's lying aversion measure from the die-rolling task might however still be blurred to some extent by their luck. For instance,

consider a subject, who has not observed any 5s but reported truthfully in the first two rounds of the die-rolling task, but observes a 5 in the third round. In this case, she ends up with a positive payoff without any lying; however, she might have lied to earn some money if she had not observed a 5 in the last round. In this case, she would be classified as someone with a lying aversion of 1, whereas she could have instead been classified as someone with a lying aversion of 0. We therefore re-did our analyses (i) by including the number of 5s observed by a trustee as an independent variable, and (ii) by restricting the data to the subset of trustees who did not observe any 5s, in order to eliminate this type of noise. Tables 3A.5 and 3A.6 report results controlling for the number of 5s observed by trustees, for trustee returns and trustor expectations, respectively. Tables 3A.7 and 3A.8 report results including only the trustees who never observed a 5, for trustee returns and trustor expectations, respectively.

The dependent variables in our analysis, trustee returns and trustor expectations, are between 0 and 1. Tables 3A.9 and 3A.10 report results for the fractional response model specification, for trustee returns and trustor expectations, respectively; as a robustness check using a method specifically developed for bounded dependent variables. Finally, Tables 3A.11 and 3A.12 report results excluding outliers who sent/expected back more than 50%, for trustee returns and trustor expectations, respectively.

	(1) Ratio	(2) Ratio	(3) Ratio	(4) Ratio	(5) Ratio	(6) Ratio
C	0.0379		0.0389		0.0366	
	(0.313)		(0.301)		(0.331)	
Multiplier	-0.0601*		-0.0538*		-0.0635**	
	(0.053)		(0.081)		(0.045)	
(C=1)#(M=1)	0.0291		0.0123		0.0299	
	(0.606)		(0.810)		(0.616)	
Lying av. continuous	0.0826***					
	(0.009)					
(C=1)#(M=1)#(LA)	-0.0544					
	(0.296)					
(C=0)#(M=1)#(1-LA)		-0.147***				
		(<0.001)				
Lying av. 1 vs. not 1			0.0601**			
			(0.029)			
(C=1)#(M=1)#(LA=1)			-0.0404			
			(0.376)			
(C=0)#(M=1)#(LA=0)				-0.101***		
				(<0.001)		
Lying av. 0 vs. not 0					0.0707**	
					(0.019)	
(C=1)#(M=1)#(LA=1)					-0.0444	
					(0.366)	
(C=0)#(M=1)#(LA=0)						-0.146***
						(<0.001)
Constant	0.252***	0.300***	0.268***	0.299***	0.254***	0.295***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
N	272	272	272	272	272	272
adj. R^2	0.045	0.045	0.034	0.032	0.040	0.038

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.1: OLS regression results on trustee returns with the alternative comparisons of trustees' lying aversion measures. Columns 1-2 include the continous measure of lying aversion, Columns 3-4 include the dummy variable measuring whether the trustee's lying aversion is 1 or whether it is lower than 1, and Columns 5-6 include the dummy variable measuring whether the trustee's lying aversion is 0 or whether it is higher than 0. This implies that Columns 3-4 separate individuals as those who never overreported and those who overreported at least once, and Columns 5-6 separate individuals as those who always overreported and those who reported truthfully at least once. In Columns 1-2, LA is the actual level of the lying aversion measure. In Columns 3-6, LA=1 and LA=0 indicate the values of the indicator variable for the corresponding lying aversion comparison.

	(1) Ratio	(2) Ratio	(3) Ratio	(4) Ratio	(5) Ratio	(6) Ratio
C	-0.0265	-0.0253	-0.0298	-0.0287	-0.0272	-0.0261
	(0.287)	(0.319)	(0.226)	(0.253)	(0.278)	(0.309)
Multiplier	-0.00679	-0.00567	-0.00679	-0.00567	-0.00679	-0.00567
	(0.270)	(0.388)	(0.270)	(0.388)	(0.270)	(0.388)
(C=1)#(M=1)		0.000438		0.000690		-0.00333
		(0.983)		(0.969)		(0.880)
Lying av. continuous	0.150***	0.152***				
	(<0.001)	(<0.001)				
(C=1)#(M=1)#(LA)		-0.00500				
		(0.866)				
Lying av. 1 vs. not 1			0.125***	0.127***		
			(<0.001)	(<0.001)		
(C=1)#(M=1)#(LA=1)				-0.00654		
				(0.798)		
Lying av. 0 vs. not 0					0.121***	0.120***
					(<0.001)	(<0.001)
(C=1)#(M=1)#(LA=1)						0.00176
						(0.953)
Constant	0.244***	0.242***	0.272***	0.270***	0.250***	0.250***
	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
N	536	536	536	536	536	536
adj. R^2	0.141	0.138	0.119	0.115	0.103	0.099

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.2: OLS regression results on trustor expectations with the alternative comparisons of trustees' lying aversion measures. Columns 1-2 include the continous measure of lying aversion, Columns 3-4 include the dummy variable measuring whether the trustee's lying aversion is 1 or whether it is lower than 1, and Columns 5-6 include the dummy variable measuring whether the trustee's lying aversion is 0 or whether it is higher than 0. This implies that Columns 3-4 separate individuals as those who never overreported and those who overreported at least once, and Columns 5-6 separate individuals as those who always overreported and those who reported truthfully at least once. In Columns 1-2, LA is the actual level of the lying aversion measure. In Columns 3-6, LA=1 indicates the value of the indicator variable for the corresponding lying aversion comparison.

	(1) Ratio	(2) Ratio	(3) Ratio
Lying av.	0.0288	0.0405	
	(0.281)	(0.226)	
C	0.0254	0.0253	
	(0.317)	(0.532)	
Multiplier	-0.0588**	-0.0600*	
	(0.017)	(0.079)	
SVO angle	0.00398***	0.00397***	0.00413***
	(<0.001)	(<0.001)	(<0.001)
Female	-0.0514**	-0.0528**	-0.0501*
	(0.050)	(0.049)	(0.053)
Age	0.00115	0.00123	0.00107
	(0.723)	(0.707)	(0.739)
Experience1	-0.0187*	-0.0179	-0.0178*
	(0.087)	(0.111)	(0.099)
Experience2	0.00966	0.00812	0.00950
	(0.520)	(0.597)	(0.523)
Europe	-0.0597**	-0.0603**	-0.0575**
	(0.033)	(0.035)	(0.030)
Econ	-0.0192	-0.0183	-0.0179
	(0.498)	(0.523)	(0.511)
(C=1)#(M=1)		0.0233	
		(0.709)	
(C=1)#(M=1)#(LA=1)		-0.0426	
		(0.412)	
(C=1)#(M=1)#(LA=0)			-0.125***
			(0.001)
Constant	0.293***	0.285***	0.299***
	(0.001)	(0.002)	(<0.001)
N	222	222	222
adj. R^2	0.134	0.128	0.146

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.3: OLS regression results on trustee returns including control variables. SVO angle is the SVO angle of the subjects based on their choices in the SVO task. Female is a dummy variable that takes the value of 1 for females. Age is the age of the subject. Experience1 and Experience2 are subjects' experience level in experiments in the CREED lab and outside, respectively. Europe is a dummy variable indicating European nationality. Econ is a dummy variable indicating whether the subject majors in a department in the Economics and Business Faculty.

	(1) Ratio	(2) Ratio
Lying av.	0.143***	0.144***
	(<0.001)	(<0.001)
C	-0.0255	-0.0238
	(0.343)	(0.394)
Multiplier	-0.00842	-0.00676
	(0.206)	(0.378)
Lying av. trustor	0.0120	0.0121
	(0.744)	(0.744)
SVO angle	0.00124	0.00124
	(0.241)	(0.241)
Female	0.00415	0.00419
	(0.883)	(0.882)
Age	0.00341**	0.00341**
	(0.038)	(0.038)
Experience1	-0.0234*	-0.0234*
	(0.063)	(0.065)
Experience2	-0.0308	-0.0308
	(0.159)	(0.161)
Europe	-0.0137	-0.0137
	(0.697)	(0.698)
Econ	-0.0191	-0.0190
	(0.643)	(0.645)
(C=1)#(M=1)		-0.00144
		(0.945)
(C=1)#(M=1)#(LA=1)		-0.00341
		(0.909)
Constant	0.206**	0.204**
	(0.024)	(0.028)
N	438	438
adj. R^2	0.175	0.171

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.4: OLS regression results on trustor expectations including control variables. SVO angle is the SVO angle of the subjects based on their choices in the SVO task. Female is a dummy variable that takes the value of 1 for females. Age is the age of the subject. Experience1 and Experience2 are subjects' experience level in experiments in the CREED lab and outside, respectively. Europe is a dummy variable indicating European nationality. Econ is a dummy variable indicating whether the subject majors in a department at the Economics and Business Faculty.

	(1) Ratio	(2) Ratio	(3) Ratio
Lying av.	0.0639**	0.0779**	
	(0.014)	(0.015)	
C	0.0278	0.0168	
	(0.279)	(0.684)	
Multiplier	-0.0619**	-0.0741**	
	(0.016)	(0.038)	
No. of 5s	0.00694	0.00770	0.00824
	(0.708)	(0.678)	(0.656)
(C=1)#(M=1)		0.0503	
		(0.424)	
(C=1)#(M=1)#(LA=1)		-0.0523	
		(0.325)	
(C=1)#(M=1)#(LA=0)			-0.149***
			(<0.001)
Constant	0.263***	0.260***	0.294***
	(<0.001)	(<0.001)	(<0.001)
N	222	222	222
adj. R^2	0.038	0.033	0.041

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.5: OLS regression results on trustee returns controlling for the number of 5s observed by the trustee.

	(1)	(2)
	Ratio	Ratio
Lying av.	0.144***	0.145***
	(<0.001)	(<0.001)
C	-0.0306	-0.0289
	(0.267)	(0.313)
Multiplier	-0.00842	-0.00676
	(0.202)	(0.375)
No. of 5s	0.00543	0.00546
	(0.756)	(0.756)
(C=1)#(M=1)		-0.000837
		(0.968)
(C=1)#(M=1)#(LA=1)		-0.00453
		(0.880)
Constant	0.250***	0.249***
	(<0.001)	(<0.001)
N	438	438
adj. R^2	0.144	0.140

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.6: OLS regression results on trustor expectations controlling for the number of 5s observed by the trustee.

	(1) Ratio	(2) Ratio	(3) Ratio
Lying av.	0.0730**	0.0888**	
	(0.043)	(0.041)	
C	0.0396	0.0000379	
	(0.265)	(0.999)	
Multiplier	-0.0579*	-0.101**	
	(0.095)	(0.024)	
(C=1)#(M=1)		0.111	
		(0.186)	
(C=1)#(M=1)#(LA=1)		-0.0524	
		(0.463)	
(C=1)#(M=1)#(LA=0)			-0.187***
			(<0.001)
Constant	0.251***	0.263***	0.298***
	(<0.001)	(<0.001)	(<0.001)
N	120	120	120
adj. R^2	0.037	0.035	0.068

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.7: OLS regression results on trustee returns, restricted to the group of trustees who never observed a 5.

	(1)	(2)
	Ratio	Ratio
Lying av.	0.191***	0.189***
	(<0.001)	(<0.001)
C	-0.0770**	-0.0710*
	(0.041)	(0.067)
Multiplier	-0.0134*	-0.00725
	(0.085)	(0.444)
(C=1)#(M=1)		-0.0171
		(0.503)
(C=1)#(M=1)#(LA=1)		0.00946
		(0.817)
Constant	0.252***	0.250***
	(<0.001)	(<0.001)
N	236	236
adj. R^2	0.253	0.247

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.8: OLS regression results on trustor expectations, restricted to the group of trustees who never observed a 5.

	(1) Ratio	(2) Ratio	(3) Ratio
Lying av.	0.194**	0.237**	
	(0.013)	(0.014)	
C	0.0844	0.0468	
	(0.260)	(0.684)	
Multiplier	-0.183**	-0.227**	
	(0.014)	(0.033)	
(C=1)#(M=1)		0.165	
		(0.384)	
(C=1)#(M=1)#(LA=1)		-0.154	
		(0.335)	
(C=0)#(M=1)#(LA=0)			-0.510***
			(0.002)
Constant	-0.632***	-0.636***	-0.529***
	(<0.001)	(<0.001)	(<0.001)
N	222	222	222

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.9: Fractional response regression results on trustee returns.

	(1) Ratio	(2) Ratio
Lying av.	0.419*** (<0.001)	0.418*** (<0.001)
C	-0.0885* (0.071)	-0.0839 (0.229)
Multiplier	-0.0238 (0.627)	-0.0192 (0.802)
(C=1)#(M=1)		-0.0118 (0.927)
(C=1)#(M=1)#(LA=1)		0.00463 (0.968)
Constant	-0.673*** (<0.001)	-0.675*** (<0.001)
N	438	438

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.10: Fractional response regression results on trustor expectations.

	(1) Ratio	(2) Ratio	(3) Ratio
Lying av.	0.0494**	0.0623**	
	(0.035)	(0.027)	
C	0.0104	-0.0133	
	(0.648)	(0.696)	
Multiplier	-0.0351	-0.0594*	
	(0.124)	(0.073)	
(C=1)#(M=1)		0.0701	
		(0.217)	
(C=1)#(M=1)#(LA=1)		-0.0428	
		(0.397)	
(C=1)#(M=1)#(LA=0)			-0.132***
			(0.001)
Constant	0.255***	0.260***	0.282***
	(<0.001)	(<0.001)	(<0.001)
N	216	216	216
adj. R^2	0.017	0.016	0.047

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.11: OLS regression results on trustee returns, excluding the outliers who sent back more than 50% of the available amount.

	(1)	(2)
	Ratio	Ratio
Lying av.	0.164***	0.170***
	(<0.001)	(<0.001)
C	-0.00710	-0.00635
	(0.703)	(0.748)
Multiplier	-0.00610	-0.00546
	(0.380)	(0.520)
(C=1)#(M=1)		0.0110
		(0.616)
(C=1)#(M=1)#(LA=1)		-0.0225
		(0.414)
Constant	0.195***	0.192***
	(<0.001)	(<0.001)
N	404	404
adj. R^2	0.285	0.283

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3A.12: OLS regression results on trustor expectations, excluding the outliers who expected back more than 50% of the available amount.

3A.4 Instructions

Our instructions can be found below. The subjects received the instructions on their computer screens. In the trust game, each subject got instructions from the perspective of their role in the trust game on their screen. Subjects were also handed out printed copies of the general instructions for the die-rolling task and the trust game. Unless stated otherwise, the instructions were the same for both treatments. The instructions for the die-rolling task are based on Gross et al. (2018), and the instructions for the SVO task are based on Böhm et al. (2018).

3A.4.1 Die-rolling instructions

Part 1 of the experiment consists of 3 rounds of a die rolling task.

- In each round, you will see a random roll of a die on the computer screen.
- Your task is to report the outcome of the die roll that you saw on the screen.
- Your payoff will be determined by the result that you report.
- You will be paid for all of your reports. Specifically, for each round, ...
  - ... you will earn 75 points if you report a 5.
  - ... you will earn 0 points otherwise.

Figure 3A.1: A screenshot of the die-roll task with the videos created by Kocher et al. (2018).

3A.4.2 Trust game instructions

In Part 2 of the experiment, you will interact with two other participants. In this interaction, there are three different roles: *Participant A, Participant B, and Participant C*

The order of events is as follows:

1. Participant A is given 50 points. Participant A observes the die rolls and the reports of Participant B and Participant C from Part 1 of the experiment. Participant A then chooses to send the 50 points either to Participant B or to Participant C. Until the end of Part 2, Participant B and Participant C will not learn who was picked by Participant A.

2. Upon sending, the 50 points are multiplied by 2 or by 4. The multiplier is chosen randomly and both multipliers are equally likely. Participant B and Participant C both observe the multiplier. Participant A does not observe the multiplier.

3. Participant B and Participant C both choose how many points they want to send back to Participant A, in case they are chosen. Participant B and Participant C can choose to send back any amount between 0 points and the full amount available, which is 100 points if the multiplier is 2, and 200 points if the multiplier is 4.

4. **(Only in the communication treatment)** Together with the amount to be sent back, Participant B and Participant C will also choose a message about the multiplier. They can choose to say to Participant A that the multiplier was 2 or that it was 4.

5. The choices of the participant that Participant A selected are implemented. Participant A observes the amount sent back by the participant that s/he chose. **(Only in the communication treatment)** Participant A also observes the message sent by the participant that s/he chose.

6. **(Only in the communication treatment)** Participant A then sends a reply to the chosen participant. S/he can either reply that s/he believes, or that s/he does not believe the message sent by this participant.

7. Participant B and Participant C learn who was chosen by Participant A. **(Only in the communication treatment)** The participant that was chosen by Participant A then observes the reply that Participant A sent to her/him.

The payoffs in Part 2 are determined as follows:

- Participant A earns the amount sent back by the participant that s/he chose.
- The participant that Participant A chose earns the amount s/he received minus the amount s/he sent back to Participant A.
- The participant that was not chosen earns an amount that is determined by a random draw. The range for this random draw depends on the multiplier. Specifically, ...

- ... if the multiplier is 2, the non-chosen participant earns an amount ranging between 0 and 100 points, with increments of 1 point. Any of these possible payoffs is drawn with equal probability.
- ... if the multiplier is 4, the non-chosen participant earns an amount ranging between 0 and 200 points, with increments of 1 point. Any of these possible payoffs is drawn with equal probability.

Belief elicitation instructions In the next page, you will be asked to predict how many points Participants B and C will choose to send back to you. Both for Participant B and for Participant C, we will ask you to predict what that participant will send back in case that the multiplier is 2, and in case that the multiplier is 4. So, you will submit four predictions in total.

Based on your predictions about each participant, you will have a chance of winning a prize of 50 points. The closer your prediction is to the actual choice of the participant, the higher your chance of winning the prize will be.

More specifically, given the actual value of the multiplier, we will compare the amounts that Participant B and Participant C (would have) sent back with the amounts that you predicted they would send back for that value of the multiplier. We then use a formula to determine your chance of winning the prize. The formula is designed such that your chance of winning the prize becomes higher, the closer your prediction is to the actual choice of the participant.

Note that the formula ensures that you maximize your chance of winning money by stating your true beliefs regarding the choices of Participant B and Participant C for each level of the multiplier. It is not important that you understand the formula in detail. What matters is for you to know that you maximize your chances of winning by truthfully reporting your best estimate regarding the other participants' choices. If you under- or overstate your true beliefs, you will reduce your chances of winning the prize.

However, if you want to take a closer look at the formula, we have prepared some more detailed information for you. Below you can choose to proceed straight to the prediction task, or to first review the additional information on the formula.

Do you want to review the details of the formula?

[IF YES]

How a given answer translates to your chance of winning the prize of 50 points is based

on a formula. The formula is designed to make sure that you maximize your chance of winning if you truthfully report your best estimates of the amounts you expect Participant B and Participant C to send back for each level of the multiplier.

Suppose that the multiplier is 2, Participant B chooses to send you R_1 points and Participant C chooses to send you S_1 points. The variables r_1 and s_1 are your reports for this case - your reports on how much you think Participants B and C will choose to send back to you in case the multiplier is 2. The winning probabilities $p_{1,B}$ for your prediction on Participant B and $p_{1,C}$ for your prediction on Participant C are then given by:

$$p_{1,B} = 1 - \left(\frac{|r_1 - R_1|}{100}\right)^{1/2} \qquad \text{and} \qquad p_{1,C} = 1 - \left(\frac{|s_1 - S_1|}{100}\right)^{1/2}$$

Similarly, for the case where the multiplier is 4, the winning probabilities $p_{2,B}$ for your prediction on Participant B and $p_{2,C}$ for your prediction on Participant C are then given by:

$$p_{2,B} = 1 - \left(\frac{|r_2 - R_2|}{200}\right)^{1/2} \qquad \text{and} \qquad p_{2,C} = 1 - \left(\frac{|s_2 - S_2|}{200}\right)^{1/2}$$

All of the p-functions attain the value of 1 if, and only if, your guess is equal to the actual amount sent by (respectively) Participant B or Participant C. Moreover, using some calculus, one can show that each of the p-functions is strictly increasing below the actual amount sent by (respectively) Participant B or Participant C, and it is strictly decreasing above that amount. So, while you do not know the actual amount each participant decided to send back for a given amount of the multiplier, the best thing you can do to make your probability of winning as high as possible, is to submit the average amount that you expect a participant like Participant B, or like Participant C, to return.

Below you can find a simple example. Suppose that the multiplier is 2, and that you believe that in this case, Participant B would choose to send you 25 points with probability 0.1, 33 points with probability 0.8, and 41 points with probability 0.1. In the table below, you can find the probability of winning the prize of 50 points for different beliefs you can choose to report about Participant B's choice for this case.

For example, if you report that you expect Participant B to send back 25 points, the probability that you win the prize is 73.37%. As you can see, you maximize your chance of winning by reporting the average amount that you expect Participant B to return.

	Report #1	Report #2	Report #3	Report #4	Report #5
Hypothetical report r_1	12	25	33	41	56
Expected winning probability $p_{1,B}$	54.35%	73.37%	94.34%	73.37%	52.19%

3A.4.3 Social value orientation task instructions

- In Part 4, you will be making a series of 6 decisions about allocating resources between yourself and a randomly selected participant.
- The participant you will be matched with will be a different person than the participants you were matched with in Part 2.
- In every decision, a slider bar will represent the allocations that are available to you. You select your most preferred allocation by marking the corresponding position on the slider bar.
- At the end of the experiment, we will randomly pick one of the decisions from Part 4 for payment.
- You will be paid twice, once being in the *sender role* and once being in the *receiver role*:
  - In the *sender role*, you get the points you decided to keep on the selected decision according to your own allocation decision. A randomly selected participant will be in the receiving role for your allocation decision.
  - In the *receiver role*, you will receive points based on the allocation decision of a second randomly chosen participant for that decision. That participant will be different from the person that received points based on your allocation decision and all other participants you were previously matched with in the experiment.

3A.4.4 Survey items

Please answer the following questions:

- What is your age?
- What is your gender?
- What are you studying?
- Where are you from?

- How many experiments have you participated at CREED (excluding this experiment)? Please select the right category below.
- How many experiments have you participated at other institutes than CREED? Please select the right category below.
- Reminder: In Part 1 of the experiment, you did a die rolling task for 3 rounds. In each round, you observed a random roll of a die on the computer screen and reported the outcome of the die roll. You can see the rolls you observed and the outcomes you reported in the table below.

  [REMINDER ON PART I BEHAVIOUR]

  How did you make your choices in Part 1? Could you describe the method(s) you followed?

- Reminder: In Part 2 of the experiment, you interacted with two other participants. In this interaction, Participant A chose either Participant B or Participant C, and sent 50 points to that participant. These points were then multiplied by 2, or by 4. Participants B and C observed the multiplier and chose an amount to send back to Participant A.

  (The communication treatment) Together with the amount to be sent back, Participants B and C also chose a message about the multiplier. The choices of the participant that was chosen by Participant A were implemented. Participant A then observed the amount and the message sent by the participant that s/he chose.

  (The no communication treatment) The choice of the participant that was chosen by Participant A was implemented. Participant A then observed the amount sent by the participant that s/he chose.

  [REMINDER ON PART II ROLE AND BEHAVIOUR]

  How did you make your choices in Part 2? Could you describe the method(s) you followed?

Chapter 4

The evolution of morality and the role of commitment[1]

4.1 Introduction

There is an extensive theoretical literature on the evolution of cooperation. Most papers in this literature (including our own) present models in which individuals play prisoners' dilemmas, or public goods games, and look for ways in which cooperation can outperform defection. If we paint the mechanisms at work with a broad brush, then, in most of those models, cooperation evolves because of population structure (which often means that it can be seen as kin selection) or because of repeated interactions between players, with partner choice coming in third at a respectable distance.

These models can be elegant and technically gratifying, but the match between what evolves in these models and the empirical evidence for human cooperation in the real world is not overwhelming. One of the ways in which it is less than spectacular is that it does not give a good answer to the question why humans cooperate more than other species. Our population structure is not that different from other primates – relatedness within groups of human hunter gatherers is similar to that of chimpanzees or gorillas – and our interactions are also not more repeated. One theory that points to a possible human-specific cause is cultural group selection, which suggests that cultural inheritance creates a population structure that differs from the one in which genetic inheritance takes place. We will discuss this in more detail in Sections 4.2 and 4.5, along with other things related to the cross-species evidence.

[1]This chapter is based on Akdeniz and van Veelen (2021).

Here, we suggest another possible explanation, namely, that the difference between humans and other species is not caused by differences in population structure or repetition rates, but by humans playing different games. Humans are a social technological species; our niche requires us to make a living in ways that involve planning ahead and working together. This opens doors for opportunistic behaviours that do not exist in other species. Typical strategic situations for humans therefore may be better described by games with a time component, like the ultimatum game or the trust game. In games that consist of a sequence of choices it is possible for cooperation to unravel if individuals behave opportunistically, while cooperation can be sustained if players can commit to not doing that.

In this paper, we will go over a few examples to illustrate how that makes for a proper different mechanism for the evolution of what is usually called pro-social behaviour, and that we will sometimes call "rationally irrational" behaviour if we want to stress the difference between what is fitness maximizing *ex ante* and what would be fitness maximizing *ex post*. The core of the mechanism is that not behaving selfishly reduces your fitness, but being committed to not behaving selfishly can increase your fitness. The reason for why this works is that being committed to not behaving selfishly can have an effect on how other individuals, with their own interest at heart, then behave towards you. This does not require population structure or positive relatedness between individuals, nor does it need interactions to be repeated. It can very well work through partner choice, but commitment does not need the freedom to choose your partner, as being committed can also have an advantageous effect on the behaviour of existing partners.

The idea that the purpose of our moral sentiments is to allow us to credibly commit to otherwise irrational behaviours is by no means new. It is the central premise of the book *Passions Within Reason* by Robert Frank (1988), which in turn refers to *The Strategy of Conflict* by Thomas Schelling (1960) as a source of inspiration (see also Frank, 1987, Hirshleifer, 1987, Schelling, 1978, and Quillien, 2020). In the first chapter of *Evolution and the Capacity for Commitment*, Randolph Nesse (2001) also identifies commitment as a mechanism that is different from repetition and population structure, as do other authors in the book, including Frank (2001) and Hirshleifer (2001). Moreover, the literature on the role of reputation in ultimatum games (Nowak et al., 2000) or in games with punishment (Brandt et al., 2003; dos Santos et al., 2011; 2013; dos Santos and Wedekind, 2015; Hauert et al., 2004; Hilbe and Traulsen, 2012; Sigmund et al., 2001) also fits with this idea, because knowing who is and who is not committed is a prerequisite for commitment to evolve. However, even though the idea of commitment has been around for a while, it

is hardly ever used to interpret the empirical evidence – with exceptions, such as Smith (2005) – and it is almost always absent in overviews of mechanisms for the evolution of cooperation – again with exceptions, such as Sterelny (2012).

Over the last thirty-odd years, a theoretical and an empirical literature have developed alongside each other, without too much emphasis on possible discrepancies between the two. Besides the modest cross-species predictive power of much of the theory, one of the other ways in which theory and empirical data do not match concerns the nature of the pro-social behaviour. In models with prisoners' dilemmas or public goods games, and population structure, for instance, what evolves is a willingness to forego fitness for the benefit of another individual, as long as these benefits to the other are sufficiently high to outweigh the costs to oneself. Not all deviations from simple selfishness that are observed in experiments, however, fit that mold – even if they all travel under the same banner in the empirical literature. Rejecting offers in the ultimatum game, for instance, is hardly accurately described as cooperative or pro-social. Rejections would be pro-social, if they increased the fitness of the other player, but that is not what they do; they reduce the fitness for both players involved.

If commitment evolves, it therefore does not necessarily advance the common good; it can do that, as we will see, in games like the trust game, but in games like the ultimatum game, it just helps individuals secure a larger share of a fixed-size pie. Indeed, even commitment that hurts the common good can evolve. While the particulars of the deviations from simple selfishness that empirical studies find are at odds with what can evolve in models with prisoners' dilemmas or public goods games, they do align with what a theory that looks at the benefits of commitment would predict, as we will see in more detail in Section 4.4. A theory of commitment thereby not only covers the presence (or absence) of good, but it predicts good as well as evil to be part of human nature. In this paper, we further argue that a theory of commitment aligns better with a number of other aspects of human nature, such as our taste for revenge, our preoccupation with sincerity, and the existence of "hypothetical reciprocity", that is, a sensitivity to whether others *would* have done the same for you, over and above what others actually did.

The remainder of the paper is organised as follows. In Sections 4.2 and 4.3 we take a look at theories for the evolution of cooperation. In Section 4.4 we review how well the empirical evidence for humans fits the different mechanisms, and in Section 4.5 we consider the cross species evidence.

4.1.1 A note on terminology

It is not always possible to choose labels that are concise and consistent with all of the literature. We will use *cooperation* first of all for behaviour that benefits someone else. In the literature, sometimes this is subdivided into mutualistic cooperation (or cooperation with direct benefits, or byproduct mutualism), and costly cooperation. That can be a useful distinction, but if we are looking for an explanation for a behaviour that at least momentarily comes with a fitness cost to the agent, then whether it is one or the other depends on the explanation. When we consider different possible explanations, the most concise term therefore will just be cooperation, without qualifiers. More generally, in games other than the prisoners' dilemma and the public goods game, one can identify (combinations of) behaviours that can be described as cooperative, but we will regularly refer to those in more descriptive standard terms.

We will use *altruism* to describe the willingness to give up payoffs, or fitness, to the benefit of another individual. This describes a preference, or a pattern of behaviour, that deviates from what in the literature is described as selfish money-maximizing, and that we will refer to as *simple selfishness*.

4.2 Models for the evolution of cooperation

Before we discuss the role of commitment in the evolution of human cooperation, we will briefly review the existing models in which commitment is not possible. This will be useful for when we compare how well the empirical data match models with and without commitment. Most of the literature without commitment focuses on prisoners' dilemmas, and, to a lesser extent, on public goods games.

4.2.1 The prisoners' dilemma

The prisoners' dilemma is usually – and with good reason – seen as the purest, most distilled description of the problem of cooperation. It has two players. Both can choose between cooperation (C) and defection (D). Their payoff, or fitness, depends on the combination of their choices; if both of them cooperate, they receive a payoff that is regularly referred to as R for reward; if both defect, they receive a payoff that is usually referred to as P for punishment; and if one defects and the other cooperates, the usual names for their payoffs are temptation (T) for the defector, and the sucker's payoff (S) for the cooperator. This is conveniently represented in a payoff matrix.

$$\begin{bmatrix} & C & D \\ C & R & S \\ D & T & P \end{bmatrix}$$

There are two properties that are required for this to be an actual prisoners' dilemma. The first is that for both players, playing D must be better than playing C, whatever the other player does. That means that $T > R$ and $P > S$. The second property is that mutual cooperation has to be better than mutual defection, or, in other words, $R > P$. These two properties make the prisoners' dilemma an interesting game, because together they imply that there is a tension between the players' individual interests – which is to defect – and their collective interests – which is for both to cooperate.

4.2.2 The public goods game

In the standard public goods game, players can choose how much to contribute to a public good. For the individual, the benefits of the public good are assumed to be lower than the costs of contributing, and therefore it in everyone's individual interest not to contribute. Players, however, also benefit from each other's contribution to the public good, and therefore we can assume that the joint benefits are higher than the individual costs of contributing. This makes it in the collective interest for everyone to contribute everything.

The public goods game is therefore a generalized version of the prisoners' dilemma; it allows for 2 or more players, and it allows players to also choose intermediate levels of cooperation, rather than just giving them a binary choice.

In the standard public goods game, every additional contribution to the public good increases the benefits to everyone by the same amount. Other versions of the public goods game allow the benefits to also depend on the joint contributions in more interesting ways than just linearly (Archetti and Scheuring, 2012; Palfrey and Rosenthal, 1984).

4.2.3 Why cooperate in prisoners' dilemmas

The explanations for the evolution of cooperation can be classified in three broad categories; repetition; population structure; and partner choice.

Repeated interactions When prisoners' dilemmas are played repeatedly, this changes the game. Players now have the opportunity to reward cooperative behavior, and retali-

ate against defection. If the probability of another interaction is high enough, and both players reciprocate, cooperation can become the self-interested thing to do. There is an extensive literature on the large variety of equilibria that this "shadow of the future" creates (Fudenberg and Levine, 2008; Fudenberg and Maskin, 1986; Mailath and Samuelson, 2006), and their relative stability (Axelrod and Hamilton, 1981; Bendor and Swistak, 1995; García and van Veelen, 2016; van Veelen and García, 2019).

There is no doubt that repetition matters, and that humans have evolved reciprocity. Experimental evidence indicates that people understand that others will reciprocate, and that repetition therefore changes incentives (Dal Bó, 2005; Dal Bó and Fréchette, 2018). The remarkable thing, however, is that people sometimes also cooperate, help others, and think it is wrong to be selfish, when interactions are not repeated. One possible explanation for this is that that most of our everyday interactions are repeated, and the rarity of real one-shot encounters means that it is not worth differentiating (Delton et al., 2011). There are theoretical objections against that argument, as an easy way around this would be to defect in the first round, and only start cooperating when the game turns out to be repeated (see Jagau and van Veelen, 2017, for a more general and precise version). Moreover, it is somewhat hard to reconcile the idea that people have a hard time differentiating between repeated and one-shot games with the finding that people can and do differentiate rather accurately between repeated games with high and with low probabilities of repetition (Dal Bó, 2005; Dal Bó and Fréchette, 2018). In addition, although the rarity of one-shot interactions (in the distant past) is a possibility, it is not an established fact. Something to consider when thinking about repetition rates is that even if interactions happen between people that know each other, and that are very likely to meet again, major opportunities for helping each other out (or for doing something bad, like selling someone out) may only present themselves once in a blue moon. If high stakes games are few and far between, that means that the effective repetition rate for those may be too low to evolve reciprocity, even if players interact with low stakes more regularly (Jagau and van Veelen, 2017).

Population structure Population structure encompasses any deviation from a setup in which individuals are matched randomly for playing a prisoners' dilemma or a public goods game. For example, interactions can happen locally on networks (Allen et al., 2017; Lieberman et al., 2005; Ohtsuki et al., 2006; Santos and Pacheco, 2005; Santos et al., 2008; Taylor et al., 2007a), or within groups (Akdeniz and van Veelen, 2020; Luo, 2014; Simon et al., 2013; Traulsen and Nowak, 2006; Wilson and Wilson, 2007). In many such models, local dispersal causes neighbouring individuals, or individuals within the same group, to

have an increased probability of being identical by descent, and when they do, one can also see this as kin selection operating (Hamilton, 1964a;b; Kay et al., 2020).

One complication here is that on networks, for example, individuals may compete as locally as they have their opportunities for cooperation. If they do, then the cancellation effect prevents the evolution of cooperation (Taylor, 1992a;b; Wilson et al., 1992). Positive relatedness is therefore not enough. What is required for the evolution of cooperation is a discrepancy between how local cooperation is, and how local competition is (or a discrepancy between how related individuals are to those they cooperate with, and how related they are to their competitors). Because overcoming the cancellation effect is essential, and not always included in descriptions of what is needed for kin selection to work, Box 1 elaborates on this.

Some of these models allow for an interpretation with genetic transmission as well as an interpretation with cultural transmission. Others are explicitly one or the other. With respect to genetic transmission, one thing that is hard to square with the evolution of prosociality in humans is that people also cooperate with, and care for others, with whom they are not genetically related. This is at odds with the fact that, within this category of models, positive relatedness is a necessary, but, because of the cancellation effect, not even a sufficient condition for the evolution of altruism or costly cooperation. Some researchers have therefore suggested that what seems to be costly cooperation, or altruism, in public goods games in the lab, really is a mirage, caused by subjects being confused rather than pro-social (Burton-Chellew et al., 2016; Burton-Chellew and West, 2013). While their results suggest an interesting possibility, Camerer (2013) points to methodological flaws in Burton-Chellew and West (2013), and to a variety of ways in which an explanation based on confusion would be inconsistent with a host of other results (see also Andreoni, 1995, and Bayer et al., 2013). An explanation based on selfish, but confused subjects, is moreover at odds with what we observe in simpler experiments, in which there is no game, and all subjects have to do, is make choices that affect how much money they get themselves, and how much money someone else gets (Andreoni and Miller, 2002). Absent any other moving parts, this is the most straightforward setting to test for pro-social preferences, and here we do find that a sizable share of subjects is not simply selfish.

With respect to cultural transmission, many models show how cooperation could evolve, but not all models provide reasons why the details of such models match human population structure particularly well. One exception is cultural group selection, which suggests that conformism and norms make groups more homogeneous than they would otherwise be, and more homogeneous behaviorally than they are genetically (Bell et al., 2009; Handley

and Mathew, 2020). This then allows for group beneficial norms and costly cooperation to be selected. For group selection, cultural or not, it is relevant though that there is also a cancellation effect at the group level, which makes the evolution of costly cooperation harder, but not impossible (Akdeniz and van Veelen, 2020). We will return to cultural group selection in Section 4.4 and in Section 4.5, where we will also revisit payoff-biased imitation in general.

Partner choice Partner choice is a relatively small category (Barclay, 2004; 2013; Barclay and Willer, 2007; Baumard et al., 2013; McNamara et al., 2008; Melis et al., 2006; Sherratt and Roberts, 1998; Sylwester and Roberts, 2010). Here, the idea is that, if we can select with whom we play the game, then we can select cooperative traits in each other. This is also one of the two channels through which commitment can evolve, and we therefore return to this category below.

Mix and match Population structure, repeated interactions, and partner choice are very broad categories, but even then, the boundaries are not set in stone. Partner choice for instance can be seen as an endogenous source of population structure. Also some models combine ingredients from different categories, such as repetition and partner choice (Aktipis, 2004; Fujiwara-Greve and Okuno-Fujiwara, 2009; Izquierdo et al., 2014; 2010), or repetition and population structure (van Veelen et al., 2012).

4.3 Ultimatum games, trust games, backward induction and commitment

In order to understand the role of commitment, it helps to look at sequential games. This is what we will do below, and we will also introduce what *subgame perfection* is, and how *backward induction* works.

4.3.1 The ultimatum game

One classic example of a sequential game is the ultimatum game (Güth et al., 1982). This game is played between a proposer and a responder. The proposer makes an proposal to the responder regarding the distribution of a given amount of money, say 4 euros, between them. The responder can then accept or reject that proposal, and in cases where she rejects, neither player gets any money. If the proposer proposes, for instance, 3 for herself and 1 for the responder, then the responder chooses between, on the one hand, accepting and getting 1, and, on the other hand, rejecting and getting 0.

Box 1: The cancellation effect. One common, good intuition for how kin selection works, is that there can be a selective advantage for a gene that makes its carrier help other individuals, that are relatively likely to carry the same gene. Even if that help reduces the fitness of the helper, it can increase the expected number of copies of that gene in the next generation, through the help to these others. In the first decades after Hamilton (1964a;b), this intuition was thought to imply – understandably – that altruism can evolve, as soon as the possible helper and the possible recipient are related; for every $r > 0$, there is a benefit b and a cost c, such that $rb > c$. Therefore, when reproduction is local, and neighbours are related, one would expect altruism to evolve. Wilson et al. (1992) and Taylor (1992a;b) showed that this implication is not correct. The reason is that reproduction being local not only means that, if individuals have the opportunity to help their neighbour, they are related to the possible recipient, but it often also implies that competition happens between individuals that are close by, and therefore related too. If that is the case, then if I help my neighbour, the additional offspring that he or she gets goes at the expense of his or her neighbours (including me), and while I carry the gene for sure, in this scenario, also the other neighbours are related, and are therefore relatively likely to carry the same gene. This reduction in how much extra offspring of a related individual contributes to more copies of the gene in the next generation is called the *cancellation effect*. If the opportunities for cooperation are as local as competition is, cancellation is complete, and altruism does not evolve, regardless of the benefits and costs.

What is needed for altruism, or costly cooperation, to evolve, is that competition happens between individuals that are less related than those that have the opportunity for cooperation. In models with local dispersal and local interaction, that would require the opportunities for cooperation to occur more locally than the competition (see examples in Section 7 in van Veelen et al., 2017a). The need for this discrepancy is also the reason why kin recognition is effective for making kin selection happen. If competition happens between siblings, the cancellation effect would also prevent the evolution of altruism between them. However, during most of our life history, we compete with siblings and non-siblings alike. Therefore, if we recognize our siblings, and seek them out for (mutual) cooperation, this circumvents the cancellation effect.

Once a proposal has been made, the remainder of the game is called a *subgame*. There is a subgame for every possible proposal that the proposer can make. If we assume that

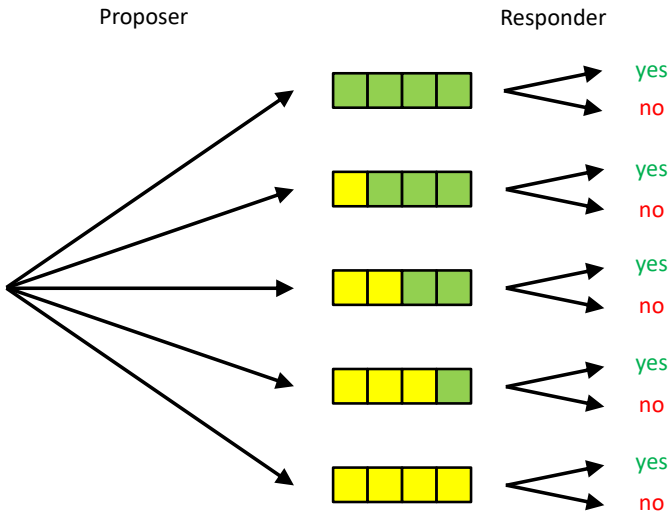


Figure 4.3.1: A simple version of the ultimatum game. The proposer chooses between proposals in which, from bottom to top, she gets 4, 3, 2, 1, and 0 herself, and the responder, also from bottom to top, gets 0, 1, 2, 3, and 4. For every proposal, the responder chooses whether or not to accept it. If the responder can commit to, for instance, rejecting the bottom two proposals, the proposer is best off proposing an equal split.

proposals can only be made in whole euros, then there is a subgame that starts after the proposer proposed 4 for herself and 0 for the responder; one that starts after the proposer proposed 3 for herself and 1 for the responder; and so on (see Figure 4.3.1). *Subgame perfection* now requires that that in any of these subgames, a Nash equilibrium is played, that is, that both players maximize their payoffs, given what the other does.

In all of these subgames, what the Nash equilibrium is, is simple. There is only one player that has any decision to make, and that is the responder. She always earns more by accepting rather than rejecting, unless the proposal is for her to receive 0, in which case she gets nothing either way.

Subgame perfection also assumes that in earlier rounds, players correctly anticipate their own future behavior and that of the other player in the different scenarios that could unfold. This means that the proposer anticipates that all proposals will be accepted, with the possible exception of the proposal in which the responder gets nothing. That leaves us with two subgame perfect Nash equilibria. In the first, the responder accepts every possible proposal, and the proposer, anticipating that all proposals will be accepted, proposes 4 for herself and 0 for the responder. In the second subgame perfect equilibrium,

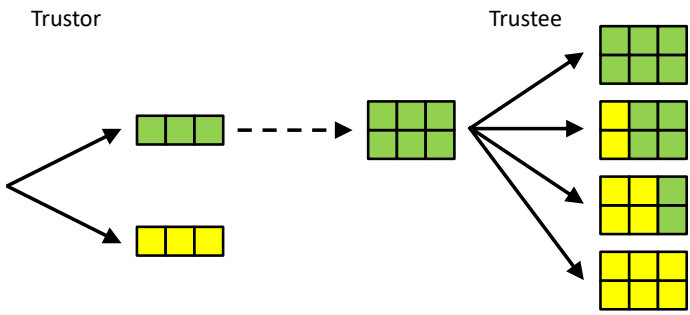


Figure 4.3.2: A simple version of the trust game. The trustor chooses whether or not to entrust the trustee with 3 euro. These 3 euros are doubled when entrusted to the trustee, who then gets to decide how much to send back; 0, 2, 4, or all 6 euros, from top to bottom. If the Trustee can commit to sending back 4, the Trustor is best off entrusting the Trustee with the money. Compared to the subgame perfect Nash equilibrium with selfish preferences, in which the Trustee does not return any money, and the Trustee does not send any money, this will be better for both.

the responder accepts every proposal, except for the one in which she gets 0, which she rejects. The proposer anticipates this, and proposes 3 for herself and 1 for the responder. (Here we assume that players do not randomize. If we allow them to randomize, we would get more subgame perfect equilibria, but in none of those does the responder ever get more than 1).

The process by which we find the subgame perfect Nash equilibria, i.e., start at the end of the game, determine what the equilibrium behavior will be when the players arrive at this point, and then work back towards the beginning of the game, under the assumption that players correctly anticipate their behavior in later stages, is called *backward induction*. This process also plays a role later in our argument, where we will see that the purpose of commitment is to alter the course of backward induction.

4.3.2 The trust game

Another classic example of a sequential game is the trust game (Berg et al., 1995), which is played between a trustor and a trustee. In this game, the trustor can choose an amount of money to send to the trustee. For simplicity, here we let the trustor choose between two options only: sending all (3) or nothing (0). In the original trust game, a range of values

is allowed for, but this makes it hard to visualize, hence the simplification. The amount that the trustor decides to send to the trustee then is multiplied by 2, and the trustee can choose how much of this multiplied amount of money she sends back to the trustor. Here, the options are: send back nothing; send back 2; send back 4; and send back all 6 euros (see Figure 4.3.2).

In this simple version of the trust game, there is only one proper subgame, which we arrive at when the trustee sends the 3 euros over (i.e., the trustee receives 6 euros). If she does, then the trustee maximizes how much she can keep, if she, in turn, sends back nothing. The subgame perfect Nash equilibrium of this game, therefore, is for the trustee to send back nothing, if the 6 euros come her way, and for the trustor, anticipating that the trustee will send back nothing, to just hold on to the 3 euros herself and send nothing.

What makes this game interesting, is that, like the prisoners' dilemma, there is a combination of choices that would leave both players better off than in the subgame perfect Nash equilibrium; if the trustor chooses to send the 3 euros over, and the trustor sends back 4, both will end up with a higher payoff; the trustor will have 4 instead of 3 euros, and the trustee will have 2 instead of 0.

4.3.3 Commitment

In both of these games, players can benefit from being able to commit to behaviour that one could describe as "rationally irrational", in the sense that the behaviour itself is not fitness maximizing, but being able to commit to it is.

In the ultimatum game, if the proposer knows that the responder will accept anything, then the proposer will propose 4 for herself, and 0 for the responder. If, on the other hand, the responder is committed to rejecting offers in which she gets less than, say 2, and the proposer knows this is the case, then it will be in the proposer's own best interest to accommodate this, and propose 2 for herself and 2 for the responder. Therefore, when possible, it is advantageous for the responder to commit to as high as possible a minimum amount that she would accept. The reason is that by doing so, she can change the behaviour of the proposer, or, in other words, she can alter the course of backward induction. A way to commit would be that when the proposer chooses to make a disadvantageous proposal, the responder actually *prefers* to walk away with nothing, provided that the responder also receives 0.

A similar commitment issue is central to the trust game. If the trustee is able to commit to sending back 4, and the trustor knows this, then the trustor should send the money over

– to their mutual benefit. As before, the benefit to the trustee of being able to commit to sending back money (a fitness reducing behaviour) is that, in doing so, it changes the behaviour of the trustor in ways that are fitness increasing. A way to commit to this would be to *prefer* to send back money, and to feel bad about not doing so.

The ability to commit can help an individual in two different ways. First, when matched to a given partner, commitment can influence the behaviour of that partner. In the ultimatum game, committing to rejecting (very) disadvantageous proposals can induce the proposer to make more generous proposals. In the trust game, committing to sending back money can induce the trustor to send money in the first place. It is however also possible that individuals can choose who they play the game with. If there are two possible trustees, and one trustor, and one of the possible trustees has a seemingly irrational preference for sending back a sizeable share, and the other does not, then the trustor should pick the irrational trustee, who then benefits from being picked. For the ultimatum game, on the other hand, partner choice works in the opposite direction, as proposers would prefer to interact with responders that reject less (Fischbacher et al., 2009).

Of course, this all assumes that commitment is, in fact, possible, and that others can figure out who is and who is not committed. A possible reaction to the idea of commitment therefore would be: "I understand that it would be beneficial to be able to commit to something that, when the time comes, runs against your interests, but I don't believe that one can." That raises a perfectly valid point. If a committed type has established itself, a mutant that seems committed, but is not, would have an advantage in the presence of noise or heterogeneity. Our suggestion, however, is to set aside the issue of credible commitment for now, and instead take a look at how people actually behave. We believe that the empirical evidence shows that evolution has found a way to make us prefer rejecting unfair proposals (Güth et al., 1982; Henrich et al., 2001; 2006; Oosterbeek et al., 2004) – which makes our behaviour different from chimpanzees (Jensen et al., 2007) – and that it has made us want to send back money after being entrusted with it (Alós-Ferrer and Farolfi, 2019; Berg et al., 1995; Johnson and Mislin, 2011). We also think our taste for revenge suggests we have managed to commit to punishment, our quest for sincerity suggests we have managed to commit to caring for each other for better or worse, and that even a preference for conditional cooperation in prisoners' dilemmas and public goods games can be a symptom of commitment. After assessing whether or not the empirical evidence is consistent with this notion of commitment, in these and other games, we can perhaps decide that the more important evolutionary question for humans is "how on earth did we manage to commit?" and not "why do we cooperate in prisoners' dilemmas". As

a matter of fact, we will suggest that answering the former may actually help us answer the latter.

4.4 Behaviour in the lab

Many papers in the theoretical literature refer to the behaviour to be explained in general terms, like (human) cooperation or prosociality. Many papers in the empirical literature, on the other hand, are not specific about the evolutionary mechanism being tested, and tend to aim more at characterizing the behaviour itself accurately. As a result, there is not always a well-trodden path between different parts of the theory and different parts of the empirical evidence. In this section, we will try to establish such links and show that, in many cases, there is some space left between predictions and empirical evidence in the absence of commitment. We begin with a detailed examination of the ultimatum game. Following this, we continue with a less detailed survey of other relevant games.

4.4.1 The ultimatum game

Selection without commitment The first possibility to consider is that in our evolutionary history, we have played sufficiently many games with the strategic structure of an ultimatum game, for us to assume that the behaviour we see in this game actually evolved for playing this game – but without the further assumption that commitment is possible. In a simple model with selection only, then, the relatively straightforward result is that responders evolve to accept all proposals in which they get positive amounts, while there is no selection pressure for or against accepting proposals in which they get nothing. Given this, proposers evolve to offer to the responder the smallest positive amount, or zero. A more precise version is given in the appendix, but this is the benchmark in the literature; a subgame perfect equilibrium with simply selfish money-maximizing preferences is selected. This is clearly not in line with what subjects do in the lab (Güth et al., 1982; Oosterbeek et al., 2004).

There is the possibility to move away from this outcome, either when there is noise, or when there are mutations. Gale et al. (1995) make the point that mistakes with smaller consequences may happen more frequently than more costly mistakes, and that, for the ultimatum game, this can make a difference. Rejecting a proposal in which you will receive almost nothing anyway is not very costly. In contrast, if responders already accept very disadvantageous proposals, then making a proposal that allocates even less to the responder, and therefore is rejected, is a much costlier mistake. Something similar applies

to mutations; genuinely costly mutations will be selected away pretty fast, or pretty surely, while less costly ones linger for much longer, or have a fair chance of not being weeded out (Rand et al., 2013). The relative abundance of not-so costly mistakes or mildly disadvantageous mutations can then change the selection pressure, and, in this case, move offers to responders upwards.

Rand et al. (2013) explicitly allow for a genetic as well as a cultural interpretation. There are complications with both. With a genetic interpretation, one could summarize the problem by saying that with weak selection, the model has no predictive power, but with higher intensities of selection, the model requires unreasonably high mutation rates in order to push the offers in the mutation-selection equilibrium up to the levels we observe in the lab. This is especially true if we replace their global, and biased, mutation process with a local, and much less biased version (Akdeniz & van Veelen, 2022).

With a cultural interpretation, the assumption is that, in choosing their strategy, individuals aim for high payoffs, and in doing so, they are more likely to imitate strategies with high payoffs than they are to imitate strategies with low payoffs. In the mutation-selection equilibrium, strategies that reject offers that are currently hardly ever made only experience a small loss in expected payoffs, and therefore they can be relatively abundant, while the mild selection pressure against them still balances against the inflow due to mutations. However, the assumption that individuals are trying to maximize their payoffs, and only fail to do so in matches that do not occur often enough to constitute enough of a selection pressure, is at odds with how good humans are at understanding incentives. In the lab, subjects are well aware that when they reject, this is bad for how much money they walk away with; it is just that they are willing to accept that in order to get even with the proposer. We will return to the issue of payoff-biased cultural transmission and strategic savvy in Section 4.5.

Spillover from evolution in prisoners' dilemmas Another option is to assume that deviations from selfishness evolved for behaviour in other games, like the prisoners' dilemma, and that we bring those preferences along when we play the ultimatum game. This implies that our behaviour is maladaptive, and that games like the ultimatum game were not relevant enough in our evolutionary history to tailor our behaviour to. This possibility would be consistent with the approach by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) in economics, where it should be noted that neither of these original papers claim evolutionary explanations, rather they simply aim at finding a model that is consistent with play across different games.

The deviations from simple selfishness that Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) describe, and that work for the ultimatum game, go by the name *inequity aversion*. This is a willingness to give up payoff to benefit the other, when the other has less than you (advantageous inequity aversion), combined with a willingness to give up payoff to hurt the other, when you have less than the other (disadvantageous inequity aversion). Rejections in the ultimatum game can then be explained by responders having sufficiently strong disadvantageous inequity aversion. Because this has become more or less the standard in economics, we elaborate a little more on this in Box 2, where we also show how this can be represented in pictures.

There are two problems with this approach. The first is that what evolves in models with population structure, or kin selection models, is not inequity aversion. What evolves in such models is altruism for positive relatedness (Hamilton, 1964a;b), or perhaps spite for negative relatedness (Hamilton, 1970). What does not evolve is altruism when ahead, and spite when behind, directed towards one and the same person with whom relatedness is just one number. We make this point a bit more formally in Box 3, but the short version is that if the prediction of a model comes in the form of Hamilton's rule (van Veelen et al., 2017a), then how much of their own fitness individuals are willing to give up for how much fitness for the other should not depend on whether the individual making this decision is ahead or behind (van Veelen, 2006). Because the explanation of the behaviour in the ultimatum game depends mainly on the disadvantageous part of the inequity aversion leading responders to reject unequal offers, this could perhaps be salvaged by assuming that people are across the board spiteful. This, however, is at odds with behaviour in other games, including the trust game, as we discuss below, and also with behaviour in situations where they can simply trade money for themselves for money for others (Andreoni and Miller, 2002).

The second problem with this approach is that it assumes that how we evaluate trade-offs between our own fitness and the fitness of the other, is fixed, and therefore independent of the strategic details of the game and independent of the behaviour of the proposer. That is how the model in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) is set up, and this is also how it should be, if these preferences have evolved for games like the prisoners' dilemma, and we just carry them over to games like the ultimatum game. The assumption of fixed preferences is not consistent, however, with the way in which behaviour in the ultimatum game compares to the behaviour displayed in some famous altered versions of it. Fow example, when the proposal is generated by a computer, responders do not reject quite as much as they do when the proposal is generated by the person they are playing

with (Blount, 1995). Also, when an unequal split is proposed, but the only other option was for the proposer to propose an even more unequal split, the rejection rate is lower than when an unequal split is proposed but the proposer also had the option to offer the equal split (Falk et al., 2003). Both differences should not be there if rejections are being driven by proposals falling short of a fixed threshold for acceptance, generated by a fixed level of disadvantageous inequity aversion. It is also worth noting that if responder behaviour has evolved with the purpose of influencing the behaviour of the proposer, as our commitment-based explanation suggests, then rejections *should* be contingent on how much room to maneuver the proposer has. Another finding that speaks against an explanation based on inequity averse preferences, is that, in cases where the responder can only reject to receive her own share of the proposal, and rejection therefore *increases* inequity, some responders reject nevertheless (Yamagishi et al., 2009).

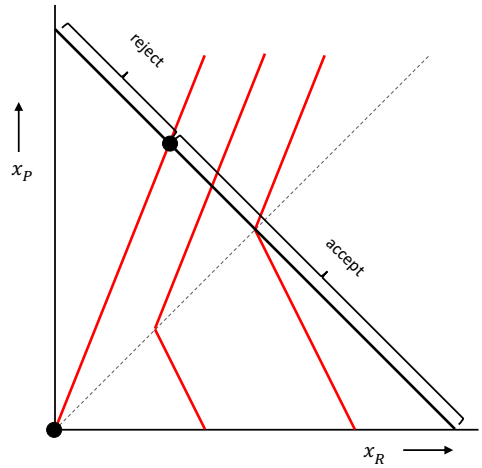
Just to be prevent misconceptions, we do not deny that there are (many) people that have a preference for equal outcomes over unequal ones; see, again, Andreoni and Miller (2002). All we claim here is that the notion of inequity aversion has to be stretched a bit too much in order to match the empirical evidence for the ultimatum game.

Box 2: Fehr-Schmidt inequity averse preferences and the ultimatum game. If x_P is the amount of money for the proposer, and x_R is the amount of money for the responder, then a responder who has Fehr-Schmidt inequity averse preferences attaches utilities to combinations of x_R and x_P as follows:

$$u\left(x_R, x_P\right) = \begin{cases} x_R - \alpha\left(x_P - x_R\right), & \text{if } x_P \geq x_R \\ x_R - \beta\left(x_R - x_P\right), & \text{if } x_R \geq x_P \end{cases}$$

The higher the utility, the more this responder likes the combination of x_R and x_P. The distaste for disadvantageous inequity is measured by α, which, if the proposer has more, is multiplied by how much more the proposer has. The dislike of advantageous inequity is measured by β, which, in cases where the responder has more, is multiplied by how much more the responder has. These preferences can be represented by *indifference curves*, which are contour lines, connecting points with equally high utility. In the figure below, with the amount of money for the responder on the horizontal axis, and the amount of money for the proposer on the vertical axis, and where we chose $\alpha = \frac{2}{3}$ and $\beta = \frac{1}{3}$, those are the red kinked lines. The responder is indifferent between combinations of money amounts (x_R, x_P) on

one and the same indifference curve, and likes combinations more to the right better than combinations more to the left.



In the ultimatum game, the proposer can propose combinations anywhere on the black 45 degree line, where the money amounts add up to a fixed sum. The responder then chooses between that proposal and $(0, 0)$, which is the origin in this picture. When choosing between accepting and rejecting, a responder with these inequity averse preferences would reject a range of very unequal proposals, and accept all other proposals. A proposer that also has Fehr-Schmidt inequity averse preferences would maximize his or her utility by choosing the point where the responder barely accepts (barely prefers the proposal over both getting 0), unless the proposer has a $\beta > \frac{1}{2}$. If she does – which means that she is very averse to inequity when ahead – she would propose an equal split.

Box 3: Hamilton's rule does not suggest inequity aversion. If the prediction of a model can be summarized by Hamilton's rule, then cooperation, or altruism, will evolve if

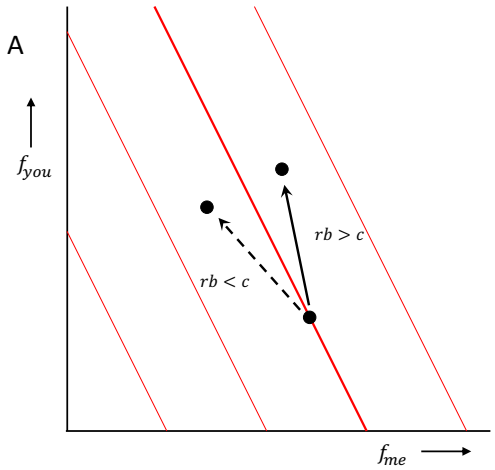
$$rb > c,$$

where r is the relatedness between donor and recipient, or between the two players of the prisoners' dilemma, b is the benefit to the recipient, or the other player, and

c is the cost to the donor, or the one player. This can be interpreted as a rule that, for a given behaviour, with given costs and benefits, predicts whether or not that behaviour will be selected. We can however also assume that we face a variety of opportunities to help, or a variety of prisoners' dilemmas, with a range of b's and c's. If we do, then we can also think of this as a prediction that separates those we will choose to cooperate in, from those in which we will not (see panel A, with $r = \frac{1}{2}$, and van Veelen, 2006). That implies that our preferences would have a uniform level of altruism, that is independent of whether one is ahead or behind

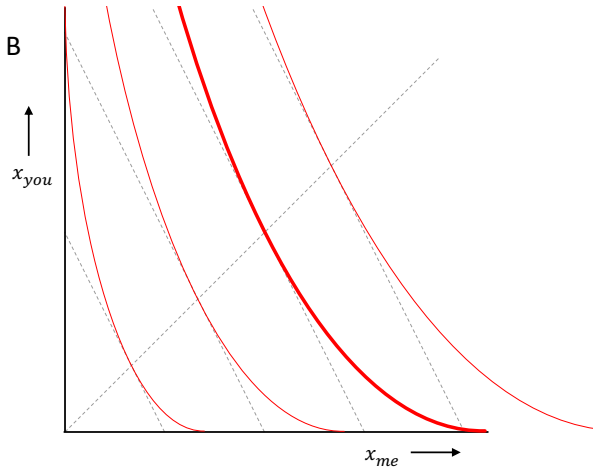
$$u_{me}\left(f_{me}, f_{you}\right) = f_{me} + \alpha f_{you},$$

where $\alpha = r$, and where f_{me} and f_{you} are the fitness of the donor, or the one player, and the fitness of the recipient, or the other player, respectively. Indifference curves therefore should be tilted straight lines, and the higher relatedness is, the more tilted they should be. //



Here, the variables are fitnesses, and the b and c therefore are both expressed in fitness terms. Many decisions we take, however, (including decisions in the lab) are in terms of money, food, or other resources. If additional amounts of those contribute more to fitness when individuals have little of them, and less when they already have a lot, then the straight lines in fitness terms turn into curved lines in money terms (panel B). One could call those preferences inequity averse in money terms, because how much resource they are willing to give up in order to give the other a fixed benefit, depends on how equal or unequal the status quo is. However,

this still does not lead to the disadvantageous inequity aversion in Fehr and Schmidt (1999), where individuals are willing to give up resources of their own to *reduce* the amount that the other has, if the other has more.



Group-beneficial norms Cultural group selection provides a reason why group beneficial norms can spread. When different groups have different norms, groups with norms that are more group-beneficial outcompete groups with less group-beneficial ones almost by definition. To account for why upholding a group beneficial norm beats not upholding any norm, additional assumptions need to be made about the individual costs of maintaining the norm, the group benefits, and the details of the cultural group structure.

For the ultimatum game, one could assume that responders who reject are upholding a norm of equality. This is not group-beneficial in money terms; instead, all that the norm does, in the standard version of the ultimatum game, is change how a fixed amount of money is distributed. It can however be group-beneficial in fitness terms, because receiving additional money, calories, or whatever it is that helps survive and reproduce, typically contributes more to fitness when you only have little of it than when you already have a lot. Reducing inequality, and shifting resources from the rich to the poor, can therefore increase efficiency in fitness terms.

One problem with this approach is that the efficiency of the norms that are enforced by rejecting unequal proposals is a possibility, but not a given. In Kagel et al. (1996), the proposals are in terms of chips, and these chips are either worth 3 to the responder and 1 to the proposer, or vice versa. If norms are meant to increase efficiency, then they should

make people transfer more (or everything) if chips are worth more to the responder, and less (or nothing) if chips are worth more to the proposer. In the experiment, the opposite happens (see also Schmitt (2004) for more self-serving aspects of fairness norms in ultimatum games).

Also, if we think of real-life examples, there is a spectrum of settings in which people "reject proposals" they deem inappropriate. On one end of the spectrum, there may be sharing norms that increase joint fitness by redistributing assets. On the other end of the spectrum, however, there are mafia bosses, who reject proposals by killing earners that bring envelopes that are too light, or by destroying businesses that do not cough up enough protection money. Criminal activities typically decrease the size of the pie (burglars benefits less from stolen goods than the damage they inflict on those that they steal from) and extortion can easily make money flow towards criminals that are much richer than their victims. The norm that they enforce therefore shrinks the size of the pie in monetary terms, and, on top of that, makes its division more unequal. Here it is worth noticing that the one thing that is consistent across the spectrum, is that being committed to rejection increases how much proposers are willing to fork over to responders.

Another thing to keep in mind is that the core difference between this explanation and our commitment-based explanation is where the benefits accrue. In both explanations, rejections are bad for fitness, but in our explanation, being committed to rejection is actually good for the fitness of that same individual, whereas with group-beneficial norms, the benefits of upholding the norm accrue to future responders within the same group. We will return to this issue when we discuss games with punishment.

Again, we are not saying that there is no role for cultural group selection, or for the evolution of norms, it is just that, all by itself, it is an uneasy fit for rejecting, or engaging in destructive behaviour if you do not get your "fair" share, across the spectrum of social settings where such behaviours occur.

Repeated interactions Yet another possibility is to assume that there is no such thing as a one-shot ultimatum game, and what we see people do in one-shot games is an extrapolation of behaviour that has evolved for repeated versions, where players take turns in being a responder and a proposer (see papers in the review by Debove et al., 2015). This is discussed in Section 4.2.3 for the prisoners' dilemma. For the ultimatum game, there is an additional consideration, which is that, when the roles alternate, equilibria in which the proposer gets the whole pie every other day are almost as good as equilibria in which both get half the pie every day. Unlike repeated prisoners dilemmas, the behaviour that

is enforced here is only marginally more efficient.

Selection with commitment In an overly simplified model, one can assume that responders can commit to rejections, and proposers can tell the difference between committed and uncommitted responders. If we further assume that proposers simply wish to maximize their payoffs, this would turn the tables between proposers and responders. Proposers now will always want to match the minimal acceptable offer of the responder, and responders with ever higher demands will be selected (see the appendix, and Güth and Yaari, 1992).

The assumption that proposers can detect commitment is, of course, crucial. If committed responders do not get better proposals than uncommitted responders, then the only difference is that they sometimes leave money on the table, and that sets in motion the cascade of ever lower thresholds and ever lower proposals that we started Section 4.4.1 with. This could be countered if committed responders sometimes get better proposals. The importance of proposers knowing who is and who is not committed, led Nowak et al. (2000) to describe the evolution of higher thresholds and higher offers in their model as the result of reputation. This is also how Debove et al. (2016) classify the mechanism. While this is a defensible choice, an equally reasonable alternative, and the one that we suggest, is that reputation simply facilitates the flow of information that is required for commitment to work.

The assumption that proposers can detect commitment, and that responders can commit, are also related. Given a choice between being committed and not being committed to rejecting unfair proposals, the first will obviously be better for responders, provided that proposer can detect committed players. Of course, it would be even better for a responder if proposers *think* she has a high threshold for accepting the proposal, when she does not in reality. A mutant that does everything to suggest that she is committed, but is not, undermines the credibility of the signal when it increases in frequency. One should bear in mind, though, that if we allow for pretenders, then a population of committed rejecters and matching proposers is not an equilibrium anymore (because of the mutants that fake their commitment), but neither is a population where there is no commitment at all. One way to summarize the direction of selection, therefore, is that there will be a never-ending tug of war between proposers, the truly committed, and those who are faking it.

In terms of preferences, a crucial difference between an explanation with commitment and an explanation where preferences are shaped by evolution in prisoners' dilemmas, is that here, preferences depend on what the first mover does. This possibility was previously

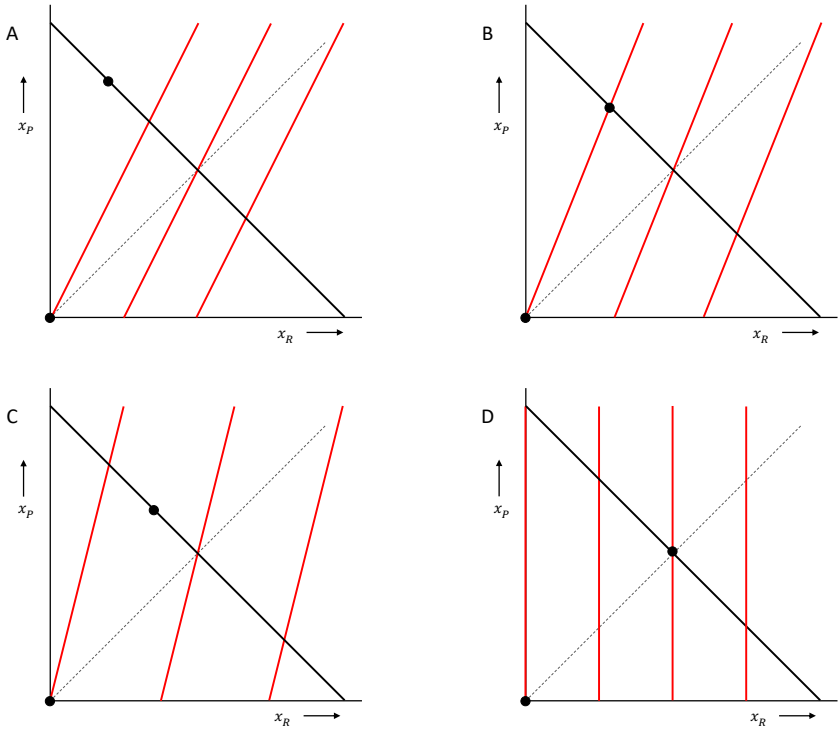


Figure 4.4.1: Preferences that depend on what the other did. Following Hirshleifer (1987; 2001) and Cox et al. (2008), we can let the preferences of the responder depend on the options that the proposer made her choose between (where the proposer's "menu of menus" also matters). The menu in panel A is less generous than the menu in panel B, which in turn is less generous than the menus in panel C and D. This would make the responder sufficiently angry to reject the proposal in panel A, barely accept it in panel B, accept it in panel C, and happily accept it in panel D; see also van Leeuwen et al. (2018).

suggested by Hirshleifer (1987; 2001), who applied it to a sequential version of the prisoners' dilemma or Hawk Dove game. Cox et al. (2008) formulate a beautifully general approach to how preferences can change as a result of the menu of options that an earlier mover chooses to give to a later mover. Figure 4.4.1 illustrates this for the ultimatum game.

Observing the benefits in experiments One of the questions that could be addressed with experiments, is whether there is an individual advantage to being committed to rejection. Many lab experiments, however, do not allow for subjects to learn about each other, for instance by observing past behaviour. In the absence of a channel for proposers

to find out who is and is not committed, only the costs of being committed will show in such experiments. One exception is a study by Fehr and Fischbacher (2003), which includes an ultimatum game in which proposers are able to see what the responder they are matched with accepted or rejected in past interactions with others. This comes with a complication, because not only does this allow proposers to find out who is and who is not committed to rejection, but it also opens the door for responders to strategically inflate their reputation for being a tough responder. This is precisely what happened: in the treatment with reputation, acceptance thresholds were higher. In the treatment without reputation, however, the acceptance threshold was not 0 (as we also know from other experiments with ultimatum games). This is consistent with some subjects being truly committed, and one could even say that trying to inflate your perceived level of commitment is only worth it if there is also real commitment around. Also, it has been shown that people do better than chance when trying to guess who did and who did not reject an unfair offer in the mini-ultimatum game, when all they can go on is pre-experiment pictures of the subjects (van Leeuwen et al., 2018). This suggests that nature has found a way for us to spot commitment to some degree. Here, it is important to know that it is not necessary to always and unfailingly detect the truly committed; it is enough if being (more) committed sometimes results in a better proposal.

External validity How we can explain the evolution of behaviour we observe in the lab is only a good question if the behaviour in the lab is representative of behaviour outside the lab, and if the people displaying it in the lab are representative of people in general. For both of these steps, one can have reservations. Levitt and List (2007) argue that the setting of a lab exaggerates all behaviours that can be described as a norm – including behaviour in the ultimatum game. Also Gurven and Winking (2008) and Winking and Mizer (2013) suggest that results from the lab are optimistic about pro-social behaviour outside the lab. As for the second step, Henrich et al. (2010) show that western, educated, industrialized, rich, and democratic (WEIRD) subjects are at an extreme end of the spectrum in many domains. One of the examples, based on Henrich et al. (2001; 2005; 2006), is behaviour in the ultimatum game, where WEIRD subjects have higher average thresholds for accepting, and make on average higher offers than almost any of 15 small-scale societies that were investigated. Because growing up in WEIRD societies is evolutionarily new, this most likely makes the typical lab results not representative. It is, however, important to note that these are mostly differences in degree, and that they do not suggest the total absence of the idea of an unfair offer in non-WEIRD populations.

4.4.2 The trust game

Some of the reasons why model predictions and empirical evidence do not match perfectly for the ultimatum game also apply to the trust game. If we assume that the behaviour in the trust game evolved for the trust game, but without assuming that trustees can commit, then trustees should send back nothing. This is not what trustees do (Alós-Ferrer and Farolfi, 2019; Berg et al., 1995; Johnson and Mislin, 2011). If we assume that inequity averse (or maybe altruistic) behaviour evolved for other games, and that we carry those preferences over to the trust game, then there are, again, two complications. In Cox (2004) there are three versions of the trust game, two of which we will focus on here: the standard trust game, which differs from our simplified version, in that the trustor can send any amount between 0 and 10, which then gets tripled, and the trustee can send back any share of the tripled amount; and a version in which trustees face the same decision, but the trustor is made inactive, and the budget that the trustee decides over is generated by taking observations from the first treatment. In this second treatment "trustees" do send money "back" (in quotation marks, because the money they have was not really sent to them by anyone), which suggests that they do have preferences over how the money is divided that are not simply selfish. However, they behave significantly differently between treatments, and send back more in the first treatment, when their trustor is the one responsible for the budget they can divide. This difference should not be there if this behaviour evolved, for instance, through population structure in games like the prisoners' dilemma. Also, as noted before, such models generate altruism, or spite, but not inequity aversion. Here, that could be mended by doing away with the disadvantageous inequity aversion, but it is obviously not possible to assume people are across the board spiteful when interpreting their behaviour in the ultimatum game, and across the board altruistic, when interpreting their behaviour in the trust game (see also Figure 4.4.1).

In the trust game, sending back money can be seen as a reward for behaviour that increases joint fitness; the more the trustor sends, the larger the pie. The individual that receives the benefit, however, is the trustee herself, so there is no need to invoke group selection for efficient norms. If we assume that the reason why trustees send back money is that being committed to doing that makes trustors send over more, then that does facilitate mutually beneficial cooperation, but the reason it evolves is that it is beneficial for the trustee.

Of course, as before, being committed to sending back money has to be observable to some degree in order to evolve.

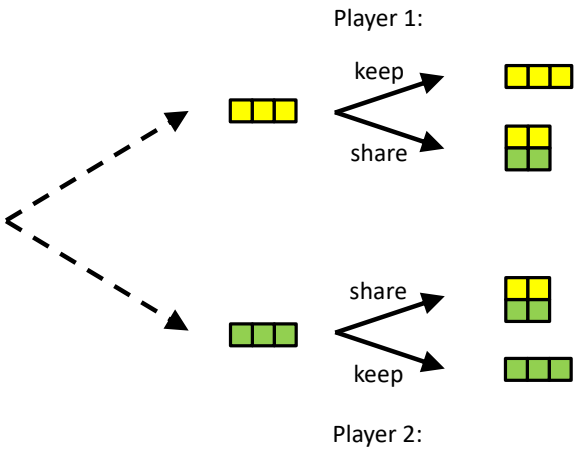


Figure 4.4.2: A simple version of the insurance game. Both players can be lucky or unlucky and the probabilities with which that happens are the same for both. If you are lucky, you have three, if you are unlucky you have zero. If both are lucky, or both are unlucky (not depicted here), there is no use for helping. If one is lucky, and the other is not, then helping will typically cost the lucky one less than it benefits the unlucky one. Ex post, after the dice are cast, it is better not to help, but if both would be able to commit to helping when the situation is uneven, this would, ex ante, be better for both.

In the lab, the trust game is usually played without communication. Situations in real life with a similar structure, however, often involve some communication, which allows trustees to make promises. As suggested by Frank (1987; 1988), a promise can work as an on-switch for commitment. Ellingsen and Johannesson (2004) studied a social dilemma called the "*hold up problem*", which is a combination of the trust game and the ultimatum game. Player 1 can invest 60 kronor, or keep it. If invested, the 60 kronor turn into 100 kronor. Player 2 then proposes a split, which Player 1 can accept or reject. Ellingsen and Johannesson (2004) found that threatening to reject low offers works to get higher offers, and also that the possibility to make a threat increases the share of Player 1's that invest. However, allowing Players 2 to make promises works even better; they keep their promises, and even more Player 1's invest. Observations in experiments without communication can be viewed, therefore, as a lower bound on the capacity to commit.

4.4.3 The insurance game

We would like to illustrate that commitment can also explain behaviour or phenomena that are less well-researched, such as our preoccupation with sincerity, and why we value genuine caring more than opportunistic helping. To do so we introduce another game, which one could call the "insurance game", or the "friendship game". In this game, there are two players that are either lucky or unlucky. In this simple version, lucky means you get three, unlucky means you get zero. If one is lucky, and the other is not, then the lucky one can help the unlucky one, in which case both will end up with two. The idea behind this is that sharing is more beneficial for the unlucky player than it is costly for the lucky one (see Figure 4.4.2).

In this game, it is always better not to share when you happen to be lucky, and the other one is not. However, if both players can commit to sharing, they would both be better off on average. If players that can commit are able to recognize each other, or even better, single each other out, and play this game amongst themselves, they would do better than those that would never share and always keep what they have.

In a population playing such a game, there would therefore be two related selection pressures. The first is a selection pressure to commit to sharing by genuinely caring for the other, which helps being chosen as a partner or friend. The second is a selection pressure to recognize genuine altruism, and distinguish it from fake displays of affection. Of course there is a tension that remains, as the best option would be to be chosen as a partner or friend, be on the receiving end of sharing if you are unlucky yourself, and the other is not, but refuse to share when the tables are turned. However, this tension is the whole reason why commitment would be needed in the first place, and it seems that the existence of sincere altruism and true love, as well as our preoccupation with distinguishing genuine care from opportunistic behaviour, indicates that evolution might have found a way to help us commit at least to a certain degree. It also makes sense that friendship and love typically converge to being symmetric partnerships, in the sense that people tend to end up being each others' friends, and if people stop liking us, we tend towards liking them less too.

Again, one could think of this as an extrapolation of reciprocity, which evolved in the context of repeated interactions, and there is of course no doubt that reciprocity has evolved in humans. However, it is important to realize that not only do we pay people back, and say "you did the same for me", but we also engage in hypothetical reciprocity, and say "you *would have done* the same for me" in such cases where we help a friend who

has not had the opportunity to help us, and probably never will. The latter would be consistent with the idea of evolved commitment in the insurance game, and that might be a better explanation than the idea of a maladaptive spillover from the repeated prisoners' dilemma. There are also instances like the Maasai concept of *osotua*, which serves to tie people together, and involves giving each other gifts only when in need, even if this turns out to make the gift-giving structurally asymmetric (Cronk, 2007).

If the insurance game is played repeatedly, and if helping a friend who is dealt a bad hand today increases her capacity for helping you in the future, then being committed to helping can also be in one's own self interest in a more direct way (Eshel and Shaked, 2001). Provided that both parties are already committed to helping each other, then that help can be a great investment in receiving help in the future, not because you are investing in the other's *willingness* to help (as in standard models of reciprocity in repeated games), but because you are investing in the other's *ability* to help, assuming the other's commitment is already there. A friend who you know would save your life, for instance, would not be around anymore to do that if you did not save hers, and hence it might be worthwhile taking a risk to do just that.

4.4.4 Prisoners' dilemmas and public goods games

We have looked at reasons why predictions from models with prisoners' dilemmas (without commitment) do not match deviations from simple selfishness in games like the ultimatum game or the trust game. However, even if we look at how humans actually play one-shot prisoners' dilemmas and public goods games, there are some peculiarities that are at odds with the standard explanations without commitment. Although some people are selfish and opportunistic, the majority are conditional cooperators in public goods games (Fischbacher et al., 2001) or prisoners' dilemmas (Charness et al., 2016). Many are happy to cooperate if the other one cooperates too, but if the other one defects, most people prefer to defect as well. It seems therefore that evolution did not just make us indiscriminate cooperators or indiscriminate defectors – which is the menu of phenotypes in many models of evolution in the literature. Instead, evolution seems to have given a decent share of us the ability to commit to not defecting, as long as we are sufficiently sure that the other will not defect either.

Conditional cooperation can, again, be interpreted as a spillover from repeated games, where reciprocal strategies can evolve, that stop cooperating if the other does not also cooperate (Delton et al., 2011). It is important to realize, however, that cooperation in prisoners' dilemmas can also evolve without repetition, or population structure. What is

needed in this scenario with commitment, is the ability to tell who is (also) committed to cooperation, provided that the other one cooperates too, or, in public goods games, provided sufficiently many others cooperate too. For cooperation to actually happen, knowing that the other will cooperate as well is also needed, because between two conditional cooperators, this becomes a coordination game with two equilibria; one where both play C; and one where both play D.

If conditional cooperators can seek each other out for cooperation, then the mechanism at work would be partner choice, which would result in endogenous population structure. This mechanism does not require cooperation to be conditional, it just needs cooperators to prefer to be matched with other cooperators, and to know how to spot them (Frank, 1988; 1994; Frank et al., 1993). However, also without partner choice, but with the ability to tell if others are also conditional cooperators, conditional cooperation can evolve. In this case, conditional cooperators will cooperate if they happen to be matched with each other, but defect if they meet defectors. Provided that conditionally cooperative players can tell sufficiently often whether they are playing with another conditional cooperator, that would give them a selective advantage (Akdeniz, Graser & van Veelen, *in preparation*).

There are two more ways in which cooperation can evolve in prisoners' dilemmas through commitment. The first is that in a sequential version of the prisoners' dilemma, a commitment to rewarding cooperation with cooperation can evolve in the same way it can in the ultimatum or trust game; the second mover would commit to rewarding cooperation with cooperation, and that would make it in the interest of the first mover to cooperate rather than defect (Hirshleifer, 1987). The second is that also in simultaneous move, but non-linear continuous versions of the prisoners' dilemma, commitment can induce the other player to contribute more (see examples in the appendix, based on Alger and Weibull, 2012).

4.4.5 Games with punishment

It has been widely recognized that punishment can sustain cooperation (Fehr and Gächter, 2002). This observation is regularly followed by the realization that this is an incomplete explanation. While punishment may explain why there is cooperation, we would still need a reason why there is punishment, especially if punishment is costly (Brandt et al., 2006; Fehr and Gächter, 2002; Fowler, 2005; Hauert et al., 2007; Mathew and Boyd, 2009). One explanation for the existence of costly punishment is group selection. This is also a candidate to explain cooperation without the option to punish, but here it can be combined with the idea that, when established, punishment might be cheaper than the cooperation

it enforces (Boyd et al., 2003). Higher order punishment might be even cheaper (Fehr and Fischbacher, 2003; Henrich and Boyd, 2001), but people do not really seem to use it (Kiyonari and Barclay, 2008). Another explanation is the existence of the possibility to opt out of the public goods game, at a payoff that is higher than the payoff one gets if everyone defects. Models with this option predict cycles, and populations can spend sizable shares of their time in states where everyone cooperates and everyone punishes defectors (Brandt et al., 2006; Garcia and Traulsen, 2012; Hauert et al., 2007; Mathew and Boyd, 2009).

The premise of punishment as an incomplete explanation of cooperation, however, overlooks the possibility that, even if punishment is costly, being committed to punishing may already be beneficial for the individual (dos Santos et al., 2011; 2013; dos Santos and Wedekind, 2015; Hilbe and Traulsen, 2012). This would imply that the possible benefits to others might not be the reason why we punish, nor do we need the game to be voluntary. To help make sure that we identify the possible advantages that commitment brings, it is perhaps helpful to realize that a prisoners' dilemma or public goods game with the option to punish really is a different game than the prisoners' dilemma or the public goods game without punishment. With the option to punish, being committed to punishment might change the course of backward induction, and make it in the other players' best interest to cooperate (Hauert et al., 2004; Sigmund et al., 2001). If the commitment to punish makes others cooperate often enough, then this can outweigh the costs of punishment when others defect, or the remaining deficit between individual costs and individual benefits may be so small that it only takes a little bit of population structure to make the benefits to others outweigh the deficit (Brandt et al., 2003). Of course, as always, this requires that commitment, in this case to punishment, can be recognized.

Terminology Unfortunately, not all terminology in this area of research is neutral. Both 2nd and 3rd party punishment in non-repeated interactions are sometimes referred to as *altruistic*. The idea behind this label is that punishing a defector after she has defected on me might induce her to cooperate in later interactions with other individuals (Fehr and Fischbacher, 2003). This makes the punishment beneficial to the next person she interacts with, but not to me, and hence it is called altruistic. Also in 3rd party interactions, the idea is that those that benefit from the punishment are those that the wrongdoer will interact with in the future. When the mechanism behind the evolution of punishment is that commitment changes other people's behaviour, 2nd order punishment, however, does not have to be altruistic, because the real reason why one would be committed to punish defections could also be to avoid being defected on oneself. In experiments where

participants have no way of learning whether someone is committed to punishment, this might fail to work, and only the collateral benefits to future interactants might show. In such cases, the design of the experiment therefore eliminates the benefits to oneself of being committed to punishment. Similarly, with respect to 3rd party punishment, the commitment might not exist to benefit the *next* person that the wrongdoer meets, but to protect the *current* person she interacts with. This perspective is also more in line with the way in which Bernhard et al. (2006) find 3rd party punishment to be parochial. If the purpose of 3rd party punishment is to better the behaviour of 1st parties in future in-group interactions, then 3rd parties should punish when all three belong to the same group, or maybe when the 3rd party and the 1st party belong to the same group. Instead, they find that the chances that an unfair choice by a 1st party is punished are determined by whether or not the 3rd party and the 2nd party belong to the same group, which suggests a commitment to stand up for fellow group members.

Heterogeneity In the prisoners' dilemma or the public goods game with punishment, the ability to commit can only make a difference if there are opportunistic others around, who will cooperate when they think they are matched with a committed punisher, or with too many committed punishers. Opportunism on the other hand only pays if not everyone is (equally) committed to punishment, and there is something to be opportunistic about. The presence of these types therefore only makes sense if they coexist.

Extrapolation A recurrent explanation for behaviour in one-shot games is that it is an extrapolation of behaviour that evolved for repeated games. One of the core points of this paper is that deviations from simple selfishness in one-shot games may, in fact, have evolved for one-shot games. There might even be some extrapolation going on in the other direction. In Dreber et al. (2008), subjects played a repeated game, in which the options were not only to cooperate or to defect, but there was also an additional punishment option. In equilibria of the standard repeated prisoners' dilemma where both players cooperate (for instance when both play Tit-for-Tat) defection is already used as a form of punishment. The extra punishment option here is one in which the player that uses it pays a cost (which makes is more expensive than defection), and for that extra buck, you get that the other player is hurt more. The fact that some subjects go for this punishment option, to their own detriment, and in spite of the fact that defection already is a bad enough deterrent, suggests that they may bring some revengeful sentiments to these repeated games that originally evolved for one shot games, so that players end up punishing harder than they need to, and more than is good for them.

4.5 Other species

If we consider evolutionary explanations for human morality, or deviations from selfishness, then it is not only important that they give reasons for why humans evolved to be moral, or pro-social, but also why other species did not (Mathew et al., 2013), or at least not to the same extent. Some authors argue that the more closely related primates have a proto-morality (Brosnan and De Waal, 2003; Brosnan et al., 2005; Burkart et al., 2007), others put more emphasis on the discontinuity between human and nonhuman minds (Penn et al., 2008), including their pro-social behaviour (Silk, 2009), but even with a margin of error around where other primates stand, there is no doubt that humans are unique in the extent and complexity of their morality (Call and Tomasello, 2008; Tomasello et al., 2003). This implies that it would be interesting to determine the selection pressure(s) on humans that made them different (Melis and Semmann, 2010; Silk and House, 2011).

4.5.1 Population structure

The classical ingredients in explanations for the evolution of cooperation are population structure and repetition, and these two ingredients are indeed present in the human ecology. Humans, however, are not unique in living in (group) structured populations, nor are we special in interacting repeatedly. Many species live in groups, including other primates; see for instance Wilson and Wrangham (2003) for group structures in chimpanzees. Langergraber et al. (2011) moreover show that the level of genetic differentiation in nonhuman primate populations comes close to those observed in human groups, and also other studies report levels of genetic differentiation that are similar between humans and gorillas (Scally et al., 2013) and between humans and a variety of great apes (Fischer et al., 2006).

As discussed in Section 4.2.3, cultural inheritance can make groups more homogeneous behaviourally than they would otherwise be, and more than they are genetically (Bell et al., 2009; Handley and Mathew, 2020). This creates a population structure that is unique to humans. In Section 4.2.3 we mentioned one caveat – the cancellation effect at the group level, which applies to group selection models in general. In Section 4.5.3 we will mention another, which applies to all models with payoff-biased cultural transmission.

4.5.2 Repetition

Repeated interactions with the same partner also occur in many animal species, especially those characterized by group living. Clutton-Brock (2009) indicates that, despite this,

there is not all that much behaviour outside humans that qualifies as genuinely reciprocal, with individuals that pay costs now, and that expect to receive benefits in the future, especially when the future is not immediate. His explanation for the absence of reciprocity in other species is that reciprocity requires that the parties involved are able to make detailed arrangements for exchanges in the future, and that this requires, amongst other things, language. Stevens and Hauser (2004) also argue that cognitive constraints are the likely reason for why we do not see much reciprocity in non-humans animals compared to humans. This is definitely something that we agree with, and we actually think that our capacity to work out cooperative arrangements that require time to mature, and "*establish the intentions and expectations of the parties involved regarding the nature and timing of exchanges*", as Clutton-Brock (2009) puts it, is a key piece of information on what makes humans different. Language, theory of mind, and morality are three things worth investing in, if you want thrive in the human niche. The absence of a human-like talent for language and theory of mind in other animals therefore is not so much an exogenous constraint, as their presence in our species is an indication of what we specialize in.

4.5.3 Our niche

One way in which humans are special is the way in which we make a living – and the incidence of commitment problems that this generates. That is not to say that there are no commitment problems elsewhere in nature, for which evolution may or may not have found solutions too, but it is not controversial to say that our niche involves acquiring food in ways that require more complex cooperation, and more planning ahead than other species. Our technologically more elaborate, more information intensive, and collaborative way to make a living opens doors for opportunistic behaviour that remain closed in other species. If our morality is shaped to solve problems that do not exist in other species, or at least not to the same extent, then this also explains why we would be unique in our morality.

Language and planning ahead together The way we make a living comes with a few faculties that stand out (Tomasello, 2009). Humans are technological. There is evidence of some tool making in other animals, but it is nowhere near human levels (Seed and Byrne, 2010; see also Shumaker et al., 2011, for an extensive review of animal tool use). Humans also plan ahead, and we can delay gratification. Many of our collective efforts also require detailed coordination and planning ahead together. Language allows us to do this, and it is not strange to assume that this is one of the reasons why we talk (besides other reasons for why we have the rich language that we have; see for instance Miller, 2000).

Language facilitates planning ahead together, and such plans can create commitments problems that can be solved by deviations from simple selfishness. The role of language in morality, however, does not stop there. Language also allows us to make promises, which we have already seen can activate commitment in the hold up game (Ellingsen and Johannesson, 2004), but it can do so more generally (Vanberg, 2008). Also when people agree on a way to divide the different parts of a job, they all commit to doing their part, which becomes their responsibility, even if they do not solemnly swear, but just say OK. Not doing something that was your responsibility will subsequently be frowned upon much more than not doing the same thing when it was not your responsibility.

Some collective efforts, moreover, may have parts of the job that will not be observed by everyone. This creates what economists call asymmetric information; some parties are better informed than others. With language, person A can tell person B what she saw person C do, but even with that possibility, information asymmetry may persist, especially if no one saw what person C did. The better informed party then can choose between lying or telling the truth. While telling the truth can be disadvantageous, depending on what the truth is, being committed to telling the truth can be advantageous. Lying aversion, or honesty, therefore can also be a solution to a commitment problem (Heintz et al., 2016, Akdeniz, Jagau, Shalvi & van Veelen, *in preparation*).

Theory of mind and backward induction Besides language and planning ahead, humans are also exceptionally good at theory of mind, which means that we attribute desires and beliefs to others that may differ from our own. Being able to put yourself in someone else's shoes, and understanding the strategic consequences of different behaviours, also seem to be prerequisites for the type of cooperation that humans engage in. Much of the evolutionary game theory concerning the evolution of cooperation is, however, neutral (at best) on whether individuals understand the game they are playing, and on attributing goals, beliefs and intentions to others. As mentioned in Section 4.2.3, many models with population structure allow for an interpretation with either genetic or cultural transmission (Allen et al., 2017; Lieberman et al., 2005; Ohtsuki et al., 2006; Santos and Pacheco, 2005; Santos et al., 2008; Taylor et al., 2007a). In the latter case, individuals typically update their behaviour based on the payoffs that others get. Assuming that individuals resort to copying successful others suggests a limited understanding of the game. If they would understand the game, they would base their decisions on comparisons between what their payoffs are if they do A, and what their payoff are if they do B (given what they expect the other players to do). Copying successful others is something that you would only do if you do not understand the game, and the best you can do is to generally

assume that those that get high payoffs must be doing something right. In fact, not really understanding the game is actually a prerequisite for cooperation to evolve in this case. If individuals would understand the game, and make decisions, based on counterfactuals (i.e., on comparisons between their payoffs and what their payoff would have been, had they behaved differently), they would never cooperate in a prisoners' dilemma – unless there is another mechanism at work that makes them deviate from selfishness.

One such mechanism is classical kin selection – which for instance can make siblings help each other, fully aware of the individual costs. Another such mechanism is commitment. This mechanism actually requires theory of mind and an understanding of the game being played. If proposers in the ultimatum game cannot put themselves in the shoes of their responders, it would be futile for responders to try to change the course of backward induction by developing an angry button (van Leeuwen et al., 2018). If trustors cannot read their trustee, then there is no amount of nice or dependable that will ever generate trust. Theory of mind, therefore, is a prerequisite for the suggested solutions to commitment problems, while it stands in the way of explanations based on payoff-biased imitation.

4.6 Conclusion

There is a number of deviations from simple selfishness in humans that do not make sense, except in the light of commitment. The recurrent theme is that these deviations are bad for fitness, but being committed to them can be good. This is true for rejections in the ultimatum game, for sending back money in the trust game, for truly caring for each other in the insurance game, and for punishing defections in prisoners' dilemmas or public goods games with the option to punish. The empirical evidence does not match the explanations for human pro-sociality that are based on population structure or repetition, or, more generally, on models for the evolution of cooperation in prisoners' dilemmas. The evolution of commitment can be mutually beneficial, as it is in the trust game, the insurance game, or the prisoners' dilemma with punishment. In the ultimatum game, on the other hand, commitment to rejections is neutral with respect to the greater good, and in other instances that tend to blackmail, it can even hurt the common good. Although the idea of commitment as a mechanism for the evolution of cooperation has been around for a while (Frank, 1987; 1988; Hirshleifer, 1987; Nesse et al., 2001), it is hardly ever referred to when interpreting the empirical evidence.

Also the cross-species evidence suggests that repetition or population structure would not predict the differences between species that we see. What is different about humans is

the technological, social niche that we occupy. This goes hand in hand with us playing games that are different from the games other animals play. In the games that we play, individuals can benefit from being committed to deviations from simple selfishness. The language and theory of mind that we need for coordinating our way of making a living, is also necessary for commitment to have an effect – while theory of mind and understanding the game stand in the way of explanations with population structure in combination with payoff-biased cultural transmission. The importance of this observation can hardly be overstated.

In his book *The Righteous Mind*, Jonathan Haidt (2012) describes six moral foundations. As a way to summarize the mechanisms that he considers for their evolution, he describes humans as "90% chimp and 10% bee". The chimp part is a metaphor that represents the selfish part of human nature, while the bee part stands for those parts of human nature that seem designed to promote the functioning of the group. He thereby takes a position in the polarized debate on the levels of selection, siding with those who see a substantial role for group selection in human evolution.

While we do not want to deny the possibility that group selection has played a role in our evolution, we think it is important to recognize that the empirical evidence aligns with an explanation in which many ingredients of morality have evolved as a solution to a variety of commitment problems. A focus on the role of commitment helps organise and make sense of the rich catalogue of human morality. Within the Care/Harm dimension – perhaps the most prominent of Haidt's moral foundations – it helps understand why we care so much for sincerity, why truly caring exists, and why there is such a thing as responsibility. Thinking of honesty as a commitment to telling the truth helps understand why Honesty/Dishonesty, which was not originally included, should be a separate dimension (Graham et al., 2015; Hofmann et al., 2014; Purzycki et al., 2018). For understanding human morality, it really helps to not only think of prisoners' dilemmas or public goods, but also look at games in which the behaviour of others depends on our own willingness to walk away from bad deals, on our intent to reward trust, and on our taste for revenge. If the sincerity of our altruism, and the honesty of our heart has an effect on what other people do, then this effect on others might just be what our moral sentiments are for.

Appendix

There are two parts in the appendix. We start with the replicator dynamics for the ultimatum game, first without the possibility to commit, and then with the possibility to commit, where commitment is perfectly observable. In the second part, we illustrate how commitment can also work in one-shot simultaneous move games. These illustrations are based on Alger and Weibull (2012), and they also show that commitment can either advance the common good, or work against it.

4A.1 Replicator dynamics for the ultimatum game

4A.1.1 Without commitment

Consider an ultimatum game, where the proposer suggests a way to split n euros, and the responder accepts or rejects. In this version, proposals can only be made in whole euros, so the strategy set is not a continuum.

The proposer's choice is represented by i, which is how many euros she allocates to the responder in her proposal. That means there are $n+1$ strategies, and that the proposal would be $(n-i, i)$, for $i = 0, \ldots, n$, where the first number refers to how much the proposer would get, and the second to how much the responder would get. The frequencies with which these strategies are present in the proposer population are given by x_i, for $i = 0, \ldots, n$. Since these are frequencies, they must add up to 1; $\sum_{i=0}^{n} x_i = 1$.

For the responders, we assume that if they reject a proposal in which they get i euros, they also reject proposals in which they get less than i euros. Responders could in principle also play strategies for which this is not true, but this assumption keeps things relatively manageable, without fundamentally changing the dynamics. This implies that a strategy for the responder can be represented by j, which indicates that she accepts all proposals in which she gets at least j, for $j = 0, \ldots, n$. The frequencies with which these strategies are present in the responder population are given by y_j, for $j = 0, \ldots, n$. These are also frequencies, and must add up to 1; $\sum_{j=0}^{n} y_j = 1$.

The average payoff to proposer strategy i is how much she allocates to herself in her proposal, which is $n-i$, times the probability that the proposal is accepted. This probability is the share of responders that start accepting at i or less, making the payoff to proposer strategy i equal to $(n-i)\sum_{j=0}^{i} y_j$.

The payoff to responder strategy j is 0 if she meets a proposer who proposes i, and i is

less than her threshold j, and i if she meets a proposer who proposes i, and i is larger than or equal to her threshold j. That makes the average payoff $\sum_{i=j}^{n} i \cdot x_i$.

Lower thresholds beat higher thresholds for responders The intuition that selection always favours responders with lower thresholds follows directly from the fact that in any instance in which responders reject, they can increase their expected payoff by switching to accepting. In other words, it is never worse to accept more, $\sum_{i=j}^{n} ix_i \geq \sum_{i=k}^{n} ix_i$ if $j \leq k$; and if there are proposers that make proposals that are currently rejected, it is strictly better to accept more, $\sum_{i=j}^{n} ix_i > \sum_{i=k}^{n} ix_i$ if $j < k$ and $\sum_{i=j}^{k-1} x_i > 0$. Therefore the payoff to responders with thresholds 0 and 1 are the highest, and the payoffs to responders with threshold n are the lowest.

Proposers Which proposer strategies are doing better than average, and which are doing worse than average, depends on the state of the responder population. Between proposing i and proposing $i-1$ for the responder, the latter is better if $(n-(i-1))\sum_{j=0}^{i-1} y_j \geq (n-i)\sum_{j=0}^{i} y_j$, or, in other words, if how much you gain by allocating more to yourself on proposals that get accepted either way, or $\sum_{j=0}^{i-1} y_j$, is less than how much you lose by having proposals rejected that otherwise would be accepted, or $(n-i)y_i$.

If we start with a population where all strategies are present (so $x_i > 0$ for all $i = 0, ..., n$, and $y_j > 0$ for all $j = 0, ..., n$), then ever lower thresholds will evolve in responders, and as they do, for every $i > 1$, there will always come a point in time where proposing $i-1$ is better, because $\sum_{j=0}^{i-1} y_j$ inevitably gets large enough compared to $(n-i)y_i$.

4A.1.2 With perfectly observable commitment

Now assume, as before, that responder strategies can still be characterized by their threshold j, but, unlike before, assume that this threshold is visible to proposers. That means that proposer strategies now turn to ways to respond to what they see. We assume that if proposers match a responder threshold j, they will also match a responder threshold below j. Of course there is a richer space of possibilities for proposer strategies now, but, again, this keeps things relatively simple, without fundamentally changing the dynamics. A proposer strategy therefore is characterized by a value i, which indicates that she will match all thresholds $j \leq i$, and not match thresholds $j > i$, to which she makes proposals that will be rejected.

This turns the tables. The average payoff to responder strategy j is her threshold times the probability that a proposer will match it, which makes $j\sum_{i=j}^{n} x_i$. The payoff to proposer strategy i is 0 if she meets a responder with strategy $j > i$, and $n-j$ if she meets a

responder with strategy $j \leq i$, so the average payoff to a proposer with strategy i is $\sum_{j=0}^{i}(n-j)y_j$.

In the case without commitment, responders with lower thresholds j always got higher average payoffs. With perfectly observable commitment, on the other hand, proposers with a higher i always get higher average payoffs, because in any case in which they do not match the responder's threshold, they can increase their payoffs by switching to matching it.

For responders, switching from a threshold j to a threshold $j+1$ is better if how much they gain on interactions in which their threshold would be matched either way, $\sum_{i=j+1}^{n} x_i$, is larger than how much they loose on interactions in which the proposer will stop matching the threshold, which is jx_j. With proposers getting ever more accommodating, this will start being true at some point, and hence the responders end up following the proposers to ever higher thresholds.

All of this is the mirror image of the situation without commitment. The difference between the two situations is of course that in the case without commitment by the responders, proposers cannot reconsider their proposal if it is rejected, while in the other case, responders can reconsider their intent to reject. It will therefore be harder for responders to commit to rejection than it is, by the nature of the game, for proposers to stick to their proposal.

4A.2 Commitment in simultaneous move games

Also in simultaneous move games, commitment can evolve. The principle is the same as with sequential move games. An individual that is altruistic ends up taking an action that is not fitness maximizing, given what the other player does. But what the other player does, might depend of your level of altruism, even if the other player is selfish. In public goods games, the return to the public good for the other player might increase, if your contribution increases. The benefit of committing to giving more than one would otherwise, lies in the increase in contribution that brings about in the other. Also the opposite is possible; individuals can evolve spite, if committing to not contributing helps force your partner to pick up the tab, and step up her contribution.

In order to illustrate this, we go to the framework of Alger and Weibull (2012), where players are endowed with preferences, which can be altruistic, selfish, or spiteful. Players choose an action from a continuum. Which action they choose, depends on their prefer-

ences, and on what they expect the other player to do. A Nash equilibrium between two players with given preferences is a combination of actions, for which both maximize their utility (they follow their preferences), given the action of the other. Selection then acts on preferences, where preferences that result in higher fitnesses, or material payoffs, for the player that has them, have a selective advantage over preferences that result in lower material payoffs for the player that has them. In this framework, there are therefore two levels; behaviour is determined by preferences, and preferences are selected on the basis of the material payoffs they result in.

One would perhaps expect that this would always lead to preferences that simply align with maximizing the material payoff to oneself, but we will see that this is not the case. Alger and Weibull (2012) find that for games with strategic complements, altruism can evolve, and for games with strategic substitutes, spite can evolve. This can then be combined with assortment, which can add extra altruism, but here, we just focus on the commitment part, which we illustrate with two examples.

In order for commitment to work, we of course need to assume that commitment is recognized, and therefore we assume that the preferences are common knowledge; both players know their own preferences, and they know the preferences of the other player.

4A.2.1 Example 1: altruism for strategic complements

Consider a symmetric 2-player game, with the following fitness function, or material payoffs, for player 1:

$$\pi_1(x, y) = 4(xy)^{\frac{1}{2}} - x^2$$

Here, x is the action, or strategy, of player 1, y is the action of player 2, and $\pi_1(x, y)$ denotes the material payoffs to player 1 for this combination of actions. These material payoffs may differ from the utilities that different combinations of x and y may give the players. The game is symmetric, so the material payoffs to player 2 are $\pi_2(x, y) = \pi_1(y, x) = 4(xy)^{\frac{1}{2}} - y^2$.

Figure 4A.1A depicts these material payoffs. For the red lines, we fixed the action y of player 2, varied the action x of player 1, and plotted the corresponding material payoffs for both players; for player 1 on the horizontal axis, and for player 2 on the vertical axis. If player 1 increases x, then that always increases the material payoff of player 2. The effect on her own material payoffs depends on the current combination of x and y. For $x < y^{1/3}$, increasing x also increases the material payoff of player 1. For $x > y^{1/3}$, increasing x decreases her own material payoff. For the four red lines, y is fixed at 1, $1\frac{1}{6}$, $1\frac{1}{3}$, and $1\frac{1}{2}$,

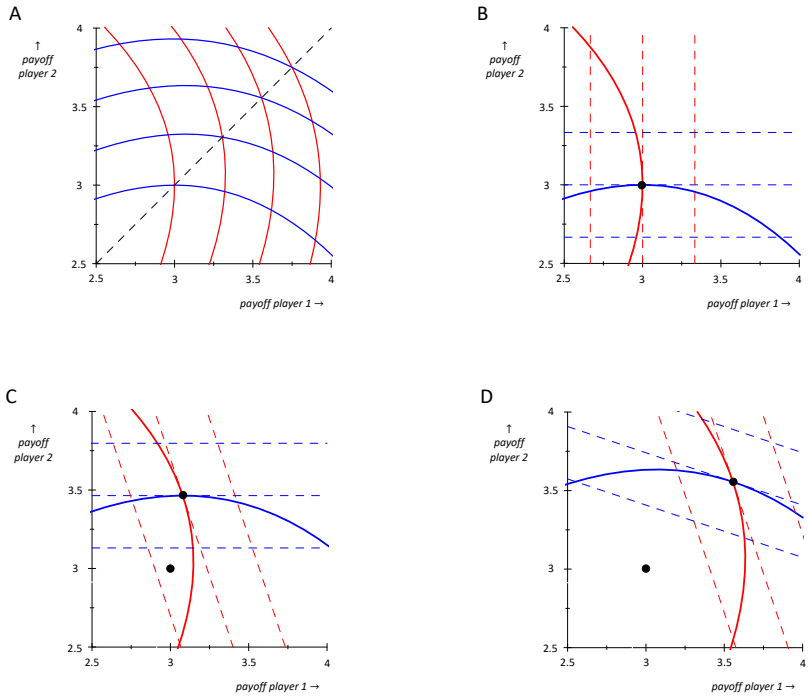


Figure 4A.1: Commitment to altruism in games with strategic complements. (A) Given a choice for y by player 2, player 1 can choose x's that result in material payoffs on a red curve. Given a choice for x by player 1, player 2 can choose y's that result in material payoffs on a blue curve. If both players are selfish, and maximize their own material payoffs, (B) depicts the Nash equilibrium between them. If player 1 is altruistic, and player 2 is selfish, (C) depicts the Nash equilibrium between them. Player 1 now ends up with higher material payoffs than in (B), because her altruism induces player 2 to increase y. Ever higher levels of altruism evolve, until further increases in altruism do not lead to higher material payoffs. (D) depicts the Nash equilibrium between two individuals that have the equilibrium level of altruism.

respectively.

The blue lines do the same, but from the perspective of player 2. We fixed the action x of player 1, varied the action y of player 2, and plotted the corresponding material payoffs for both players. For the four blue lines, y is fixed at 1, $1\frac{1}{6}$, $1\frac{1}{3}$, and $1\frac{1}{2}$, respectively, and player 2 maximizes her own material payoffs at intermediate values of y.

If both players are selfish, their utilities are determined only by how much material payoff they get themselves. A selfish utility function for player 1 would be

$$u_1(x, y) = \pi_1(x, y),$$

while for player 2, it would be the mirror image. In Figure 4A.1B, this is represented by indifference curves, which are vertical straight lines for player 1, and horizontal straight lines for player 2. Maximizing player 1's material payoff, given an action of player 2, would amount to finding the rightmost point on a red curve, and maximizing player 2's material payoff, given an action of player 1, would amount to finding the highest point on a blue curve. In a Nash equilibrium between two selfish players, they would both maximize their own material payoff, given the action of the other.

If a player is altruistic, it would attach a positive weight to the material payoff of the other player. For player 1, an altruistic utility function would be

$$u_1(x, y) = \pi_1(x, y) + \alpha_1 \pi_2(x, y).$$

In this example, if player 2 remains selfish, but player 1 changes to an altruistic preference (for instance, one with $\alpha_1 = \frac{1}{3}$, as in Figure 4A.1C), it will prefer to increase its x, as long as the increase in material payoffs to the other player is at least three times the decrease in material payoffs to herself. Because of the strategic complementarity, this increase in x will induces player 2, who is still selfish, to increase y. In the equilibrium between an altruistic player 1 and a selfish player 2, player 1 gets a material payoff that is higher than the material payoff that a selfish player 1 would get (see Figure 4A.1C). The selfish player 2 it is matched with gets even higher payoffs, but that is not hat matters; what matters is how a selfish player 1 and an altruistic player 1 compare, when both meet a selfish player 2. Given that the altruistic player 1 does better, altruism can invade.

Mutants with increased levels of altruism can invade, and will take over, as long as the resident has an altruism level below $\frac{1}{3}$. Past that point, even more altruistic mutants start getting lower material payoffs. At the equilibrium level of altruism, neither of the players would want to change their behaviour, given their preferences (Fig 4A.1D), and evolution would not change their preferences.

4A.2.2 Example 2: spite for strategic substitutes

Consider a symmetric 2-player game, with the following material payoff function for player 1:

$$\pi_1(x, y) = 8(x + y)^{\frac{1}{2}} - \sqrt{2}x^2$$

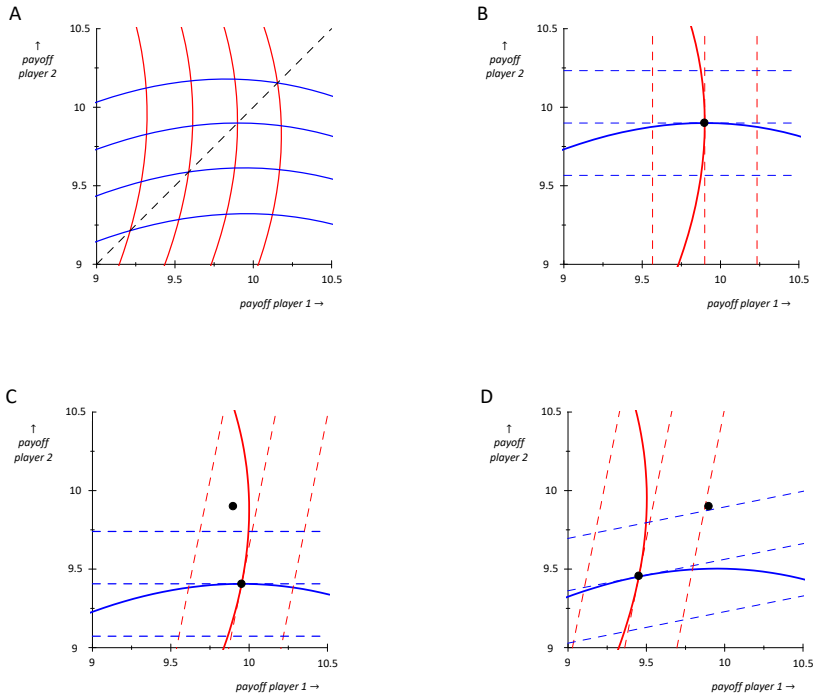


Figure 4A.2: Commitment to spite in games with strategic substitutes. (A) Given a choice for y by player 2, player 1 can choose x's that result in material payoffs on a red curve. Given a choice for x by player 1, player 2 can choose y's that result in material payoffs on a blue curve. If both players are selfish, and maximize their own material payoffs, (B) depicts the Nash equilibrium between them. If player 1 is spiteful, and player 2 is selfish, (C) depicts the Nash equilibrium between them. Player 1 now ends up with higher material payoffs than in (B), because her spite induces player 2 to increase y. Ever higher levels of spite evolve, until further increases in spite do not lead to higher material payoffs. (D) depicts the Nash equilibrium between two individuals that have the equilibrium level of spite.

Here, x is the action, or strategy, of player 1, and y is the action of player 2. The game is symmetric, so the material payoffs to player 2 are $\pi_2(x,y) = \pi_1(y,x) = 8(x+y)^{\frac{1}{2}} - \sqrt{2}y^2$.

Figure 4A.2A depicts these material payoffs. For the red lines, we fixed the action y of player 2, varied the action x of player 1, and plotted the corresponding material payoffs for both players; for player 1 on the horizontal axis, and for player 2 on the vertical axis. If player 1 increases x, then that always increases the material payoff of player 2. The effect

on her own material payoffs depends on the current x and y. For low x, increasing x also increases the material payoff of player 1. For high x, increasing x further decreases her own material payoff. For the four red lines, y is fixed at 0.8, 0.9, 1, and 1.1, respectively.

The blue lines do the same, but from the perspective of player 2. We fixed the action x of player 1, varied the action y of player 2, and plotted the corresponding material payoffs for both players. For the four blue lines, y is fixed at 0.8, 0.9, 1, and 1.1, respectively, and player 2 maximizes her own material payoffs at intermediate values of y.

If both players are selfish, their utilities are determined only by how much material payoff they get themselves. A selfish utility function for player 1 would be

$$u_1(x, y) = \pi_1(x, y),$$

while for player 2, it would be the mirror image. In Figure 4A.2B, this is represented by indifference curves, which are vertical straight lines for player 1, and horizontal straight lines for player 2. Maximizing player 1's material payoff, given an action of player 2, would amount to finding the rightmost point on a red curve, and maximizing player 2's material payoff, given an action of player 1, would amount to finding the highest point on a blue curve. In a Nash equilibrium between two selfish players, they would both maximize their own material payoff, given the action of the other.

If a player is spiteful, it would attach a negative weight to the material payoff of the other player. For player 1, a spiteful utility function is the same as an altruistic utility function, but with a negative altruism parameter α:

$$u_1(x, y) = \pi_1(x, y) + \alpha_1 \pi_2(x, y).$$

In this example, if player 2 remains selfish, but player 1 changes to a spiteful preference (for instance, one with $\alpha_1 = -\frac{1}{5}$, as in Figure 4A.2C), it will prefer to decrease its x, as long as the decrease in material payoffs to the other player is at least five times the decrease in material payoffs to herself. Because of the strategic substitutability, this decrease in x will induces the other player, who is still selfish, to make up for that, and increase y. In the equilibrium between a spiteful player 1 and a selfish player 2, player 1 gets a material payoff that is higher than the material payoff that a selfish player 1 would. Given that the spiteful player 1 does better, spite can invade.

Mutants with increased levels of spite can invade, and will take over, as long as the resident has an α above $-\frac{1}{5}$. Past that point, even more spiteful mutants start getting

lower material payoffs. At the equilibrium level of spite, neither of the players would want to change their behaviour, given their preferences (Fig 4A.2D), and evolution would not change their level of spite.

4A.2.3 Math notes for example 1

Assume that player 1 has altruism level α_1, and player 2 has altruism level α_2. That implies that player 1 maximizes her utility if the derivative of her utility to x is zero:

$$\begin{aligned}
\frac{d\left(\pi_1\left(x,y\right)+\alpha_1\pi_2\left(x,y\right)\right)}{dx} &= 0 \\
2\left(1+\alpha_1\right)\left(\frac{y}{x}\right)^{\frac{1}{2}}-2x &= 0 \\
\left(1+\alpha_1\right)\left(\frac{y}{x}\right)^{\frac{1}{2}} &= x \\
\left(1+\alpha_1\right)y^{\frac{1}{2}} &= x^{\frac{3}{2}} \\
\left(1+\alpha_1\right)^{\frac{2}{3}}y^{\frac{1}{3}} &= x
\end{aligned}$$

The contribution x of player 1 is increasing in her level of altruism α_1, and it is also increasing in the contribution y of the other player.

Similarly, player 2 maximizes her utility if

$$\left(1+\alpha_2\right)^{\frac{2}{3}}x^{\frac{1}{3}}=y$$

In a fixed point (x,y), where both maximize their utility given the choice the other, both of these need to hold. That makes the equation for x

$$\begin{aligned}
\left(1+\alpha_1\right)^{\frac{2}{3}}\left(1+\alpha_2\right)^{\frac{2}{9}}x^{\frac{1}{9}} &= x \\
\left(1+\alpha_1\right)^{\frac{2}{3}}\left(1+\alpha_2\right)^{\frac{2}{9}} &= x^{\frac{8}{9}} \\
\left(1+\alpha_1\right)^{\frac{3}{4}}\left(1+\alpha_2\right)^{\frac{1}{4}} &= x
\end{aligned}$$

Similarly, in Nash equilibrium, player 2 plays

$$\left(1+\alpha_1\right)^{\frac{1}{4}}\left(1+\alpha_2\right)^{\frac{3}{4}}=y$$

This leads to material payoffs to player 1, as functions of their altruism levels:

$$4\left(\left(1+\alpha_1\right)\left(1+\alpha_2\right)\right)^{\frac{1}{2}}-\left(1+\alpha_1\right)^{\frac{3}{2}}\left(1+\alpha_2\right)^{\frac{1}{2}}$$

Now we can set the derivative to α_1 to zero, to see which level of altruism maximizes fitness, or material payoffs.

$$\begin{aligned}
2\left(\frac{1+\alpha_2}{1+\alpha_1}\right)^{\frac{1}{2}} - \frac{3}{2}\left(1+\alpha_1\right)^{\frac{1}{2}}\left(1+\alpha_2\right)^{\frac{1}{2}} &= 0 \\
2\left(1+\alpha_1\right)^{-\frac{1}{2}} &= \frac{3}{2}\left(1+\alpha_1\right)^{\frac{1}{2}} \\
2 &= \frac{3}{2}\left(1+\alpha_1\right) \\
\alpha_1 &= \frac{4}{3} - 1 = \frac{1}{3}
\end{aligned}$$

In this case, the optimal level of altruism for player 1 is independent of the level of altruism that player 2 has. That makes $\alpha = \frac{1}{3}$ the evolutionary stable equilibrium level of altruism.

4A.2.4 Math notes for example 2

Assume that player 1 has altruism level α_1, and player 2 has altruism level α_2. That implies that player 1 maximizes her utility if

$$\begin{aligned}
\frac{d\left(\pi_1\left(x,y\right)+\alpha_1\pi_2\left(x,y\right)\right)}{dx} &= 0 \\
4\left(1+\alpha_1\right)\left(x+y\right)^{-\frac{1}{2}} - 2\sqrt{2}x &= 0 \\
2\left(1+\alpha_1\right)\left(x+y\right)^{-\frac{1}{2}} &= \sqrt{2}x \\
4\left(1+\alpha_1\right)^2\left(x+y\right)^{-1} &= 2x^2 \\
2\left(1+\alpha_1\right)^2 &= x^2\left(x+y\right)
\end{aligned}$$

We will leave this an implicit solution, but from the equation, we can see that the contribution x of player 1 is increasing in her level of altruism α_1, and decreasing in the contribution y of the other player.

Similarly, player 2 maximizes her utility if

$$2\left(1+\alpha_2\right)^2 = y^2\left(x+y\right)$$

In a fixed point (x, y), where both maximize their utility given the choice of the other,

both of these need to hold, and therefore

$$\begin{aligned}\frac{2\left(1+\alpha_1\right)^2}{2\left(1+\alpha_2\right)^2} &= \frac{x^2\left(x+y\right)}{y^2\left(x+y\right)}\\ \frac{1+\alpha_1}{1+\alpha_2} &= \frac{x}{y}\\ y &= \left(\frac{1+\alpha_2}{1+\alpha_1}\right)x\end{aligned}$$

That makes the equation for x

$$\begin{aligned}2\left(1+\alpha_1\right)^2 &= x^2\left(x+\left(\frac{1+\alpha_2}{1+\alpha_1}\right)x\right)\\ 2\left(1+\alpha_1\right)^2 &= x^3\left(\frac{2+\alpha_1+\alpha_2}{1+\alpha_1}\right)\\ 2\left(\frac{\left(1+\alpha_1\right)^3}{2+\alpha_1+\alpha_2}\right) &= x^3\\ \left(1+\alpha_1\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}} &= x\end{aligned}$$

Similarly, in Nash equilibrium, player 2 plays

$$\left(1+\alpha_2\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}} = y$$

This leads to material payoffs to player 1, as functions of their altruism levels:

$$\begin{aligned}8\left(\left(1+\alpha_1\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}+\left(1+\alpha_2\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}\right)^{\frac{1}{2}}-\sqrt{2}\left(\left(1+\alpha_1\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}\right)^2 &=\\ 8\left(\left(2+\alpha_1+\alpha_2\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}\right)^{\frac{1}{2}}-\sqrt{2}\left(\left(1+\alpha_1\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}\right)^2 &=\\ 8\left(\left(2+\alpha_1+\alpha_2\right)^{\frac{2}{3}}\left(2\right)^{\frac{1}{3}}\right)^{\frac{1}{2}}-\sqrt{2}\left(\left(1+\alpha_1\right)\left(\frac{2}{2+\alpha_1+\alpha_2}\right)^{\frac{1}{3}}\right)^2 &=\\ 2^{19/6}\left(2+\alpha_1+\alpha_2\right)^{\frac{1}{3}}-2^{7/6}\left(1+\alpha_1\right)^2\left(2+\alpha_1+\alpha_2\right)^{-\frac{2}{3}} &=\\ 2^{7/6}\left[4\left(2+\alpha_1+\alpha_2\right)^{\frac{1}{3}}-\left(1+\alpha_1\right)^2\left(2+\alpha_1+\alpha_2\right)^{-\frac{2}{3}}\right] &=\end{aligned}$$

Now we can set the derivative to α_1 to zero, to see which level of altruism maximizes

fitness, or material payoffs.

$$\frac{4}{3}\left(2+\alpha_1+\alpha_2\right)^{-\frac{2}{3}}-2\left(1+\alpha_1\right)\left(2+\alpha_1+\alpha_2\right)^{-\frac{2}{3}}+\frac{2}{3}\left(1+\alpha_1\right)^2\left(2+\alpha_1+\alpha_2\right)^{-\frac{5}{3}}=0$$

Because this is symmetric, there will be an equilibrium where $\alpha_1=\alpha_2$, so we can rewrite this as

$$\begin{aligned}
\frac{4}{3}\left(2+2\alpha\right)^{-\frac{2}{3}}-2\left(1+\alpha\right)\left(2+2\alpha\right)^{-\frac{2}{3}}+\frac{2}{3}\left(1+\alpha\right)^2\left(2+2\alpha\right)^{-\frac{5}{3}}&=0\\
\frac{4}{3}*2^{-\frac{2}{3}}\left(1+\alpha\right)^{-\frac{2}{3}}-2*2^{-\frac{2}{3}}\left(1+\alpha\right)^{\frac{1}{3}}+\frac{2}{3}*2^{-\frac{5}{3}}\left(1+\alpha\right)^{\frac{1}{3}}&=0\\
\frac{4}{3}\left(1+\alpha\right)^{-\frac{2}{3}}-2\left(1+\alpha\right)^{\frac{1}{3}}+\frac{1}{3}\left(1+\alpha\right)^{\frac{1}{3}}&=0\\
\frac{4}{3}\left(1+\alpha\right)^{-\frac{2}{3}}-\frac{5}{3}\left(1+\alpha\right)^{\frac{1}{3}}&=0\\
4\left(1+\alpha\right)^{-\frac{2}{3}}-5\left(1+\alpha\right)^{\frac{1}{3}}&=0\\
4-5\left(1+\alpha\right)&=0\\
\alpha&=-\frac{1}{5}
\end{aligned}$$

Chapter 5

Evolution and the ultimatum game: Why do people reject unfair offers?[1]

5.1 Introduction

Humans are not just selfish. When deciding what to do, we do not only look at how our behaviour affects ourselves, but we also take into account the consequences of our actions for others. How we came to deviate from straightforward selfishness is one of the bigger questions in human evolution.

One of the classical games in which we see deviations from selfish money-maximizing behaviour is the ultimatum game (Güth et al., 1982). This game is played between a proposer and a responder. The proposer makes a proposal how to distribute a given amount of money between herself and the responder. The responder then accepts or rejects the proposal. If she rejects, neither player gets any money. For responders, the selfish money-maximizing choice would be to accept any proposal in which she gets a positive amount of money. That, however, is not what we find in lab experiments (Güth et al., 1982; Oosterbeek et al., 2004), where low offers regularly get rejected.

In this paper, we will review existing models from evolutionary game theory that aim at explaining this behaviour. We will also create new and improved versions of those models, describe their predictions in greater detail, and compare these predictions with

[1]This chapter is based on the working paper by Akdeniz and van Veelen (2022).

the existing empirical evidence. This means that the paper will make a series of points, but we hope that the multitude of observations does not conceal the importance of each individual one.

5.1.1 Mutation-selection equilibria, bias, and the asymmetry argument

Two well-known models from the literature are Gale et al. (1995) and Rand et al. (2013). Both of these models describe mutation-selection equilibria. The ingredient that these models aim to capture is that not all suboptimal behaviours are equally costly. Rejecting a proposal in which the responder gets 1 euro and the proposer gets 9 costs the responder 1 euro. On the other hand, if offering 2 euros would have been accepted, then proposing 1 euro for the responder and 9 for oneself, and having this proposal rejected, costs the proposer 8 euros. In this example, one could therefore say that both the proposer and the responder made a mistake, and that the mistake made by the proposer is much more costly than the mistake made by the responder.

The concept of a mutation-selection equilibrium assumes that mutation creates a constant inflow of suboptimal strategies. In the ultimatum game, the asymmetry in how bad these mutations are for the fitness of their carriers then translates to an asymmetry in how long it takes for selection to eliminate them, and an asymmetry in how much they hurt the fitness of those they meet on their way out. Mutant proposers and mutant responders therefore ripple through the population differently, and that ends up having a nontrivial effect on what we should expect to see if mutation and selection balance in equilibrium.

A problem with both Gale et al. (1995) and Rand et al. (2013), however, is that both these models have global, and therefore biased mutations. As a result, the deviations from selfish, money maximizing behaviour that they find are primarily driven by the bias in the mutations, and not so much by the asymmetry in how costly different suboptimal behaviours are. Because mutation bias is not a good basis for an explanation, we redo both models, with mutations that are local instead of global. Switching from global to local mutations reduces the bias to a minimum, and changes the results significantly. In Section 5.2 we do this for the model in Rand et al. (2013), and in Section 5.3 we do this for the model in Gale et al. (1995).

The original papers, understandably, only focus on predictions regarding the average offer and the average threshold, below which responders start rejecting. We also look at other aspects of the prediction, such as, for instance, the relation between the average offer and

the average threshold, the within population variance, and the variance across time, or across populations. In the model from Rand et al. (2013) low intensities of selection push offers and thresholds up, thereby getting them closer to levels found in experiments. In Section 5.2 we show that this comes at a cost, and that is that weakening selection also deteriorates the match between model predictions and empirical findings on these other dimensions.

5.1.2 Quantal Response Equilibria, learning dynamics, and the asymmetry argument (again)

Instead of looking at the ultimate level immediately, one can also look at a more proximate level first, and ask the question if the conclusion that humans deviate from selfish money-maximizing behaviour is justified. An alternative interpretation of the behaviour observed in the lab could be that individuals are in fact selfish, but that they are not perfectly informed, or otherwise not perfectly aware of what it is that they should do in order to earn as much money as they can. In order to do that, we calculate the Quantal Response Equilibria (McKelvey and Palfrey, 1995) for the ultimatum game (Yi, 2005), under the assumption of selfishness. This is a concept from classical game theory, that can also be seen as resulting from learning dynamics with noise. Also here, as with mutation-selection equilibria, the asymmetry argument is relevant, and also here, one can crank up the noise to get the average offer and the average threshold up quite a bit. We do however observe that other characteristics of the equilibrium distribution provide a poor match with the empirical evidence. In one of the two types of Quantal Response Equilibria, the distribution of MAO's is predicted to be downward sloping; the higher the MAO, the rarer they should be. This is not confirmed by the data, which suggests that there is more to human behaviour in the ultimatum game than everyone trying to get high monetary payoffs, but not knowing exactly what to do to get them.

This also carries over to explanations at the ultimate level. An equivalent argument about the mismatch between the shape of the distribution and the experimental evidence actually implies that every model that is only based on the asymmetry argument, and that does not include a pathway through which rejecting has actual fitness benefits, is to be rejected too. This therefore also applies to the models in Gale et al. (1995) and Rand et al. (2013) as well as our de-biased versions of them.

5.1.3 Commitment, and a unified model

Another well-known model for the evolution of behaviour in the ultimatum game is Nowak et al. (2000). In this model, rejecting itself is still bad for fitness, but accepting lower offers than others do, can also lead to getting lower offers than others – provided that proposers have a way of finding out how low they can go and still have their offer accepted. This means that the model specifies a pathway through which being the accepting type can actually be bad for fitness, and being the rejecting type can be good for fitness. That makes this model different from the models in Gale et al. (1995) and Rand et al. (2013), and their de-biased versions, in which there is never a fitness advantage to being the rejecting type.

We provide a version of the model from Nowak et al. (2000) which lifts some exogenously imposed restrictions on what strategies individuals can and cannot use. Our version is moreover a general model, in the sense that it contains our de-biased version of the model from Rand et al. (2013) as a special case. This helps illustrate the interaction between commitment (Akdeniz and van Veelen, 2021; Frank, 1987; 1988) and the asymmetry in costliness of mistakes.

5.2 Mutation-selection equilibria: Rand et al. (2013)

5.2.1 The simulation model in Rand et al. (2013)

Rand et al. (2013) consider a finite population model, in which 100 individuals play ultimatum games in both roles. Every individual has a strategy that specifies the offer they make in the role of proposer, as well as their minimal acceptable offer (MAO) in the role of responder. These offers and thresholds range from 0 to 1, and in the simulations we will be focusing on, they do so continuously. Each generation, every individual plays the ultimatum game with every other individual, once as a proposer and once as a responder. The resulting payoff is the average of the payoffs over all 99 pairings (in which a total of 198 games are played).

The population is updated according to a Moran process. One agent is picked at random to die, and individual $i \in \{1, ..., 100\}$ is picked with probability proportional to $exp\,(w\pi_i)$ to reproduce, where w is the intensity of selection, and π_i is the average payoff of individual i. Mutations happen at rate u at reproduction; with probability $1-u$, the new individual inherits the strategy from the reproducing individual, and with probability u, the new individual carries a randomly selected strategy. The distribution from which the mutant

is drawn is independent of the strategy before mutation; both the new offer and the new MAO are always drawn from a uniform distribution on $[0, 1]$. We will refer to this as global mutation.

The average trait value of the mutant is always $\frac{1}{2}$ – which is the value halfway the interval out of which the mutants are drawn – regardless of the trait value before mutation. Selection always works in favour of low values of the MAO's, and for low values of the MAO, it works in favour of low offers. That means that selection pulls these values towards the bottom of the interval $[0, 1]$. Therefore, when selection is at work, the average mutant has a higher offer and a higher MAO than the average offer and the average MAO in the population. In other words, mutation is biased, and mutants will result in increased offers and MAO's more often than they result in decreased offers and MAO's.

5.2.2 Our version

There are two inconsequential differences between their simulations and ours. The first is that we use a Wright-Fisher process instead of a Moran process. The Wright-Fisher process is computationally more efficient, but other than that, it perfectly reproduces the findings in Rand et al. (2013) for global mutations. The second inconsequential difference is that Rand et al. (2013) have co-occurring mutations; if an individual mutates, then both a new offer and a new MAO are drawn. Our version of the model has independent mutations. At any reproduction event, the offer mutates with probability u, and so does the MAO. That means that with probability u^2 mutations of the offer and of the MAO co-occur, and with probability $2u\,(1-u)$ only one of them mutates. Also this does not make much of a difference (see Supplementary Material 1.4 for details).

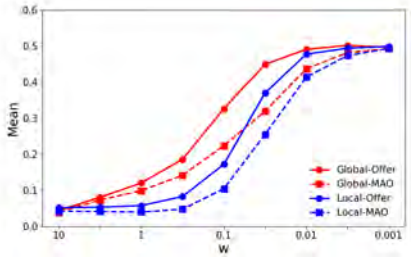
The important, and consequential difference is that in our version mutations are not global. Instead, mutations are changes with respect to the old trait value. That means that if a mutation of the offer happens, and the old offer is p, then the new offer is $p+\Delta p$, where Δp is drawn from a uniform distribution on $[-0.1, 0.1]$. There are two exceptions. The first is a result of the fact that we do not allow for offers below 0. Therefore, if $p + \Delta p < 0$, the new offer is 0. Similarly, we also do not allow for offers over 1, and therefore, if $p+\Delta p > 1$, the new offer is 1. This implies that mutations are unbiased for trait values in $[0.1, 0.9]$, and become a little biased if they drop below 0.1 or go over 0.9 (in which case the bias is still very small compared to the bias with global mutations in Rand et al., 2013). The same procedure applies to the MAO.

5.2.3 Global versus local mutation

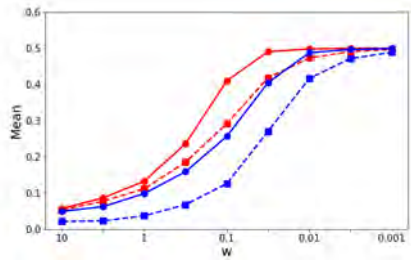
In Figure 5.2.1 we compare the results from Rand et al. (2013) with the results for our version. That makes this figure the counterpart of Figure 2 in Rand et al. (2013). For our figure, we did however choose to invert the horizontal axis. Their Figure 2 has low intensities of selection on the left and high intensities on the right. We do the opposite. The reason for that is that we want to make it clear that the benchmark, on the left, is the situation where responders accept all positive offers, and proposers offer nothing or close to nothing. The models investigate ways to arrive at dynamics that push the average offer and the average MAO up from 0, and we want it to be clear that reducing the intensity of selection does exactly that in both versions.

Lower offers, lower MAO's From the simulations, we learn that there are two important differences between global and local mutations. The first is that with the bias significantly reduced, the average offers and MAO's stay low for longer, and require further reduced intensities of selection to reach the same average offers and average MAO's. Here it is important to note that on the right hand side of the graph, with low intensities of selection, average offers and average MAO's end up at 0.5 in both versions. The reason why this eventually happens, and why that would happen for a range of reasonable modeling choices, is that 0.5 is halfway the parameter space, and therefore this must be the average over time in the limit of weak selection, where payoffs cease to matter. In Section 5A.2 we will discuss in more detail why allowing for arbitrarily low intensity of selection, while focusing on how far offers and MAO's can be pushed up on average, limits the predictive power of the model on other criteria.

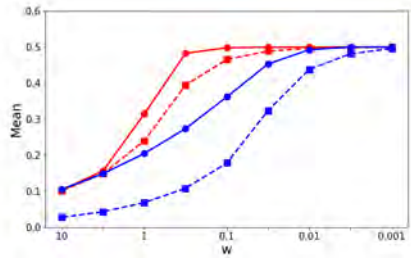
Gap vs. no gap The second difference is there for all mutation rates in Figure 5.2.1, but it is the most visible for $u = 0.1$ (Figure 5.2.1C). Here we see that on the left side of the graph, at high intensities of selection, there is no perceptible gap between the average offer and the average MAO for global mutation, while there is a very visible gap for local mutations. The latter is consistent with the asymmetry argument. Given that there is a consistent inflow of *local* mutations, proposers benefit from creating some space between their offer and the average MAO in the population; this way they reduce the risk that their offer is rejected by a responder with an above average MAO. Responders always get higher payoffs if they accept, and therefore they always benefit from moving their MAO down. The closer they get to an MAO of 0, however, the less of a difference a further decrease in their MAO makes. Also, if proposers increase their offers, that reduces the selection pressure against low MAO's. Both sides therefore want to create some distance, but since MAO's cannot drop below 0, that will result in mutations moving both averages



(A)



(B)



(C)

Figure 5.2.1: Global versus local mutations. In red the average offers and MAO's for the model in Rand et al. (2013), which has global, co-occurring mutations. In blue the same, but for local, independent mutations. Both the average offers and the average MAO's are higher with global, and therefore biased mutations, and lower with local, and therefore much less biased mutations. In order to get offers, or MAO's, up to average levels found in experiments, one would have to move to lower intensities of selection with local mutations than one would with global mutations. Section 5A.2 explains why that is problematic. The mutation rate is 0.001 in panel A, 0.01 in panel B, and 0.1 in panel C.

up from 0, with a gap in between.

If we then start on the left hand side of the graph, and move a little to the right, then first the effect of reducing the intensity of selection is that this keeps mutants around for longer. With local mutations, this creates a wider distribution of offers and MAO's, which selects for strategies that on average keep more distance. This, in turn, leads to higher

offers and MAO's due to the asymmetry in selection pressure. On the left end of the graph, we therefore see a widening gap, and an increase in offers and MAO's. Later on, when selection gets even weaker, and we get closer to the right end of the graph, everything just becomes noise. That causes both average offers and average MAO's to approach 0.5, which closes the gap.

With global mutations, on the other hand, there is hardly any gap at first. Here, the moving up of the offers and MAO's as selection gets weaker is the result of the bias in mutations balancing against ever weaker selection. The absence of a gap in the beginning therefore is understandable, because with *global* mutations, the equilibrium distribution of MAO's away from the mode is much more spread out. This implies that moving away from where most MAO's are does not make enough of a difference for the probability to have one's offer accepted, and that makes the reason to move away, that is there with local mutations, vanish.

Both differences – there being a gap versus there not being a gap at the left end of the graph, and the overall difference in average offers and average MAO's – indicate that with local mutations, the dynamics are mainly driven by the asymmetry in fitness effects, while the dynamics with global mutations are primarily driven by bias in the mutations. The latter is not a good basis for an explanation of deviations of selfishness.

5.2.4 Predictions for weak selection

In Section 5.2.3, we have seen that both with global and with local mutations, lowering the intensity of selection allows the average offer and the average MAO's to move away from 0, and towards $\frac{1}{2}$. There are however limitations to how observations about the averages for weak selection can be interpreted meaningfully. To see the reasons why, we will look a bit more closely at the dynamics in the absence of selection.

Reason 1: going against selection by shutting selection down When the intensity of selection is 0, the dynamics in the model by Rand et al. (2013) are only driven by mutations. That implies that with global mutations, what we are seeing is the result of a sequence of random draws from a uniform distribution on $[0, 1]$, where no value of the draw is more likely to survive for longer and reproduce more than any other. Therefore, if we let the simulation run long enough, and we choose the intensity of selection to be 0, we will see the average offer and the average MAO converge to $\frac{1}{2}$ (which is the midpoint of the interval $[0, 1]$, and the expected value of the uniform distribution over it). By choosing a sufficiently weak intensity of selection, one can moreover get these averages

anywhere between 0 – the limit for unfettered selection – and $\frac{1}{2}$ – the limit for unfettered mutation.

All of this implies that the fact that it is possible to get the average offer or the average MAO up to any value between 0 and 0.5, by choosing a sufficiently low intensity of selection, is not necessarily informative about selection – other than that selection always points towards lower offers and lower MAO's. Selection pulls both of them down, and if one reduces selection strength ever more, one can reduce by how much both are dragged down. The observation that one can find parameter choices for which the averages in the simulations match averages from experiments therefore is a somewhat arbitrary result of the fact that the average trait value in the strategy set is $\frac{1}{2}$, and not a reflection of what selection does to the strategies in this set. By definition of what happens in the limit of weak selection, and what happens in the limit of strong selection, deviations of which we try to explain, the model covers everything between offers and MAO's being 0, and the equal split.

A more general probabilistic symmetry argument, given in Supplementary Material 2.1, also applies to the version with local mutations. In this case, the results are driven much less by bias in the mutation, and much more by the asymmetry in costliness of mistakes, but the fact that also here *any* average offer below 0.5 can be reached by choosing a sufficiently low intensity of selection is an artefact of the fact that the neutral process, with mutation only, finds itself in the middle of the strategy space on average.

Reason 2: averages over the population and time versus averages over the population There is also a second reason why not too much should be made of the fact that one get the average offer and the average MAO in the simulations to match average offers and average MAO's from lab experiments, if that requires choosing low intensities of selection. That reason has to do with the fact that the averages reported for the simulations are averages over populations and over time, and the averages in lab experiments are only averages over a given population. It is important to stress that these two are not the same. There are different ways in which the average in a population in a lab experiment can be the same as the average over the population and over time in the simulations, while other aspects of the simulations generate a remarkable mismatch with the empirical evidence.

Figure 5A.1A displays how the average offer and the average MAO within the population change over time in part of a run with relatively infrequent mutation, and weak selection. Mutations there are global and co-occurring, as they are in Rand et al. (2013). Figure

5A.1B is a snapshot, which illustrates that most of the time, the population is at fixation, or close to it. The variance within the population therefore is almost always 0 or close to 0. Over time, however, the offers and MAO's are highly variable; Figure 5A.1C indicates that they are quite literally all over the place.

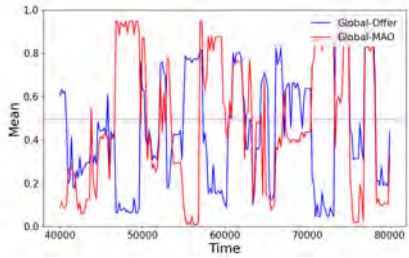
When considering the results from lab experiments, we can assume that different populations are undergoing the same, or similar dynamics, and that implies that we may treat experiments in different populations as equivalent to different moments in time in the same simulation. If we do that, then the within population variance in experiments is much too large, and the between population variance in experiments is much too small to match the simulations, even if we can find model parameters for which the average over time of the average over the population in the simulations match the average for a sample from a population at a given moment in time.

Reason 3: lack of correlation between offers and MAO's Another remarkable observation is that the offer and the MAO in these simulations are almost completely uncorrelated (this is also visible in Figure 5A.1A). As a consequence, the average offer within the population is sometimes higher than the average MAO within the population, but almost equally often it is the other way around. Only when also averaged over time, is the average offer a bit above the MAO, but that masks that they move almost completely independently. Therefore, under weak selection, we should expect to find the average offer to be lower than the average MAO almost as often as the other way around. That is at odds with what is found in for instance cross-cultural experiments, where the offers in any given population are not independent of the income-maximizing offer in that population (Henrich et al., 2001; 2005; 2006). The lack of correlation between offers and MAO's in the simulations therefore is a remarkable mismatch with the empirical data.

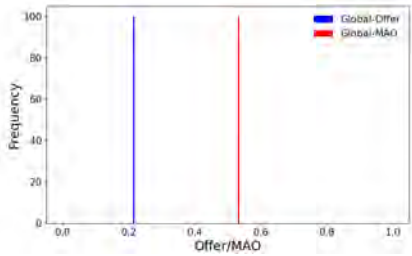
Supplementary Material 2 shows that these mismatches are not confined to the combination of global and infrequent mutation. Whether mutations are global and co-occurring, as in Rand et al. (2013), or local and independent, as in our version, and whether mutations are frequent or infrequent, when selection is weak, averages over time from the simulations may coincide with averages from lab experiments, but predictions from the model that are not aggregated over time are not in line with the empirical evidence.

5.2.5 Mutation rates

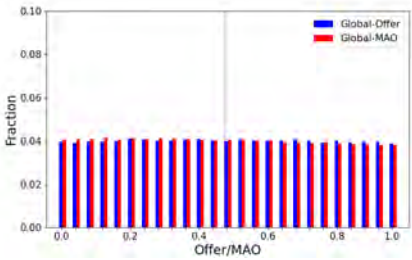
The model in Rand et al. (2013) has two variables that can tilt the balance between mutation and selection; the intensity of selection, and the mutation rate. Decreasing the



(A)



(B)



(C)

Figure 5.2.2: Weak selection. The top panel shows how the average offer and the average MAO in the population change over time for part of a run with an intensity of selection of $w = 0.001$. Mutations are global and co-occurring, and the mutation rate is $u = 0.001$. In the neutral process, the average offer and the average MAO move completely independently. Here, with weak selection, they move almost completely independently (although the timing of changes to both coincides because of co-occurring mutations). The middle panel is a snapshot during the run. The bottom panel gives the average distribution over time, where we collected strategies within intervals of length 0.04. This average distribution is very close to the uniform distribution from which the mutants are drawn. The average over time of the average offers (MAO's) is a horizontal red (blue) line in panel A, and a vertical red (blue) line in panel C.

intensity of selection and increasing the mutation rate both make mutations overwhelm selection against rejecting positive offers. The reasons above point to limitations we encounter if we use the intensity of selection to push up average offers and MAO's for a given

mutation rate. That still leaves the door open for increasing the mutation rate as a way to push offers and MAO's up.

A natural next question is therefore what a reasonable mutation rate is. For global mutations, it is important to realize that the "upward force" as a result of the bias scales up with the mutation rate. With global mutations, at a mutation rate of 1, individuals with high fitness still reproduce more than individuals with low fitness, but all selection is washed out completely by the bias. That means that, whether or not 1 is a realistic mutation rate, what is unrealistic for sure is that the force that is pushing the offers and MAO's up is just the bias in the mutations going in the other direction than selection.

For local mutations, on the other hand, there is only a little bit of bias around the edges (close to 0 and 1). That means that if dynamics take the offers and MAO's in the population up to intermediate levels, then even the moderate amount of bias that is there for trait values close to 0 disappears (instead of scaling up). The argument against high mutation rates with global mutations therefore does not apply with local mutations. One can moreover decide not to interpret the mutation rate too literally. There may be alternative genetic architectures that maintain the same variance within the population with much lower mutation rates. With sexual reproduction, for instance, no offspring is an exact copy of either of the parents. It is however important to realize that, for a given intensity of selection, with local mutations, it is *not* possible to push the average offer and MAO up to any level between 0 and 1. At 1, the highest possible mutation rate, these averages are not at $\frac{1}{2}$, but somewhere strictly (and, depending on the intensity of selection, possibly substantially) below $\frac{1}{2}$. All of this is discussed in more detail in Supplementary Material 1.3, where we fix intensities of selection, and let mutation rates vary.

5.2.6 WEIRD people

Another consideration that suggests we should allow for a margin of error when comparing averages from simulations and averages from experiments, is that those experiments tend to be done with WEIRD subjects, and the environment that makes us WEIRD is evolutionarily new. This is a point made by Henrich et al. (2010). One of the examples they point to is behaviour in the ultimatum game, and this is based on Henrich et al. (2006). In this study, they find that the income maximizing offers in two WEIRD populations (Emory students and rural Missouri) are relatively high compared to 13 non-WEIRD populations – and for the income maximizing offer to be high, there needs to be a relatively large share with a relatively large MAO. With WEIRD people having relatively high MAO's, experiments with WEIRD people therefore may set a bar that is a bit higher than

necessary.

5.2.7 The shape of the distribution

Section 5.4 discusses a possible explanation of the data from lab experiments based on noise (instead of deviations from selfishness). This explanation is rejected by the empirical evidence, and this rejection is based on properties of the distribution other than the average offer or the average MAO. This mismatch between the empirical evidence and the predictions of the noise-based Quantal Response model also carries over to mutation-selection equilibria. It is helpful to first look at what one could consider to a be a somewhat more proximate explanation in order to understand what the prediction is, and why that would also follow from a mutation-selection model. Therefore, we will postpone this point to the end of Section 5.4. It may be good though to point to the fact that the mutation-selection equilibrium has another prediction in store, and to the fact that this one does not pertain to the average offer and the average MAO.

5.2.8 Summarizing

The results in Rand et al. (2013) are for a large part driven by bias in the mutations. If we un-bias the mutation process by replacing global mutations with local mutations, average offers and average MAO's in the simulations drop significantly. We can still get these averages up to levels found in experiments, but in order to do that, we have to choose really low intensities of selection. The fact that one can always do that, is, first of all, a somewhat gratuitous result of the fact that the intensity of selection can serve as a slider that can put us anywhere between the middle of the strategy space, and the point where selection alone would take us. Moreover, as we lower the intensity of selection, we may get the average offer and MAO (over time and over the population) closer to the averages (over the population) in experiments, but other characteristics of the prediction move away from what we observe – including the fact that for really low intensities of selection, the average offer and the average MAO are almost uncorrelated over time.

5.3 Mutation-selection equilibria: Gale et al. (1995)

Another paper that describes mutation-selection equilibria in the ultimatum game is Gale et al. (1995). While Rand et al. (2013) allow for an interpretation with genetic transmission as well as social learning, Gale et al. (1995) explicitly focus on the latter. There are also some technical differences. The model in Rand et al. (2013) has a finite population, for

which they run stochastic simulations. Gale et al. (1995) on the other hand assume an infinitely large population, for which they calculate deterministic replicator dynamics. The strategy space in the main part of Rand et al. (2013) is continuous. The strategy space in Gale et al. (1995), on the other hand, is discrete; individuals can choose offers or MAO's only with certain, fixed increments. There are also some subtle differences concerning how mutation events and reproduction events relate.

These differences in modelling details come with differences in results. We will describe some of those differences here, and, in more detail, in the Supplementary Material. The similarities, however, are more important, and more prominent, than the differences. We will therefore first reproduce their main set of equations, and discuss what we see in equilibrium.

5.3.1 The model in Gale et al. (1995)

In Gale et al. (1995), the size of the pie is 40, but it is clear that one can choose any integer for size. We will therefore let n denote the amount to be divided. In Gale et al. (1995), proposers can offer $i = 1, ..., n$ to the responder; they can only offer integer numbers equal to or smaller than the pie size, but not including 0. Responders are characterized by an MAO, which is denoted by j, and which also ranges from 1 to n in steps of 1.

The differential equations that describe the dynamics are then given by

$$\dot{x}_i = (1-\delta)\left(\pi_{i,P} - \overline{\pi}_P\right) x_i + \delta\left(\frac{1}{n} - x_i\right)$$

for proposers, where x_i is the share of proposers that propose i, $\dot{x}_i$ is its time derivative, δ is the mutation rate, $\pi_{i,P}$ is the payoff of proposers that propose i, and $\overline{\pi}_P$ is the average payoff in the proposer population, and by

$$\dot{y}_j = (1-\delta)\left(\pi_{j,R} - \overline{\pi}_R\right) y_j + \delta\left(\frac{1}{n} - y_j\right)$$

for responders, where y_j is the share of responders with an MAO of j, $\dot{y}_j$ is its time derivative, $\pi_{j,R}$ is the payoff of responders with an MAO of j, and $\overline{\pi}_R$ is the average payoff in the responder population. The payoffs to different types of proposers depend on the composition of the responder population, and the payoffs to different types of reponders depend on the composition of the proposer population. In their paper, Gale et al. (1995) allow for the mutation rates to differ between the proposer and responder populations, but we will start with their default case, where they are the same.

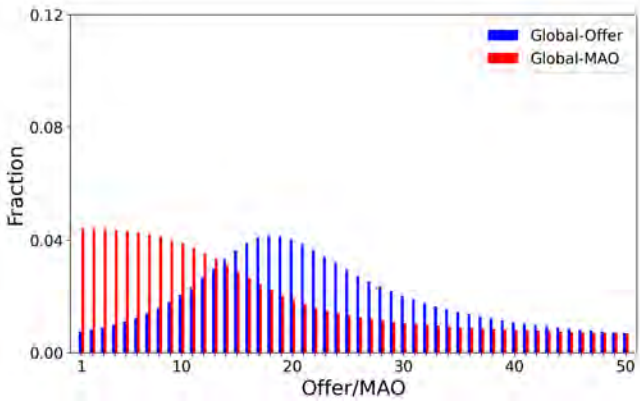


Figure 5.3.1: Mutation-selection equilibrium in Gale et al. (1995). The original model has global mutations, and this mutation-selection equilibrium has a mutation rate δ of 0.15. The thick tails of the distributions are a symptom of the bias in the mutation. With local mutations, the tails are much less thick (see Figure 5.3.2).

We would like to keep the models of Gale et al. (1995) and Rand et al. (2013) as comparable as possible. Some of the simulation results from the model in Rand et al. (2013) are represented by frequencies of strategies in intervals of finite size; see for instance Figure 5A.1B and C. In order to be as close as possible to that way of representing properties of simulation runs, we adjust the spacing of the strategies a little – which does not affect the equations above; the change only induces a minor change in how the payoffs are calculated. Instead of having proposer strategy i propose i, we choose for strategy i to propose the midpoint of the interval $[i-1, i]$, which is $i - \frac{1}{2}$. Similarly, we let responder strategy j have an MAO of $j - \frac{1}{2}$. That means we still have n strategies for both roles, but now we are not treating one end of the range from 0 to n differently; the smallest offer now is $\frac{1}{2}$ up from 0, and the largest is $\frac{1}{2}$ down from n, while before, 0 was excluded and n was included. This change is not consequential for what the mutation-selection equilibria look like.

Without mutations, at $\delta = 0$, almost all starting populations will converge to a population state where all responders have the lowest possible MAO and all proposers make the lowest possible offer. With mutations, that need not be the case. Mutations in Gale et al. (1995) are again global, as they introduce all MAO's and all offers at the same rate. This means that introducing mutations will by definition increase the average offer and the average MAO above 0 as a result of the bias. There are obviously also asymmetries in how fast suboptimal strategies are selected away, which creates the patterns in the offers and MAO's

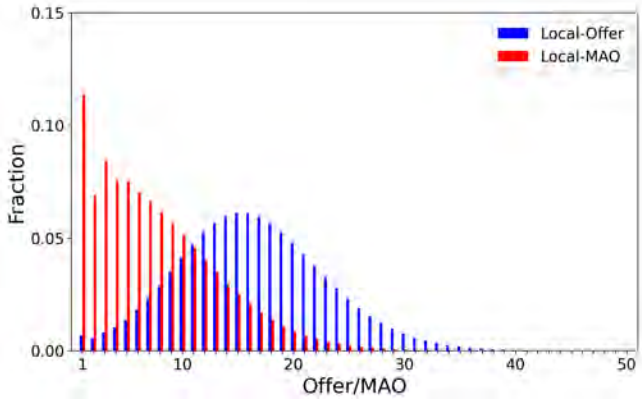


Figure 5.3.2: Mutation-selection equilibrium in Gale et al. (1995) with local mutations. This mutation-selection equilibrium has a mutation rate δ of 0.75. The tails are much thinner than with highly biased, global mutations. The spike at 1 and the dip at 2 are part of a dampening wave pattern caused by the remaining bias in mutations at the edges of the strategy space.

in Figure 5.3.1, but the main force behind the deviations from 0 with global mutations is the bias.

5.3.2 Our version

Because mutation bias is still not a good basis for an explanation, we also made a version of Gale et al. (1995) with local instead of global mutations. Local mutations work in a similar way as in our version of the model from Rand et al. (2013) with local mutations. A mutation induces a change in the offer, and this change can be up to a fixed number of steps to the right, or to the left, where all changes within that range are equally likely (with exceptions similar to those for our local mutations for Rand et al., 2013, if those changes would lead to offers or MAO's below 0 or over n). An example of a mutation-selection equilibrium with local mutations is given in Figure 5.3.2. Comparing the mutation-selection equilibria in Figures 5.3.1 and 5.3.2, we see that with local mutations, it takes much higher mutation rates to get to the same levels of average offers and MAO's, and that with local mutations, that happens without the thick tails that are symptomatic of the fact that with global mutations, higher mutation rates imply more upward push from the bias.

5.3.3 Finite versus infinite populations, and multiplicity of equilibria

In Gale et al. (1995), a mutation-selection equilibrium is a population state, characterized by a combination of frequencies of different strategies, for which the dynamics indicate no net change due to the combination of mutation and selection. The population state depicted in Figure 5.3.1 is such an equilibrium. These equilibria are moreover stable, in the sense that at least nearby population states move towards it, and sometimes there is even global convergence. What the authors seem to have overlooked, however, is that for one and the same combination of parameters, there can be multiple mutation-selection equilibria. In the Supplementary Material, we show that this is the case for low mutation rates. If the mutation rate is low enough, then there are multiple mutation-selection equilibria, at which almost all proposers make the same offer, and with a range of options for what that offer is. For higher mutation rates, there is just one, globally attracting, mutation-selection equilibrium.

The finite population dynamics in Rand et al. (2013) on the other hand are noisy, and not deterministic. The population will therefore keep moving around, and a mutation-selection equilibrium becomes a distribution over population states that reflects that some population states are visited (much) more often than others. By letting simulations run for a long time, we can figure out properties of this distribution of states, such as the average offer or the average MAO. This distribution is always unique, also if mutation rates are low enough for the infinite population version from Gale et al. (1995) to have multiple equilibria. The noise in Rand et al. (2013) would then make the population visit these different equilibria, and states close to them, over time.

In the Supplementary Material, we compare Gale et al. (1995) and Rand et al. (2013) by choosing versions of the latter with increasing population sizes. We find that infinite population models are not a great approximation for finite population dynamics with small or even moderately sized populations.

5.3.4 Unequal mutation rates and Quantal Response

The setup in Gale et al. (1995) does allow for the possibility that mutation rates differ between proposers and responders. This is more reasonable for social learning than it would be for genetic transmission. With social learning, one could argue that if not much is at stake, there is less incentive to try to retain what you have learned. This kind of control over mutation rates make the agents more sophisticated than they are in the

default version of the model, in which mutation rates are the same for both roles in the game. It also makes agents more sophisticated than they are in the model of Rand et al. (2013), where mutation rates are the same for offers and for MAO's.

The motivation that Gale et al. (1995) give for the unequal mutation rates is strikingly similar to the motivation given for the definition of a Quantal Response Equilibrium (McKelvey and Palfrey, 1995). A Quantal Response Equilibrium does not describe a mutation-selection equilibrium, so conceptually these are two different things, but both do have in common that they are ways in which the asymmetry in costliness of mistakes shapes how noise ripples through the population. In Quantal Response Equilibria, this noise in caused by perception error, or otherwise failures to maximize, and in mutation-selection equilibria the mutations are the sourse of the noise. The next section discusses Quantal Response Equilibria for the ultimatum game, and one important thing that we will see there, is that there is a whole set of models, including Quantal Response Equilibria and mutation-selection equilibria, that predict types of distributions that are not in line with the empirical evidence. We will make this general observation once we have also looked at Quantal Response Equilibria.

5.4 Quantal Response Equilibria

In this section, we will try to see if one can explain human behaviour in the ultimatum game without assuming that people deviate from selfishness. Instead, we assume that people are in fact selfish, but that they are also limited in their understanding of what it is that they need to do in order to maximize their fitness, or something that translates to fitness, like money. This imperfect understanding is formalized by the game-theoretic notion of a Quantal Response Equilibrium (QRE, McKelvey and Palfrey, 1995), which can be described as a statistical version of a Nash equilibrium, where suboptimal behaviour is not ruled out, but only assumed to be unlikely. The reason why this can be interesting for the question how behaviour in the ultimatum game has evolved, is that Quantal Response Equilibria can emerge as the result of a variety of learning dynamics. Like the mutation-selection equilibria discussed in Section 5.2, the QRE for the ultimatum game is shaped by the asymmetry in how costly mistakes are. After working our way through the details of the different types of QRE's, we will see that the empirical evidence actually rejects that humans play a QRE in which they try to maximize how much money they earn.

Looking at QRE's and comparing them to the empirical evidence is first of all interesting, because it helps rule out that people are selfish after all, and that their behaviour in the

ultimatum game is just the result of not knowing exactly how to maximize their payoff. The deviations from selfishness we observe in experiments therefore cannot be explained away by people making mistakes. On top of this, the discrepancies between the empirical evidence on the one hand and the predictions of a combination of QRE and selfishness on the other also carry over to a larger class of evolutionary models at the ultimate level. Any dynamical model in which the reason why rejections are still present in equilibrium is that there is some source of noise that keeps introducing suboptimal behaviour, while selection keeps selecting against it, turns out to be inconsistent with the empirical evidence. That includes models that are in principle also open to an interpretation in which individuals evolve a preference for rejecting, as is the case for all mutation-selection equilibria discussed in the previous sections. This is an important, consequential observation, because it rules out a whole category of models that aim to explain the behaviour in the ultimatum game; all models that do not include a mechanism through which an actual fitness benefit is associated with rejecting proposals, do not explain behaviour in the ultimatum game. Before being able to articulate what the prediction is, and how that is refuted by the empirical evidence, it will be helpful to work through the technical details of the QRE, and first answer the more proximate question whether the behaviour can be reconciled with selfishness after all.

5.4.1 Quantal Response Equilibria and learning dynamics

The idea behind a Quantal Response Equilibrium (QRE, McKelvey and Palfrey, 1995, Goeree et al., 2016) is that players are imperfectly informed about the consequences of different behaviours. This concept does not assume anything about whether people are selfish or not; it can be combined with any type of preference, be it selfish, pro-social, anti-social, or inequity averse. Here, however, we will combine QRE with selfish preferences. The predictions therefore will be the result of a combination of selfish preferences and the Quantal Response model. Because of the fact that we do assume selfish preferences, words like "payoffs" will in fact coincide with money amounts when we use them below.

The defining property of a QRE is that strategies that result in high payoffs are played with higher probability than strategies that earn the agent lower payoffs. In the standard specification, the difference in probabilities is determined using a rationality parameter λ. The higher this rationality parameter, the larger the difference between these probabilities, and in the limit of $\lambda \to \infty$, only strategies that get the highest payoff are played. As a result, Quantal Response Equilibria become Nash equilibria in the limit of $\lambda \to \infty$. One reason why players might not be infinitely, or perfectly rational, is that increasing one's

λ might not be free. At some point, getting better at recognizing which actions lead to high payoffs might not be worth the additional costs of boosting this capacity.

There are different ways to model individual behaviour that would imply dynamics that justify using the notion of a QRE. One such way is if individuals observe the payoffs in the population with a little bit of noise. This implies that their idea of what actions would get them the highest payoffs is mostly accurate, but due to the noise, they may sometimes think that the best they can do is choose an action that does not in fact come with the highest expected payoff. This is more likely to happen for actions that are close to optimal, for which it takes only a tiny shock to make it seem as if this is the optimal choice. In the resulting "*perturbed best response dynamics*" (Alós-Ferrer and Netzer, 2010; Hofbauer and Sandholm, 2002; Sandholm, 2010), individuals play what they think is the optimal thing to do against the current state of the population. If the noise follows a certain distribution, then this perturbed best response dynamic becomes the *logit response dynamics*, which can bring populations playing a game to a (logit) QRE.

A second way would be to assume instead that players adjust their behaviour locally, where they take their current strategy as their point of departure, and tend to adjust it in the direction in which payoffs increase. This is then combined with some amount of noise in how they adjust. The balance between those two factors implies that if payoffs increase steeply in one direction, individuals are most likely to adjust their behaviour in the right direction, and, in expectation, by a lot, whereas if payoff differences are small, then noise makes it more likely that they misdirect their adjustment. The resulting dynamics also converge to a QRE (Anderson et al., 2004). We do not focus on either of these two dynamic justifications; we just want to point to the fact that there is a variety of dynamic justifications for the concept of a QRE.

There are two versions of the QRE that we can apply to the ultimatum game; the *Agent*-QRE and the *Normal form*-QRE. There is a rationale behind those names, but it is not important for our purposes, and below we will just describe what they are for the ultimatum game.

5.4.2 Agent-QRE

If an offer in which the responder would get x is made, the responder chooses between, on the one hand, accepting, and getting x by doing so, and, on the other, rejecting and getting 0. In an *Agent*-QRE, that means that the responder is more likely to accept than to reject – unless $x = 0$ – and that this gap grows as x increases. For $x = 0$, there is

no payoff difference, and therefore she accepts with 50% chance. In the logit specification of an *Agent*-QRE, the probability that the responder accepts depends on the offer x as follows:

$$P\left(accept|x\right) = \frac{e^{\lambda \cdot x}}{e^{\lambda \cdot x} + e^{\lambda \cdot 0}} = \frac{e^{\lambda x}}{e^{\lambda x} + 1}$$

The formula itself is not too important, but for the comparison with the empirical evidence, it is important to observe that this would indeed predict that all positive proposals are more likely to be accepted than they are to be rejected, while the proposal $x = 0$ would have to be accepted half of the time. This is also illustrated by the red lines in Figure 5.4.1, that plot how the acceptance rates would depend on the proposal x for different rationality parameters λ.

Which offer would maximize the earnings for the proposer depends on what responders do. More precisely, what the best offer is, depends on the way in which the probability with which the responder accepts, changes with the offer that is made. For rationality parameters λ between 0 and 2, the probability with which the responder accepts is so insensitive to the proposal, that proposers are best off just proposing nothing for the responder and everything for themselves.[2] This proposal will then be accepted with 50% probability. Increasing the offer does increase the probability with which the proposal is accepted a bit, but not enough to offset the reduction of the share of the pie when it is accepted. Therefore, for low λ's, the offer with the highest payoffs to the proposer, and therefore with the highest probability in the QRE, is 0 (see Figure 5.4.1A).

For λ's larger than 2, what the best response is first increases with λ. This can be seen in Figure 5.4.1B, where the peak of the blue graph has moved to the right. Later, for even higher λ's, changes in responder behaviour push the offer with the highest expected payoff back down again, which can be seen in Figure 5.4.1C and 5.4.1D, where the position of the peak moves back to the left, and gets ever closer to 0 as λ gets ever larger.

One can also see this from the formula for the density of offers made by the proposer in the logit specification for the *Agent*-QRE.

$$\frac{e^{\lambda(1-x)\frac{e^{\lambda x}}{e^{\lambda x}+1}}}{\int_0^1 e^{\lambda(1-y)\frac{e^{\lambda y}}{e^{\lambda y}+1}} dy}$$

[2]This can be found by taking the derivative of the expected earnings. These expected earnings are the amount the proposer gets if the offer is accepted (which is $1-x$) times the probability with which it is accepted (which is $\frac{e^{\lambda x}}{e^{\lambda x}+1}$). The derivative of $(1-x)\frac{e^{\lambda x}}{e^{\lambda x}+1}$ to x is negative for all $x \in (0,1)$ for $0 \leq \lambda \leq 2$, while for all $\lambda > 2$, there is one – and only one – x within the interval $(0,1)$ for which this derivative is 0.

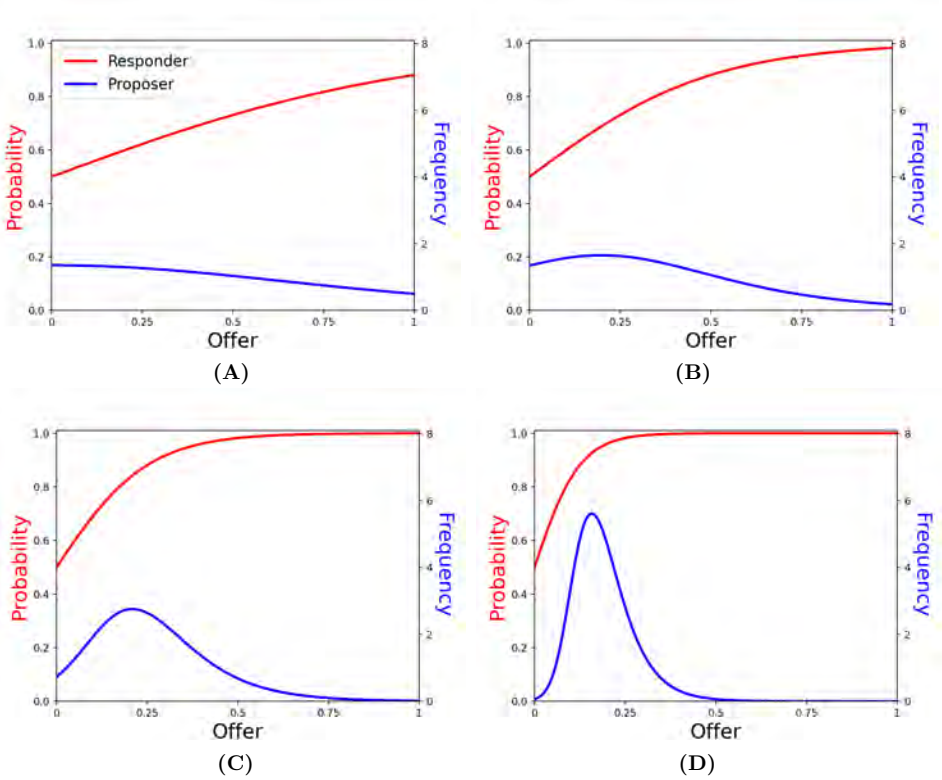


Figure 5.4.1: ***Agent*-QRE.** The red lines represent the probability with which the responder accepts an offer in which she gets x. The higher x, the larger the difference in payoff between accepting and not accepting, and therefore the higher the probability of acceptance. How strong the probability to accept responds to the payoff difference depends on the rationality parameter λ, which is $2, 4, 8$ and 16 in panels A, B, C and D, respectively. The blue lines represent the probability distribution over the offers made by the proposers in the QRE (just to be sure: this makes it is a different type of line than the red line is). The red line always starts at 0.5; the proposal in which the responder gets nothing is accepted with 50% chance.

The exponent in the numerator is λ times the expected payoff to the proposer of offering x. The density therefore peaks at the point where this expected payoff is maximized. High λ's moreover make for larger differences between the density for strategies with low expected payoffs and high expected payoffs. The peak therefore gets ever higher as λ increases, while the position of the peak, which is determined by the behaviour of responders, first moves to the right, and then back to the left.

5.4.3 Comparison empirics

For the comparison with the empirical evidence, we focus on responder behaviour, for which we pool data from different studies that use the direct-response method together (see Figure 5.4.2). The following studies are included: Andersen et al. (2011); Barmettler et al. (2012); Bornstein and Yaniv (1998); Cameron (1999); Carpenter et al. (2005a;b); Croson (1996); Forsythe et al. (1994); Lightner et al. (2017); Ruffle (1998); Slonim and Roth (1998). For each experiment, we extract the data for the standard ultimatum game, and discard the data for other treatments that vary certain aspects. Offers are calculated proportional to the total amount available in the ultimatum game in order to standardize the behavior across different experiments. Because it is not universally agreed upon whether stakes size matters, we also exclude the observations for the largest stakes in Andersen et al. (2011); Cameron (1999); Slonim and Roth (1998), while the Supplementary Material contains a version where we do include all stake sizes. The Supplementary Material also contains a version where we use data obtained with the strategy method to calculate rejection rates and compare them to the predictions of the *Agent*-QRE.

To test whether the predictions of the *Agent*-QRE fit the experimental evidence, we ran a logistic regression. With a logistic function, the probability that the offer is accepted is given by

$$P\left(accept|x\right) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

In the *Agent*-QRE, the acceptance probability of an offer in x is

$$P\left(accept|x\right) = \frac{e^{\lambda x}}{e^{\lambda x}+1} = \frac{1}{1+e^{-\lambda x}}$$

This means that this probability in the *Agent*-QRE is nested in the logit specification, because the only difference between the two specifications is the intercept term β_0. The intercept is moreover the critical term, especially at the offers of 0, since the probability of accepting an offer of 0 *Agent*-QRE is

$$P\left(accept|0\right) = \frac{1}{2}$$

whereas, for the logistic function including the intercept, it is

$$P\left(accept|0\right) = \frac{1}{1+e^{-\beta_0}}$$

Depending on whether β_0 is statistically significantly different from 0 or not, we can

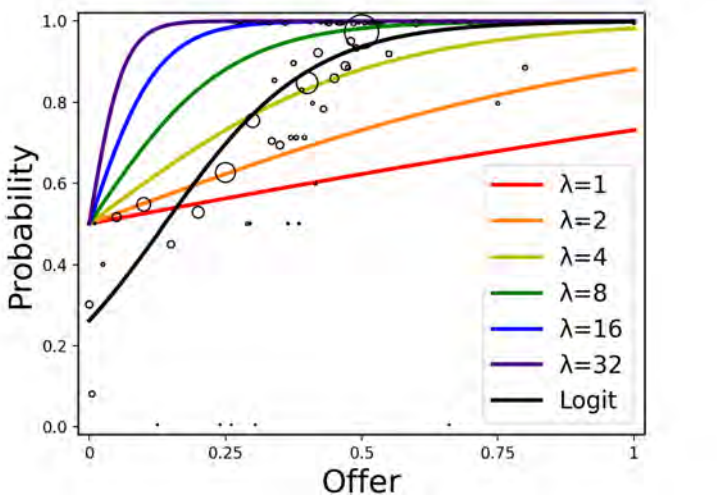


Figure 5.4.2: Acceptance/rejection rates in *agent*-QRE vs. empirical acceptance/rejection rates. The coloured lines are the acceptance rates in the *agent*-QRE for different λ's. The circles indicate acceptance rates for different proposals, pooling data from a number experiments together. Their size reflects the number of observations for that offer. The black line is the fitted acceptance rate as a function of the offer for a logit regression. Here, we use data obtained with the direct-response method, and we exclude some treatments with high stakes. In the Supplementary Material we include all stake sizes, and we compare the predictions of responder behaviour in the *Agent*-QRE with acceptance rates based on data obtained with the strategy method.

therefore directly say something about the probability of accepting an offer of 0 being different from $\frac{1}{2}$.

We fit logistic functions with and without the intercept to test between the two specifications, and our results show a highly statistically significant coefficient on the intercept (p-value < 0.001). This indicates that the logistic regression including the intercept provides a significantly better fit than the *Agent*-QRE (see the Supplementary Material for details). The estimated coefficient on β_0 moreover has a negative sign, resulting in an estimated acceptance probability of an offer of 0 that is below 50%;

$$P\left(accept|0\right) = \frac{1}{1+e^{-\beta_0}} < \frac{1}{2}$$

as $e^{-\beta_0} > 1$.

What this implies for the *Agent*-QRE is that the empirical acceptance rates are not consistent with the prediction of the QRE under the assumption that individuals are purely

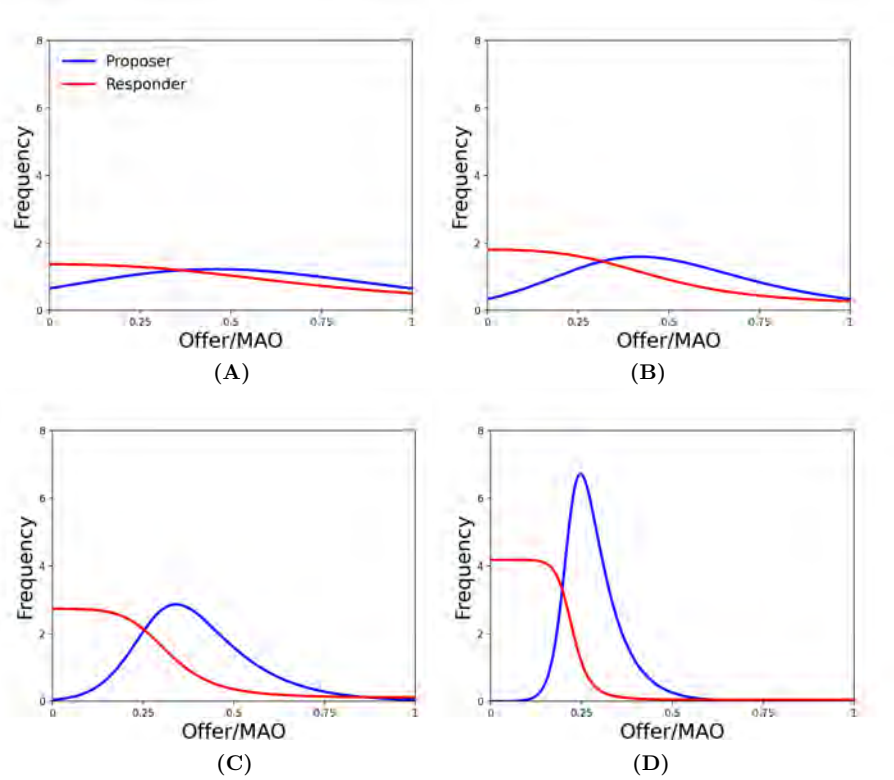


Figure 5.4.3: ***Normal form*-QRE.** The red lines represent probability distributions over MAO's for the responder, and the blue lines represent probability distributions over the offers made by the proposers, both in one and the same QRE. The rationality parameter, λ, is $2, 4, 8$ and 16 in panels A, B, C and D, respectively. The red graphs are all decreasing; in *Normal form*-QRE, lower MAO's occur with higher frequency than higher MAO's. The blue graphs have ever higher peaks, that start in the middle, and move ever more to the left.

focused on maximizing their monetary payoff. Offers of 0 are accepted in significantly less than 50% of the cases, and there is an interval of low offers, for which subjects reject more often than they accept. The fact that there is such an interval is inconsistent with the idea that monetary payoffs are the only determinant of rejecting behaviour. Instead, it is consistent with subjects balancing the money they would get from accepting the offer against something else, which is best described as the joy of rejecting an unfair offer, or an aversion to accepting it.

5.4.4 Normal form-QRE

In the *Normal form*-QRE, we assume that proposers choose a proposal between 0 and 1, and responders choose an MAO between 0 and 1. That means that instead of considering responder strategies for each proposal separately, we consider strategies for the whole spectrum of possible offers. Moreover, the strategies we consider all have a natural, simple form; they have a threshold, above which they accept all offers, and below which they reject all offers. Many models reduce the strategy set this way, including the models in Gale et al. (1995) and Rand et al. (2013), as well as our version of the latter, all of which we discussed in Sections 5.2 and 5.3. The experimental evidence moreover suggests that this is not an unreasonable simplification; many people reject low offers and accept high offers, and switch from one to the other at some point in between.

In this setup, a QRE is a combination of distributions, one for the proposer and one for the responder. What the payoffs to different strategies for the proposer are, depends on the distribution of MAO's of responders, and vice versa. In equilibrium, strategies with higher expected payoffs are chosen with higher probabilities, and strategies with lower payoffs are chosen with lower probabilities, and this is true both for proposers and for responders.[3]

The *Agent*-QRE and the *Normal form*-QRE are not the same. In the *Normal form*-QRE, the proposer strategy that maximizes expected payoff starts at 0.5 for $\lambda = 0$, and then only moves down. Therefore, if we look at Figure 5.4.3, we see that the peak starts in the middle, and always moves to the left (besides also becoming ever higher). This is different from how the distribution of proposer strategies changes with the increase of λ in the *Agent*-QRE, where the position of the peak starts at 0, first moves to the right, and then back to the left[4] (while the red lines are just not comparable, because they represent

[3]Just for completeness: if $f(x)$ is the distribution of proposals and $g(x)$ is the distribution of MAO's, then the following needs to be true for the combination of them to be a *Normal form*-QRE for the ultimatum game:

$$f(x) = \frac{e^{\lambda \int_0^x (1-x)g(y)dy}}{\int_0^1 e^{\lambda \int_0^z (1-z)g(y)dy} dz}$$

$$g(y) = \frac{e^{\lambda \int_y^1 xf(x)dx}}{\int_0^1 e^{\lambda \int_z^1 xf(x)dx} dz}$$

[4]In the *Agent*-QRE, what responders do, changes with λ, but only directly, because a higher λ gives more weight to strategies with higher payoffs. Which responder strategies would result in what payoffs is not changing with λ, because in the *Agent*-QRE, these are calculated for any given proposal, which, if your partner just made it, is happening with probability 1. In the *Normal form*-QRE, the expected payoffs that different responder strategies generate do depend on what proposers do. The distribution of

responder strategies in different ways).

These technicalities are not unimportant, but for the comparison with what subjects in labs do, what matters most is that the frequency with which players choose different MAO's decreases with the MAO; in the *Normal form*-QRE, an MAO of 0 is chosen the most, an MAO of 1 is chosen the least, and in between, if $0 < x < y < 1$, then x is chosen more often than y.

5.4.5 Comparison empirics

For the comparison with the empirical evidence, we focus on responder behaviour, for which we pool data from different studies that use the strategy method together (see Figure 5.4.4). The following studies are included: Bader et al. (2021); Bahry and Wilson (2006); Benndorf et al. (2017); Chew et al. (2013); Demiral and Mollerstrom (2020); Inaba et al. (2018); Keuschnigg et al. (2016); Peysakhovich et al. (2014). For each experiment, we extract the data for the standard ultimatum game, and discard the data for other treatments that vary certain aspects.

The majority of studies that use the strategy method restrict the subjects to strategies that can be characterized with an MAO. They ask their participants to submit a number, and if the offer they get is less than that number, it is rejected, and if it is higher than or equal to it, it is accepted. There are however a few exceptions; Bahry and Wilson (2006) and Bader et al. (2021); Keuschnigg et al. (2016) ask participants to submit their accept/reject decisions for each possible offer separately. Participants in these studies therefore have the flexibility to switch between accepting and rejecting more than once, as opposed to the single switched point imposed by the method of submitting an MAO. For these studies we include participants who never switched, who switched only once (who start with rejecting and switch to accepting at a certain offer level), and those who switched twice (once from rejecting to accepting in the first half of the strategy space for offers, and another time from accepting to rejecting in the second half of the strategy space). We included this last group of subjects as well, as it seems that also rejecting hyper-fair offers is not a mistake, but aligns with an existing, consistent preference. In this case, we take the first switching point as their MAO. We do exclude other participants who do not fall into one of these categories. If the participant accepted all offers, we take their MAO to be 0. MAO's are calculated proportional to the total amount available in the ultimatum game in order to standardize the behavior across different experiments (see the Supplementary Material for

what responders do in the *Normal form*-QRE therefore depends on λ in an additional way, because λ also has an effect on what proposers do.

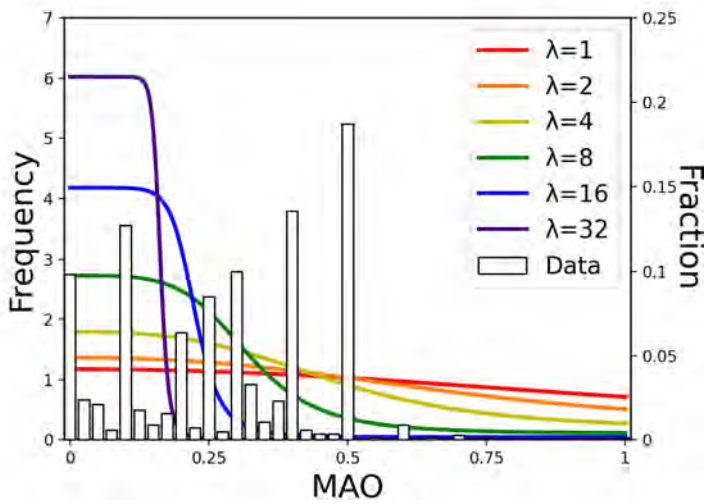


Figure 5.4.4: MAO's in *Normal form*-QRE vs. empirical MAO's. The coloured lines are the MAO's for different λ's. The bars indicate the frequencies of different MAO's in experiments. Because subjects gravitate towards round numbers, and because the increments subjects can choose also differ between experiments, we group the MAO's together as follows; the first bar is the frequency of MAO's of exactly 0, the second bar is the frequency of MAO's strictly between 0 and 0.05, the third bar is the frequency of MAO's of exactly 0.05, the fourth bar is the frequency of MAO's strictly between 0.05 and 0.1, and so on. Here, we use data obtained with the strategy method.

more details).

Figure 5.4.4 shows that the distribution of MAO's in experiments does not follow the pattern predicted by the *Normal form*-QRE. It is clear that the frequency of MAO's is not a decreasing function of the MAO, and as a simple indication of this, we can consider all MAO's below 0.25 on the one hand, and all MAO's above 0.25 up to, and including 0.5. The first interval, $[0, 0.25)$, contains fewer observations than the second one, $(0.25, 0.5]$, which is at odds with the distribution being a decreasing function.

5.4.6 Evolutionary dynamics in general

The evolutionary game theory models in the literature fit the setup of the *Normal form*-QRE perfectly. In Gale et al. (1995), Rand et al. (2013), and in our de-biased versions of both, there is a population of proposers that are characterized by their offers, and a population of responders that are characterized by their MAO's. Also there are similarities in the predicted distributions of offers and MAO's. We will therefore focus on how the mismatch for the *Normal form*-QRE carries over to evolutionary explanations at the

ultimate level.

A good first observation is that the mutation-selection equilibria in Gale et al. (1995), Rand et al. (2013), as well as in our de-biased version, all have the property that the equilibrium distribution of MAO's goes from frequent to infrequent as the MAO's go from low to high. In other words, the density is the highest at 0, and then it decreases, until it is the lowest at an MAO of 1. That is a straightforward consequence of the fact that rejecting proposals is bad for fitness, and therefore having a lower MAO is always better than having a higher MAO.

The simple version of the original evolutionary question regarding human behaviour in the ultimatum game is: if rejecting is always bad for fitness, why do we observe rejections at all? At first sight, one might think that the mutation-selection models of Gale et al. (1995) and Rand et al. (2013) offer an escape from the iron logic that rejecting can only be bad for fitness, and should be selected against. Depending on parameter values, the average MAO in mutation-selection equilibrium can after all be sizable, and even if we de-bias the model, as we did in Section 5.2, the average MAO in equilibrium can still be pushed up to non-negligible amounts by choosing high mutation rates and, in Rand et al. (2013), low intensities of selection. In mutation-selection equilibria, or models with noise in general, we should however realize that the presence of across-the-board selection against rejecting does not imply that the average MAO in the population should be 0. The only thing that it implies, is that lower MAO's are favoured by selection over higher MAO's, and therefore higher MAO's should be observed less often than lower ones. The fact that the data do not align with that prediction, implies that no model with mutation-selection equilibria, in which selection works against rejections, can explain the rejecting of positive offers in humans. That remains true for all models in which rejecting is only bad for fitness. The observation that 0 is not the most common MAO in humans (far from it) therefore rejects all models that do not open up channels through which rejecting proposals can also bring fitness benefits.

5.4.7 Implications, great and small

The first implication of the comparisons of QRE's and observed behaviour in the lab is totally intuitive and unsurprising. People really deviate from selfishness, and what we observe in the lab is not some mirage caused by noise rippling through a population of selfish individuals asymmetrically.

The way this carries over to models that aim at giving ultimate explanations for human

behaviour in the ultimatum game is less straightforward, and probably a bit more surprising, but therefore not less logically sound. Models in which all that happens is that some source of noise is added to the dynamics, without introducing a selective pressure that actually *favours* rejecting behaviour, are also at odds with the empirical evidence. These models do not predict that everyone in the population should have an MAO of 0, but they do predict that 0 should be the most common MAO (or, more generally, they predict that the higher the MAO, the less frequently it should be observed). That is clearly at odds with the empirical evidence.

5.5 Commitment

Nowak et al. (2000) propose a model for the evolution of behaviour in the ultimatum game in which the mechanism why rejections evolve is commitment (see also Frank, 1988, and Akdeniz and van Veelen, 2021). The rejecting itself is still bad for fitness, but their model opens a door through which being committed to rejections can be good for fitness. In their model, much the same as in other models, responders are characterized by a minimal acceptable offer (MAO), which is a threshold below which they reject proposals. Unlike other models, Nowak et al. (2000) allow proposers to sometimes observe the behaviour in past interactions of individual responders, and if they see that the responder accepted a proposal below what they would offer without observing, they lower their offer to what they know this responder apparently accepts. As a result, having a lower than average MAO, while leading to fewer costly rejections if unobserved, now has the disadvantage of also leading to worse offers, in case a player is observed to accept them.

In this section, we present a slightly upgraded version of the simulation model in Nowak et al. (2000). This illustrates a few core properties of this mechanism. The first is that, obviously, having a high MAO should sometimes lead to getting a higher offer for the mechanism of commitment to work. An individual's MAO must therefore be recognized from time to time, and proposers should sometimes do something with that information. On the other hand, the MAO does not always have to be recognized, and it does not have to be recognized perfectly, in order for commitment to work. A modest individual effect, by which those with higher MAO's on average get somewhat better offers, can still move the whole population to a state in which proposers serve their own interests by making sizable offers, even in cases where they do not have any information about the particular responder they are matched with.

Some of the differences between our version and the original have to do with restrictions

on the admissible strategies that Nowak et al. (2000) impose. While these restrictions are not necessarily unreasonable, we felt that it is better to see if and when strategies evolve that satisfy them, rather than imposing them exogenously. Our version of Nowak et al. (2000) is also a generalization of the version of Rand et al. (2013) that we presented in Section 5.2. This allows us to explore the relative effects of asymmetry and commitment, and it allows us to illustrate the power of the combination of them. Also, it is aesthetically nice to have a unified model.

5.5.1 The simulation model in Nowak et al. (2000)

In the simulation model in Nowak et al. (2000), each individual is defined by a default offer p and a minimal acceptable offer q. In any given interaction, the proposer will offer whatever is smaller; her own p value, or the lowest amount that she knows was accepted by the responder during previous interactions. In addition, there is a small probability that the proposer will offer her value p minus some random number between 0 and 0.1. This makes sure that even if everyone in the population has the same p for a number of subsequent generations, there still will be observations for a range of lower proposals. Strategies are restricted to those with values for p and q that add up to a number that does not exceed 1; $p + q \leq 1$.

5.5.2 Our version

Our version first of all abstracts away from the way in which players find out about the MAO's of their partners. In Nowak et al. (2000), players sometimes observe past behaviour, from which they can make inferences about the MAO. We just assume that there is a fixed probability with which individuals know what the MAO of their partner is, and with the remaining probability they do not. The pathway could be reputation, but it can also be that people have other ways of recognizing individual differences in attitudes before playing.

Because Nowak et al. (2000) use reputation as a way for proposers to get information on the MAO of their partners, the mechanism behind the evolution of rejections here is sometimes classified as reputation (see for instance Debove et al., 2016, or Henrich et al., 2010). This is a defensible and understandable choice. What we would like to emphasize, though, is that in a population that is playing the ultimatum game, there are interesting incentives concerning communication. Proposers would like to be informed about the MAO of the responder they are matched with, so that they can maximize how much they can keep without getting their proposal rejected. Responders with high MAO's would

like the proposer they are matched to know what their MAO is. Responders with below average MAO's however would like the proposer not to find out what their true MAO is. This partial alignment of the incentives for successful communication suggests that also without reputation, one could imagine some exchange of information to be established. Experimental evidence suggests that also in the absence of reputation, humans do indeed pick up on cues that help them predict rejecting behaviour with some success (van Leeuwen et al., 2018).

Therefore, what we want to stress is that one can also see commitment as the underlying mechanism for the evolution of rejecting behaviour; rejecting itself is still bad for fitness, but being committed to rejections is good, because it results in better offers (Akdeniz and van Veelen, 2021; Frank, 1988). This does require that others are able to identify, to some degree, who is committed. Reputation is one of the pathways to do that, but since it is not the only one, we abstract away from how it is that proposers tell different responders apart, and just include a parameter that represents the degree to which they can.

The second way in which our version is different, is that our proposers are characterized by two variables instead of one. One variable is their default offer, which they make if they are uninformed. The other is the maximum MAO they are willing to match as a proposer if they are informed about the MAO of the responder. Rather than assuming that proposers always match the MAO they observe, in our version, what they do with this information also evolves.

Also, in Nowak et al. (2000), proposers never propose more than their default proposal p, and only propose less than p if they know that will be accepted too. We allow for the possibility that proposers evolve to match the MAO of an opponent, also if it lies above their default offer p. The reason is that we think it is important to model the advantage it brings to be committed to an MAO that is above average, at least as much as it is important to allow being more accommodating than average to be exploited and selected against.

A fourth way in which our simulations are different, is that we do not assume that individuals sometimes lower their offer with a random amount, as in Nowak et al. (2000). This is not needed, because we abstracted away from the mechanism by which proposers are sometimes informed about the MAO of the responder they are matched with. We do have mutations on all traits, including the offer without observing, the same way as in our version with local mutations of Rand et al. (2013).

Finally, we do not impose the restriction that the default offer and the MAO should add up

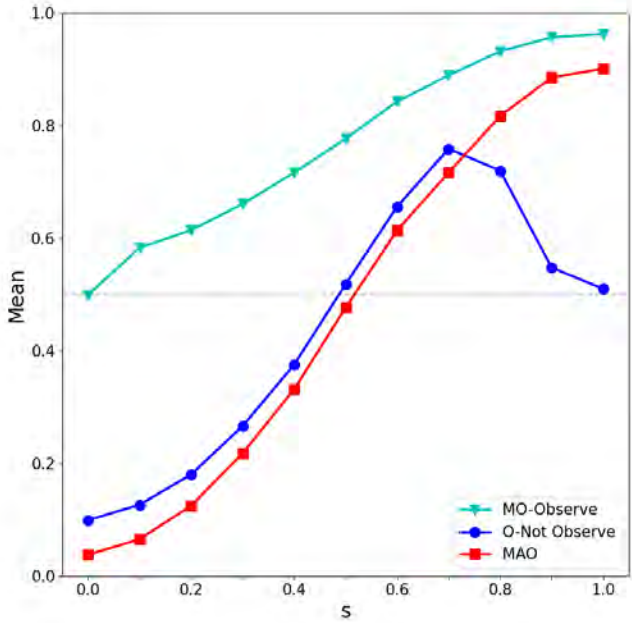
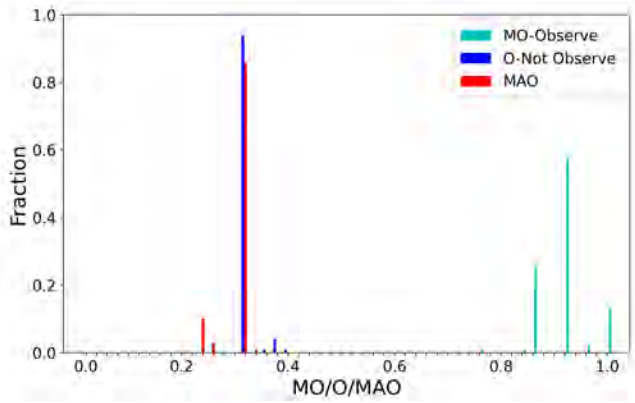


Figure 5.5.1: Local mutations and partial observability I. Individuals have three traits; the maximum offer (MO) they will make as a proposer to match the MAO of the responder, if observed; the offer (O) they make if they do not; and their MAO as a responder. The averages of these traits change with the probability with which proposers observe the MAO of the responder, which ranges from $s = 0$ (no observability) to $s = 1$ (full observability). Other parameter values are fixed at $u = 0.01$ and $w = 1$.

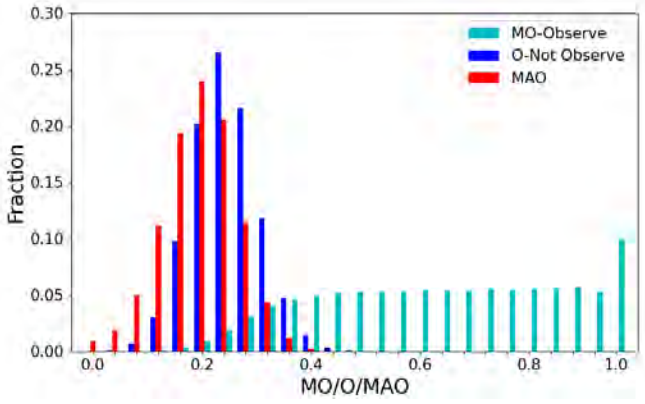
to a number that does not exceed 1. Our individuals can be endowed with any combination of those, as long as both are between 0 and 1. We do think that there are reasons why the offer and the MAO have not evolved to values larger than 0.5, but we prefer not to impose restrictions on the set of admissible strategies in order to rule out values above 0.5. Also, for understanding the working of the model, it will actually be instructive to allow values for p and q that both are larger than 0.5, even if we see reasons why these would not evolve (see also section 3.2.1 in Debove et al., 2016, where they point to the consequences of this restriction).

5.5.3 Results

Without observability, the model is the same as our version of Rand et al. (2013), but with local mutations. At $s = 0$, at the left end of Figure 5.5.1, the MAO and the offer



(A)



(B)

Figure 5.5.2: Local mutations and partial observability II. Panel A is a snapshot of the population, indicating the distribution of maximum offers (MO) they will make as a proposer to match the MAO of the responder; the offers (O) they make if they do not observe the responder's MAO; and their MAO as a responder. Panel B averages these across time, and thereby represents average distributions. Parameter values are fixed at $u = 0.01$, $w = 1$, and a probability of observing responder's MAO of $s = 0.3$. The spike at 1.0 and the dip just before 0.1 are part of a dampening wave pattern caused by the mutations being a little biased at the edges, where mutations beyond 1 are not possible.

without observing therefore are the same as they are for local mutations in Figure 5.2.1B at $w = 1$. The MO (the maximum offer they will make as a proposer to match the MAO of the responder, if observed) is a trait without fitness consequences if MAO's are never

observed. That implies that at $s = 0$ it will drift neutrally within the interval $[0, 1]$, and will be 0.5 on average.

At $s = 1$ it is the offer without observing that becomes irrelevant, and will be 0.5 on average. Full observability moreover turns the tables on proposers and responders, because now the MAO of the responder is a given to proposers, who serve their own interest best by matching all positive MAO's. The situation at $s = 1$ therefore is the mirror image of the situation at $s = 0$, where the role of the offer without observing at $s = 0$ is played by the MAO at $s = 1$, and the role of the MAO at $s = 0$ is played by the MO at $s = 1$.

In between, we see that increasing the observability shifts the balance between the costs of being committed to rejecting low offers, which mainly occur when not observed, and the benefits, which only occur when observed. This pushes the average MAO up. The average offer without observing follows suit, because even though rejecting low offers evolves for when the MAO is observed, they are still a fact of life when not observed.

This is still a mutation-selection equilibrium, in which local mutations would flatten all distributions in the absence of selection, while selection can lift certain frequencies within the distributions up. When observability increases, selection for positive offers without observing becomes weaker. Up to $s = 0.7$, selection on the offers without observing the MAO keeps it above the average MAO within the population at all times. From $s = 0.8$ onward, with ever less selection countering the flattening of the distribution of offers without observing, the flattening sometimes wins, and sends the average offer without observing roaming below the average MAO, while selection at other times manages to temporarily stabilize the average offer without observing above the average MAO.

5.5.4 Model limitations

In this model, as in Nowak et al. (2000), we treat observability as an exogenous parameter. This is useful for illustrating how commitment works, but because of the partially aligned, partially misaligned interests between proposers and responders with respect to communicating the MAO's of responders, the observability is more likely to be endogenous, and subject to evolution itself.

One can also stack levels of observability on top of each other. On top of the probability with which proposers see the MAO of the responder, one could also introduce the probability with which the responder observes the MO of the proposer, and introduce the minimal MO she is willing to adjust her MAO to. While setting the first observability to 1 turns the tables to the benefit of responders, setting this second observability to 1 would

turn the tables back to the benefit of proposers. We do not think this would be a particularly useful modelling exercise. We do however believe that there is a good reason why offers and MAO's would not exceed 0.5. The very nature of the ultimatum game makes it easier for proposers to commit than it is for responders. That implies that in a commitment tug-of-war between proposers and responders, we would expect that proposers will structurally find themselves on the shorter end of the stick.

5.6 Summary, discussion, reflection

In this paper, we have looked at a few prominent models from the evolutionary game theory literature that aim at explaining human behaviour in the ultimatum game. Gale et al. (1995) and Rand et al. (2013) both describe properties of mutation-selection equilibria, without a mechanism by which rejecting unfair proposals would get a selective advantage. We find that in both of them, the main driver of the results is bias in the mutations. This is not a good basis for an explanation. We made versions of both with local instead of global mutations. This minimizes the bias, and makes sure that the results are driven, not primarily by bias, but by the asymmetry in how costly mistakes are for proposers and responders. The reduction in bias makes average offers and average MAO's go down much more than the effect of the asymmetry makes them go back up again. The net change from global to local mutations therefore comes with significantly lower average offers and average MAO's. While the versions with local mutations capture the effect of the asymmetry in costliness of mistakes much better than the originals with global mutations, they still assume that rejecting is always bad, which is reflected by the fact that the mutation-selection equilibrium is characterized by higher MAO's always occurring less frequently than lower MAO's.

We also looked at Quantal Response Equilibria under the assumption that individuals are selfish. This noisy version of the Nash equilibrium, where people make mistakes in maximizing their payoff, and are more likely to make smaller mistakes than larger ones, comes in two versions. The first is characterized by probabilities of accepting that start at 50% for the offer of 0, and increases from there onward. The second is characterized by an equilibrium distribution in which an MAO of 0 has the highest density, an MAO of 1 the lowest, and the density always decreases as the MAO goes up everywhere in between. Both predictions are not confirmed by the experimental data, where we use data from existing experiments that use the direct response method to go with the first, and data from experiments that use the strategy method to go with the second prediction. The empirical evidence therefore rejects that the behaviour in the lab is the result of selfish

people being imperfect at maximizing the amount of money they earn.

The second mismatch, between the observed shape of the distribution of MAO's and what Quantal Response would predict if people were selfish, also carries over to models at the ultimate level, as long as these models maintain the assumption that rejections can only be bad for fitness. This includes the mutation-selection equilibria in Gale et al. (1995), in Rand et al. (2013), and those in our de-biased versions of them.

The last model we looked at is Nowak et al. (2000). The mechanism at work there is commitment. If proposers have a way of finding out what the MAO of responders is, then having a higher MAO helps getting higher offers. The act of rejecting itself therefore is still bad for fitness, but being the rejecting type is good for fitness. We made an upgraded version of their model, that avoids making assumptions that rule out certain strategies a priori. Our version of Rand et al. (2013) also becomes a special case of our version of Nowak et al. (2000).

5.6.1 Inequity aversion and the mismatch hypothesis

The three papers we have focused on here are the best known evolutionary game theory models from the literature on this topic. They are however not the only ways in which one could try to explain human behaviour in the ultimatum game. One other possibility would be to assume that humans have inequity averse preferences (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) that have evolved for playing other games, and that we inadvertendly bring to the ultimatum game too. That would imply that the rejecting behaviour in the ultimatum game is maladaptive.

Akdeniz and van Veelen (2021) show that there are two weak links in this argument. The first is that in most models for the evolution of deviations from selfishness, these "other games" are prisoner's dilemmas, and in those models, altruism evolves, or maybe spite, but not inequity aversion. The second is that this would imply that rejection rates should not depend on who makes the proposal – the person that the money is to be split with, or a computer – and that it should not depend on what the menu of possible proposals is that the proposer can choose from. Blount (1995) find that the first is not true, Falk et al. (2003) find that the second is violated (while an explanation with commitment would predict that whether or not we reject should depend on whether or not the other player is in fact responsible for an unfair proposal).

5.6.2 Other explanations

Akdeniz and van Veelen (2021) we also discuss why the fairness norm in the ultimatum game is not really group-beneficial – thereby ruling out a group selection argument – and they argue that repeating an ultimatum game would also not help explaining the behaviour that we find. Because those arguments are made elsewhere, we do not repeat them here. We also do not aim at making an exhaustive review of all existing models; there is already an excellent overview of the literature (Debove et al., 2016), and to complement that, we limited ourselves to discussing a few prominent ones in more depth.

5.6.3 Deviations from selfishness in general

There are many ways in which humans deviate from simply selfish money-maximizing behaviour. Studying deviations from selfishness in the ultimatum game therefore is part of a broader endeavour, that also tries to explain deviations from selfishness in other games. This gives another argument against asymmetry-based models – while mutation bias-based explanations are hardly ever a good option. For these other games, it is much more straightforward to see that asymmetry-based explanations could never work. Behaviour in the trust game can not be explained with models based on asymmetry in the costliness of mistakes; trustees not sending back money and trustors not trusting is very stable, also with mutations or noise. Also behaviour in the prisoners' dilemma or in the public good game, with or without punishment, cannot be explained on the basis of asymmetry. Here the simple reason is that these games are just not asymmetric. As an explanation for the human sense of fairness in general, therefore, asymmetry-based explanations would need to be combined with other mechanisms for deviations from selfishness in other games. That makes for instance commitment as a mechanism more parsimonious, because that gives an explanation of deviations from simple selfishness in a much wider variety of games (Akdeniz and van Veelen, 2021; Frank, 1988). That is not to say that asymmetries are irrelevant (they are not) but it makes it even more unlikely that asymmetry is the core driver of rejections in the ultimatum game.

Appendix

In the Appendix, we discuss a few things in more detail. Section 5A.1 compares Rand et al. (2013) with our version with local mutations. Section 5A.2 describes why choosing arbitrarily weak selection is problematic. Section 5A.3 describes the model in Gale et al. (1995) as well as our version with local mutations, and illustrates the possibility of multiple equilibria for the former. Section 5A.4 illustrates the link and the differences between the models in Gale et al. (1995) and Rand et al. (2013). Section 5A.5 discusses some details of Quantal Response Equilibria, and how their predictions are compared to experimental data.

5A.1 Finite population models

5A.1.1 The model in Rand et al. (2013)

The model in Rand et al. (2013) has a finite population, in which 100 individuals play ultimatum games in both roles. Every individual has a strategy that specifies the offer they make in the role of proposer, and their MAO in the role of responder. These offers and thresholds range from 0 to 1. Each generation, every individual plays the ultimatum game with every other individual, once as a proposer and once as a responder. The resulting payoff is the average of the payoffs over all 99 pairings.

The population is updated according to a Moran process. One agent is picked at random to die, and individual $i \in \{1, ..., 100\}$ is picked with probability proportional to $exp\,(w\pi_i)$ to reproduce, where w is the intensity of selection, and π_i is the average payoff of individual i. Mutations happen at rate u at reproduction; with probability $1-u$, the new individual inherits the strategy from the reproducing individual, and with probability u, the new individual carries a randomly selected strategy. If a mutation happens, both the new offer and the new MAO are drawn from a uniform distribution on $[0, 1]$.

5A.1.2 Our versions

There is one general, inconsequential difference between their simulations and ours, and that is that we use a Wright-Fisher process instead of a Moran process. The Wright-Fisher process is computationally more efficient, but other than that, it perfectly reproduces the findings in Rand et al. (2013) for global mutations. The more important difference is that in our version, mutations are local. We consider two local alternatives for the mutation process.

Local, co-occurring mutations In the first one, mutations on both dimensions (offer and MAO) are co-occurring, as they are in Rand et al. (2013). That means that if a mutation happens, then both a new offer and a new MAO are drawn. The only difference with Rand et al. (2013) is that they are drawn from a local distribution, instead of a global one. If the old offer is p, then the new offer is $p + \Delta p$, where Δp is drawn from a uniform distribution on $[-0.1, 0.1]$. There are two exceptions. If $p + \Delta p < 0$, the new offer is 0. Similarly, if $p + \Delta p > 1$, the new offer is 1. The same procedure applies to the MAO.

Local, independent mutations In the second version, mutations in the offer or the MAO happen independently. At any reproduction event, the offer mutates with probability u, and so does the MAO. That means that with probability u^2 mutations of the offer and of the MAO co-occur, with probability $2u\,(1-u)$ only one of them mutates, and with probability $(1-u)^2$ neither of the two mutates. Mutations still happen locally, as described above.

The differences between these two versions are relatively small (see Section 5A.1.4). Because the second version is more elegant, this is the one that we use here and in the main text.

5A.1.3 Global versus local mutation

The first question that Rand et al. (2013) answer for their model, and that we answer for ours, is: which combinations of the intensity of selection and the mutation rate put the average offer and the average MAO in the range of the averages in empirical findings. There are two ways to rephrase that question, or to visualize the answer. The first is: for a given mutation rate, how low would the intensity of selection have to be in order to get the offers and MAO's up to levels found in experiments. The second is: for a given intensity of selection, how high would the mutation rate have to be in order to get offers and MAO's up to the levels found in experiments.

For the figures in the main text, we took the first approach: we fixed a mutation rate, and considered a variety of intensities of selection. This way these figures indicate how far we would have to reduce the intensity of selection in order to push average offers and average MAO's up to values in the range found in experiments. Here, we complement that with the second approach. In Figure 5A.1, below, the intensities of selection are fixed, and we look at the average offers and MAO's for a variety of mutation rates. For reasons explained in the main text, and in Section 5A.2 of the supplementary material, we would like to stay away from the limit of weak selection, and therefore we choose the three larger

intensities of selection that feature in Figure 1 in the main text.

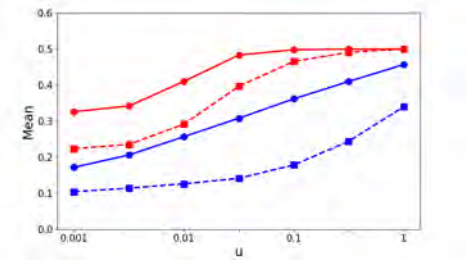
In this representation, with fixed intensities of selection and increasing mutation rates, the simulations suggest the same as Figure 1 in the main text does, and maybe even more strongly so. If we compare the version from Rand et al. (2013), with global, biased mutations, to our version with local, and therefore much less biased mutations, then the average offers and the average MAO's are significantly lower in the latter. We also see that with local mutations, the average offer and the average MAO do not always reach the averages from experiments, even at the maximum mutation rate, where everybody always mutates.

With global mutations, both average offers and average MAO's inevitably get to 0.5 as mutation rates increase. The reason is that when mutations are global, then at $u = 1$, when both the offer and the MAO mutate at every reproduction event, it becomes irrelevant who is reproducing. The parents therefore stop passing on any (genetic) information; every new individual is a mutant, and all mutants are drawn from the same distribution, regardless of what the parents are. Therefore, at $u = 1$, on both dimensions, the population at any point in time just becomes a collection of independent random draws from $[0, 1]$.

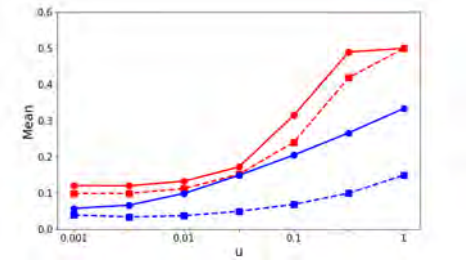
With local mutations, average offers and average MAO's do not necessarily get to 0.5 as the mutation rate increases. In this case, parents still pass on genetic information, because even if everyone mutates, these mutations are drawn from a distribution that is centered around the trait value of the parent. The trait value of the parent matters for payoffs, and therefore for the expected number of offspring, and that makes it possible for the average offer and the average MAO to stay below 0.5, even if the mutation rate is 1.

This illustrates that one can also push the average offers and the average MAO's up by increasing the mutation rate. It also illustrates that there are limits to how far one can push them up, and for moderate to high intensities of selection, even a mutation rates of 1 does not push them up high enough.

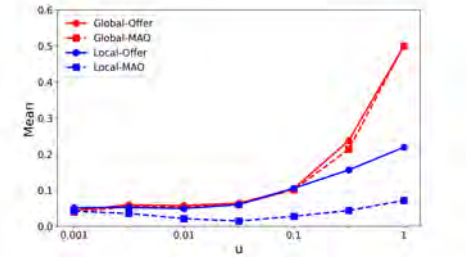
It is possible to model the genetics underlying the behaviour differently. If we for instance assume that there is a number of different loci that all can increase or decrease the offer or the MAO by a little bit, and we assume sexual reproduction, then it is possible that also with lower mutation rates, one can sustain similar levels of variation in the population, and thereby push the offers and MAO's up by the same amount. It should however be noted that this would naturally make mutations local, and therefore with global mutations, where the effect of the bias scales up with the mutation rate, there is less space to think of reasons why high mutation rates make sense.



(A)



(B)

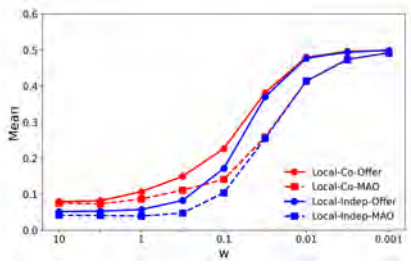


(C)

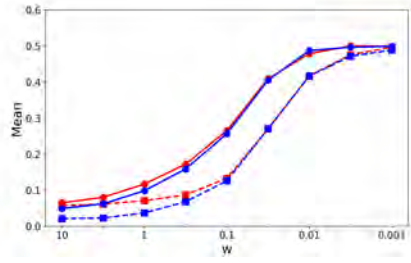
Figure 5A.1: Global versus local mutations. In red the average offers and MAO's for the model in Rand et al. (2013), which has global, co-occurring mutations. In blue the same, but for local, independent mutations. The intensity of selection is 0.1 in panel A, 1 in panel B, and 10 in panel C. For $w = 0.1$ one can still get to the averages observed in experiments, but with local mutations it requires very high mutation rates. For $w = 1$ and $w = 10$, even a mutation rate of $u = 1$ is not high enough for local mutations.

5A.1.4 Co-occurring versus independent mutations

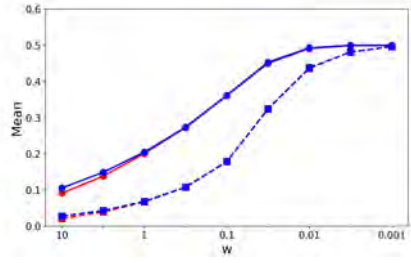
Figure 5A.2 shows that only for a combination of strong selection and a low mutation rate is there a modest difference between local, co-occurring mutations and local, independent mutations. For higher mutation rates and/or weaker selection, this difference disappears. Because there is no real reason why mutations would co-occur, we chose to use the version with independent mutations. The comparison here is done to make sure that the lion



(A)



(B)



(C)

Figure 5A.2: Co-occurring versus independent mutations. In red the average offers and MAO's with local, co-occurring mutations, and in blue the same, but for local, independent mutations. The mutation rate is 0.001 in panel A, 0.01 in panel B, and 0.1 in panel C.

share of the difference between simulations with the model from Rand et al. (2013) and simulations with ours is due to replacing global, and therefore biased mutations with local, and therefore much less biased ones, and not to switching from co-occurring to independent mutations.

5A.2 Weak selection

In the main text, we have seen that for a fixed mutation rate, we can always push average offers and the average MAO's up, from 0, to any point between 0 and 0.5, by reducing the intensity of selection. We have also seen that there are limitations to how much the intensity of selection can be reduced, and still produce a meaningful prediction. Here we will make that argument a bit more precisely and elaborately.

5A.2.1 Probabilistic symmetry

When the intensity of selection is 0, the dynamics in the model by Rand et al. (2013) become symmetric, in the sense that any transition from one population state to the other is equally likely as its mirror image. More precisely, if p_i denotes the offer of player i in the role of proposer, and q_i is the MAO of player i in the role of responder, then a population state is characterized by vectors $\mathbf{p} = [p_1, ..., p_N]$ and $\mathbf{q} = [q_1, ..., q_N]$, where N is the population size. Symmetry means that a transition from population state $(\mathbf{p}, \mathbf{q})$ to population state $(\mathbf{p}', \mathbf{q}')$ is equally likely as its mirror image, going from population state $(\mathbf{1} - \mathbf{p}, \mathbf{1} - \mathbf{q})$ to population state $(\mathbf{1} - \mathbf{p}', \mathbf{1} - \mathbf{q}')$, where $\mathbf{1}$ is a vector of 1's. This symmetry implies that if we average the population states over time, we will find a symmetric distribution. The average offer over this distribution therefore will be 0.5, and the average MAO will also be 0.5, and both of these are a consequence of the fact that 0.5 is halfway the strategy set that the population moves around in – with probabilistic symmetry.

All of this implies that the fact that it is possible to get the average offer or the average MAO up to any value between 0 and 0.5 by choosing a sufficiently low intensity of selection is not necessarily something that reflects anything to do with selection. Selection pulls both of them down, and if one reduces selection ever more, then one can reduce how much both are dragged down. The fact that one can get them to average at values arbitrarily close to 0.5 by almost eliminating selection, however, is a somewhat arbitrary result of the shape of the strategy set, and not of what selection does to the strategies in it.

In the main text we illustrate this by looking at a simulation run for global and infrequent mutation, and, of course, weak selection. Here in the supplementary material we will also consider global and frequent, local and infrequent, and local and frequent mutation.

5A.2.2 Weak selection, global mutation, low mutation rate

The left hand side of Fig. 5A.1 depicts a few aspects of a run with global mutation, a low intensity of selection ($w = 0.001$), and a low mutation rate ($u = 0.001$). Panel A shows

how the average offer and the average MAO change over time for a part of a simulation run. Panel C gives a snapshot of the distribution at some moment in time, and here we find both traits to be at fixation, as is expected to be the case for most of the time with such a low mutation rate. Panel E averages these distributions (like the one given in panel C) across time, producing the average distribution over time. As is to be expected, this is quite close to the uniform distribution on $[0, 1]$, which is the distribution that all mutants come from.

The fact that the average offer and the average MAO move around quite a bit over the course of a run limits the predictive power of the model for this combination of low intensity of selection and low mutation rate. Any average that we find in experiments would be close to the average in the simulations at some points in time, but it would be far away from the averages that the simulations produce at many other points in time. Also, at most points in time, there is not much variation; the variation in panel E is generated by the variability across time, not by the variation at any moment in time. The prediction of this model therefore is that we should observe a close to monomorphic population, where the probability with which we would observe a certain average is the result of a draw from the uniform distribution. The fact that the MAO of everyone in the population is regularly also above the offers of everyone in the population (almost 50% of the time) also implies that if we really believe in weak selection, we should also conclude that if we now find the average MAO to be below the average offer, then this is just a coincidence, and it could also have been the other way around. That would make it very unlikely that between different populations they would correlate, and that the first always turns out to be below the second (Henrich and Boyd, 2001; Henrich et al., 2001; 2006).

5A.2.3 Weak selection, global mutation, high mutation rate

The right hand side of Fig. 5A.1 depicts the same aspects for a run, also with global mutation, and also with a low intensity of selection ($w = 0.001$), but with a high mutation rate ($u = 0.1$). Here, the averages in the population do not move around as much, and the shape of the distribution at any point in time is relatively close to the distribution of the inflow of mutants, which is a uniform distribution over $[0, 1]$. Given the low intensity of selection, this makes sense. With much less variability over time, this produces a much sharper prediction: the distribution should be close to uniform on $[0, 1]$ at all times. This does not match the empirical evidence either, because the distributions that we find in experiments typically are not that close to uniform. Moreover, as before, the average offer and the average MAO move close to independently, and this does not predict the average

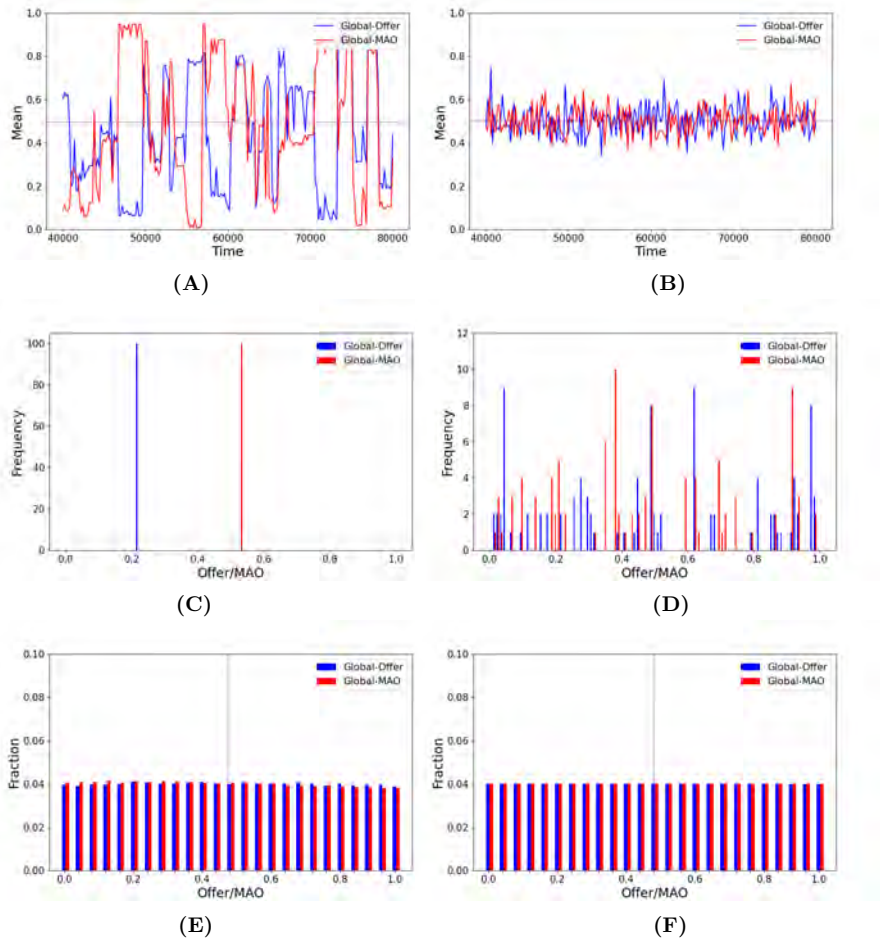


Figure 5A.1: Global, co-occurring mutations, w = 0.001, and u = 0.001 (left) and u = 0.1 (right). The top panels give the average offer and MAO over time for a part of the run. The middle panels give the distribution of strategies at some random moment in the simulation run. The bottom panels give the average distribution over time, where we bundled strategies within intervals of length 0.04 together. The average offer of the average distribution and the average MAO of the average distribution are horizontal lines in panel A and B, and vertical lines in panel E and F.

offer to be above the average MAO.

5A.2.4 Weak selection, local mutation, low mutation rate

With weak selection, local mutations, and low mutation rates, the populations are typically also close to monomorphic, as they are with global mutations in combination with low

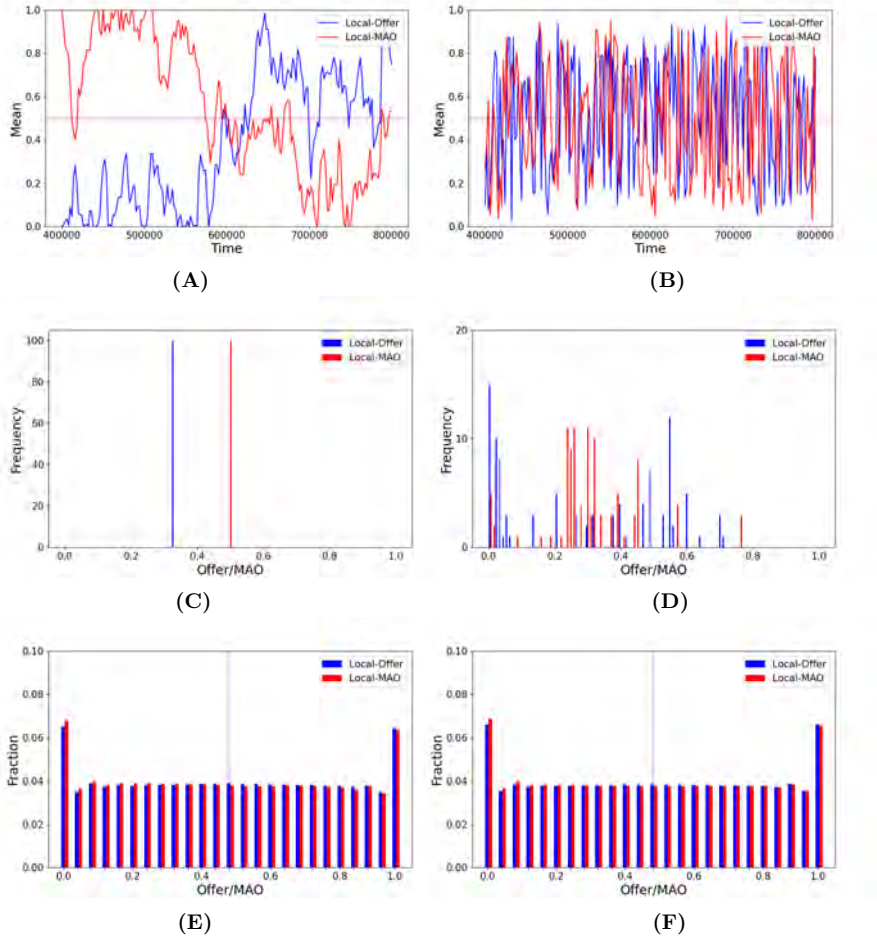


Figure 5A.2: Local, independent mutations, w = 0.001, and u = 0.001 (left) and u = 0.1 (right). The top panels give the average offer and MAO over time for a part of the run. The middle panels give the distribution of strategies at some random moment in the simulation run. The bottom panels give the average distribution over time, where we bundled strategies within intervals of length 0.04 together. The average offer of the average distribution and the average MAO of the average distribution are horizontal lines in panel A and B, and vertical lines in panel E and F.

mutation rates. Over time, they also move around quite a bit, and, similar to global mutations, not in synchrony. What is different is that changes in the averages come in much smaller steps, as a result of the mutations being local, and therefore the averages move around much slower (see Fig. 5A.2A, where time runs 10 times faster than in Fig. 5A.1A). The distribution over time is not the same as the "distribution that all mutants come from", as it is with global mutations, because with local mutations, there is no

such thing as a constant mutant distribution. Because the average is a random walk, restricted to $[0, 1]$, the distribution still ends up looking like a uniform distribution, with some deviations at and close to the boundaries (see Fig. 5A.2E). These boundaries are a bit stickier than other monomorphic states; trait values 0 and 1 collect more incoming mutations, that otherwise would have gone below 0 or over 1, and once the population is temporarily absorbed in one of the boundaries, and everyone has trait value 0, or everyone has trait value 1, leaving requires a mutant with the right sign. The population therefore spends some extra time at these extreme points.

5A.2.5 Weak selection, local mutation, high mutation rate

With weak selection and high mutation rates, the mutations being local rather than global allows random effects to persist for longer, because mutations are not biased towards the middle anymore – except for trait values at or close to 0 or 1. Deviations from the average over time therefore have much more amplitude than they do with global mutations (notice that also here, time is running 10 times faster in Fig. 5A.2B than it is in Fig 5A.1B). Compared to low mutation rates (Fig. 5A.2D vs. Fig. 5A.2C), the distribution at any given point in time is much less concentrated, and over time, the population moves around faster (Fig. 5A.2B vs. Fig. 5A.2A), but otherwise, also here the average (over time) of the averages (over the population) is a uniform distribution, with some deviations at the edges (see Fig. 5A.2F).

5A.2.6 Variance over time

In order to have an indication of how stable or unstable the distributions are over time, we can calculate the variance in average offers, or the variance in average MAO's, over time. If $\overline{p}^t$ is the average offer in the population at time t, and $\overline{\overline{p}} = \frac{1}{T}\sum_{t=1}^{T}\overline{p}^t$ is the average over time of these averages over the population, then

$$\frac{1}{T}\sum_{t=1}^{T}\left(\overline{p}^t - \overline{\overline{p}}\right)^2$$

is the variance across time. Simulations with a low variance have more predictive power than simulations with a high variance.

As a benchmark of something that has no predictive power, one could use the variance that would go with randomly drawing a new average offer or MAO from a uniform distribution

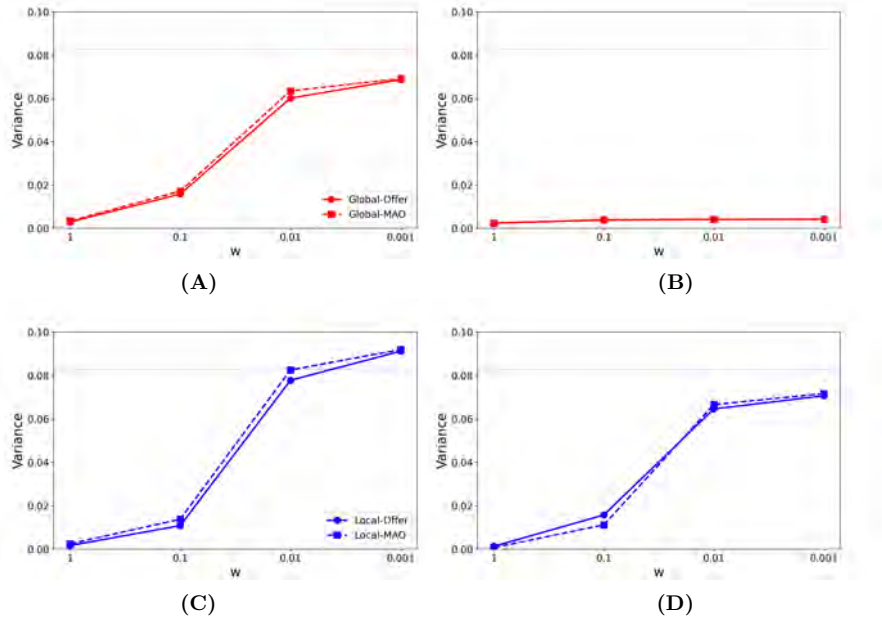


Figure 5A.3: Variances over time for weak selection. With global, and rare mutations ($u = 0.001$), the variance over time gets very high for $w = 0.01$ and $w = 0.001$ (A). With global, and frequent mutations ($u = 0.1$), the average is very stable, and the variance stays low (B). With local, and rare mutations, the variance over time gets high again, even a bit higher than the variance one would get from independent draws from the uniform distribution – indicated in the figures by straight horizontal lines (C). With local, and frequent mutations, the variance also gets very high for weak selection (D).

on $[0, 1]$ every period. In that case, the variance is

$$\int_0^1 (x - 0.5)^2 = \frac{1}{3}\left[(x - 0.5)^3\right]_0^1 = \frac{1}{12} \approx 0.083$$

Calculating these variances for the simulations paints a picture that is in line with what one would expect from Fig. 5A.1 and Fig. 5A.2; whether mutations are global and rare; local and rare; or local and frequent, variances get very high when selection gets weak (see Fig. 5A.3). Because the edges of the interval $[0, 1]$ are temporarily absorbing for local mutations, the variances there become even higher than $\frac{1}{12}$ when mutations are rare. The variance only remains low for mutations that are both global and frequent. In this case, the distribution at any point in time will be close to the distribution that the mutants come from.

This implies that with weak selection, the simulations are literally all over the place. For global and infrequent mutations, they are extremely variable over time; and for global

and frequent mutations, they are extremely variable at any moment in time. We should, however, not have a model with global, and therefore biased mutations anyway – as we have seen that it is this bias that drives the results – and if we choose local, and therefore much less biased mutations, the averages are, again, all over the place for weak selection, this time regardless of the mutation rate. That implies that the trick to push average offers and average MAO's up by lowering the intensity of selection goes at the expense of predictive power; any population average that one would find at some point in time would literally be equally likely under the model.

5A.3 Infinite population models

5A.3.1 The model in Gale et al. (1995)

First we repeat the equations for the model in Gale et al. (1995). The amount to be divided is denoted by n. The share of proposers that propose i is denoted by x_i, for $i = 1, ..., n$, and the share of responders with an MAO of j is denoted by y_j, for $j = 1, ..., n$. The mutation-selection dynamics are given by

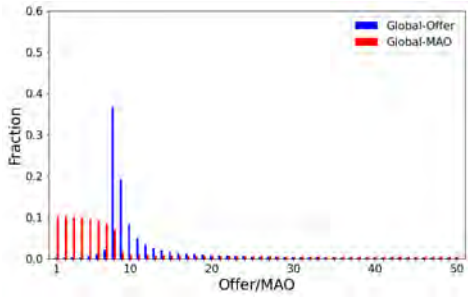
$$\dot{x}_i = (1 - \delta)(\pi_{i,P} - \overline{\pi}_P) x_i + \delta \left(\frac{1}{n} - x_i \right)$$

for proposers using strategies $i = 1, ..., n$, where $\dot{x}_i$ the time derivative of x_i, δ is the mutation rate, $\pi_{i,P}$ is the payoff of proposers that propose i, and $\overline{\pi}_P$ is the average payoff in the proposer population, and by

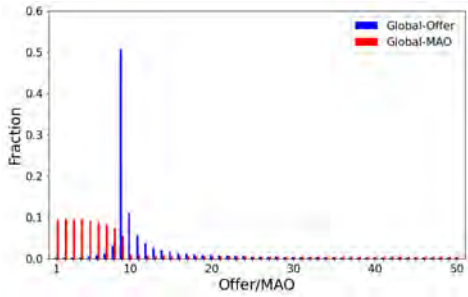
$$\dot{y}_j = (1 - \delta)(\pi_{j,R} - \overline{\pi}_R) y_j + \delta \left(\frac{1}{n} - y_j \right)$$

for responders that use strategies $j = 1, ..., n$, where $\dot{y}_j$ is the time derivative of y_j, $\pi_{j,R}$ is the payoff of responders with an MAO of j, and $\overline{\pi}_R$ is the average payoff in the responder population.

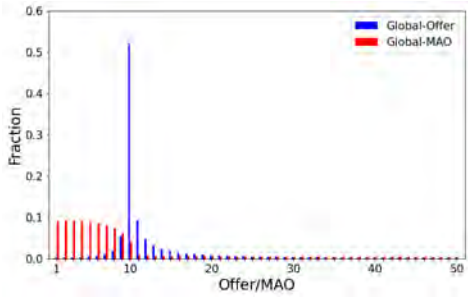
The first term on the right hand side reflects selection, the second term reflects mutation. Global mutation means that all strategies have the same inflow due to mutation (it is $\frac{\delta}{n}$ for all strategies) and an outflow that is proportional to the current shares (it is δx_i for proposers and δy_j for responders). Gale et al. (1995) allow for the mutation rate δ to differ between proposers and responders, but we will first assume that they are the same.



(A)



(B)



(C)

Figure 5A.1: Multiple equilibria for Gale et al. (1995). With $\delta = 0.05$ and global mutation, there are 3 mutation-selection equilibria. The most frequent strategy for proposers in those equilibria ranges from $i = 8$ (A) to 10 (C), making the predominant offer in those equilibria range from $7\frac{1}{2}$ (A) to $9\frac{1}{2}$ (C).

5A.3.2 Our small changes to the version with global mutations

The offers and MAO's in the original model run from 1 to n, and exclude 0. When we compare this model, which has a discrete strategy space, with the model from Rand et al. (2013), that has a continuous strategy space, it can be nice to make strategies in the former comparable to strategies within an interval in the latter. We therefore shifted all proposals to the left by $\frac{1}{2}$; instead of having proposer strategy i propose i, we choose for

strategy i to propose the midpoint of the interval $[i-1, i]$, which is $i - \frac{1}{2}$. Similarly, we let responder strategy j have an MAO of $j - \frac{1}{2}$. This only affects the equations above indirectly, in the sense that the payoff calculations now involve slightly shifted offers and MAO's. Below, we will sometimes still just refer to those strategies as strategy i or j, because that is shorter, but sometimes we will explicitly refer to the offer, and then we write $i - \frac{1}{2}$, or to the MAO, in which case we write $j - \frac{1}{2}$.

We also normalize the payoffs, so that the maximum payoff is 1 and the minimum payoff is 0. With normalization, one can see n, not as an indicator of the pie size, but as an indicator of how finely one unit can be subdivided. This helps comparing the results to simulations from Rand et al. (2013), which have a fixed amount of 1 to be divided in the ultimatum game.

5A.3.3 Multiple equilibria

Figure 5A.1 shows a variety of equilibria for the same combination of n and δ. All of those equilibria are similar, in that most proposers are making the same offer, with fewer proposers making higher offers, and even fewer making lower offers. Most responders have MAO's somewhere between the smallest possible MAO and the offer that most proposers make, and very few have larger MAO's. The different equilibria are characterized by what the most frequent offer is; for $n = 50$ and $\delta = 0.05$ – the parameters from Figure 5A.1 – there are equilibria where the most frequent MAO is $7\frac{1}{2}$ (A), $8\frac{1}{2}$ (B), or $9\frac{1}{2}$ (C).[5]

The first parameter combinations in Figure 5A.2 have multiple mutation-selection equilibria. The other three have unique, globally attracting mutation-selection equilibria. With the limited computing power of 1995, Gale et al. may have missed the possibility that the population might converge to different states depending on the starting point.

5A.3.4 Our version with local mutations

Also here, global mutations are biased, which is not a good basis for an explanation. Therefore, as we did with the model in Rand et al. (2013), we also made a version with local, and therefore much less biased mutations. In the local version, if an individual mutates that currently plays strategy i, then the new strategy becomes any strategy from $i-k$ to $i+k$, all with equal probability – provided that these changes do not make the offer

[5]For replicator dynamics with a continuous strategy space, one would expect a spectrum of equilibria, with positive point mass at some offer for proposers. These would be stable to perturbations that are small in the variational distance, but not to perturbations that are small in the Prohorov metric (see Van Veelen and Spreij, 2009).

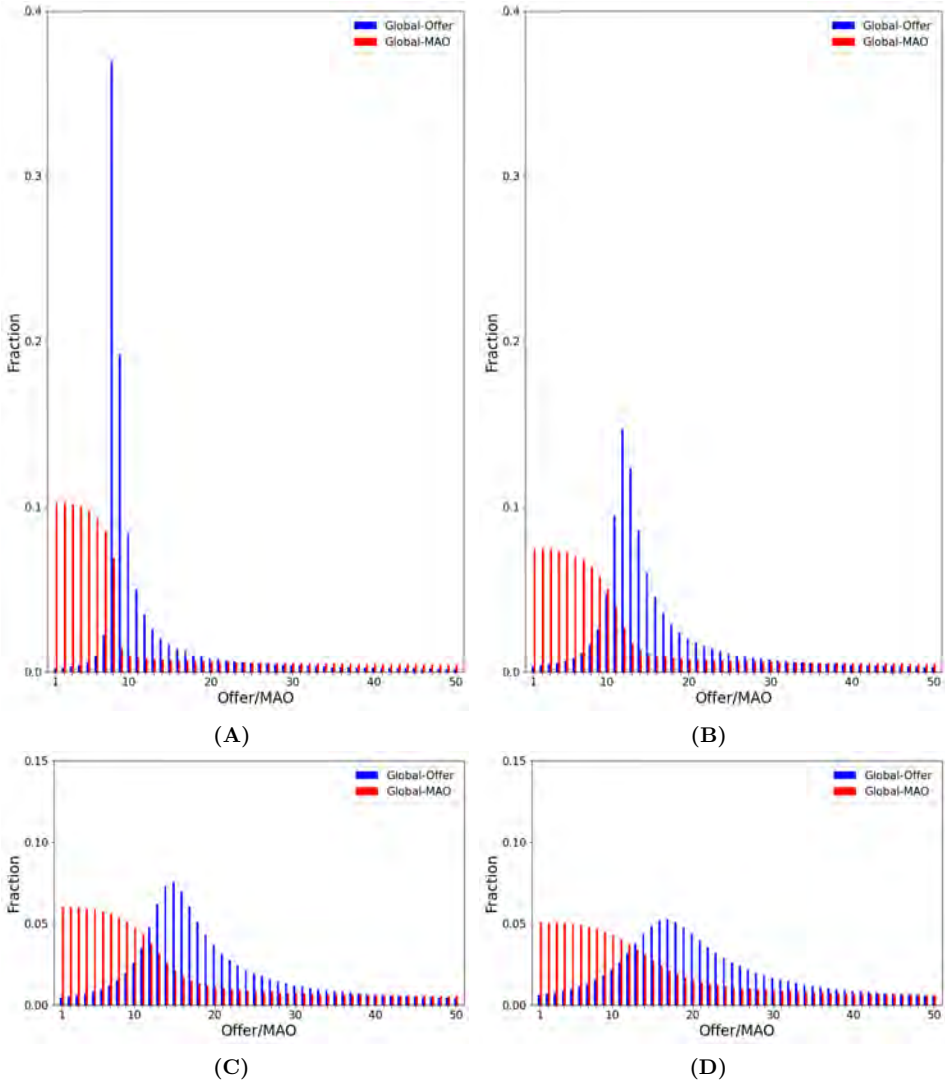


Figure 5A.2: Mutation-selection equilibria in Gale et al. (1995) with global mutations. For $\delta = 0.05$ there are multiple equilibria (see Figure 5A.1). We picked the first one for panel A. For $\delta = 0.075$ (B), $\delta = 0.1$ (C), and $\delta = 0.125$ (D), there is a unique, globally attracting mutation-selection equilibrium. The fat tails are a symptom of the bias in the mutations.

or MAO drop below 0 or go over 1. The latter is guaranteed not to occur if $k < i \leq n - k$. If $i \leq k$, the mutant becomes any strategy from 2 to $i + k$ with probability $\frac{1}{2k+1}$, and strategy 1 with the remaining probability, and if $> n - k$, then the mutant becomes any strategy from $i - k$ to $n - 1$ with probability $\frac{1}{2k+1}$, and strategy n with the remaining

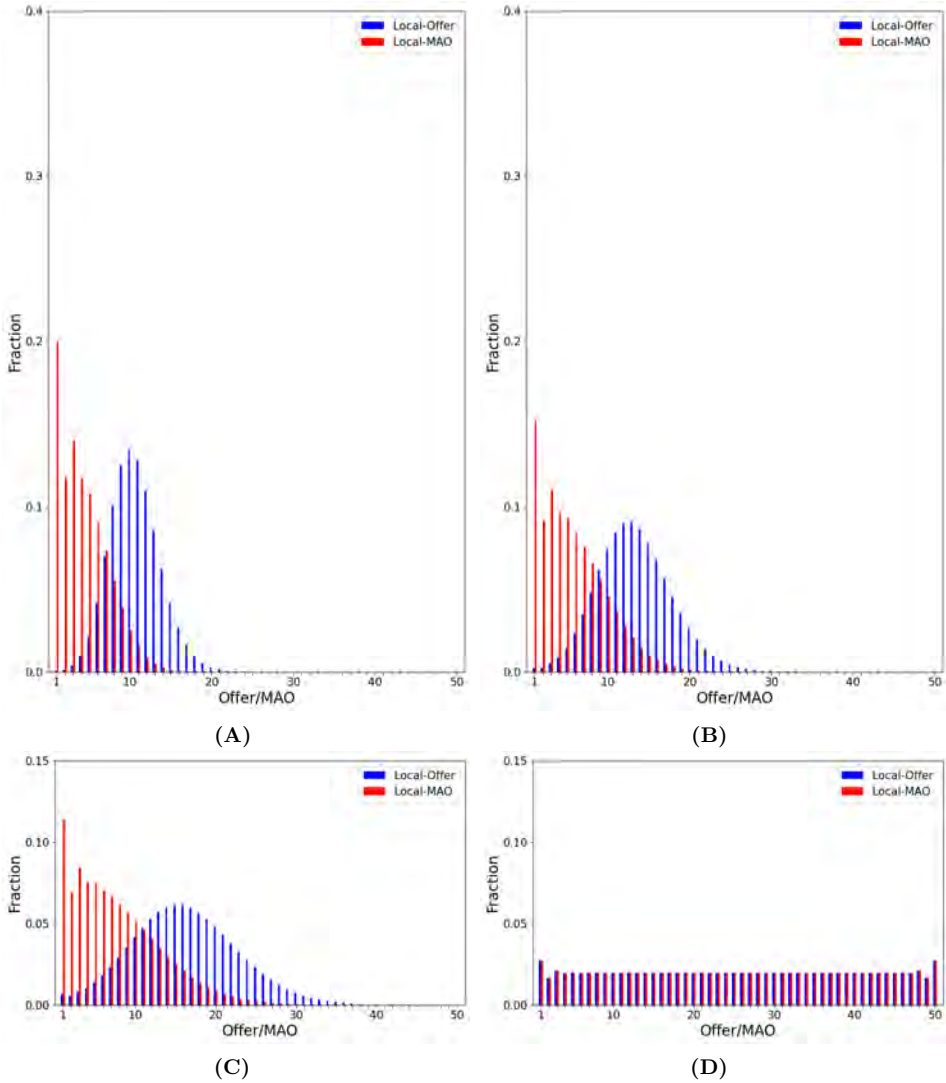


Figure 5A.3: Mutation-selection equilibria in Gale et al. (1995) with local mutations. Mutation rates are $\delta = 0.25$ (A), $\delta = 0.5$ (B), $\delta = 0.75$ (C), and $\delta = 1$ (D). The wave pattern at the boundaries of the strategy space is caused by the remaining bias in mutations at the boundaries, where mutations to strategies below 0 or above 50 are ruled out.

probability.

For $n = 50$ and mutations that take a mutant a maximum of $k = 2$ steps to the right or to the left, that means that the equations for the replicator dynamics for proposers

become

$$\dot{x}_1 = (1-\delta)\left(\pi_{1,P} - \overline{\pi}_P\right) x_1 + \delta\left(-\frac{2}{5}x_1 + \frac{2}{5}x_2 + \frac{1}{5}x_3\right)$$
$$\dot{x}_2 = (1-\delta)\left(\pi_{2,P} - \overline{\pi}_P\right) x_2 + \delta\left(\frac{1}{5}x_1 - \frac{4}{5}x_2 + \frac{1}{5}x_3 + \frac{1}{5}x_4\right)$$

for the first two, then

$$\dot{x}_i = (1-\delta)\left(\pi_{i,P} - \overline{\pi}_P\right) x_i + \delta\left(\frac{1}{5}x_{i-2} + \frac{1}{5}x_{i-1} - \frac{4}{5}x_i + \frac{1}{5}x_{i+1} + \frac{1}{5}x_{i+2}\right)$$

for strategies 3 to 48, and

$$\dot{x}_{49} = (1-\delta)\left(\pi_{49,P} - \overline{\pi}_P\right) x_{49} + \delta\left(\frac{1}{5}x_{47} + \frac{1}{5}x_{48} - \frac{4}{5}x_{49} + \frac{1}{5}x_{50}\right)$$
$$\dot{x}_{50} = (1-\delta)\left(\pi_{50,P} - \overline{\pi}_P\right) x_{50} + \delta\left(\frac{1}{5}x_{48} + \frac{2}{5}x_{49} - \frac{2}{5}x_{50}\right)$$

for the last two. For responders, this is

$$\dot{y}_1 = (1-\delta)\left(\pi_{1,R} - \overline{\pi}_R\right) y_1 + \delta\left(-\frac{2}{5}y_1 + \frac{2}{5}y_2 + \frac{1}{5}y_3\right)$$
$$\dot{y}_2 = (1-\delta)\left(\pi_{2,R} - \overline{\pi}_R\right) y_2 + \delta\left(\frac{1}{5}y_1 - \frac{4}{5}y_2 + \frac{1}{5}y_3 + \frac{1}{5}y_4\right)$$

for the first two, then

$$\dot{y}_j = (1-\delta)\left(\pi_{j,R} - \overline{\pi}_R\right) y_j + \delta\left(\frac{1}{5}y_{j-2} + \frac{1}{5}y_{j-1} - \frac{4}{5}y_j + \frac{1}{5}y_{j+1} + \frac{1}{5}y_{j+2}\right)$$

for strategies 3 to 48, and

$$\dot{y}_{49} = (1-\delta)\left(\pi_{49,R} - \overline{\pi}_R\right) y_{49} + \delta\left(\frac{1}{5}y_{47} + \frac{1}{5}y_{48} - \frac{4}{5}y_{49} + \frac{1}{5}x_{50}\right)$$
$$\dot{y}_{50} = (1-\delta)\left(\pi_{50,R} - \overline{\pi}_R\right) y_{50} + \delta\left(\frac{1}{5}y_{48} + \frac{2}{5}y_{49} - \frac{2}{5}y_{50}\right)$$

for the last two.

Figure 5A.3 shows mutation-selection equilibria for global mutation and a variety of mutation rates. Many observations made when comparing global and local mutations for Rand et al. (2013) can also be made here. The most obvious one is that when comparing Figures 5A.2 and 5A.3, we see that also here, all else equal, average offers and MAO's are lower

with local than with global mutation (note that mutation rates in Figure 5A.3 are higher than in Figure 5A.2).

Because mutants have to remain within the strategy space, we assumed that mutations to strategies below 0 are replaced with mutations to 0, and mutations to strategies above n are replaced by mutations to n. That means that 0 and n have extra incoming mutations, while, in our case, with a maximum change in strategy of 2 due to mutation, strategies 1 and $n-1$ only have a reduced amount of incoming mutations, since strategies below 0 or above n that could mutate to 1 and $n-1$, respectively, do not exist. In equilibrium, this creates a spike at 1, a valley at 2, and those also ripple through the frequencies towards the middle. In panel (D) we see the same at the top end of the strategy space.

Finally, a difference between Rand et al. (2013) with local mutations and Gale et al. (1995) with local mutations, is that we have seen that even at a mutation rate of 1, the average offer and the average MAO are not $\frac{1}{2}$ in Rand et al. (2013) with local mutations. In Figure 5A.3 we see that for the version of Gale et al. (1995) with local mutations, this is not true, and both averages are in fact $\frac{1}{2}$. This is caused by the difference in how reproduction events and mutations relate in both models. In Rand et al. (2013), mutations happen at reproduction. That means that at a mutation rate of 1, reproductions still happen, but at every one of those, a mutation occurs. With local mutations, that means that the trait value of the offspring is still correlated with the trait value of the parents (which is not true for global mutations). In Gale et al. (1995), mutations happen not at reproduction. Instead, the mutation rate reflects how many mutation events occur relative to the number of reproduction events. That means that here, at a mutation rate of 1, there are only mutations, and reproduction is just not happening.

5A.4 Link between Rand et al. (2013) and Gale et al. (1995)

For infinitely large population dynamics, if a population is in equilibrium, it does not move. In finite population dynamics, also in equilibrium, the population moves around, but visits some states (much) more often than others. With finite population dynamics, the equilibrium therefore is a distribution over population states. As the population size increases, the noise decreases, and the variation in population states across time goes down. In the limit of infinitely large populations, the dynamics become deterministic.

In order to investigate the link between the finite population model in Rand et al. (2013)

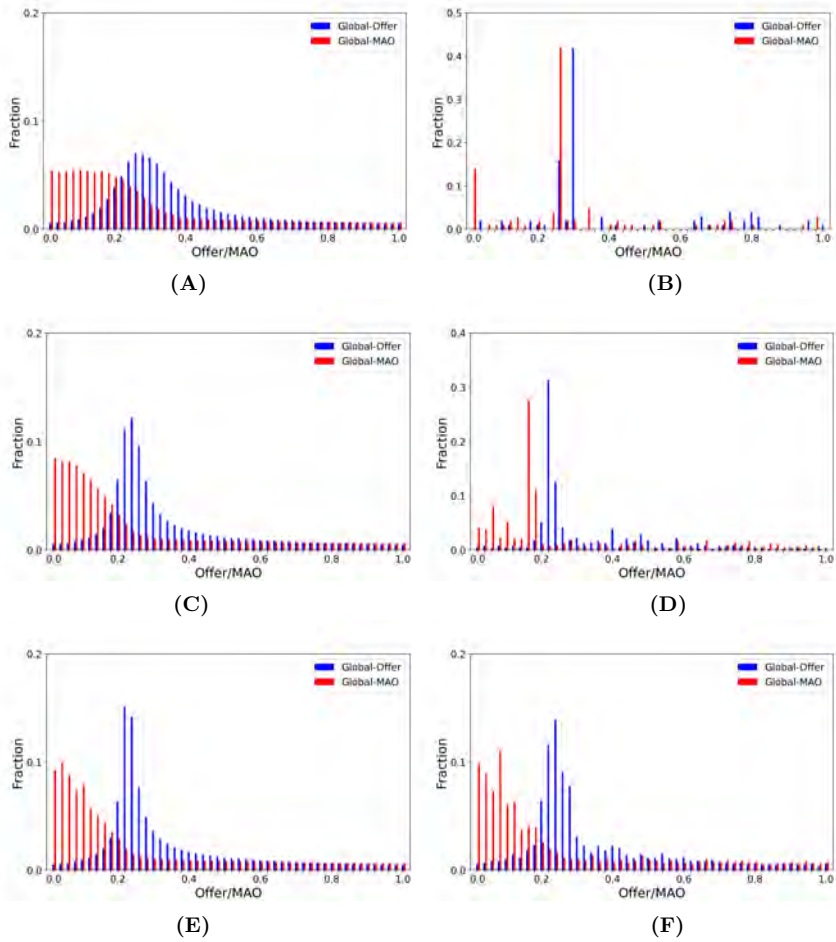


Figure 5A.1: Rand et al. (2013) for different population sizes. The mutation rate is 0.125, the intensity of selection is 1, and the population size is 100 (top panels), 1,000 (middle panels), and 10,000 (bottom panels). Panels A, C, and E show the average distributions of offers and MAO's, where the average is taken over time. Panels B, D, and F show snapshots of offers and MAO's. The scaling on the horizontal axes in B and D is different from the other panels. Because running additional generations becomes rather expensive at a population size of 10,000, the distribution in E is a bit noisier than in A and C.

and the infinite population model in Gale et al. (1995), we ran simulations with increasing population size for Rand et al. (2013). When we compare snapshots from the population with the average distribution (over time), we find that the difference between these two does indeed decrease with population size – which is an indication that the population does indeed move around less. For global mutation and a population size of 100, the difference

between the snapshot and the average distribution (which averages these snaphots over time) is very large, and also for a population size of $1,000$ it is still considerably different, and only at a population of $10,000$, they come close. The characteristics of the average distribution therefore are not perfectly representative for the average characteristics of the distribution at any given point in time, although much more at $10,000$ than the other population sizes. The variance within the population at any moment in time is also smaller than the variance in the average distribution. This difference also goes down, but is quite substantial at 100 and $1,000$. One could therefore say that the infinite population model in Gale et al. (1995) is only a good approximation of the finite population model in Rand et al. (2013) for quite large finite populations. The discrepancies are however smaller for our versions of the two with local mutations (not depicted).

5A.5 Quantal Response Equilibria

5A.5.1 Predictions and data

Data We use the data from the papers listed in Table 5A.1 in our main text and in the supplementary material. In addition to these, we use the data from the meta-analysis of Tomlin (2015) in Figure 5A.1.[6] For both the direct-response method and the strategy method we calculate the outcome variable proportional to the total amount available in the ultimatum game to make observations from different studies comparable as much as allowed by the grid used in the experiments. All of the studies use real monetary stakes.

In studies that use the direct-response method we consider how often an offer is accepted out of the number of times that offer is made, as an estimate of its acceptance probability. In studies that use the strategy method we analyze the MAO's reported by participants, or their accept/reject decisions for each possible offer level, as an estimate of the acceptance probability for every possible offer.

Logistic regression with and without the intercept Since the *Agent*-QRE is equivalent to the logit specification without the intercept, we ran logit regressions with and without the intercept to test between the two specifications. Table 5A.2 presents the results. As can be seen from Column (2), the coefficient on the intercept is highly statistically significant. In line with this, the AIC, BIC, Pearson's χ^2 criteria, and the likelihood-ratio (LR) test all indicate that the introduction of the intercept term in Column (2) improves

[6]Since the dataset in Tomlin (2015) does not include information on stake sizes, we excluded those observations in our analysis for the main text.

Author, Year	# Obs. all stakes	# Obs. low/medium	DR vs. STR
Andersen et al. (2011)	458	325	DR
Bader et al. (2021)	485	–	STR
Bahry and Wilson (2006)	288	–	STR
Barmettler et al. (2012)	100	–	DR
Benndorf et al. (2017)	98	–	STR
Bornstein and Yaniv (1998)	20	–	DR
Cameron (1999)	202	165	DR
Carpenter et al. (2005a;b)	107	–	DR
Chew et al. (2013)	207	–	STR
Croson (1996)	56	–	DR
Demiral and Mollerstrom (2020)	283	–	STR
Forsythe et al. (1994)	67	–	DR
Inaba et al. (2018)	121	–	STR
Keuschnigg et al. (2016)	487	–	STR
Lightner et al. (2017)	42	–	DR
Peysakhovich et al. (2014)	576	–	STR
Ruffle (1998)	44	–	DR
Slonim and Roth (1998)	820	570	DR

Table 5A.1: Alphabetical list of empirical papers whose data we use. The second column shows the number of observations after eliminating the treatments using the non-standard versions of the ultimatum game, for studies with a variety of stake sizes, the third column shows the number of observations after excluding also the treatments with the largest stake sizes in studies testing the stake size effects, and the last column shows whether the experimental design uses the direct-response (DR) or the strategy (STR) method.

	(1) P(accept)	(2) P(accept)
Offer	4.778*** (p<0.001)	7.463*** (p<0.001)
Intercept		-1.035*** (p<0.001)
Observations	1496	1496
Log-likelihood	-600.78098	-580.5621
AIC	1203.562	1165.124
BIC	1208.873	1175.745
Pearson's χ^2	190.56	302.69
LR-test	40.44 (p<0.001)	

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5A.2: Logistic regression results with and without the intercept.

the model fit significantly.

With and without large stakes In the main text, we compared the predictions of the *Agent*-QRE with rejection rates we calculated by pooling data from experiments that use the direct-response method. Because it is not universally agreed upon whether stake size matters, we excluded the observations for the largest stakes in the papers by Andersen et al. (2011); Cameron (1999); Slonim and Roth (1998).[7] If we include all stake sizes, the pattern is similar (see Figure 5A.1).

Strategy method data for *Agent*-QRE In the main text, we compared the predictions of the *Agent*-QRE with rejection rates that we calculate by pooling data from experiments that use the direct response method. This is the more natural thing to do, but one can also construct rejection rates from experiments that use the strategy method. The majority of studies that use the strategy method restrict subjects to strategies that can be characterized with an MAO; they exclude strategies for which there exist two offers, where the higher one is rejected, and the lower one is accepted. This restriction by construction implies that we will find that rejection rates are never decreasing in the offer. The prediction that rejection rates increase therefore cannot be tested with the experiments that use the strategy method in this way. The prediction that all acceptance rates should be above 50% can be tested, and would be rejected with data from experiments

[7] In experiments that use three levels of stake sizes we exclude the largest one; in experiments that use four levels of stake sizes we exclude the largest two. As Carpenter et al. (2005a) uses stake sizes of only \$10 and \$100, we include both stakes in our analysis in the main text.

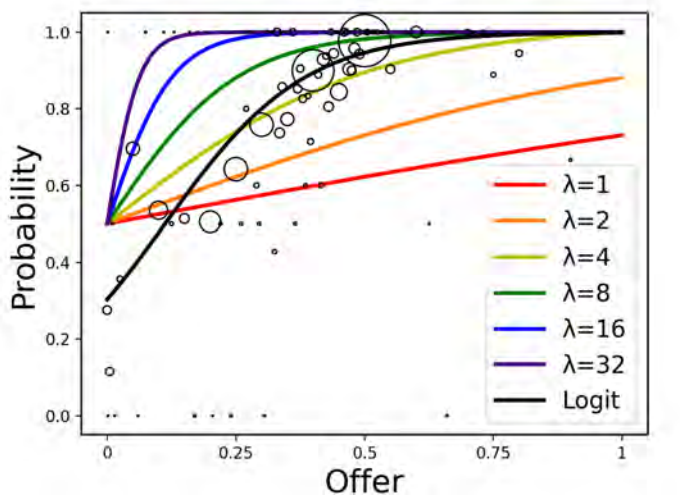
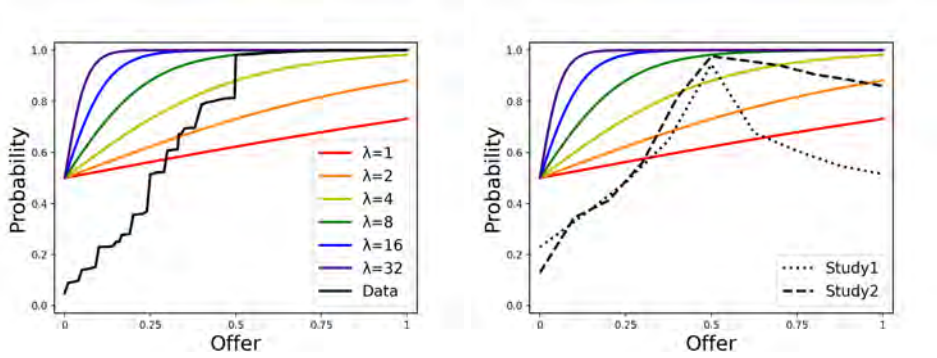


Figure 5A.1: Acceptance/rejection rates in *agent*-QRE vs. empirical acceptance/rejection rates. The coloured lines are the acceptance rates in the *agent*-QRE for different λ's. The circles indicate acceptance rates for different proposals, pooling data from a number experiments together. Their size reflects the number of observations for that offer. The black line is the fitted acceptance rate as a function of the offer for a logit regression. Here, we use data obtained with the direct-response method, including all stake sizes.

that use the strategy method – as it is with experiments that use the direct-response method (see the left panel in Figure 5A.2).

Three of the studies in our sample (Bader et al., 2021; Bahry and Wilson, 2006; Keuschnigg et al., 2016) do not restrict subjects to submitting an MAO. They instead ask their participants to submit their accept/reject decisions for each possible offer level, which allows them to freely switch between accepting and rejecting. For these studies we include the participants who never switch, who switch only once (those that start with rejecting and switch to accepting at a certain offer level), and those who switch twice (from rejecting to accepting in the first half of the strategy space of offers, and from accepting back to rejecting in the second half of the strategy space). We exclude participants who do not fall into one of these categories. The number of observations given in Table 5A.1 represents the number of observations after excluding these participants. Using the data from these studies we can test both the prediction that rejection rates monotonically increase in the offer and the prediction that all acceptance rates should be above 50%; and both would be rejected (see the right panel in Figure 5A.2).



(A) **(B)**

Figure 5A.2: Acceptance/rejection rates in *agent*-QRE vs. empirical acceptance/rejection rates under the strategy method. The coloured lines are the acceptance rates in the *agent*-QRE for different λ's. The black line in panel A indicates acceptance rates for different proposals, pooling data from a number experiments together that use the strategy method and ask subjects to submit an MAO. The two dotted lines in panel B indicate acceptance rates for different proposals, for the three experiments that use the strategy method and ask subjects to submit their accept/reject decision for each possible offer within their grid. Study1 is Bahry and Wilson (2006), and Study2 combines the data from Keuschnigg et al. (2016) and Bader et al. (2021) as they have an identical design.

5A.5.2 *Agent*-QRE and learning models

In the *Agent*-QRE, a higher offer is accepted with higher probability than a lower offer. The underlying assumption is that the noise in the perception of what the payoff-maximizing thing to do is, is the same for all proposals. Combined with the fact that the payoff difference between accepting and rejecting gets larger when proposals increase, and therefore selection against rejecting also becomes stronger, this leads to the probability of accepting the offer being larger for larger offers and smaller for smaller offers.

The assumption of constant noise can be realistic, but one can also imagine that there are models for which this does not hold. For instance, one can also assume that the noise is higher for proposals that are made less frequently, and, depending on how the increased noise balances against the increased payoff difference, the rejection rate of a higher offer, that is made less frequently, could also end up being higher than that of a lower offer, that is made more frequently.

Also more in general, for models of selection that fit the setup of an *Agent*-QRE, where responses to different offers evolve separately, the property that higher offers get lower

rejection rates may not universally hold. If we think of a model where strategies for different proposals do indeed evolve independently, then one could imagine that there is more selection happening for proposals that are made more frequently. If the mutation rate for responses to different offers is the same, one can imagine population states for which high offers are made so infrequently, that selection against rejections there is weak, and the rejection rate ends up being higher than that of a lower offer, that is made more frequently, and where there is more selection undoing the effect of mutations.

It is good to keep in mind, though, that this is only a detail, and perhaps a reason to prefer the *Normal form*-QRE over the *Agent*-QRE, but not an escape from the fact that the data would reject these models too. While the property of the *Agent*-QRE (higher x are always accepted with higher probability) might not carry over to all learning models or mutation-selection models that treat strategies for all offers separately, the property that acceptance rates should all be larger than 50% for all positive offers does. It is this property that is clearly violated by the data.

5A.5.3 Acknowledgements for data sharing

We are very grateful to Donna Bahry, Rick Wilson; Volker Benndorf, Claudia Moellers, Hans-Theo Normann; Lisa Cameron; Jeffrey Carpenter, Eric Verhoogen, Stephen Burks; Soo Hong Chew, Richard Ebstein, Songfa Zhong; Rachel Croson; Elif Demiral, Johanna Mollerstrom; Robert Forsythe, Joel Horowitz, Gene Savin, Martin Sefton; Marc Keuschnigg, Felix Bader, Johannes Bracher, Bastian Baumeister, Roger Berger; Bradley Ruffle; Robert Slonim, Alvin Roth; and Damon Tomlin, for sharing their data with us, and to Steffen Andersen, Seda Ertaç, Uri Gneezy, Moshe Hoffman, John List; Franziska Barmettler, Ernst Fehr, Christian Zehnder; Gary Bornstein, Ilan Yaniv; Misato Inaba, Yumi Inoue, Satoshi Akutsu, Nobuyuki Takahashi, Toshio Yamagishi; Aaron Lightner, Pat Barclay, Edward Hagen; Alexander Peysakhovich, Martin Nowak, and David Rand, for making their data publicly available.

Summary

The concept of *Homo economicus* constitutes a valuable benchmark for modelling economic environments. However, it regularly fails to capture human behaviour, especially in settings including interpersonal interactions. In lab experiments and field observations, it has been consistently shown that humans are not really *Homo economicus*, that they are not profit-maximizing, rational, and selfish agents, and that they knowingly and willingly share their endowments with individuals whose identities they do not and will never know. Why and under what conditions people act prosocially have therefore occupied the minds of economists, as well as scientists from other fields, for a long time.

A crucial prerequisite for understanding human prosociality is investigating its origins, and in doing so, looking for mechanisms that can explain (i) how our sociality evolved to the extent it did, and (ii) why it did not evolve to the same extent in other species. My dissertation, therefore, focuses on examining the roots of human ultra-sociality and looks for a combination of uniquely human mechanisms that would explain the evolution of our moral sentiments. In doing so, it combines methods and insights from various disciplines, since understanding the human mind and behaviour unavoidably requires an interdisciplinary approach.

Chapter 2 uses formal evolutionary modelling to examine a common critical assumption in the group selection literature. It theoretically studies how relaxing the assumption of global between-group competition affects the conditions under which cooperation can evolve by group selection. The analytical solutions and the simulation results show that the evolution of cooperation can be substantially hindered under non-global competition. This has important implications for the empirical literature on the topic, as it shows that (explicitly or implicitly) assuming global competition likely yields biased estimates. Given the importance of group selection models for explaining a wide range of social behaviours in human and other animal populations, it is crucial to take the bias generated by the global competition assumption into account and adjust one's estimates based on the competition

structure in their relevant study population.

Chapter 3 presents the results of a lab experiment that investigates how honesty can bring advantages in partner choice. In this chapter, we examine whether honesty can cause commitment to acting prosocially in a situation with asymmetric information, and whether this commitment is anticipated by others and creates a preference for honest partners. Our results show that this is indeed the case. We also test whether the connection between honesty and prosociality goes through the impact of communication, where honest individuals act prosocially to avoid lying to their partners. We find no evidence for this path. Instead, our findings suggest that honest individuals might have a hard time justifying selfish behaviour to themselves in the first place.

Chapters 4 and 5 review the literature on the evolution of social preferences, and argue that, without considering commitment as a mechanism, one cannot fully explain the empirical observations from lab experiments. Chapter 4 builds on the work of Frank (1987; 1988). It goes through the most common theoretical models on the evolution of cooperation, checks whether the predictions of these models match the behaviour in the lab, and argues that commitment is the key to explaining the empirical data. Chapter 5 also studies the role of commitment in social interactions, but restricts the attention to fairness preferences in the ultimatum game. It summarizes the most common models in the literature that explain fair behaviour in the ultimatum game with noise or mistakes, and updates them by allowing for more general mutation structures. It then compares the model predictions to empirical observations from several ultimatum game experiments. The conclusion is again that a model with commitment performs best in fitting the data. What Chapters 4 and 5 aim at doing is a shift in focus in the literature towards studying what role commitment plays in human (social) interactions, and if and how we detect true commitment.

Türkçe Özet (Summary in Turkish)

Homo economicus (İktisadi insan) kavramı, bir başka deyişle, insanların sadece kâr maksimizasyonu güden, rasyonel ve bencil varlıklar oldukları varsayımı, ekonomik modellemede kullanılan değerli bir ölçüttür. Ancak, özellikle bireylerarası etkileşimleri içeren durumlarda, *Homo economicus*, insan davranışlarını öngörmek konusunda genellikle yetersiz kalır. Sosyal davranışların incelendiği laboratuar deneylerinde ve saha çalışmalarında, insanların *Homo economicus* olmadığı, ellerindeki imkanları bilerek ve isteyerek (tanımadıkları) diğer bireylerle paylaştıkları düzenli olarak gözlemlenmektedir. Bu gözlemler, sosyal davranışların tabiatı konusuna dikkat çekmiş, ve insanların neden ve hangi koşullar altında yardımsever davranışlar sergiledikleri iktisatçıların ve diğer biliminsanlarının zihinlerini meşgul edegelmiştir.

Bireylerin yardımsever ve sosyal eğilimlerini anlamak bu davranışların kökenlerini araştırmaktan geçer. Bunu yaparken, iki noktaya özellikle dikkat etmek gerekmektedir: (i) sosyal davranışların evrimsel tarihimiz boyunca nasıl geliştiğini incelemek ve (ii) yardımsever eğilimlerin diğer canlılarda neden aynı ölçüde gelişmediğini açıklayabilmek. Bu nedenle, bu tez, sosyal davranışların kökenlerinin insana özgü mekanizmalarla açıklanmasına odaklanmaktadır. İnsan zihnini ve davranışını anlayabilmek kaçınılmaz olarak disiplinlerarası bir yaklaşım gerektirdiğinden, bu tezde, farklı disiplinlerin araştırma yöntemleri ve öğretileri sentezlenmektedir.

İkinci ünite, teorik modelleme yöntemlerini kullanarak, grup seçilimi literatüründe sıkça kullanılan bir varsayımı irdeler. Bu ünitede, gruplar arası rekabetin küresel düzlemde gerçekleştiği varsayımının gevşetilmesinin sosyal davranışlarının evrimine elverişli koşulları nasıl etkilediği teorik olarak incelenmektedir. Sunulan analitik çözümler ve simülasyon verileri, küresel rekabete kıyasla yerel rekabetin sosyal davranışların evrimini önemli ölçüde engelleyebileceğini göstermektedir. Bu sonuç, küresel rekabet varsayımı üzerine kurulu

grup seçilimi alanındaki verisel çalışmalar için büyük önem teşkil etmektedir. Grup seçilim modellerinin sosyal davranışları açıklamadaki önemi göz önüne alındığında, bu ünite, bu alanda yapılan çalışmalarda küresel rekabet varsayımının sonuçlarının hesaba katılması ve ilgili populasyona dair rekabet yapısının göz önünde bulundurulmasının zaruriyetini göstermektedir.

Üçüncü ünite, dürüstlüğün partner seçiminde ne tür avantajlar sağlayabileceğini araştıran bir laboratuvar deneyini raporlar. Bu ünitede, bireylerin bilgiye erişimlerinin asimetrik olduğu durumlarda, dürüstlüğün adaletli davranmayı taahhüt edip etmediği, bu taahhüdün diğer partnerler tarafından öngörülüp öngörülmediği ve dürüst partnerler için bir tercih yaratıp yaratmadığı incelenmektedir. Yapılan deneyin sonuçlarının bu hipotezleri desteklediği gösterilmektedir. Bu ünitede, dürüst bireylerin partnerlerine yalan söylemekten kaçınmak için adil davrandıp davranmadıkları hipotezi de test edilmektedir. Deneyin sonuçları, dürüstlük ve adalet arasındaki bağlantının, dürüst bireylerin partnerlerine yalan söylemekten kaçınmasından değil, bu bireylerin adaletsiz davranmayı en başta kendilerine karşı meşrulaştırmakta zorlandıklarından ötürü olabileceğini önermektedir.

Dördüncü ve Beşinci ünite, sosyal davranışların evrimi üzerine olan literatürdeki teorik ve deneysel çalışmaları gözden geçirir. Bu ünitelerde, mevcut literatürün göz ardı ettiği taahhüt mekanizmasının ele alınmadan bireylerin sosyal eğilimlerinin tam olarak açıklanamayacağı vurgulanır. Dördüncü ünite, Robert Frank'ın çalışmalarına dayanmaktadır (Frank, 1987; 1988). Bu ünitede, sosyal davranışların evrimini açıklamayı hedefleyen mevcut teorik modeller gözden geçirilmekte, bu modellerin öngörülerinin laboratuvar gözlemleriyle eşleşmediği gösterilmekte, ve taahhüt mekanizmasının sosyal davranışları açıklamanın anahtarı olduğu ileri sürülmektedir. Beşinci ünite, ültimatom oyununa odaklanarak, sosyal etkileşimlerde taahhüt mekanizmasının rolünü spesifik bir bağlamda detaylı bir şekilde incelemektedir. Bu ünitede, ilk olarak, literatürdeki ültimatom oyunlarında gözlenen adil davranışları bireylerin hata yapma yatkınlıklarıyla açıklayan modeller özetlenmekte ve bu modeller daha genel mutasyon yapılarıyla güncellenmektedir. Ardından, bu modellerin öngörüleri ültimatom oyunu deney verileriyle karşılaştırılmakta, ve modeller ile veriler arasında bir uyuşmazlık olduğu gösterilmektedir. Son olarak, taahhüt mekanizmasının deneylerdeki verilerle en iyi şekilde bağdaştığı ve ültimatom oyunundaki adil davranışları en başarılı şekilde açıkladığı gösterilmektetir. Bu bağlamda, bireysel ilişkilerdeki taahhüt yeterlilikleri göz önüne alınmadan insan davranışlarının tam olarak kavranılamayacağı tartışılmaktadır. Dördüncü ve Beşinci ünitede amaçlanan, mevcut literatürü taahhüt mekanizmasının sosyal davranışlar ve etkileşimlerdeki rolü üzerine daha çok odaklanmaya davet etmektir.

Nederlandse Samenvatting (Summary in Dutch)

Het concept van de *Homo economicus* is een waardevolle maatstaf voor het modelleren van economische omgevingen. Het slaagt er echter regelmatig niet in om menselijk gedrag uit te leggen, vooral in omgevingen met interpersoonlijke interacties. In laboratoriumexperimenten en veldwaarnemingen is consequent aangetoond dat mensen niet echt een *Homo economicus* zijn, dat ze geen winst maximaliserende, rationele en egoïstische agenten zijn, en dat ze willens en wetens hun middelen delen met individuen wiens identiteiten ze niet kennen. Waarom en onder welke omstandigheden mensen prosociaal handelen, heeft daarom economen en wetenschappers uit andere vakgebieden lang beziggehouden.

Een cruciale voorwaarde voor het begrijpen van menselijke prosocialiteit is het onderzoeken van de oorsprong ervan, en het daarbij zoeken naar mechanismen die kunnen verklaren (i) hoe onze socialiteit in die mate evolueerde en (ii) waarom ze niet in dezelfde mate evolueerde in andere diersoorten. Mijn proefschrift richt zich daarom op het onderzoeken van de wortels van de menselijke ultrasocialiteit en zoekt naar een combinatie van uniek menselijke mechanismen die de evolutie van onze moraliteit zouden kunnen verklaren. Daarbij combineert het methodes en inzichten uit verschillende disciplines, aangezien het begrijpen van menselijk gedrag onvermijdelijk een interdisciplinaire aanpak vereist.

Hoofdstuk 2 gebruikt formele evolutionaire modellering om een veelvoorkomende belangrijke aanname in de literatuur over groepsselectie te onderzoeken. Het bestudeert theoretisch hoe het loslaten van de aanname van wereldwijde concurrentie tussen groepen de voorwaarden beïnvloedt waaronder samenwerking kan evolueren door groepsselectie. De analytische oplossingen en de simulatieresultaten tonen aan dat de evolutie van samenwerking aanzienlijk kan worden belemmerd door niet-globale concurrentie. Dit heeft belangrijke implicaties voor de empirische literatuur over het onderwerp, aangezien dit hoofdstuk laat zien dat het (expliciet of impliciet) aannemen van wereldwijde concurrentie

waarschijnlijk vertekende schattingen oplevert. Gezien het belang van groepsselectiemodellen voor het verklaren van sociaal gedrag in menselijke en andere dierenpopulaties, is het cruciaal om rekening te houden met de vertekening die wordt veroorzaakt door de aanname van wereldwijde concurrentie en om schattingen aan te passen op basis van de concurrentiestructuur van de relevante studiepopulatie.

Hoofdstuk 3 presenteert de resultaten van een laboratoriumexperiment dat onderzoekt hoe eerlijkheid voordelen kan opleveren bij partnerkeuze. In dit hoofdstuk onderzoeken we of eerlijkheid kan leiden tot de neiging om prosociaal te handelen in een situatie met asymmetrische informatie, en of deze neiging wordt verwacht door anderen en een voorkeur creëert voor eerlijke partners. Onze resultaten laten zien dat dit inderdaad het geval is. We testen ook of het verband tussen eerlijkheid en prosocialiteit via communicatie gaat, waarbij eerlijke individuen prosociaal handelen om te voorkomen dat ze tegen hun partners liegen. We vinden geen bewijs voor dit pad. In plaats daarvan suggereren onze bevindingen dat eerlijke individuen het in de eerste plaats moeilijk vinden om egoïstisch gedrag voor zichzelf te rechtvaardigen.

Hoofdstukken 4 en 5 geven een overzicht van de literatuur over de evolutie van sociale voorkeuren en beargumenteren dat, zonder commitment als een mechanisme te beschouwen, men de empirische observaties van laboratoriumexperimenten niet volledig kan verklaren. Hoofdstuk 4 bouwt voort op het werk van Frank (1987; 1988). Het doorloopt de meest gangbare theoretische modellen over de evolutie van sociale voorkeuren, gaat na of de voorspellingen van deze modellen overeenkomen met het gedrag in het lab, en beargumenteert dat commitment de sleutel is om de empirische gegevens te verklaren. Hoofdstuk 5 bestudeert ook de rol van commitment in sociale interacties, maar beperkt de aandacht tot rechtvaardigheidsvoorkeuren in het ultimatumspel. Het vat de meest voorkomende modellen in de literatuur samen die rechtvaardig gedrag in het ultimatumspel door middel van ruis of fouten verklaren, en updatet ze door meer algemene mutatiestructuren toe te staan. Vervolgens worden de modelvoorspellingen vergeleken met empirische waarnemingen van verschillende ultimatumspel experimenten. De conclusie is wederom dat een model met commitment het beste presteert in het omschrijven van de data. Wat de hoofdstukken 4 en 5 beogen te doen, is een verschuiving van de focus in de literatuur naar het bestuderen van welke rol commitment speelt in menselijke (sociale) interacties, en of en hoe we echte commitment detecteren.

Bibliography

Akdeniz, A. and van Veelen, M. (2020). The cancellation effect at the group level. *Evolution*, 74(7):1246–1254.

Akdeniz, A. and van Veelen, M. (2021). The evolution of morality and the role of commitment. *Evolutionary Human Sciences*, 3:e41.

Akdeniz, A. and van Veelen, M. (2022). Evolution and the ultimatum game: Why do people reject unfair offers? *in preparation*.

Aktipis, C. A. (2004). Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology*, 231(2):249–260.

Alger, I. and Weibull, J. W. (2012). A generalization of Hamilton's rule—love others how much? *Journal of Theoretical Biology*, 299:42–54.

Allen, B., Lippner, G., Chen, Y.-T., Fotouhi, B., Momeni, N., Yau, S.-T., and Nowak, M. A. (2017). Evolutionary dynamics on any population structure. *Nature*, 544(7649):227–230.

Allen, B. and McAvoy, A. (2019). A mathematical formalism for natural selection with arbitrary spatial and genetic structure. *Journal of mathematical biology*, 78(4):1147–1210.

Allen, B. and Nowak, M. A. (2012). Evolutionary shift dynamics on a cycle. *Journal of Theoretical Biology*, 311:28–39.

Allen, B. and Tarnita, C. E. (2014). Measures of success in a class of evolutionary models with fixed population size and structure. *Journal of Mathematical Biology*, 68:109–143.

Allen, B., Traulsen, A., Tarnita, C. E., and Nowak, M. A. (2012). How mutation affects evolutionary games on graphs. *Journal of Theoretical Biology*, 299:97–105.

Alós-Ferrer, C. and Farolfi, F. (2019). Trust, games, and beyond. *Frontiers in Neuroscience*, 13:887.

Alós-Ferrer, C. and Netzer, N. (2010). The logit-response dynamics. *Games and Economic Behavior*, 68(2):413–427.

Andersen, S., Ertaç, S., Gneezy, U., Hoffman, M., and List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review*, 101(7):3427–39.

Anderson, D. E., DePaulo, B. M., and Ansfield, M. E. (2002). The development of deception detection skill: A longitudinal study of same-sex friends. *Personality and Social Psychology Bulletin*, 28(4):536–545.

Anderson, S. P., Goeree, J. K., and Holt, C. A. (2004). Noisy directional learning and the logit equilibrium. *The Scandinavian Journal of Economics*, 106(3):581–602.

Andersson, O. and Wengström, E. (2012). Credible communication and cooperation: experimental evidence from multi-stage games. *Journal of Economic Behavior & Organization*, 81(1):207–219.

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review*, 85(4):891–904.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Aoki, K. and Nozawa, K. (1984). Average coefficient of relationship within troops of the japanese monkey and other primate species with reference to the possibility of group selection. *Primates*, 25(2):171–184.

Archetti, M. and Scheuring, I. (2012). Game theory of public goods in one-shot social dilemmas without assortment. *Journal of Theoretical Biology*, 299:9–20.

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.

Bader, F., Baumeister, B., Berger, R., and Keuschnigg, M. (2021). On the transportability of laboratory results. *Sociological Methods & Research*, 50(3):1452–1481.

Bahry, D. L. and Wilson, R. K. (2006). Confusion or fairness in the field? Rejections in the ultimatum game under the strategy method. *Journal of Economic Behavior & Organization*, 60(1):37–54.

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior*, 25(4):209–220.

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, 34(3):164–175.

Barclay, P. and Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, 274(1610):749–753.

Barmettler, F., Fehr, E., and Zehnder, C. (2012). Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory. *Games and Economic Behavior*, 75(1):17–34.

Baumard, N., André, J.-B., and Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(1):59–78.

Bayer, R.-C., Renner, E., and Sausgruber, R. (2013). Confusion and learning in the voluntary contributions game. *Experimental Economics*, 16(4):478–496.

Bell, A. V., Richerson, P. J., and McElreath, R. (2009). Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proceedings of the National Academy of Sciences*, 106(42):17671–17674.

Bendor, J. and Swistak, P. (1995). Types of evolutionary stability and the problem of cooperation. *Proceedings of the National Academy of Sciences*, 92(8):3596–3600.

Benndorf, V., Moellers, C., and Normann, H.-T. (2017). Experienced vs. inexperienced participants in the lab: Do they behave differently? *Journal of the Economic Science Association*, 3(1):12–25.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Bernhard, H., Fischbacher, U., and Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442(7105):912–915.

Bhattacharya, P., Nielsen, K., and Sengupta, A. (2020). Timing of communication. *The Economic Journal*, 130(630):1623–1649.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2020). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. *Available at SSRN 3326146*.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2):131–144.

Blume, A., DeJong, D. V., Kim, Y., and Sprinkle, G. B. (1998). Experimental evidence on the evolution of meaning of messages in sender-receiver games. *American Economic Review*, 88(5):1323–1340.

Blume, A. and Ortmann, A. (2007). The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria. *Journal of Economic Theory*, 132(1):274–290.

Böhm, R., Theelen, M. M. P., Rusch, H., and Van Lange, P. A. M. (2018). Costs, needs, and integration efforts shape helping behavior toward refugees. *Proceedings of the National Academy of Sciences*, 115(28):7284–7289.

Boles, T. L., Croson, R. T., and Murnighan, J. K. (2000). Deception and retribution in repeated ultimatum bargaining. *Organizational Behavior and Human Decision Processes*, 83(2):235–259.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Bornstein, G. and Yaniv, I. (1998). Individual and group behavior in the ultimatum game: are groups more "rational" players? *Experimental Economics*, 1(1):101–108.

Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314(5805):1569–1572.

Bowles, S. (2009). Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science*, 324(5932):1293–1298.

Boyd, R. (1982). Density-dependent mortality and the evolution of social interactions. *Animal Behaviour*, 30:972–982.

Boyd, R., Gintis, H., Bowles, S., and Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, 100(6):3531–3535.

Boyd, R. and Richerson, P. J. (2009). Culture and the evolution of human cooperation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533):3281–3288.

Brandt, H., Hauert, C., and Sigmund, K. (2003). Punishment and reputation in spatial public goods games. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 270(1519):1099–1104.

Brandt, H., Hauert, C., and Sigmund, K. (2006). Punishing and abstaining for public goods. *Proceedings of the National Academy of Sciences*, 103(2):495–497.

Brosnan, S. F. and De Waal, F. B. (2003). Monkeys reject unequal pay. *Nature*, 425(6955):297–299.

Brosnan, S. F., Schiff, H. C., and De Waal, F. B. (2005). Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society B: Biological Sciences*, 272(1560):253–258.

Burkart, J. M., Fehr, E., Efferson, C., and van Schaik, C. P. (2007). Other-regarding preferences in a non-human primate: Common marmosets provision food altruistically. *Proceedings of the National Academy of Sciences*, 104(50):19762–19766.

Burton-Chellew, M. N., El Mouden, C., and West, S. A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences*, 113(5):1291–1296.

Burton-Chellew, M. N. and West, S. A. (2013). Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences*, 110(1):216–221.

Cai, H. and Wang, J. T.-Y. (2006). Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36.

Call, J. and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5):187–192.

Camerer, C. F. (2013). Experimental, cultural, and neural evidence of deliberate prosociality. *Trends in Cognitive Sciences*, 17(3):106–108.

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from indonesia. *Economic Inquiry*, 37(1):47–59.

Carpenter, J., Verhoogen, E., and Burks, S. (2005a). The effect of stakes in distribution experiments. *Economics Letters*, 86(3):393–398.

Carpenter, J. P., Burks, S., and Verhoogen, E. (2005b). Comparing students to workers: The effects of social framing on behavior in distribution games. In *Field experiments in economics*. Emerald Group Publishing Limited.

Casella, A., Kartik, N., Sanchez, L., and Turban, S. (2018). Communication in context: Interpreting promises in an experiment on competition and trust. *Proceedings of the National Academy of Sciences*, 115(5):933–938.

Charness, G., Rigotti, L., and Rustichini, A. (2016). Social surplus determines cooperation rates in the one-shot prisoner's dilemma. *Games and Economic Behavior*, 100:113–124.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chew, S. H., Ebstein, R. P., and Zhong, S. (2013). Sex-hormone genes and gender difference in ultimatum game: Experimental evidence from china and israel. *Journal of Economic Behavior & Organization*, 90:28–42.

Clots-Figueras, I., González, R. H., and Kujal, P. (2016). Trust and trustworthiness under information asymmetry and ambiguity. *Economics Letters*, 147:168–170.

Clutton-Brock, T. (2009). Cooperation between non-kin in animal societies. *Nature*, 462(7269):51–57.

Cohn, A. and Maréchal, M. A. (2018). Laboratory measure of cheating predicts school misconduct. *The Economic Journal*, 128(615):2743–2754.

Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. *The adapted mind: Evolutionary psychology and the generation of culture*, 163:163–228.

Cottrell, C. A., Neuberg, S. L., and Li, N. P. (2007). What do people desire in others? a sociofunctional perspective on the importance of different valued characteristics. *Journal of Personality and Social Psychology*, 92(2):208.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260–281.

Cox, J. C., Friedman, D., and Sadiraj, V. (2008). Revealed altruism. *Econometrica*, 76(1):31–69.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica*, pages 1431–1451.

Cronk, L. (2007). The influence of cultural framing on play in the trust game: A maasai example. *Evolution and Human Behavior*, 28(5):352–358.

Crosetto, P., Weisel, O., and Winter, F. (2019). A flexible z-tree and otree implementation of the social value orientation slider measure. *Journal of Behavioral and Experimental Finance*, 23:46–53.

Croson, R. T. (1996). Information in ultimatum games: An experimental study. *Journal of Economic Behavior & Organization*, 30(2):197–212.

Crow, J. F. and Aoki, K. (1984). Group selection for a polygenic behavioral trait: estimating the degree of population subdivision. *Proceedings of the National Academy of Sciences*, 81(19):6073–6077.

Dai, Z., Galeotti, F., and Villeval, M. C. (2018). Cheating in the lab predicts fraud in the field: An experiment in public transportation. *Management Science*, 64(3):1081–1100.

Dal Bó, P. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American Economic Review*, 95(5):1591–1604.

Dal Bó, P. and Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Debove, S., André, J.-B., and Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society B: Biological Sciences*, 282(1808):20150392.

Debove, S., Baumard, N., and André, J.-B. (2016). Models of the evolution of fairness in the ultimatum game: a review and classification. *Evolution and Human Behavior*, 37(3):245–254.

Delton, A. W., Krasnow, M. M., Cosmides, L., and Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, 108(32):13335–13340.

Demiral, E. E. and Mollerstrom, J. (2020). The entitlement effect in the ultimatum game–does it even exist? *Journal of Economic Behavior & Organization*, 175:341–352.

DePaulo, B. M. (1994). Spotting lies: Can humans learn to do better? *Current directions in psychological science*, 3(3):83–86.

DePaulo, B. M. and Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology*, 74(1):63.

dos Santos, M., Rankin, D. J., and Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, 278(1704):371–377.

dos Santos, M., Rankin, D. J., and Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution*, 67(8):2446–2450.

dos Santos, M. and Wedekind, C. (2015). Reputation based on punishment rather than generosity allows for evolution of cooperation in sizable groups. *Evolution and Human Behavior*, 36(1):59–64.

Dreber, A., Rand, D. G., Fudenberg, D., and Nowak, M. A. (2008). Winners don't punish. *Nature*, 452(7185):348–351.

Drouvelis, M. and Sonnemans, J. (2017). The endowment effect in games. *European Economic Review*, 94:240–262.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, 32(11):3612–3628.

Durrett, R. (2008). *Probability models for DNA sequence evolution. Probability and its applications, 2nd Edition.* Springer, New York.

Ellingsen, T. and Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, 114(495):397–420.

Eshel, I. and Shaked, A. (2001). Partnership. *Journal of Theoretical Biology*, 208(4):457–474.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, 41(1):20–26.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Fischbacher, U., Fong, C. M., and Fehr, E. (2009). Fairness, errors and the power of competition. *Journal of Economic Behavior & Organization*, 72(1):527–545.

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fischer, A., Pollack, J., Thalmann, O., Nickel, B., and Pääbo, S. (2006). Demographic history and genetic differentiation in apes. *Current Biology*, 16(11):1133–1138.

Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, 102(19):7047–7049.

Frank, R. H. (1987). If *homo economicus* could choose his own utility function, would he want one with a conscience? *American Economic Review*, 77(4):593–604.

Frank, R. H. (1988). *Passions Within Reason: The strategic role of the emotions.* WW Norton & Co, New York.

Frank, R. H. (1994). Group selection and "genuine" altruism. *Behavioral and Brain Sciences*, 17(4):620–621.

Frank, R. H. (2001). Cooperation through emotional commitment. In Nesse, R., editor, *Evolution and the capacity for commitment*, volume 3 of *The Russell Sage Foundation series on trust*, pages 57–76. Russell Sage Foundation, New York.

Frank, R. H., Gilovich, T., and Regan, D. T. (1993). The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology*, 14(4):247–256.

Fudenberg, D. and Levine, D. (2008). *A long-run collaboration on long-run games*. World Scientific.

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–554.

Fujiwara-Greve, T. and Okuno-Fujiwara, M. (2009). Voluntarily separable repeated prisoner's dilemma. *The Review of Economic Studies*, 76(3):993–1021.

Gale, J., Binmore, K. G., and Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8(1):56–90.

Garcia, J. and Traulsen, A. (2012). Leaving the loners alone: Evolution of cooperation in the presence of antisocial punishment. *Journal of Theoretical Biology*, 307:168–173.

García, J. and van Veelen, M. (2016). In and out of equilibrium I: Evolution of strategies in repeated games with discounting. *Journal of Economic Theory*, 161:161–189.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Gneezy, U. and Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, 112(2):631–645.

Goeree, J. K., Holt, C. A., and Palfrey, T. R. (2016). Quantal response equilibrium. In *Quantal Response Equilibrium*. Princeton University Press.

Grafen, A. (1983). Natural selection, kin selection and group selection. In Kreps, J. and Davies, N., editors, *Behavioural Ecology, 2nd edn.*, pages 62–84. Blackwell.

Grafen, A. (2007). An inclusive fitness analysis of altruism on a cyclical network. *Journal of Evolutionary Biology*, 20(6):2278–2283.

Graham, J., Meindl, P., Koleva, S., Iyer, R., and Johnson, K. M. (2015). When values and behavior conflict: Moral pluralism and intrapersonal moral hypocrisy. *Social and Personality Psychology Compass*, 9(3):158–170.

Gross, J., Leib, M., Offerman, T., and Shalvi, S. (2018). Ethical free riding: When honest people find dishonest partners. *Psychological Science*, 29(12):1956–1968.

Gurven, M. and Winking, J. (2008). Collective action in action: Prosocial behavior in and out of the laboratory. *American Anthropologist*, 110(2):179–190.

Güth, W., Mugera, H., Musau, A., and Ploner, M. (2014). Deterministic versus probabilistic consequences of trust and trustworthiness: An experimental investigation. *Journal of Economic Psychology*, 42:28–40.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388.

Güth, W. and Yaari, M. (1992). An evolutionary approach to explain reciprocal behavior in a simple strategic game. In Witt, U., editor, *Explaining Process and Change–Approaches to Evolutionary Economics*, pages 23–34. University of Michigan Press, Ann Arbor.

Haidt, J. (2012). *The Righteous Mind: Why good people are divided by politics and religion.* Vintage.

Hamilton, W. D. (1964a). The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16.

Hamilton, W. D. (1964b). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1):17–52.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, 228(5277):1218–1220.

Hamilton, W. D. (1971). Selection of selfish and altruistic behaviour in some extreme models. In Eisenberg, J. and Dillon, W., editors, *Man and beast: comparative social behaviour.*

Handley, C. and Mathew, S. (2020). Human large-scale cooperation as a product of competition between cultural groups. *Nature Communications*, 11(1):1–9.

Hanna, R. and Wang, S. (2017). Dishonesty and selection into public service: Evidence from india. *American Economic Journal: Economic Policy*, 9(3):262–90.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396(6706):41–49.

Hanski, I. (1999). *Metapopulation ecology.* Oxford University Press.

Hauert, C., Haiden, N., and Sigmund, K. (2004). The dynamics of public goods. *Discrete and Continuous Dynamical Systems–Series B*, 4(3):575–587.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., and Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, 316(5833):1905–1907.

Heintz, C., Karabegovic, M., and Molnar, A. (2016). The co-evolution of honesty and strategic vigilance. *Frontiers in Psychology*, 7:1503.

Henrich, J. and Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208(1):79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6):795–855.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., et al. (2006). Costly punishment across human societies. *Science*, 312(5781):1767–1770.

Hilbe, C. and Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific Reports*, 2(1):458.

Hirshleifer, J. (1987). On the emotions as guarantors of threats and promises. In Dupré, J. E., editor, *The Latest on the Best: Essays on Evolution and Optimality*, pages 307––326. The MIT Press.

Hirshleifer, J. (2001). Game-theoretic interpretations of commitment. In Nesse, R., editor, *Evolution and the capacity for commitment*, volume 3 of *The Russell Sage Foundation series on trust*, pages 77–94. Russell Sage Foundation, New York.

Hofbauer, J. and Sandholm, W. H. (2002). On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294.

Hofmann, W., Wisneski, D. C., Brandt, M. J., and Skitka, L. J. (2014). Morality in everyday life. *Science*, 345(6202):1340–1343.

Howell, P. P. (1952). Observations on the Shilluk of the Upper Nile: The laws of homicide and the legal functions of the "Reth". *Africa*, 22(2):97–119.

Inaba, M., Inoue, Y., Akutsu, S., Takahashi, N., and Yamagishi, T. (2018). Preference and strategy in proposer's prosocial giving in the ultimatum game. *PloS One*, 13(3):e0193877.

Izquierdo, L. R., Izquierdo, S. S., and Vega-Redondo, F. (2014). Leave and let leave: A sufficient condition to explain the evolutionary emergence of cooperation. *Journal of Economic Dynamics and Control*, 46:91–113.

Izquierdo, S. S., Izquierdo, L. R., and Vega-Redondo, F. (2010). The option to leave: Conditional dissociation in the evolution of cooperation. *Journal of Theoretical Biology*, 267(1):76–84.

Jagau, S. and van Veelen, M. (2017). A general evolutionary framework for the role of intuition and deliberation in cooperation. *Nature Human Behaviour*, 1(8):1–6.

Jensen, K., Call, J., and Tomasello, M. (2007). Chimpanzees are rational maximizers in an ultimatum game. *Science*, 318(5847):107–109.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

Kagel, J. H., Kim, C., and Moser, D. (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior*, 13(1):100–110.

Kay, T., Keller, L., and Lehmann, L. (2020). The evolution of altruism and the serial rediscovery of the role of relatedness. *Proceedings of the National Academy of Sciences*, 117(46):28894–28898.

Keuschnigg, M., Bader, F., and Bracher, J. (2016). Using crowdsourced online experiments to study context-dependency of behavior. *Social Science Research*, 59:68–82.

Kiyonari, T. and Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, 95(4):826.

Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? we lie! why? experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.

Krakauer, D. C. and Pagel, M. (1995). Spatial structure and the evolution of honest cost-free signalling. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 260(1359):365–372.

Kriss, P. H., Nagel, R., and Weber, R. A. (2013). Implicit vs. explicit deception in ultimatum games with incomplete information. *Journal of Economic Behavior & Organization*, 93:337–346.

Langergraber, K., Schubert, G., Rowney, C., Wrangham, R., Zommers, Z., and Vigilant, L. (2011). Genetic differentiation and the evolution of cooperation in chimpanzees and humans. *Proceedings of the Royal Society B: Biological Sciences*, 278(1717):2546–2552.

Leib, M., Köbis, N., Soraperra, I., Weisel, O., and Shalvi, S. (2021). Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin*, 147(12):1241.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

Lieberman, E., Hauert, C., and Nowak, M. A. (2005). Evolutionary dynamics on graphs. *Nature*, 433(7023):312.

Lightner, A. D., Barclay, P., and Hagen, E. H. (2017). Radical framing effects in the ultimatum game: the impact of explicit culturally transmitted frames on economic decision-making. *Royal Society Open Science*, 4(12):170543.

Lohmann, R. I. (2014). A cultural mechanism to sustain peace: How the Asabano made and ended war. *Anthropologica*, pages 285–300.

Luo, S. (2014). A unifying framework reveals key properties of multilevel selection. *Journal of Theoretical Biology*, 341:41–52.

Luo, S. and Mattingly, J. C. (2017). Scaling limits of a model for selection at two scales. *Nonlinearity*, 30(4):1682.

Mailath, G. J. and Samuelson, L. (2006). *Repeated Games and Reputations*. Oxford University Press, Oxford.

Malécot, G. (1948). *Les Mathématiques de l'Hérédite*. Masson et Cie, Paris.

Maréchal, M. A., Cohn, A., Ugazio, G., and Ruff, C. C. (2017). Increasing honesty in humans with noninvasive brain stimulation. *Proceedings of the National Academy of Sciences*, 114(17):4360–4364.

Mathew, S. and Boyd, R. (2009). When does optional participation allow the evolution of cooperation? *Proceedings of the Royal Society B: Biological Sciences*, 276(1659):1167–1174.

Mathew, S., Boyd, R., and van Veelen, M. (2013). Human cooperation among kin and close associates may require enforcement of norms by third parties. In Richerson, P. J. and Christiansen, M. H., editors, *Cultural evolution: Society, technology, language, and religion*, pages 45–60. MIT Press.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38.

McNamara, J. M., Barta, Z., Fromhage, L., and Houston, A. I. (2008). The coevolution of choosiness and cooperation. *Nature*, 451(7175):189–192.

Melis, A. P., Hare, B., and Tomasello, M. (2006). Chimpanzees recruit the best collaborators. *Science*, 311(5765):1297–1300.

Melis, A. P. and Semmann, D. (2010). How is human cooperation different? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1553):2663–2674.

Miller, G. (2000). *The Mating Mind: How sexual choice shaped the evolution of human nature*. Doubleday & Co.

Molleman, L., Quiñones, A. E., and Weissing, F. J. (2013). Cultural evolution of cooperation: the interplay between forms of social learning and group selection. *Evolution and Human Behavior*, 34(5):342–349.

Murphy, R. O., Ackermann, K. A., and Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.

Nesse, R. (2001). Natural selection and the capacity for subjective commitment. In Nesse, R., editor, *Evolution and the capacity for commitment*, volume 3 of *The Russell Sage Foundation series on trust*, pages 1–44. Russell Sage Foundation, New York.

Nesse, R. et al. (2001). *Evolution and the capacity for commitment. Vol. 3 in the Russell Sage Foundation series on trust*. Russell Sage Foundation, New York.

Nowak, M. A. (2006). *Evolutionary dynamics*. Harvard University Press.

Nowak, M. A., Page, K. M., and Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, 289(5485):1773–1775.

Nowak, M. A., Tarnita, C. E., and Wilson, E. O. (2010). The evolution of eusociality. *Nature*, 466(7310):1057.

Ohtsuki, H. (2012). Does synergy rescue the evolution of cooperation? an analysis for homogeneous populations with non-overlapping generations. *Journal of Theoretical Biology*, 307:20–28.

Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502.

Ohtsuki, H. and Nowak, M. A. (2006). Evolutionary games on cycles. *Proceedings of the Royal Society of London B: Biological Sciences*, 273(1598):2249–2256.

Okasha, S. (2006). *Evolution and the levels of selection*. Oxford University Press.

Oosterbeek, H., Sloof, R., and Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188.

Palfrey, T. R. and Rosenthal, H. (1984). Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics*, 24(2):171–193.

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2):109–130.

Peysakhovich, A., Nowak, M. A., and Rand, D. G. (2014). Humans display a 'cooperative phenotype' that is domain general and temporally stable. *Nature Communications*, 5(1):1–8.

Pillutla, M. M. and Murnighan, J. K. (1995). Being fair or appearing fair: Strategic behavior in ultimatum bargaining. *Academy of Management Journal*, 38(5):1408–1426.

Pinker, S. (2015). The false allure of group selection. *The handbook of evolutionary psychology*, pages 1–14.

Potters, J. and Stoop, J. (2016). Do cheaters in the lab also cheat in the field? *European Economic Review*, 87:26–33.

Purzycki, B. G., Pisor, A. C., Apicella, C., Atkinson, Q., Cohen, E., Henrich, J., McElreath, R., McNamara, R. A., Norenzayan, A., Willard, A. K., and Xygalatas, D. (2018). The cognitive and cultural foundations of moral behavior. *Evolution and Human Behavior*, 39(5):490–501.

Queller, D. C. (1992). Does population viscosity promote kin selection? *Trends in Ecology & Evolution*, 7(10):322–324.

Queller, D. C. (1994). Genetic relatedness in viscous populations. *Evolutionary Ecology*, 8(1):70–73.

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, 492:110204.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., and Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, 110(7):2581–2586.

Regan, P. C., Levin, L., Sprecher, S., Christopher, F. S., and Gate, R. (2000). Partner preferences: What characteristics do men and women desire in their short-term sexual and long-term romantic partners? *Journal of Psychology & Human Sexuality*, 12(3):1–21.

Rich, P. and Zollman, K. J. S. (2016). Honesty through repeated interactions. *Journal of Theoretical Biology*, 395:238–244.

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., et al. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, 39.

Rousset, F. (2004). *Genetic structure and selection in subdivided populations*. Princeton University Press, Princeton, NJ.

Rousset, F. and Billiard, S. (2000). A theoretical basis for measures of kin selection in subdivided populations: finite populations and localized dispersal. *Journal of Evolutionary Biology*, 13(5):814–825.

Ruffle, B. J. (1998). More is better, but fair is fair: Tipping in dictator and ultimatum games. *Games and Economic Behavior*, 23(2):247–265.

Rusch, H. (2018). Ancestral kinship patterns substantially reduce the negative effect of increasing group size on incentives for public goods provision. *Journal of Economic Psychology*, 64:105–115.

Sánchez-Pagés, S. and Vorsatz, M. (2007). An experimental study of truth-telling in a sender–receiver game. *Games and Economic Behavior*, 61(1):86–112.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT press.

Santos, F. C. and Pacheco, J. M. (2005). Scale-free networks provide a unifying framework for the emergence of cooperation. *Physical Review Letters*, 95(9).

Santos, F. C., Santos, M. D., and Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201):213–216.

Scally, A., Yngvadottir, B., Xue, Y., Ayub, Q., Durbin, R., and Tyler-Smith, C. (2013). A genome-wide survey of genetic variation in gorillas using reduced representation sequencing. *PLoS One*, 8(6):e65066.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Schelling, T. C. (1978). Altruism, meanness, and other potentially strategic behaviors. *American Economic Review*, 68(2):229–230.

Schmitt, P. M. (2004). On perceptions of fairness: The role of valuations, outside options, and information in ultimatum bargaining games. *Experimental Economics*, 7(1):49–73.

Seed, A. and Byrne, R. (2010). Animal tool-use. *Current Biology*, 20(23):R1032–R1039.

Shalvi, S., Dana, J., Handgraaf, M. J. J., and De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Sherratt, T. N. and Roberts, G. (1998). The evolution of generosity and choosiness in cooperative exchanges. *Journal of Theoretical Biology*, 193(1):167–177.

Shumaker, R. W., Walkup, K. R., and Beck, B. B. (2011). *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press.

Sigmund, K., Hauert, C., and Nowak, M. A. (2001). Reward and punishment. *Proceedings of the National Academy of Sciences*, 98(19):10757–10762.

Silk, J. B. (2009). Social preferences in primates. In Glimcher, P. W., Fehr, E., Camerer, C. F., and Poldrack, R. A., editors, *Neuroeconomics: Decision making and the brain*, pages 269–284. Elsevier.

Silk, J. B. and House, B. R. (2011). Evolutionary foundations of human prosocial sentiments. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10910–10917.

Simon, B. (2010). A dynamical model of two-level selection. *Evolutionary Ecology Research*, 12(5):555–588.

Simon, B., Fletcher, J. A., and Doebeli, M. (2013). Towards a general theory of group selection. *Evolution*, 67:1561–1572.

Slonim, R. and Roth, A. E. (1998). Learning in high stakes ultimatum games: An experiment in the slovak republic. *Econometrica*, pages 569–596.

Smith, E. A. (2005). Making it real: Interpreting economic experiments. *Behavioral and Brain Sciences*, 28(6):832.

Sober, E. and Wilson, D. S. (1998). *Unto others: the evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Soltis, J., Boyd, R., and Richerson, P. J. (1995). Can group-functional behaviors evolve by cultural group selection?: An empirical test. *Current Anthropology*, 36(3):473–494.

Sterelny, K. (2012). *The evolved apprentice*. MIT press.

Stevens, J. R. and Hauser, M. D. (2004). Why be nice? psychological constraints on the evolution of cooperation. *Trends in Cognitive Sciences*, 8(2):60–65.

Stone, V. E., Cosmides, L., Tooby, J., Kroll, N., and Knight, R. T. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences*, 99(17):11531–11536.

Sullivan, M. S. (1994). Mate choice as an information gathering process under time constraint: implications for behaviour and signal design. *Animal Behaviour*, 47(1):141–151.

Sylwester, K. and Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, 6(5):659–662.

Taylor, P. D. (1992a). Altruism in viscous populations—an inclusive fitness model. *Evolutionary Ecology*, 6(4):352–356.

Taylor, P. D. (1992b). Inclusive fitness in a homogeneous environment. In *Proc. R. Soc. Lond. B*, volume 249, pages 299–302. The Royal Society.

Taylor, P. D., Day, T., and Wild, G. (2007a). Evolution of cooperation in a finite homogeneous graph. *Nature*, 447(7143):469–472.

Taylor, P. D., Day, T., and Wild, G. (2007b). From inclusive fitness to fixation probability in homogeneous structured populations. *Journal of Theoretical Biology*, 249(1):101–110.

Ten Brinke, L., Vohs, K. D., and Carney, D. R. (2016). Can ordinary people detect deception after all? *Trends in Cognitive Sciences*, 20(8):579–588.

Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard University Press.

Tomasello, M., Call, J., and Hare, B. (2003). Chimpanzees understand psychological states–the question is which ones and to what extent. *Trends in Cognitive Sciences*, 7(4):153–156.

Tomlin, D. (2015). Rational constraints and the evolution of fairness in the ultimatum game. *PloS One*, 10(7):e0134636.

Traulsen, A. and Nowak, M. A. (2006). Evolution of cooperation by multilevel selection. *Proceedings of the National Academy of Sciences*, 103(29):10952–10955.

van Leeuwen, B., Noussair, C. N., Offerman, T., Suetens, S., van Veelen, M., and van de Ven, J. (2018). Predictably angry–Facial cues provide a credible signal of destructive behavior. *Management Science*, 64(7):3364–3364.

van Veelen, M. (2006). Why kin and group selection models may not be enough to explain human other-regarding behaviour. *Journal of Theoretical Biology*, 242(3):790–797.

van Veelen, M. (2018). Can Hamilton's rule be violated? *eLife*, 7:e41901.

van Veelen, M. (2020). The problem with the Price equation. *Philosophical Transactions of the Royal Society B*, 375(1797):20190355.

van Veelen, M., Allen, B., Hoffman, M., Simon, B., and Veller, C. (2017a). Hamilton's rule. *Journal of Theoretical Biology*, 414:176–230.

van Veelen, M., Allen, B., Hoffman, M., Simon, B., and Veller, C. (2017b). Hamilton's rule. *Journal of Theoretical Biology*, 414:176–230.

van Veelen, M. and García, J. (2019). In and out of equilibrium II: Evolution in repeated games with discounting and complexity costs. *Games and Economic Behavior*, 115:113–130.

van Veelen, M., García, J., Rand, D. G., and Nowak, M. A. (2012). Direct reciprocity in structured populations. *Proceedings of the National Academy of Sciences*, 109(25):9929–9934.

van Veelen, M., Luo, S., and Simon, B. (2014). A simple model of group selection that cannot be analyzed with inclusive fitness. *Journal of Theoretical Biology*, 360:279–289.

Van Veelen, M. and Spreij, P. (2009). Evolution in games with a continuous action space. *Economic Theory*, 39(3):355–376.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Von Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1):1.

Vrij, A. and Mann, S. (2005). Police use of nonverbal behavior as indicators of deception.

Wade, M. J. (1978). A critical review of the models of group selection. *The Quarterly Review of Biology*, 53(2):101–114.

Walker, R. S. (2014). Amazonian horticulturalists live in larger, more related groups than hunter–gatherers. *Evolution and Human Behavior*, 35(5):384–388.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

West, S. A., Griffin, A. S., and Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2):415–432.

West, S. A., Pen, I., and Griffin, A. S. (2002). Cooperation and competition between relatives. *Science*, 296(5565):72–75.

Wiessner, P. (2019). Collective action for war and for peace a case study among the enga of Papua New Guinea. *Current Anthropology*.

Wiessner, P. et al. (2010). Youths, elders, and the wages of war in enga province, Papua New Guinea.

Wiessner, P. and Pupu, N. (2012). Toward peace: Foreign arms and indigenous institutions in a Papua New Guinea society. *Science*, 337(6102):1651–1654.

Williams, G. C. (1966). *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton: Princeton University Press.

Wilson, D. S., Pollock, G. B., and Dugatkin, L. A. (1992). Can altruism evolve in purely viscous populations? *Evolutionary Ecology*, 6(4):331–341.

Wilson, D. S. and Wilson, E. O. (2007). Rethinking the theoretical foundations of sociobiology. *Quarterly Review of Biology*, 82:327–348.

Wilson, M. L. and Wrangham, R. W. (2003). Intergroup relations in chimpanzees. *Annual Review of Anthropology*, 32(1):363–392.

Winking, J. and Mizer, N. (2013). Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, 34(4):288–293.

Wu, B., García, J., Hauert, C., and Traulsen, A. (2013). Extrapolating weak selection in evolutionary games. *PLoS Computational Biology*, 9(12):e1003381.

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., and Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences*, 106(28):11520–11523.

Yi, K.-O. (2005). Quantal-response equilibrium models of the ultimatum bargaining game. *Games and Economic Behavior*, 51(2):324–348.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. For a full list of PhD theses that appeared in the series we refer to List of PhD Theses – Tinbergen.nl. The following books recently appeared in the Tinbergen Institute Research Series:

760 A. RUSU, Essays in Public Economics

761 M.A. COTOFAN, Essays in Applied Microeconomics: Non-Monetary Incentives, Skill Formation, and Work Preferences

762 B.P.J. ANDRÉE, Theory and Application of Dynamic Spatial Time Series Models

763 P. PELZL, Macro Questions, Micro Data: The Effects of External Shocks on Firms

764 D.M. KUNST Essays on Technological Change, Skill Premia and Development

765 A.J. HUMMEL, Tax Policy in Imperfect Labor Markets

766 T. KLEIN, Essays in Competition Economics

767 M. VIGH, Climbing the Socioeconomic Ladder: Essays on Sanitation and Schooling

768 YAN XU, Eliciting Preferences and Private Information: Tell Me What You Like and What You Think

769 S. RELLSTAB, Balancing Paid Work and Unpaid Care over the Life-Cycle

770 Z. DENG, Empirical Studies in Health and Development Economics

771 L. KONG, Identification Robust Testing in Linear Factor Models

772 I. NEAMŢU, Unintended Consequences of Post-Crisis Banking Reforms

773 B. KLEIN TEESELINK, From Mice to Men: Field Studies in Behavioral Economics

774 B. TEREICK, Making Crowds Wiser: The Role of Incentives, Individual Biases, and Improved Aggregation

775 A. CASTELEIN, Models for Individual Responses

776 D. KOLESNYK, Consumer Disclosures on Social Media Platforms: A Global Investigation

777 M.A. ROLA-JANICKA, Essays on Financial Instability and Political Economy of Regulation

778 J.J. KLINGEN, Natural Experiments in Environmental and Transport Economics

779 E.M. ARTMANN, Educational Choices and Family Outcomes
780 F.J. OSTERMEIJER, Economic Analyses of Cars in the City
781 T. ÖZDEN, Adaptive Learning and Monetary Policy in DSGE Models
782 D. WANG, Empirical Studies in Financial Stability and Natural Capital
783 L.S. STEPHAN, Estimating Diffusion and Adoption Parameters in Networks New Estimation Approaches for the Latent-Diffusion-Observed-Adoption Model
784 S.R. MAYER, Essays in Financial Economics
785 A.R.S. WOERNER, Behavioral and Financial Change – Essays in Market Design
786 M. WIEGAND, Essays in Development Economics
787 L.M. TREUREN, Essays in Industrial Economics - Labor market imperfections, cartel stability, and public interest cartels
788 D.K. BRANDS, Economic Policies and Mobility Behaviour
789 H.T.T. NGUYEN, Words Matter? Gender Disparities in Speeches, Evaluation and Competitive Performance
790 C.A.P BURIK The Genetic Lottery. Essays on Genetics, Income, and Inequality
791 S.W.J. OLIJSLAGERS, The Economics of Climate Change: on the Role of Risk and Preferences
792 C.W.A. VAN DER KRAATS, On Inequalities in Well-Being and Human Capital Formation
793 Y. YUE, Essays on Risk and Econometrics
794 E.F. JANSSENS, Estimation and Identification in Macroeconomic Models with Incomplete Markets
795 P.B. KASTELEIN, Essays in Household Finance: Pension Funding, Housing and Consumption over the Life Cycle
796 J.O. OORSCHOT, Extremes in Statistics and Econometrics
797 S.D.T. HOEY, Economics on Ice: Research on Peer Effects, Rehiring Decisions and Worker Absenteeism
798 J. VIDIELLA-MARTIN, Levelling the Playing Field: Inequalities in early life conditions and policy responses
799 Y. XIAO, Fertility, parental investments and intergenerational mobility
800 X. YU, Decision Making under Different Circumstances: Uncertainty, Urgency, and Health Threat
801 G. GIANLUCA, Productivity and Strategies of Multiproduct Firms

802 H. KWEON, Biological Perspective of Socioeconomic Inequality

803 D.K. DIMITROV, Three Essays on the Optimal Allocation of Risk with Illiquidity, Intergenerational Sharing and Systemic Institutions

804 J.B. BLOOMFIELD, Essays on Early Childhood Interventions

805 S. YU, Trading and Clearing in Fast-Paced Markets

806 M.G. GREGORI, Advanced Measurement and Sampling for Marketing Research

807 O.C. SOONS, The Past, Present, and Future of the Euro Area

808 D. GARCES URZAINQUI The Distribution of Development. Essays on Economic Mobility, Inequality and Social Change

809 A.C. PEKER, Guess What I Think: Essays on the Wisdom in Meta-predictions

Economic models regularly make the assumption that individuals are rational, selfish, profit-maximizing agents. Although this provides a valuable benchmark for modeling purposes, it fails to capture actual human behavior in interpersonal interactions. Empirical studies on cooperative behaviors consistently show that most people care about the well-being of others, even if they do not know the identity of those others. This thesis contributes to the understanding of social behaviors and preferences by exploring their origins. Using a combination of formal evolutionary modeling, lab experiments and agent-based simulations, it investigates how human sociality evolved to the extent it did, while sociality in other species remained relatively restricted.

Aslıhan Akdeniz holds a BA in Economics from Boğaziçi University and an MPhil in Economics from Tinbergen Institute. In September, 2017, she joined the Department of Economics at the University of Amsterdam as a PhD candidate, where she wrote this thesis.