1. Introduction

Humans see other humans as members of groups, and they treat them differently based on their group membership. Cognitively, they classify humans as members of different groups, form stereotypes about other groups, and track group reputations. Motivationally, they form positive or negative attitudes to other groups and may discriminate against some groups. They may also reciprocate towards groups as a whole: that is, they may respond to one person's actions by acting differently towards his or her fellow group members. We have experimental evidence for group reciprocity, but so far there is no theoretical account of how it might have evolved. This paper sets out a theoretical model in which group reciprocity can evolve. The requirements are quite minimal. Individuals know other individuals' group membership, and they can recall outgroup members' past behaviour towards them. That is enough for group reciprocity to evolve in a repeated prisoner's dilemma. So, intergroup cognition alone is enough to make intergroup reciprocity evolutionarily stable.

Group reciprocity is important in several ways. Theoretically, interethnic peace may be sustained by the threat of reciprocation. This could have allowed humans to live in peace with neighbouring groups (perhaps punctuated by episodes of punitive conflict), whereas chimpanzees always attack neighbouring bands on contact. It could also enable deeper forms of intergroup cooperation, such as trade. So, the capacity for group reciprocity may be a stepping stone in the evolution of humans as a cooperative species.

Group reciprocity is also related to generalized upstream reciprocity (XXX). The literature has concluded that it is hard for generalized upstream reciprocity to evolve except in special circumstances. So, our model could explain the existence of generalized upstream reciprocity in a limited form.

Practically, group reciprocity may underlie modern conflicts, including interethnic violence horowitz1985ethnicgroups,horowitz2001thedeadly and civil wars haushofer$_both_2010.fearon1996explainingshowhowinterethnicpeacemaybesustainedbyth$

Our result comes about as follows. Groups of group-reciprocators are able to establish cooperative relationships with other groups of group-reciprocators, while they avoid cooperating with groups of selfish types. Selfish types within a group of mostly group-reciprocators get the highest payoff, but this is balanced out because most selfish types are in mostly-selfish groups, who have too few reciprocators to sustain cooperation.

We also use computations to check the robustness of our theoretical model. (XXX result preview.)

Our model shows how group reciprocity can evolve, and it also has some insights about how its form is shaped by evolutionary pressure. The most robust forms of group reciprocity have high thresholds of cooperation - that is, reciprocators only cooperate with groups a high proportion of whose members previously cooperated towards them. These high thresholds make it less likely that selfish free riders will invade, since only groups which are almost all reciprocators get the benefit of mutual cooperation. The resulting equilibrium is ”trigger happy”: it sustains a high level of cooperation, but is also sensitive to small amounts of defection.

2. Literature review

Group reciprocity is a form of ”upstream reciprocity”, where an individual who is helped or harmed by someone becomes more likely to respectively help or harm other third parties. It is thought hard for upstream reciprocity to evolve, because it does not target reciprocity in a way that might lead to stable bilateral relationships nowak2007upstream. We show that the existence of groups can make this problem easier, by allowing different *groups* to form bilateral relationships. This is true even though the groups don’t possess any power of collective action (for instance, they can’t collectively decide to sanction outgroups, or force their members to behave a certain way). Simply making individuals’ group membership visible is enough to let group reciprocity evolve, because it lets people target help on groups whose members have (mostly) been helpful.

Several experiments show evidence for group reciprocity, also known as ”vicarious revenge” lickel$_{v}icarious_{2}006, gaertner2008whenrejection, stenstrom_{r}oles_{2}008, hugh$

What we lack is a theory of how group reciprocity might evolve. While there are evolutionary theories to explain individual reciprocity and revenge mccullough2013cognitive, there is none for group reciprocity. This is a problem, because one can't simply assume that the same forces are behind the evolution of individual reciprocity and group reciprocity. In particular, group reciprocity seems to be less targeted than individual reciprocity, and it also seems to involve a public good, since a person's cooperation against a group reciprocator subsequently benefits the person's entire group. petersen2010evolutionary argue that group members ”have an interest in advertising that victimizations of its members will not go unpunished”, but this ignores the public good aspect of the problem.

fearon1996explaining introduce a repeated-game model to explain how different ethnic groups can live at peace. In their ”spiral regime”, defection by any member of ethnic group A towards a member of B leads to subsequent defection by all members of B towards members of A for a fixed number of periods. This is an infinitely repeated game with multiple equilibria. The goal is to explain institutions which support interethnic cooperation; the spiral regime is analogous to institutions like feuds.

Our theory has a different setup and motivation. We examine the evolutionary stability of different types in a finitely-repeated game. In Fearon and Laitin, cooperation is supported by the threat of collective punishment. In our model, group reciprocity is evolutionarily stable because individual free-riding is balanced against group selection. That is, cooperating with cooperative outgroups benefits one's own group by encouraging the outgroup to cooperate in return. The individual benefit of free-riding is balanced against this group benefit. When the threshold for reciprocity is high, groups which get into cycles of cooperation with outgroups contain many reciprocators and few free-riders; most free-riders are in groups below the threshold, which don't cooperate, and therefore they don't benefit much from free-riding off the reciprocal types. So, our model is designed to explain the evolution of reciprocal (”strong” reciprocal) motivations in humans, rather than the stability of institutions supporting reciprocity.

3. Model

We consider a mixed population of two types, selfish and group reciprocators (GR). At the beginning of each generation, the population randomly divides into a large number of groups of size G each. Let p denote the population share of GR types. Let p_g be the proportion of GR in group g, which is distributed binomially.

At every step t, everybody interacts with everybody. In each pair, each individual chooses between cooperation and defection. Cooperation entails a cost c to the cooperator and a benefit b to her partner. Defection carries no costs. That is, each pair plays the following Prisoner's Dilemma game:

	Cooperate	Defect
Cooperate	$b-c$	$-c$
Defect	b	0

Selfish types always defect, A GR individual i starts by cooperating, and then cooperates with all individuals belonging to group g with a probability $\phi(l_{gi})$, where l_{gi} is the proportion of individuals from group g who cooperated with individual i in round $t-1$. $\phi(\cdot)$ is monotonically weakly increasing. We consider the cutoff strategy:

$$\phi(l_{gi}) = \begin{cases} 1 & \text{if } l_{gi} \geq k \\ 0 & \text{otherwise.} \end{cases}$$

The fitness is the payoff at the limit where $t \to \infty$. Equivalently, since the game always settles to a stationary action profile, it is the average payoff of T rounds when $T \to \infty$.

Given this behaviour, GR in all groups where $p_g \geq k$ help all individuals in other groups where $p_g \geq k$ and defect against members of all other groups. GR in other groups always defect. Individuals' fitness therefore depends only on whether they are in a "supraliminal" group with $p_g \geq k$, and on their type. Let q be the proportion of supraliminal groups. Let $\bar{p}$ be the proportion of GR individuals in supraliminal groups (out of the total population in such groups).

Let $\underline{p}$ be the proportion of GR individuals in subliminal groups (out of the total population in such groups). It follows that

- Group reciprocators in supraliminal groups get a payoff of $\bar{p}qb - qc$.
- Selfish types in supraliminal groups get $\bar{p}qb$.
- Group reciprocators and selfish types in subliminal groups get 0.

After each generation, reproductive success is proportional to fitness, the total population size stays the same, and children are remixed randomly into new groups of the same size. The mean fitness of the GR type is

$$\frac{\bar{p}q(q(\bar{p}b - c))}{p}$$

and the mean fitness of selfish types is

$$\frac{(1-\bar{p})q(q\bar{p}b)}{1-p}$$

After rearranging, the mean fitness of reciprocators is higher if

$$\frac{\bar{p}-p}{1-p} \geq \frac{c}{b}. \tag{1}$$

where

$$\bar{p} = E[p_g | p_g \geq k] = \frac{1}{G}\frac{\sum_{l=kG}^{G} l\text{Binom}(l, G, p)}{\sum_{l=kG}^{G}\text{Binom}(l, G, p)}$$

The LHS of (2) is decreasing in p and is equal to the threshold k when $p = 0$.[1]

Proof. We will first prove an auxiliary lemma and then the main result.

Lemma 1. *Let $x_1, x_2, ..., x_n$ be iid Binomially-distributed variables with probability of success p, and let $x_i = 1$ denote the event where x_i fails. Then showing that $\frac{\bar{p}-p}{1-p}$ decreases in p is equivalent to showing that $\frac{p(x_1=1|S_n \leq k)}{p(x_1=1)}$ increases in p, where $S_n = x_1 + x_2 + ... + x_n$.*

[1] When the share of GR in the population is very small, the probability that $p_g = k$ conditional on $p_g \geq k$ goes to one.

Proof of Lemma ??

First note that proving that the LHS of (2) is decreasing in p is equivalent to proving that $1-\frac{\bar{p}-p}{1-p}=\frac{1-\bar{p}}{1-p}$ is increasing in p. Second, note that $1-\bar{p}$ captures the expected proportion of failures of a Binomially-distributed variable given that the proportion of successes was at least k, or, put differently, given that the proportion of failures was at most k. Finally, we can replace the expected proportion of failures with the probability of a failure. All in all, we get that proving that $\frac{1-\bar{p}}{1-p}$ is increasing in p is equivalent to showing that $\frac{p(x_1=1|S_n\leq k)}{p(x_1=1)}$ is increasing in p.

Proving the main result

Lemma **??** implies that if $x_i=1$ denotes the event where x_i fails, and S_n counts the number of failures among n trials, then we need to show that $\frac{p(x_1=1|S_n\leq k)}{p(x_1=1)}$ increases in the probability of success p. For tractability, we will now revert to the more standard notation of $x_i=1$ as denoting the event where x_i *succeeds* (s.t. S_n counts the number of successes among n trials), and show that $\frac{p(x_1=1|S_n\leq k)}{p(x_1=1)}$ *decreases* rather than increases in the probability of success p.

Proof

$$\frac{p(x_1=1|S_n\leq k)}{p(x_1=1)}=\frac{p(s_n\leq k|x_1=1)}{p(s_n\leq k)}=\frac{p(s_{n-1}\leq k-1)}{qp(s_{n-1}\leq k)+pp(s_{n-1}\leq k-1)}$$

$$=\frac{p(s_{n-1}\leq k-1)}{q(p(s_{n-1}\leq k)-p(s_{n-1}\leq k-1))+p(s_{n-1}\leq k-1)}=\frac{p(s_{n-1}\leq k-1)}{qp(s_{n-1}=k)+p(s_{n-1}\leq k-1)}.$$

Using a known result,[2] according to which

$$p(s_n\leq k)=\frac{n!}{(n-k-1)!k!}\int_0^q t^{n-k-1}(1-t)^k dt=\frac{n!}{(n-k-1)!k!}\int_0^1 q^{n-k}s^{n-k-1}(1-qs)^k ds,$$

we get that $\frac{p(x_1=1|S_n\leq k)}{p(x_1=1)}$ is decreasing in p if and only if $\frac{q\binom{n-1}{k}p^k q^{n-k-1}}{\frac{(n-1)!}{(n-k-1)!(k-1)!}\int_0^1 q^{n-k}s^{n-k-1}(1-qs)^{k-1}ds}$ is decreasing in q, i.e., if and only if $\frac{\binom{n-1}{k}p^k}{\frac{(n-1)!}{(n-k-1)!(k-1)!}\int_0^1 s^{n-k-1}(1-qs)^{k-1}ds}$ is decreasing in q. Thus, it is sufficient to show that $\int_0^1 s^{n-k-1}(\frac{1-qs}{1-q})^{k-1}\frac{1}{1-q}ds$ is in-

[2] See equations (3) and (4) in https://mathworld.wolfram.com/BinomialDistribution.html.

creasing in q. Since $(\frac{1-qs}{1-q})^{k-1}\frac{1}{1-q}$ is non-decreasing in q for every $s \in [0,1]$, the proof is complete.

□

It follows that there is a unique ESS with a positive share of group reciprocators if and only if $k > \frac{c}{b}$. Otherwise the population is homogeneously selfish in the unique ESS.