



<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Journal of Theoretical Biology*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Jansson, F. (2015)

What games support the evolution of an ingroup bias?.

Journal of Theoretical Biology, 373: 100-110

<http://dx.doi.org/10.1016/j.jtbi.2015.03.008>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-115894>

What Games Support the Evolution of an Ingroup Bias?

Fredrik Jansson

*Centre for the Study of Cultural Evolution, Stockholm University
School of Education, Culture and Communication, Mälardalen University
Institute for Analytical Sociology, Linköping University*

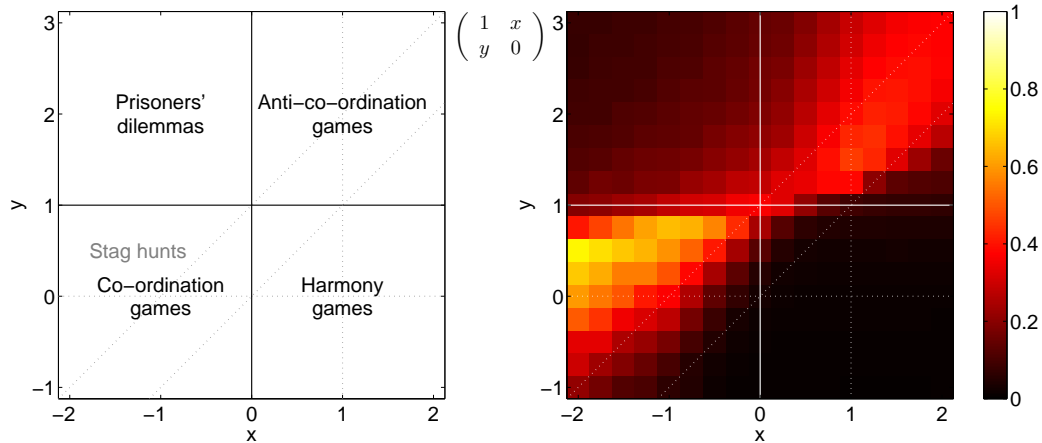
There is an increasing wealth of models trying to explain the evolution of group discrimination and an ingroup bias. This paper sets out to systematically investigate the most fundamental assumption in these models: in what kind of situations do the interactions take place? What strategic structures – games – support the evolution of an ingroup bias? More specifically, the aim here is to find the prerequisites for when a bias also with respect to minimal groups – arbitrarily defined groups void of group-specific qualities – is selected for, and which cannot be ascribed to kin selection.

Through analyses and simulations of minimal models of two-person games, this paper indicates that only some games are conducive to the evolution of ingroup favouritism. In particular, this class does not contain the prisoners' dilemma, but it does contain anti-co-ordination and co-ordination games. Contrasting to the prisoners' dilemma, these are games where it is not a matter of whether to behave altruistically, but rather one of predicting what the other person will be doing, and where I would benefit from you knowing my intentions.

In anti-co-ordination games, on average, not only will agents discriminate between groups, but also in such a way that their choices maximise the sum of the available pay-offs towards the ingroup more often than towards the outgroup. And in co-ordination games, even if agents do manage to co-ordinate with the whole population, they are more likely to co-ordinate on the socially optimal equilibrium within their group. Simulations show that this occurs most often in games where there is a component of risk-taking, and thus trust, involved. A typical such game is the stag hunt or assurance game.

Keywords ethnocentrism, minimal groups, cooperation, replicator dynamics, assurance game

Graphical Abstract Both analyses and simulations show that an ingroup bias evolves in (anti-)co-ordination games. The simulations further show that the strategy becomes particularly prevalent in stag hunts. The picture depicts, to the left, the games derived from the game matrix, in the middle, for different values of x and y . The panel to the right shows the simulated proportional prevalence of an ingroup bias for the different games when there are ten groups in the population.



1 Introduction

Human beings are often quick at dividing people into groups, implicitly or explicitly, and then let these divisions guide their behaviour towards them. More specifically, we tend to have an ingroup bias, meaning that we give preferential treatment to fellow group members.

The bias has been demonstrated in numerous settings, such as field studies and laboratory experiments (Brewer and Campbell, 1976; Kramer and Brewer, 1984; Yamagishi and Mifune, 2009; Balliet et al., 2014). The puzzle is that in some of these settings, people either lose in potential benefits from discriminating against outgroup members, or they take on net costs for helping out ingroup members when it would appear beneficial to abstain. Within small groups, apparently altruistic behaviour can often evolve by kin selection (Hamilton, 1964) or reciprocity (Trivers, 1971). In these groups, it would be more straightforward to use individual recognition rather than relying on weak group signals, and the ingroup bias observed would be an average preference based on whom people manage to co-operate with, rather than an evolved bias for how to behave beyond individual recognition. Meanwhile, people do display an ingroup bias also in situations where these mechanisms are not at work, and have shown to have preferences based purely on group signals. The bias can be triggered by minimal cues from arbitrary group definitions (Tajfel et al., 1971; Doise et al., 1972; Ahmed, 2007). What needs to be understood is thus how a bias that is activated among complete strangers has emerged. Co-operation can emerge as a spill-over effect from experiences from repeated interactions where it is rational (see e.g. Kiyonari et al., 2000; Rand et al., 2014), but it remains to explain the mechanisms that then lead to a bias towards strangers that is

dependent on minimal signals. There is evidence that the bias works on an implicit level (Otten and Wentura, 1999) and that it is regulated by the hormone oxytocin (De Dreu et al., 2011), suggesting deep biological roots. Thus, it seems reasonable to look for an adaptationist explanation.

The human species is not alone in giving preferential treatment to similar individuals. In this respect, the bias resembles the green-beard effect (Hamilton, 1964; Dawkins, 1976; Gardner and West, 2009; West and Gardner, 2010) that has been observed in less complex organisms (Keller and Ross, 1998; Queller et al., 2003). Individuals have phenotypes that other individuals can condition their behaviour on, with the result being preferential treatment towards individuals with a certain phenotype. However, the human bias stretches far beyond kin recognition, is highly flexible and applies also to cultural cues (Lindenfors, 2013). While theories on green-beards are concerned with how selective altruism can withstand invasion by cheaters (with the phenotype but without the co-operative genotype), for a bias that is activated for so many various situations as the human one, we likely need to extend the question beyond conditions for altruistic behaviour and ask, in general, in what situations does group discrimination give an evolutionary advantage, also without kin selection?

Defining situations, or interactions with strategic structures with consequences for the fitness of individuals, brings us into the realm of game theory. When accounting for selective altruism, some version of the *prisoners' dilemma* is assumed. In this situation, the ingroup bias can be formally expressed as a propensity to choose the individually costly but socially optimal co-operative strategy towards fellow group members, while choosing the individually rational defective strategy towards others.

In a one-shot game, an individual with such a bias has an evolutionary disadvantage to anyone defecting in both cases. Several evolutionary models of discriminating co-operative behaviour try to solve this by introducing elements of group selection (Wilson and Dugatkin, 1997; Eshel and Cavalli-Sforza, 1982; Bowles et al., 2003) or group conflict (Choi and Bowles, 2007; Lehmann and Feldman, 2008). The former models assume high cognitive demand, small groups and high degrees of between-group selection related to selection within the groups for free-riders to be kept at stake (although some conditions have been derived for when groups may be large, see Boyd and Richerson 1990). As for the latter models, it is controversial whether conflict is likely to have been a major mechanism in evolving an ingroup bias (Brewer and Caporael, 2006; Brewer, 1999; Brewer and Campbell, 1976; Yamagishi and Mifune, 2009; Halevy et al., 2008; Cashdan, 2001; Mäs and Dijkstra, 2014; Balliet et al., 2014). In the end, the phenomenon under study does include preferential treatment towards the ingroup, whether or not this entails hostility towards the outgroup, and a model is more parsimonious if it can explain the former without assuming the latter.

What these models, and other models taking departure in the one-shot prisoners' dilemma, have in common, is that the aim is to find the conditions under which what is played is no longer a dilemma. For example, in an infinitely repeated version of the game, the *folk theorem* states that co-operation is an equilibrium. Given the right (and sufficiently many) assumptions, the situation can be tweaked so that people play a game where co-operation is

rational within the group, while they still play the prisoners' dilemma between groups.

Before setting out to make assumptions that lead away from a social dilemma, we should know what situation to aim for. That is, what set of games support the evolution of an ingroup bias? A partial answer is those sets where we have different games for ingroup and outgroup interactions such that co-operation is rational in the former but not the latter. However, evidence suggests that the bias is activated also when the same game applies to both types of interactions. The aim here is thus to answer the question when all individuals play the same game.

Depending on the game in question, it is not always obvious how to define an ingroup bias. Let (p, q) be the probability that a random agent chooses the strategy that is most beneficial towards the partner from the ingroup (p) or the outgroup (q), respectively. We would then have group discrimination on the population level if $p \neq q$, and this would be an ingroup bias if $p > q$. Of course, what is beneficial needs to be defined, and will depend on the game. In general, this may be the socially optimal strategy, as in prisoners' dilemmas, but for other classes of games, such as *anti-co-ordination games*, where people are better off making different choices, it may be more reasonable to use another definition. We will return to making such a definition in Section 2.3.

First, theoretical analyses will be conducted to systematically define categories of two-person games that allow for group discrimination to evolve, and then, through simulations, we will find payoffs that optimally drive evolution towards an ingroup bias.

1.1 Previous Models

Previous models of the evolution of ingroup favouritism typically focus on a specific game, commonly the prisoners' dilemma.

A well-cited model was presented by Riolo et al. (2001), where agents have a visible marker on a continuum and co-operate with sufficiently similar others. The number of offspring is determined by the success of the interactions and offspring inherit marker and tolerance level, subject to mutations. The result is that co-operation is maintained within small tolerance levels, but as tolerance levels increase due to drift, mutants with lower tolerance levels invade and form new co-operative clusters consisting of their offspring. Thus, in this model, and typical for models in its wake, preferential treatment based on the marker is successful if and only if it correlates highly with relatedness, with signals being but proxies for kin recognition. Another restriction in this model is that co-operation relies on the fact that agents are not given the possibility of co-operating with no one (Roberts and Sherratt, 2002). Similar models have been developed where groups are many and small (Traulsen and Nowak, 2007), agents have different mutation rates for tags and strategies (Antal et al., 2009), or a reputation (Masuda and Ohtsuki, 2007).

In an elaborate mathematical treatment of a similar situation, Fu et al. (2012) derived conditions for when in-group favouritism emerges. There are circumstances that allow for discrimination also when full defection is an available strategy, but the crucial part is still that the number of available markers and mutation levels need to be so high that groups remain

small and shift signals before relatedness levels fall and the group can be overrun by false-bearded defectors. We should thus remain sceptical of any other accounts not acknowledging kin selection as the main drive of an ingroup bias in the prisoners' dilemma in the absence of mechanisms for direct reciprocity.

Another approach is spatially assorted interactions, where interaction partners are not drawn uniformly from the whole population, but where you have a higher probability of being matched with some agents than others, combined with a visible group tag (Jansen and van Baalen, 2006; Axelrod et al., 2004; Hammond and Axelrod, 2006). This allows for the groups to grow larger since you only interact locally in a small and constant fraction of the group. The mechanisms are still the same, with model designs that maintain a high concentration of kin within the interacting neighbourhoods. Also, when you interact mostly with your offspring, co-operation is more viable than defection, and group markers make a marginal change in the amount of co-operation taking place, by excluding neighbouring non-kin. See Read (2010) and Jansson (2013) for a further discussion on this type of models and an analysis of the model of Hammond and Axelrod (2006).

Unless we resort to group conflict (García and van den Bergh, 2011) and group selective accounts, a one-shot prisoners' dilemma will not form the breeding ground for group discrimination. There is a small, but growing, number of models explicitly based on other games. Colman et al. (2012) tested six games with specific payoffs that can be classified into versions of both dilemmas and co-ordination games. The analysis was however restricted to such small populations that inclusive fitness is at work. McElreath et al. (2003) looked at how markers can evolve as a co-ordination device also without requiring high concentration of kin. Axtell et al. (2001) showed that discrimination can also evolve in a hawk–dove game.

What this overview gives at hand is that even if a single interaction on the individual level may be a prisoners' dilemma, this is not the game that leads to group discrimination, but rather, the model must be designed so that the game that is being played cumulates to some other game. Indeed, group discrimination can emerge in models explicitly based on other games. This calls for a systematic treatment of what games are liable to preferential treatment. Such a treatment could inform modellers on specifically what strategic structure the agents in their models must face in the long run, and also contribute to the general understanding of basic adaptive mechanisms of discrimination.

1.2 Overview of the Paper

The objective here is to derive analytically the specific conditions for group discrimination to evolve with respect to all symmetric two-player games with two strategies, through replicator dynamics. We begin with a presentation on how to categorise games with respect to equilibria and then find conditions for each class of games. In order to incorporate mutations and random drift and their effects also on more than two groups, simulations are conducted for sets of payoffs that cover important subclasses of games.

2 Analysis

The full analysis can be found in the Appendix. The details are thus omitted here and the presentation focuses on the main findings.

2.1 Game Space

The presentation here follows Weibull (1995). All symmetric two-by-two games can be described by a payoff matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

where a_{ij} is the payoff to an agent choosing strategy i after an interaction with an agent choosing strategy j . In order to find evolutionary stable strategies in this game space, first note that evolutionary stable strategies (ESS) are defined in terms of payoff differences, so what matters for stability is not the absolute payoffs, but rather the differences in payoff between two strategies. By subtracting the left column, payoffs for responses to strategy 1, by a_{21} and the right column, responses to strategy 2, by a_{12} , we get a simpler normalised form

$$\begin{pmatrix} a_{11} - a_{21} & 0 \\ 0 & a_{22} - a_{12} \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}.$$

With respect to ESSs, there are only four (or three, modulo renamed strategies) classes of games: *prisoners' dilemmas* and *harmony games*, with one evolutionary stable equilibrium; *co-ordination games*, with two equilibria; and *anti-co-ordination games*, with three equilibria.

Naming strategy 1 co-operate and strategy 2 defect, all games where a_1 is negative and a_2 positive are *prisoners' dilemmas*, with defection always being a rational response, while shifting the signs makes it into a *harmony game*, an identical game, only with renamed strategies.

With positive payoffs on the diagonal, the best response is to do what the opponent does, resulting in a *co-ordination game*. Only these two pure strategies are ESSs. Apart from the ESSs, there is also a mixed strategy Nash equilibrium where strategy 1 is chosen with probability

$$\alpha = \frac{a_2}{a_1 + a_2}.$$

With negative payoffs on the diagonal, the best response is to do the opposite of what the opponent does, an *anti-co-ordination game*. In these games, α is an ESS. There are also two pure asymmetric equilibria.

2.2 Evolutionary Dynamics

We here look at the replicator dynamics for all the classes of games. For prisoners' dilemmas and harmony games, the whole population will converge to the single ESS, full co-operation and defection, respectively, in both types of interactions (assuming at least one individual with that strategy at time zero). It remains to analyse co-ordination and anti-co-ordination

games, first with respect to within-group interactions, then between-group interactions, and finally, assuming within- and between-group interactions at equilibrium, relative group sizes are computed, giving the relative success of the possible combinations of strategies.

2.2.1 Within-Group Interactions

Let $p_i(t)$ be the population share of individuals playing strategy i at time t . The population dynamics are

$$p_1' = (a_1 p_1 - a_2 p_2) p_1 p_2$$

and $p_2' = -p_1'$. There is an interior fixed point α . In co-ordination games, either of the ESSs will be reached, with the population drifting away from α , towards full dominance of strategy 1 if the initial share $p_1(0) > \alpha$ and strategy 2 if $p_1(0) < \alpha$. For anti-co-ordination games, α is instead an attractor, with a basin including all interior initial points, and the population will converge to the mixed strategy equilibrium.

2.2.2 Between-Group Interactions

Let $p_i(t)$ and $q_i(t)$ be the share of individuals in group X and Y , respectively, playing strategy i at time t . The replicator dynamics are

$$\begin{aligned} p_1' &= (a_1 q_1 - a_2 q_2) p_1 p_2, \\ q_1' &= (a_1 p_1 - a_2 q_2) q_1 q_2, \end{aligned}$$

$$p_2' = -p_1' \text{ and } q_2' = -q_1'.$$

For co-ordination games, there is a curve m that divides the strategy space into two basins of attraction. Except for strategies on the curve, which will converge to the saddle point (α, α) , both groups will converge to strategy 1 or 2 depending on whether (p_1, q_1) is above or below m , respectively. For further analyses below, note that m has tangent $q_1 = 2\alpha - p_2$ in the saddle point and that $p_1, q_1 > \alpha$ and $p_1, q_1 < \alpha$ are below and above m , respectively.

In anti-co-ordination games, if both groups employ the same frequency of strategy 1, then they will converge to the mixed strategy (α, α) . In all other cases, the population will converge to an asymmetric equilibrium. Strategies 1 and 2 will be reached by the group with the largest initial frequency.

In sum, prisoners' dilemmas and co-ordination games have the same dynamics within and between groups. Co-ordination games can result in one strategy towards the ingroup and another towards the outgroup, depending on initial conditions. Finally, in anti-co-ordination games, not only may agents have different in- and outgroup strategies, but on the group level, they will. Within the group, the only stable equilibrium is the symmetric mixed strategy, while encounters with another group allow for polarisation and specialisation. Contrasting to within-group interactions, between groups, all agents may encounter the same strategy in all interactions, for which the other strategy is the best response.

2.2.3 Mixed Interactions

It remains to investigate how groups that have arrived at one of the possible equilibria will fare against each other, that is, what will be the respective group sizes. Setting up replicator equations where both groups play equilibrium strategies and encounter both in- and outgroup members gives the following results.

As was derived above, in co-ordination games, both groups will have the same outgroup strategy, but may have converged on different ingroup strategies. If the two groups have different strategies, then the group with the highest-paying strategy will take over the whole population. If the groups have the same ingroup strategy, but this is different from the outgroup strategy, then the result depends on which of the strategies is more successful. If the outgroup strategy has a higher payoff, then the groups will converge to having the same size, while in the opposite case, the group with the largest initial population share will take over. With both payoffs being equal, all population ratios are stable.

These results are robust to assortment, that is, changing the frequency of which an agent meets an ingroup versus outgroup member, such that the probability for an ingroup member to be chosen as an interaction partner may differ from that of an outgroup member.

Within the groups, agents will on average play a mixed strategy in anti-co-ordination games, while the groups will choose different pure strategies between the groups. With the mixed strategies, the replicator dynamics are no longer invariant under a local shift of payoffs in the game matrix, so we need to revert to the original matrix and allow all four payoffs to be different in the analysis. This gives us two types of anti-co-ordination games, depending on whether one or both anti-co-ordination payoffs (a_{12} and a_{21}) are better than both co-ordination payoffs (a_{11} and a_{22}).

In games where one group would have earned more from the other group co-ordinating on their strategy than anti-co-ordinating, the former group will go extinct. These are the games where the diminishing group earns more from ingroup than outgroup interactions, and they constitute a well-known subclass of games commonly referred to as *hawk-dove games*.

In games where anti-co-ordination is always more profitable than co-ordination, both groups will survive, and the group playing strategy 1 towards the outgroup will converge to a share β of the population, where

$$\beta = \frac{(a_{12} - a_{11})(a_{12} - a_{22})}{(a_{12} - a_{11})(a_{12} - a_{22}) + (a_{21} - a_{11})(a_{21} - a_{22})}.$$

2.3 Conclusions

Let p^I and q^I denote the ratio of the respective populations of two groups playing strategy 1 towards the ingroup and p^O and q^O be the ratios towards the outgroup. To avoid fixed endpoints, assume that $p^I, q^I, p^O, q^O \notin \{0, 1\}$. For notational convenience, we will also assume that in the co-ordination games, (1, 1) is the socially optimal strategy, that is, $a_{11} > a_{22}$. Note that this still includes all co-ordination games, modulo renaming strategies. The notations α , β and m are as above. The results are compiled in Table 1.

Game	Strategies	Size of largest group	Initial conditions
PD	22	Any	Any
H	11	Any	Any
C	11	Any 1	$p^I > \alpha$ and or $q^I > \alpha, q^O > m(p^O)$
	21	$\frac{1}{2}$	$p^I, q^I < \alpha, q^O > m(p^O)$
	12	1	$p^I > \alpha$ or $q^I > \alpha, q^O < m(p^O)$
	22	Any	$p^I, q^I < \alpha, q^O < m(p^O)$
AC	$\alpha 1 / \alpha 2$	$\min(\max(\beta, 1 - \beta), 1)$	$q^O \neq m(p^O)$

Table 1: Strategies are denoted by ab , where a is the strategy towards the ingroup and b towards the outgroup. In the anti-co-ordination games, α denotes the mixed ESS strategy. In all other games, agents in both groups have the same strategy set. In the co-ordination games with the socially optimal strategy 11 for both in- and outgroups, one group overtakes the population if initially $p^I < \alpha$ or $q^I < \alpha$. Unstable nodes ($p^I = \alpha, q^I = \alpha$) and saddle points ($(p^O, q^O) = (\alpha, \alpha)$) are not included in the listing.

Neither prisoners' dilemmas nor harmony games offer any conditions for an ingroup bias, as in- and outgroup members are always treated the same.

In the co-ordination games, it suffices that any of the groups has a majority of agents choosing the strategy with the highest maximum payoff in the ingroup interactions for it to take over the whole population, while in outgroup interactions, the ratio must reach a threshold $q^O > m(p^O)$ in the population as a whole. This leads to the following.

Proposition. *In a co-ordination game, the basin of attraction for the ingroup strategy to be socially optimal at equilibrium is at least as large as that of the outgroup strategy.*

Proof. With the notation above, the socially optimal strategy is (1, 1). The ingroup strategy will converge to this if either $p^I > \alpha$ or $q^I > \alpha$, so the size of the basin of attraction is $1 - \alpha^2$. The outgroup strategy will converge to the same if and only if $q^O > m(p^O)$. We know that $q^O < m(p^O)$ if $p^O < \alpha$ and $q^O < \alpha$, and thus the area where $q^O > m(p^O)$ has a size no greater than $1 - \alpha^2$. \square

In anti-co-ordination games, almost all initial conditions result in a group distinctive strategy set, where, on the group level, agents use a mixed strategy towards the ingroup and a pure strategy towards the outgroup. This means that outgroup interactions are pareto optimal, while ingroup interactions are not always.

Define by a benevolent strategy one that maximises the sum of the payoffs that are available to the other agent. We then have the following.

Proposition. *In an anti-co-ordination game, at equilibrium, more agents will choose the benevolent strategy towards an ingroup member than an outgroup member.*

Proof. If $a_{21} > a_{11} > a_{12} > a_{22}$ or $a_{12} > a_{22} > a_{21} > a_{11}$ (one of the co-ordinated strategies is pareto optimal, also known as hawk-dove games), then the group choosing the non-benevolent outgroup strategy ("hawks") will take over the population, while the ingroup

strategy remains mixed and thus sometimes benevolent. In all other games (where coordination is never pareto optimal), the difference between the share of agents choosing strategy 1 towards the outgroup and the ingroup is

$$\begin{aligned}\alpha - \beta &= \frac{a_2}{a_1 + a_2} - \frac{(a_{12} - a_{11}) a_2}{(a_{21} - a_{22}) a_1 + (a_{12} - a_{11}) a_2} \\ &= \frac{a_2}{a_1 + a_2} - \frac{a_2}{\frac{a_{21} - a_{22}}{a_{12} - a_{11}} a_1 + a_2},\end{aligned}$$

which is positive if and only if $a_{21} - a_{22} > a_{12} - a_{11}$, which is equivalent to $a_{11} + a_{21} > a_{12} + a_{22}$. Thus, strategy 1 is more common in ingroup interactions if and only if it is benevolent. \square

The following is worth repeating from the above proof.

Corollary. *In a hawk-dove game, one group will overtake the population, playing a mixed strategy towards the ingroup and hawk towards the outgroup. Thus, agents are more inclined to play hawk towards an outgroup member.*

3 Simulations

The replicator dynamics assume several simplifications compared to previous models in the literature. The dynamics are deterministic in that they operate on selection on the group level rather than interactions among randomly selected pairs of agents and there are no mutations. We will here investigate what happens when we introduce mutations, immigration and more than two groups, by running simulations of an agent-based model.

In order to stay close to the analytical model when increasing the number of groups, agents do not distinguish between different outgroups, but have one strategy towards the ingroup, and one towards any outgroup. It would be possible to extend the number of groups to a continuous scale and add tolerance levels for accepting agents as ingroup members, but this would sacrifice model simplicity and introduce kin selection as a driving force, as discussed in Section 1.1 and Jansson (2013).

The model here largely follows the protocol of the Hammond and Axelrod (2006) model, but excludes the spatial structure that has also been shown to produce results based mainly on kin selection (Jansson, 2013).

Following a previous systematic approach to chart the spatial Hammond and Axelrod model for different games (Kaznatcheev, 2010), simulations will be run for a large enough set of payoff matrices to include various subclasses of the four games.

3.1 Description of Simulation Model

For each simulation, interactions are modelled by a specific game with payoffs held constant throughout the simulation. Different simulations are run for different payoff matrices. For each game reported on here, 500 simulations were run and the results averaged from the last

round in each simulation. One simulation, in turn, runs for 2,000 rounds and the outline of a round in a simulation is the following:

Immigration. An agent from a random group and with one random strategy (1 or 2) towards the ingroup and one towards the outgroup enters the population (which is empty in the beginning). The potential to reproduce (PTR) of all agents is (re)set to 0.12.

Interaction. For each agent, another agent is chosen at random for an interaction. The agent observes the group membership of the random partner and chooses a strategy accordingly. The agents receive the payoff of the interaction as an increase or decrease in PTR.

Reproduction. The density of the population is the number of agents divided by 2,500 (which thus caps the population size). The PTR of all agents is multiplied by one minus the density. Each agent produces one offspring with the probability of their new PTR. The offspring inherits the traits of its parent, but each of the traits group marker, ingroup strategy and outgroup strategy are subject to a mutation probability of 0.005 per trait, meaning that the marker would change into another randomly chosen marker and the strategy to the opposite strategy.

Death. Each agent has a probability of 0.1 to die.

In order to plot different games in a two-dimensional figure, note that

$$\begin{aligned} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} &= \begin{pmatrix} a_{11} - a_{22} & a_{12} - a_{22} \\ a_{21} - a_{22} & 0 \end{pmatrix} + a_{22} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \\ &= (a_{11} - a_{22}) \left(\begin{pmatrix} 1 & \frac{a_{12} - a_{22}}{a_{11} - a_{22}} \\ \frac{a_{21} - a_{22}}{a_{11} - a_{22}} & 0 \end{pmatrix} + a_{22} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right), \end{aligned}$$

provided that $a_{11} - a_{22} \neq 0$, so all games where the diagonal elements are not equal can be represented by the payoff matrix

$$\begin{pmatrix} 1 & a \\ b & 0 \end{pmatrix}.$$

To fit the magnitude of the PTR, the matrix will be multiplied by a scalar 0.06, such that the following payoff matrix will be used, for $x \in [-2, 2]$, $y \in [-1, 3]$:

$$0.06 \begin{pmatrix} 1 & x \\ y & 0 \end{pmatrix}.$$

With this payoff matrix, prisoners' dilemmas are the games where $x < 0, y > 1$, harmony games where $x > 0, y < 1$, co-ordination games where $x < 0, y < 1$ and anti-co-ordination games where $x > 0, y > 1$. This set of games can be further divided into subclasses of games depending on whether x and y are smaller or greater than 0 and 1, $y > x$ (which of the strategies gives the highest payoff when agents anti-co-ordinate), and $1 + x > y$ (which strategy is risk dominant).

3.2 Results

Figure 1 illustrates the average values, represented by colours, from 500 runs of the simulations for 17×17 different games in the range $x \in [-2, 2]$, $y \in [-1, 3]$. Simulation results

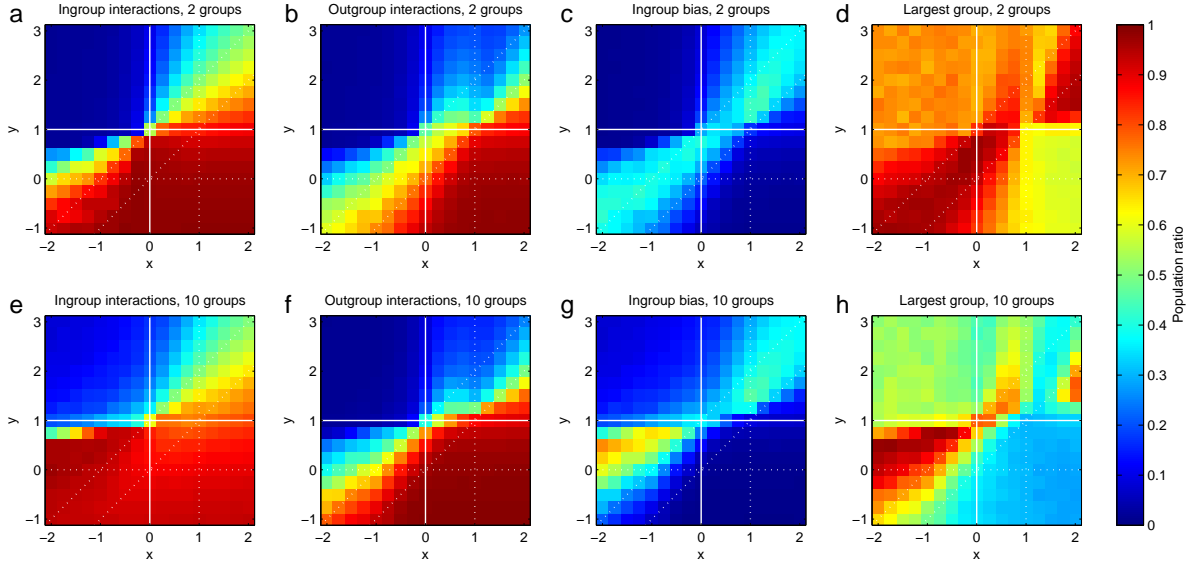


Figure 1: Simulation results. Average population ratios of strategy 1, “co-operation”, in interactions with only ingroup (panels a and e) and only outgroup members (b and f), strategy set (1, 2), “ingroup bias” (c and g), and largest group (d and h), for different games in the range $x \in [-2, 2]$, $y \in [-1, 3]$.

are given for populations consisting of two (panels a–d) and ten (panels e–f) groups. The values are the ratio of agents choosing strategy 1 (interpreted as co-operate in most games) in ingroup (a,e) and outgroup (b,f) interactions, and the prevalence of strategy set (1, 2) (interpreted as an ingroup bias in most games) (c,g). The figure also depicts the average size of the largest group in the last iteration (d,h).

In order to identify the different games more easily, each panel is divided by solid lines into four areas, each of which corresponds to the four classes of games (top left: prisoners’ dilemmas, top right: anti-co-ordination games, bottom left: co-ordination games, bottom right: harmony games). The area is further divided by dotted lines to identify subclasses of games, as described in the previous section and discussed further below. We can see the following.

First of all, we can note that prisoners’ dilemmas ($x < 0, y > 1$) result in no discrimination among two groups. With ten groups, however, the ingroup biased strategy is prevalent at low levels in the population, with a 10%–20% frequency. This is consistent with the fact that with many groups present, some of them will be small with high levels of kinship.

For the most part, harmony games ($x > 0, y < 1$) offer no ingroup bias. Remarkably, however, for the limited area $0 < x < y < 1$, about one third of the population has such a bias. It seems that here, the harmony game rather turns into a kind of harming game. Within the groups, the game remains a harmony game, but not between groups. Agents gain mutually by co-operating, but an agent that defects not only takes on a cost, but also imposes a larger cost on its partner (irrespective of whether that agent co-operates or defects). The dynamics may be quite intricate here and call for further research. A defecting group benefits against the outgroup, and its individuals will reproduce more. Meanwhile, the individual gains by

co-operating, so defecting is unsuccessful within groups, and cannot be sustained in the long run between groups, neither is it beneficial with many other groups around.

Co-ordination games ($x < 0, y < 1$) are the ones where we see the highest frequency of an ingroup bias. The highest values are attained where y is sufficiently smaller than 1 and $y > x$. A similar argument as above may be applied here. Also here, mutual co-operation has the highest payoff, but defection harms alter more than ego. Furthermore, mutual defection is a Nash equilibrium. For $y > x + 1$, defection is also risk-dominant. As y approaches 1, and the relative gain from taking the risk of co-operating diminishes, co-operation plummets also within groups, when there are only two groups available.

As the number of groups increases, however, the ingroup bias takes a more beneficial course. In the same area where co-operation is the mutually beneficial and defection the risk dominant equilibrium, agents start co-operating within the group. This is in line with the analytical result for two groups that only one group needs to reach the socially dominant equilibrium in ingroup interactions for the strategy to take over the population: the more groups, the more likely it is that at least one of them will succeed. The only area where agents with an ingroup bias constitute the majority of the population is where $x < 0$ and $x + 1 < y < 1$ (with ten groups). This is a well-studied class of games called stag hunt or assurance games. It appears that in the presence of competition from several groups, the population can avoid the risk dominant and reach the socially optimal equilibrium in interactions within the group.

For anti-co-ordination games ($x > 0, y > 1$), co-operation (playing strategy 1, the benevolent strategy in this area) is generally more likely towards the ingroup than the outgroup. For $x < y < x + 1$, playing strategy 1 towards the ingroup and 2 towards the outgroup is also the most successful strategy. This is particularly the case for $x > 1$, where anti-co-ordination is more beneficial for both parties than any co-ordination. Here we will always get anti-co-ordination between groups, with one group being better off than the other. When $y > x$, this will be the defecting group. Finally, if $y > x + 1$, then defection becomes risk dominant, and takes over also within groups, making agents less biased.

3.2.1 Convergence and Variation

The results presented above deal with average values after 2,000 rounds. Have the strategies converged to potential equilibria at this point and how do the results vary between simulations?

First, for most of the games, the average frequency of an ingroup bias has converged within 500 rounds, and the results look roughly the same as after the complete simulation. Convergence is slightly slower for borderline games, where x is close to 0 or y is close to 1, but average ratios change very little towards the end of the simulations.

Looking at each of the simulations, instead of average results, the prevalence of the ingroup bias changes by less than ten percentage units during the last 1,000 rounds in almost all of the runs of most of the games, except for borderline and co-ordination games. In the latter, ratios change more than ten percentage units in up to 50% of the runs, but this applies mainly

to games where the largest group takes over the population, with there being less selection on the outgroup strategy.

In general, there is little variation between simulation runs of two groups in the success of ingroup favouritism, with a standard deviation (often well) below 0.1, except for some borderline and co-ordination games. The average frequency is close to zero in prisoners' dilemmas and harmony games, with a standard deviation consequently on par. (Note however that the outcomes in the borderline harmony games where there is an ingroup bias vary widely.) The outcomes in anti-co-ordination games fall closely to the mean, with a narrow, approximately normal distribution. The co-ordination games, instead, have wide distributions, with standard deviations up to 0.3. All of these results, except for borderline harmony games, are consistent with the analytical findings.

The pattern is similar for ten groups, but with some increased variation in ingroup co-operation for prisoners' dilemmas. As discussed earlier, co-operation towards the ingroup can spread in small groups with a high density of kin.

3.2.2 Robustness

In the simulations presented here, all parameter values except for the payoff matrix were held constant. The parameters were set to the same values as those in the spatial Hammond and Axelrod (2006) model. Are the choices of these parameter values crucial to the results or is the model robust?

Keeping the PTR fixed, all other parameters can be modified to measure their relative impact. Increasing or decreasing the immigration rate has a similar effect to varying the mutation rate. It remains to investigate, then, whether the results are stable with respect to varying death and mutation rate.

The short answer is that extreme death and mutation rates can stimulate kin selection, but for other values, the results are robust against variation.

With two groups, the same pattern emerges independent of variable values as long as the death rate does not greatly exceed the PTR and the mutation rate is small (some percent). With a large death rate, discrimination can occur in any game due to random effects, while a high mutation rate eliminates it, since group marker no longer correlates with strategy. By decreasing the mutation rate, discrimination increases in the same games as in Figure 1.

Also with ten groups, the results are similar for small values of the parameters, except for death rates close to zero. Again, for large death rates, discrimination occurs randomly, while a high mutation rate eliminates it. For both small death rates and those slightly exceeding the PTR, discrimination is common in the prisoners' dilemma. These two extremes have a high relatedness coefficient in common: for small death rates, around one third of the population are related, and for high rates, more than one half. In the former case, since agents rarely die, the end result is highly dependent on initial conditions, for which the population consists of several small groups with high relatedness, in which ingroup discriminators are favoured. In the latter case, the explanation also spells small groups, since the high death rate maintains groups at a constant small size.

4 Conclusions

There is an increasing wealth of models trying to explain the evolution of group discrimination and an ingroup bias. This paper set out to systematically investigate the most fundamental assumption in these models: in what kind of situations do the interactions take place? What strategic structures support the evolution of an ingroup bias?

First of all, in order to study the effects of this very assumption, a minimal model with as few other assumptions as possible was constructed. In terms of the five mechanisms that can promote co-operation (Nowak, 2006), the model assumes no kin selection or reciprocity; spatial selection only in the form of group tags; and it allows for multilevel selection only in that agents benefit individually from being in a group that has co-ordinated on an optimal equilibrium. This model was analysed for the space of symmetric two-by-two games with two strategies, using replicator dynamics. This led to the general result that games of (anti-)co-ordination, with more than one equilibrium, are conducive to an ingroup bias, while games of co-operation are not. It was found that not only do games of co-ordination enable group discrimination, but on average they will make agents more prone to choose strategies that are favourable to their interlocutor when meeting an ingroup rather than an outgroup member. These results should shed some light on when and why an ingroup bias emerges. In games of co-ordination, there is a selective pressure for such a bias. It should be kept in mind that these games can take many faces. For example, repeated interactions or reputation can transform the one-shot prisoners' dilemma into a co-ordination game.

Anti-co-ordination games occur when there is a benefit of specialisation, and have a built-in mechanism for group discrimination: if agents can discriminate, then they will, on the group level, apply different strategies depending on group membership. Whether the resulting discrimination is truly an ingroup bias is a matter of definition – and it calls for a broader definition of the term beyond social dilemmas. One definition was suggested here, based on a concept of benevolent strategies. Within-group interactions did have the bias to more often lead to the strategy that maximises the sum of the payoffs available to the interlocutor.

It is perhaps not surprising that co-ordination games are also a breeding ground for group discrimination. If the exchange with certain groups is rare, then the parties may not know what to co-ordinate on and rather refrain from interacting at all, if anti-co-ordination is costly, or play a safe strategy. However, as was seen in the analysis here, discrimination is also possible at equilibrium, after ample contact. Discrimination favours the ingroup over the outgroup in that ingroup interactions are more likely to be socially optimal. Due to group competition, it suffices that one group converges to the socially optimal strategy for it to dominate the population. It should be noted that this group competition is not in the form of exploiting other groups, but rather having a selective advantage from the behaviour within the own group. Similar to ideas advanced already by Wright (1932), by a subdivision of the population into smaller groups, chances are higher that at least one of them will reach a higher peak in the fitness landscape.

The ingroup bias can also result from negative competition – competition that rather than enabling co-operation in a risky setting, increases defection when none such is to be expected.

ted in the absence of competition. When stochasticity was introduced into the model, in the form of drift, immigration and mutations, agents adopted a harming behaviour towards the outgroup in a small subset of harmony games. Group competition can take away the harmony from harmony games, but the very dynamics of this is an area for further research.

Another effect in the simulations with ten groups not accounted for in the two-group analyses is that the ingroup bias is a prevalent strategy also in prisoners' dilemmas, though at low levels. Adding more groups would increase the prevalence of the bias. This is to no surprise, however, since the ingroup bias can be accounted for by increased ingroup cooperation, which can in turn be accounted for by the fact that several of the groups will be small with high measures of kinship. These groups are not minimal, with a marker void of meaning, but an indicator of a shared gene pool.

The simulations indicate that certain co-ordination games may promote an ingroup bias more than others. Increasing the number of groups increases the range of games for which the socially optimal strategy dominates within-group interactions, while between-group interactions are not discernibly affected at all. An ingroup bias, in the sense of social optimality, occurs mainly when the strategy that may lead to the social optimum is also the strategy associated with the greatest risk. In particular, the difference between in- and outgroup strategy is largest for the class of games commonly referred to as stag hunt or assurance games, at least for the subset where the risk dominant equilibrium is not socially optimal.

Several authors have suggested that the stag hunt plays a key role in human interactions and may advance the understanding of social dilemmas (see for example Skyrms, 2004 and Kollock, 1998*a*). We have seen here that an ingroup bias may occur quite easily in these games, and there is also empirical evidence under a minimal group paradigm (Ahmed, 2007), and that expectations of trust are an important parameter (Jansson and Eriksson, Unpublished results). In experiments, arbitrary group tags are sufficient signals for people to trust each other on opting for the socially optimal strategy, and other empirical studies confirm the dynamics presented here: that intergroup competition may facilitate group co-ordination (Bornstein et al., 2002; Riechmann and Weimann, 2008).

In fact, experiments show that while being presented with a prisoners' dilemma in monetary terms, people often perceive such a situation subjectively as an assurance game (Kollock, 1998*b*; Kiyonari et al., 2000). Subjects are more liable to make this transformation when paired with an ingroup member (Kollock, 1998*b*), suggesting that people may be more likely to use a heuristic from repeated interactions in these cases, in turn suggesting that an ingroup bias might have emerged as a spill-over effect from different game structures on average within and between groups. However, the same experiments show that also outgroup interactions are commonly transformed into assurance games, and the models presented here show that we do not need to consider asymmetric games for an ingroup bias to evolve. Rather, the game that subjects play subjectively in experiments, is also the game where an ingroup bias is most liable in these models.

This may also provide explanations as to why ingroup favouritism has been observed also in prisoners' dilemmas. If an ingroup bias has evolved under circumstances where the assur-

ance game has been particularly prevalent, and people do perceive other games as assurance games, then we would in fact expect the bias to be activated also in these other games. The models presented here give conditions for when ingroup favouritism may evolve, but other processes, such as spill-over effects, could potentially elicit it under other circumstances. Experimental studies that compare the effects between different games are called for.

Future research on ingroup favouritism and group discrimination in general may be successful in adopting an increased focus on games of co-ordination where risk dominance competes with social optimality, such as the assurance game, and anti-co-ordination, rather than co-operation and defection, such as the prisoners' dilemma, without necessarily having to resort to complex social structures nor outgroup hostility.

Finally, the evidence for an innate propensity for ingroup favouritism suggests that interaction structures selecting for such a bias may have been important in our evolutionary past. With the ample indications of ingroup favouritism in society, there may also be a bias beyond biological roots, selected through cultural evolution and reflecting also current strategic situations in our everyday life. The models here investigate only symmetric two-by-two games with two strategies, but, unless people systematically play different games towards ingroup versus outgroup strangers, these can also be used as simplified models of more complex games including externalities such as reputation mechanisms. The results found here are also consistent with the empirical evidence for how people perceive interaction strategies subjectively, suggesting that games of co-ordination, in particular the stag hunt or assurance game, and anti-co-ordination games may have played, and may still play, a key role.

Acknowledgements

I am grateful to Kimmo Eriksson and Pontus Strimling for very useful discussions during the process of writing this manuscript. I would also like to thank two anonymous referees for insightful and valuable comments.

This research has been partly funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 324233.

References

- Ahmed, Ali M. (2007): Group identity, social distance and intergroup bias. *Journal of Economic Psychology* 28: 324–337
- Antal, Tibor, Hisashi Ohtsuki, John Wakeley, Peter D. Taylor and Martin A. Nowak (2009): Evolution of cooperation by phenotypic similarity. *Proceedings of the National Academy of Sciences* 106: 8597–8600
- Axelrod, Robert, Ross A. Hammond and Alan Grafen (2004): Altruism via Kin-Selection Strategies that Rely on Arbitrary Tags with which They Coevolve. *Evolution* 58: 1833–1838

- Axtell, Robert L., Joshua M. Epstein and H. Peyton Young (2001): The Emergence of Classes in a Multi-agent Bargaining Model. In Steven N Durlauf and H. Peyton Young (eds.), *Social Dynamics*, (pp. 191–211). MIT Press, Cambridge, Massachusetts
- Balliet, Daniel, Junhui Wu and Carsten K. W. De Dreu (2014): In-Group Favoritism in Cooperation: A Meta-Analysis. *Psychological Bulletin* 140(6): 1556–1581
- Bornstein, Gary, Uri Gneezy and Rosmarie Nagel (2002): The effect of intergroup competition on group coordination: an experimental study. *Games and Economic Behavior* 41: 1–25
- Bowles, Samuel, Jung-Kyoo Choi and Astrid Hopfensitz (2003): The co-evolution of individual behaviors and social institutions. *Journal of Theoretical Biology* 223: 135–147
- Boyd, Robert and Peter J. Richerson (1990): Group Selection among Alternative Evolutionarily Stable Strategies. *Journal of Theoretical Biology* 145: 331–342
- Brewer, Marilyn B. (1999): The Psychology of Prejudice: Ingroup Love or Outgroup Hate? *Journal of Social Issues* 55: 429–444
- Brewer, Marilyn B. and Donald T. Campbell (1976): *Ethnocentrism and Intergroup Attitudes: East African Evidence*. Halstead Press, New York, New York
- Brewer, Marilyn B. and Linnda R. Caporael (2006): An Evolutionary Perspective on Social Identity: Revisiting Groups. In D.T. Kenrick M. Schaller, J.A. Simpson (ed.), *Evolution and social psychology*, (pp. 143–161). Psychosocial Press, Madison, Connecticut
- Cashdan, Elizabeth (2001): Ethnocentrism and Xenophobia: A Cross-Cultural Study. *Current Anthropology* 42: 760–765
- Choi, Jung-Kyoo and Samuel Bowles (2007): The Coevolution of Parochial Altruism and War. *Science* 318(5850): 636–640
- Colman, Andrew M., Lindsay Browning and Briony D. Pulford (2012): Spontaneous similarity discrimination in the evolution of cooperation. *Journal of Theoretical Biology* 299: 162–171
- Dawkins, Richard (1976): *The Selfish Gene*. Oxford University Press, Oxford
- De Dreu, Carsten K. W., Lindred L. Greer, Gerben A. Van Kleef, Shaul Shalvi and Michel J. J. Handgraaf (2011): Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences of the United States of America* 108: 1262–1266
- Doise, Willem, Gyorgy Csepeli, Hans D. Dann, Celia Gouge, Knud S. Larsen and Alistair Ostell (1972): An experimental investigation into the formation of intergroup representations. *European Journal of Social Psychology* 2: 202–204
- Eshel, Ilan and L. Luca Cavalli-Sforza (1982): Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences of the United States of America* 79: 1331–1335
- Fu, Feng, Corina E. Tarnita, Nicholas A. Christakis, Long Wang, David G. Rand and Martin A. Nowak (2012): Evolution of in-group favoritism. *Scientific Reports* 2: 460
- García, Julián and Jeroen C. J. M. van den Bergh (2011): Evolution of parochial altruism by multilevel selection. *Evolution and Human Behavior* 32: 277–287
- Gardner, Andy and Stuart A. West (2009): Greenbeards. *Evolution* 64(1): 25–38
- Halevy, Nir, Gary Bornstein and Lilach Sagiv (2008): "In-Group Love" and "Out-Group Hate" as Motives for Individual Participation in Intergroup Conflict. *Psychological Science* 19: 405–411

- Hamilton, Willam D. (1964): The Genetical Evolution of Social Behaviour. I. *Journal of Theoretical Biology* 7: 1–16
- Hammond, Ross A. and Robert Axelrod (2006): The Evolution of Ethnocentrism. *Journal of Conflict Resolution* 50: 1–11
- Jansen, Vincent A. A. and Minus van Baalen (2006): Altruism through beard chromodynamics. *Nature* 440: 663–666
- Jansson, Fredrik (2013): Pitfalls in Spatial Modelling of Ethnocentrism. *Journal of Artificial Societies and Social Simulations* 16(3): 2
- Jansson, Fredrik and Kimmo Eriksson (Unpublished results): Cooperation and Shared Beliefs about Trust
- Kaznatcheev, Artem (2010): Robustness of Ethnocentrism to Changes in Interpersonal Interactions. *Artificial Intelligence* 31: 71–75
- Keller, Laurent and Kenneth G. Ross (1998): Selfish genes: a green beard in the red fire ant. *Nature* 394: 573–575
- Kiyonari, Toko, Shigehito Tanida and Toshio Yamagishi (2000): Social exchange and reciprocity: confusion or a heuristic? *Evolution and Human Behavior* 21: 411–427
- Kollock, Peter (1998a): Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24: 183–214
- (1998b): Transforming Social Dilemmas: Group Identity and Co-operation. In Peter Danielson (ed.), *Modeling Rationality, Morality, and Evolution*, (pp. 185–210). Oxford University Press, Oxford, United Kingdom
- Kramer, Roderick M. and Marilynn B. Brewer (1984): Effects of group identity on resource use in a simulated commons dilemma. *Journal of Personality and Social Psychology* 46: 1044–1057
- Lehmann, Laurent and Marcus W. Feldman (2008): War and the evolution of belligerence and bravery. *Proceedings of the Royal Society B* 275: 2877–2885
- Lindfors, Patrik (2013): The green beards of language. *Ecology and Evolution* 3(4): 1104–1112
- Mäs, Michael and Jacob Dijkstra (2014): Do Intergroup Conflicts Necessarily Result from Outgroup Hate? *PLoS ONE* 9(6): e97848
- Masuda, Naoki and Hisashi Ohtsuki (2007): Tag-based indirect reciprocity by incomplete social information. *Proceedings of the Royal Society B* 274
- McElreath, Richard, Robert Boyd and Peter J. Richerson (2003): Shared Norms and the Evolution of Ethnic Markers. *Current Anthropology* 44: 122–129
- Nowak, Martin A. (2006): Five Rules for the Evolution of Cooperation. *Science* 314(5805): 1560–1563
- Otten, Sabine and Dirk Wentura (1999): About the impact of automaticity in the minimal group paradigm: evidence from affective priming tasks. *European Journal of Social Psychology* 29(8): 1049–1071
- Queller, David C., Eleonora Ponte, Salvatore Bozzaro and Joan E. Strassmann (2003): Single-Gene Greenbeard Effects in the Social Amoeba *Dictyostelium discoideum*. *Science* 299: 105–106
- Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen

- Wurzbacher, Martin A. Nowak and Joshua D. Greene (2014): Social heuristics shape intuitive cooperation. *Nature Communications* 5: 3677
- Read, Dwight (2010): Agent-based and multi-agent simulations: coming of age or in search of an identity? *Computational and Mathematical Organization Theory* 16(4): 329–347
- Riechmann, Thomas and Joachim Weimann (2008): Competition as a coordination device: Experimental evidence from a minimum effort coordination game. *European Journal of Political Economy* 24: 437–454
- Riolo, Rick L., Michael D. Cohen and Robert Axelrod (2001): Evolution of cooperation without reciprocity. *Nature* 414: 441–443
- Roberts, Gilbert and Thomas N. Sherratt (2002): Does similarity breed cooperation? *Nature* 418: 499–500
- Skyrms, Brian (2004): *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge, United Kingdom
- Tajfel, Henri, Michael G. Billig, Robert P. Bundy and Claude Flament (1971): Social categorization in intergroup behavior. *European Journal of Social Psychology* 1: 149–178
- Traulsen, Arne and Martin A. Nowak (2007): Chromodynamics of Cooperation in Finite Populations. *PLoS ONE* 2: e270
- Trivers, Robert L. (1971): The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology* 46: 35–57
- Weibull, Jörgen W. (1995): *Evolutionary Game Theory*. The MIT Press, Cambridge, Massachusetts
- West, Stuart A. and Andy Gardner (2010): Altruism, spite, and greenbeards. *Science* 327(5971): 1341–1344
- Wilson, David Sloan and Lee A. Dugatkin (1997): Group Selection and Assortative Interactions. *The American Naturalist* 149: 336–351
- Wright, Sewall (1932): The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. In D.F. Jones (ed.), *Proceedings of the sixth international congress of genetics, 1*, (pp. 356–366). Brooklyn Botanic Garden, New York, New York
- Yamagishi, Toshio and Nobuhiro Mifune (2009): Social exchange and solidarity: in-group love or out-group hate? *Evolution and Human Behavior* 30: 229–237

Appendix

A Evolutionary Stable Strategies

All symmetric two-by-two games can be described by a payoff matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

where a_{ij} is the payoff to an agent choosing strategy s_i after an interaction with an agent choosing strategy s_j . In order to find evolutionary stable strategies in this game space, first note that evolutionary stable strategies (ESS) are defined in terms of payoff differences, so what matters for stability is not the absolute payoffs, but rather the differences in payoff between two strategies. By subtracting the left column, payoffs for responses to strategy s_1 , by a_{21} and the right column, responses to strategy s_2 , by a_{12} , we get a simpler normalised form

$$\begin{pmatrix} a_{11} - a_{21} & a_{12} - a_{12} \\ a_{21} - a_{21} & a_{22} - a_{12} \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}.$$

Following Weibull (1995), the game space can now be classified into only four categories, based on their ESSs, of which two can be merged into one.

Prisoners' dilemmas All games where a_1 is negative and a_2 positive have strategy s_2 as the only Nash equilibrium (NE), which is also an ESS. Naming strategy s_1 cooperate and strategy s_2 defect, these are the games where defection is always a rational response, thus referred to as prisoners' dilemmas.

Harmony games Shifting the signs, we get an identical game, only with renamed strategies. Strategy s_1 , cooperation, is the only ESS.

Coordination games With positive payoffs on the diagonal, the best response is to do what the opponent does. Pure strategies s_1 and s_2 , and a mixed one, where strategy s_1 is chosen with probability $\alpha = a_2/(a_1 + a_2)$, are all NE. However, agents with pure strategies earn the same against agents with the mixed strategy as they do, but more against themselves. Thus, only the pure strategies are ESSs.

Anti-co-ordination games With negative payoffs on the diagonal, the best response is to do the opposite of what the opponent does. Contrasting to the other categories, these games have two asymmetric equilibria apart from the symmetric mixed one, which is the same as in the co-ordination games. The mixed strategy is also an ESS. To see this, let s_α denote the mixed strategy where 1 is chosen with probability α and s_y be any strategy where 1 is chosen

with probability y . The payoff $\pi_{\alpha y}$ to s_α is independent of counterpart s_y , since

$$\begin{aligned}\pi_{\alpha y} &= \alpha y a_1 + (1 - \alpha)(1 - y)a_2 \\ &= \frac{a_2}{a_1 + a_2} y a_1 + \left(1 - \frac{a_2}{a_1 + a_2}\right) (1 - y)a_2 = \frac{a_1 a_2}{a_1 + a_2}.\end{aligned}$$

The payoff to s_y is

$$\pi_{yy} = y^2 a_1 + (1 - y)^2 a_2 \leq \pi_{\alpha y},$$

where equality is attained only when $s_y = s_\alpha$, since

$$\pi'_{yy} = 2y a_1 - 2(1 - y)a_2 = 0 \Leftrightarrow y = \frac{a_2}{a_1 + a_2} = \alpha,$$

which is a maximum, since $\pi''_{yy} = 2a_1 + 2a_2 < 0$.

B Replicator Dynamics

B.1 Within-Group Interactions

Let $x_i(t)$ be the number of individuals playing strategy s_i at time t , $p_i(t) = x_i(t)/(x_1(t) + x_2(t))$, and $p(t) = (p_1(t), p_2(t))$ be the population state vector (which gives a mixed strategy s_p), and π_{ij} be the payoff to an individual playing strategy s_i towards and individual playing s_j . Given a base fitness b and a death rate d , the population dynamics are

$$x'_i(t) = (b + \pi_{ip}(t) - d)x_i(t).$$

The proportion $p_i = x_i(t)/x(t)$ can then be simplified to

$$\begin{aligned}p'_i(t) &= \frac{x'_i(t)}{x(t)} - \frac{x_i(t)x'(t)}{x^2(t)} \\ &= \frac{(b + \pi_{ip}(t) - d)x_i(t)}{x(t)} - \frac{x_i(t)(b + \pi_{pp}(t) - d)}{x(t)} \\ &= (\pi_{ip}(t) - \pi_{pp}(t))p_i(t),\end{aligned}$$

which is thus independent of base fitness and death rate. Also the replicator dynamics are invariant to shifts of payoffs (adding a constant to any column or row will not change the difference $\pi_{ip} - \pi_{pp}$), so we can use the normalised matrix form. Given the game matrix

$$\begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix},$$

the replicator equation for the corresponding two-player game can be written as

$$\begin{aligned}p'_1 &= (a_1 p_1 - a_2 p_2)p_1 p_2 \\ &= (a_1 p_1 - a_2(1 - p_1))p_1(1 - p_1) \\ &= ((a_1 + a_2)p_1 - a_2)p_1(1 - p_1)\end{aligned}$$

and $p_2' = -p_1'$. Let $p'(t) = f(p(t))$. The fixed points are

$$f(p) = 0 \Leftrightarrow p_1 = 0 \vee p_1 = 1 \vee p_1 = \frac{a_2}{a_1 + a_2} = \alpha \text{ if } a_1 a_2 > 0$$

and $p_2 = 1 - p_1$. For determining stability of the fixed points, the derivate is

$$\begin{aligned} f'(p_1) &= (a_1 + a_2)p_1(1 - p_1) + ((a_1 + a_2)p_1 - a_2)(1 - 2p_1) \\ &= -3(a_1 + a_2)p_1^2 + 2(a_1 + 2a_2)p_1 - a_2, \end{aligned}$$

so the derivatives of the fixed points are

$$\begin{aligned} f'(0) &= -a_2, \\ f'(1) &= -a_1, \\ f'(\alpha) &= -3\frac{a_2^2}{a_1 + a_2} + 2a_2 + 2\frac{a_2^2}{a_1 + a_2} - a_2 = a_2 \left(1 - \frac{a_2}{a_1 + a_2}\right). \end{aligned}$$

Prisoners' dilemmas ($a_1 < 0, a_2 > 0$) The fixed point 0 is stable, since $f'(0) < 0$, while 1 is unstable, since $f'(1) > 0$. Thus, the population share p_1 always declines, and the population will reach the ESS.

Harmony games ($a_1 < 0, a_2 > 0$) These games are identical to prisoners' dilemmas, but with strategies renamed. The population share p_1 will always grow, towards the ESS.

Co-ordination games ($a_1 > 0, a_2 > 0$) The interval endpoints are stable ($f'(0), f'(1) < 0$), while the interior fixed point is unstable ($f'(\alpha) > 0$). The population will converge to one of the ESSs, either strategy s_1 or strategy s_2 dominance, depending on whether p_1 is greater or smaller than α , respectively.

Anti-co-ordination games ($a_1 < 0, a_2 < 0$) This case is the opposite of the co-ordination games, with unstable interval endpoints and a stable interior point. Hence, given any share $p_1 \notin \{0, 1\}$, the population will converge to the ESS α .

B.2 Between-Group Interactions

Let $x_i(t)$ and $y_i(t)$ be the number of individuals in group X and Y , respectively, playing strategy s_i at time t , and let $p_i(t) = x_i(t)/(x_1(t) + x_2(t))$ and $q_i(t) = y_i(t)/(y_1(t) + y_2(t))$ be their associated population shares. With the same payoff matrix for both groups, the replicator dynamics are

$$\begin{aligned} p_1' &= (a_1 q_1 - a_2 q_2) p_1 p_2 = ((a_1 + a_2) q_1 - a_2) p_1 (1 - p_1) =: f(p_1, q_1), \\ q_1' &= (a_1 p_1 - a_2 p_2) q_1 q_2 = ((a_1 + a_2) p_1 - a_2) q_1 (1 - q_1) =: g(p_1, q_1), \end{aligned}$$

$p'_2 = -p'_1$ and $q'_2 = -q'_1$. The fixed points are all the corners (0,0), (0,1), (1,0) and (1,1), and the interior point (α, α) when $a_1 a_2 > 0$.

The Jacobian of the system is

$$\begin{aligned} A(p_1, q_1) &= \begin{pmatrix} f'_{p_1}(p_1, q_1) & f'_{q_1}(p_1, q_1) \\ g'_{p_1}(p_1, q_1) & g'_{q_1}(p_1, q_1) \end{pmatrix} \\ &= \begin{pmatrix} ((a_1 + a_2)q_1 - a_2)(1 - 2p_1) & (a_1 + a_2)p_1(1 - p_1) \\ (a_1 + a_2)q_1(1 - q_1) & ((a_1 + a_2)p_1 - a_2)(1 - 2q_1) \end{pmatrix}. \end{aligned}$$

The Jacobians at the fixed points are thus

$$\begin{aligned} A(0, 0) &= \begin{pmatrix} -a_2 & 0 \\ 0 & -a_2 \end{pmatrix}, \quad A(0, 1) = \begin{pmatrix} a_1 & 0 \\ 0 & a_2 \end{pmatrix}, \quad A(1, 0) = \begin{pmatrix} a_2 & 0 \\ 0 & a_1 \end{pmatrix}, \\ A(1, 1) &= \begin{pmatrix} -a_1 & 0 \\ 0 & -a_1 \end{pmatrix}, \quad A(\alpha, \alpha) = \begin{pmatrix} 0 & a_1 \alpha \\ a_1 \alpha & 0 \end{pmatrix}. \end{aligned}$$

The eigenvalues of these matrices are given by the following equations:

$$\begin{aligned} |A(0, 0) - \lambda I| &= (-a_2 - \lambda)^2 = 0 && \Leftrightarrow \lambda = -a_2, \\ |A(0, 1) - \lambda I| &= (a_1 - \lambda)(a_2 - \lambda) = 0 && \Leftrightarrow \lambda \in \{a_1, a_2\}, \\ |A(1, 0) - \lambda I| &= (a_2 - \lambda)(a_1 - \lambda) = 0 && \Leftrightarrow \lambda \in \{a_1, a_2\}, \\ |A(1, 1) - \lambda I| &= (-a_1 - \lambda)^2 = 0 && \Leftrightarrow \lambda = -a_1, \\ |A(\alpha, \alpha) - \lambda I| &= (-\lambda)^2 - a_1^2 \alpha^2 = 0 && \Leftrightarrow \lambda = \pm \frac{a_1 a_2}{a_1 + a_2}. \end{aligned}$$

Since all eigenvalues are real, the fixed points are nodes and saddle points.

Prisoners' dilemmas ($a_1 < 0$, $a_2 > 0$) The point (1,1) is an unstable node, (0,1) and (1,0) are saddle points, and (0,0) is a stable node. Thus, as long as there exists a defector in any of the groups, the whole group will converge to defection, slightly faster in the group with most defectors.

Harmony games ($a_1 < 0$, $a_2 > 0$) This case is identical to the previous one, modulo renaming of strategies. The whole population will converge to full cooperation.

Co-ordination games ($a_1 > 0$, $a_2 > 0$) The points (0,1) and (1,0) are unstable nodes, (α, α) is a saddle point and both (0,0) and (1,1) are stable nodes. Except for points on the stable manifold to the saddle point, the population will converge to either of the pure equilibria. If $p_1 < \alpha$, then q_1 will decrease, and vice versa. If $p_1 > \alpha$, then q_1 will increase, and vice versa. So, if both $p_1 < \alpha$ and $q_1 < \alpha$, then the population will converge to (0,0), while if $p_1 > \alpha$ and $q_1 > \alpha$, then the population will converge to (1,1). For the other two quadrants, the outcome is not as easily predicted, but at (α, α) , the tangent to the separatrix between the two basins of attraction (the stable manifold, the only curve attracted by the saddle point) is $q_1 = 2\alpha - p_1$, since the eigenvector associated with the negative eigenvalue of $A(\alpha, \alpha)$ is $t(1, -1)$, $t \in \mathbb{R}$.

Anti-co-ordination games ($a_1 < 0, a_2 < 0$) The points (0,0) and (1,1) are unstable nodes, (α, α) is a saddle point and both (0,1) and (1,0) are stable nodes. The stable manifold of the saddle point is the line $p_1 = q_1$. If $p_1 < q_1$, then the population will converge to the point (0,1), and for $p_1 > q_1$ to the point (1,0).

B.3 Mixed interactions

Let p and $1-p$ be the proportion of agents in group X and Y , respectively. Assuming all agents play a strategy at equilibrium, all population distributions are stable in prisoners' dilemmas and harmony games, since all agents will receive equal payoffs. It remains to analyse the dynamics of co-ordination and anti-co-ordination games.

B.3.1 Co-ordination games ($a_1 > 0, a_2 > 0$)

In a co-ordination game, the population may converge to either of the two pure strategies. Within the group, the groups may converge to different equilibria, but the between-group strategy will be the same for both. Let a and c be the payoffs of the respective ingroup interactions and b the common payoff of between-group interactions. Assume that an agent meets any other agent with equal probability, irrespective of group membership. The replicator dynamics are then:

$$\begin{aligned} p' &= (pa + (1-p)b - ((1-p)c + pb))(1-p)p \\ &= ((a - 2b + c)p + b - c)(1-p)p \\ &= (a - 2b + c) \left(p + \frac{b - c}{a - 2b + c} \right) (1-p)p, \end{aligned}$$

given that a, b and c are not all equal (in which case $p' \equiv 0$). We have two cases: either the two ingroup strategies are equal, or (at least) one of the ingroup strategies is equal to the outgroup strategy.

Equal ingroup strategies Assume $a = c$. The dynamics become

$$p' = 2(b - a)(p - 1) \left(p - \frac{1}{2} \right) p.$$

Hence, the fixed points are 0, $\frac{1}{2}$ and 1. If $b > a$, then $\frac{1}{2}$ is stable, while the endpoints are unstable, and for $a > b$, we have the opposite. Thus, if outgroup interactions are more successful than those with the ingroup, then the groups will reach equal size, while if ingroup interactions have a more favourable payoff, then the initially largest group will take over the population.

One of the ingroup strategies equal to the outgroup strategy Assume $b = c$. The dynamics becomes

$$p' = (b - a)(p - 1)p^2.$$

If $b > a$, then p will converge to 0, and with $a > b$ to 1. Thus, the group with the highest payoff from ingroup interactions will take over.

Weighted interactions The assumption that agents from both groups meet with equal probability can be relaxed. Multiplying the frequency of ingroup interactions by a weight w , such that an ingroup member is w times more likely to be selected for a random interaction, gives the replicator dynamics

$$p' = \left(\frac{wpa + (1-p)b}{1 + (w-1)p} - \frac{w(1-p)c + pb}{w - (w-1)p} \right) (1-p)p.$$

For $a = c$ the equation simplifies to (calculations are omitted)

$$p' = \frac{w}{(1 + (w-1)p)(w - (w-1)p)} 2(b-a)(p-1) \left(p - \frac{1}{2} \right) p$$

and for $b = c$

$$p' = \frac{w}{1 + (w-1)p} (b-a)(p-1)p^2,$$

which both have the same fixed points with the same stability properties as for $w = 1$.

B.3.2 Anti-co-ordination games ($a_1 < 0, a_2 < 0$)

Within the groups, agents will play a mixed strategy, while the groups will choose different pure strategies between the groups. With the mixed strategies, the replicator dynamics are no longer invariant under a local shift of payoffs in the game matrix. Reverting to the original matrix, and assuming, without loss of generality, that group X plays strategy s_1 against Y , the dynamics are

$$p' = \left(p(\alpha^2 a_{11} + \alpha(1-\alpha)(a_{12} + a_{21}) + (1-\alpha)^2 a_{22}) + (1-p)a_{12} - ((1-p)(\alpha^2 a + \alpha(1-\alpha)(a_{12} + a_{21}) + (1-\alpha)^2 a_{22}) + pa_{21}) \right) (1-p)p,$$

which simplifies to (calculations are omitted)

$$p' = \frac{(a_{12} - a_{11})(a_{12} - a_{22}) + (a_{21} - a_{11})(a_{21} - a_{22})}{a_{12} - a_{22} + a_{21} - a_{11}} \cdot \left(p - \frac{(a_{12} - a_{11})(a_{12} - a_{22})}{\underbrace{(a_{12} - a_{11})(a_{12} - a_{22}) + (a_{21} - a_{11})(a_{21} - a_{22})}_{\beta}} \right) (p-1)p.$$

Note that in an anti-co-ordination game, $a_{21} - a_{11} > 0$ and $a_{12} - a_{22} > 0$. Also note that this implies that $a_{12} - a_{11}$ and $a_{21} - a_{22}$ cannot both be negative. Thus, there exists an interior fixed point β if and only if both differences are positive. To see this, we can write β as

$$\beta = \frac{a}{a+b} \Leftrightarrow \frac{b}{a} = \frac{1-\beta}{\beta},$$

which is positive if and only if $\beta \in (0, 1)$.

Let $p'(t) = f(p(t))$. The derivatives of the fixed points are

$$\begin{aligned} f'(0) &= \frac{(a_{12} - a_{11})(a_{12} - a_{22})}{a_{12} - a_{22} + a_{21} - a_{11}} < 0 \text{ if } a_{12} - a_{11} < 0, \\ f'(1) &= \frac{(a_{21} - a_{11})(a_{21} - a_{22})}{a_{12} - a_{22} + a_{21} - a_{11}} < 0 \text{ if } a_{21} - a_{22} < 0 \text{ and} \\ f'(\beta) &= -\frac{(a_{21} - a_{11})(a_{21} - a_{22})}{a_{12} - a_{22} + a_{21} - a_{11}}\beta < 0, \end{aligned}$$

which is defined for

$$\beta \in (0, 1) \Leftrightarrow a_{12} - a_{11} > 0 \wedge a_{21} - a_{22} > 0.$$

Thus, in games where one group would have earned more from the other group co-ordinating on the strategy of the former, the former group will go extinct. These are the games where the diminishing group earns more from ingroup than outgroup interactions, commonly known as hawk-dove games. In games where anti-co-ordination is always more profitable than co-ordination, both groups will survive, and the group playing strategy s_1 towards the outgroup will converge to a share β of the population.