GROUPING, IN-GROUP BIAS AND THE COST OF CHEATING

Moti Michaeli*

Abstract: The tendency of people to divide into groups and to show in-group bias – preferential treatment for insiders – is widely observed. This paper shows that it arises naturally when people incur a moral cost when defecting against cooperators, provided that this cost is concave in the number of such defections. If some people are asocial, i.e. insusceptible to the moral cost, then, under incomplete information, free-riding and cooperation can coexist within groups. Costly signaling of sociality enables groups to screen out free-riders, but its availability may decrease the welfare of *all* individuals in society. Finally, the concave moral cost is shown to be evolutionary stable with respect to an invasion by a convex mutation.

Keywords: In-group Bias, Group Formation, Costly Signaling, PD Game, Social Identity.

JEL codes: D7, D03, Z13, D64, D82, C72.

*The University of Haifa. This paper was part of my PhD dissertation at The Hebrew University and was previously circulated under the name "Group Formation, In-group Bias and the Cost of Cheating". Contact: motimich@econ.haifa.ac.il.

1 Introduction

People tend to form groups. Frequently this is accompanied by in-group bias toward members of other groups. The ubiquitous tendency to grouping has received some attention in disciplines such as anthropology (Dunbar 1993) and evolutionary biology (Wilson 1975), but the origins of in-group bias remain largely unexplored. In economics, the focus has mainly been on explaining the puzzle of cooperation rather than its restriction to small (cooperative) groups and the accompanied in-group bias. This paper aims to fill in the gap and provide a game-theoretic explanation to these group behaviors.

Existing theories that study the effect of group size on cooperation from a game-theoretic perspective (e.g., Boyd and Richerson 1988, Nowak and Sigmund 1998 and Suzuki and Akiyama 2005) often assume that agents are purely selfish hence resort to explanations that require repeated interactions to induce any cooperation, even if only within a limited group. However, there is abundant evidence in the literature that humans have other-regarding preferences (e.g., Fehr and Schmidt 1999 and Charness and Rabin 2002) and that cooperation is maintained even in one shot interactions (e.g., Marwell and Ames 1979). Thus the puzzle remains: if other-regarding preferences enable people to cooperate, why do they divide into groups that maintain cooperation only within?

In the current paper, grouping and in-group bias are jointly explained by the existence of a *moral cost of cheating* at the individual level. This moral cost can explain these group-related behaviors provided that it increases *concavely* in the number of cheated individuals, capturing the notion that people tend to be insensitive to the number of individuals who are affected by their actions. A concave moral cost is in line with recent empirical findings (see e.g. Amir et al. 2016 and Schumacher et al. 2017), reflects the psychological notion of scope neglect (see evidence surveyed in Slovic 2007) and has been argued and shown to fit moral costs in various settings (see e.g. Osborne 1995, Kamada and Kojima 2014 and Chen et al. 2017). The model shows that when people endowed with a concave moral cost interact, cooperation is bound to be limited and this stimulates the division into groups that display in-group favoritism. This in-group bias is not built upon any joint agreement between

group members to "boycott" outgroup members. Neither is it dependent on the existence of identity-related preferences. Rather, it is a pure equilibrium phenomenon based on the shape of other-regarding preferences.

In the basic model, each pair of individuals simultaneously play a one-shot Prisoner's Dilemma (PD) game. Each individual decides whether or not to cooperate with any other individual in society (which allows for discriminatory behavior). A person who engages in unilateral defection (cheating) is subject to a moral cheating cost that increases concavely in the number of individuals he cheats. This essentially implies that while the total moral disutility from cheating increases as the number of cheated partners grows, the marginal and the average disutility decrease. So while it may be unattractive to cheat one partner, it may well be attractive to cheat many partners. In fact, cheating is bound to take place once there are so many partners that the average disutility from cheating a partner falls below the material gain from unilateral defection. But the other side of the coin is that cheating can be prevented if one is mutually cooperating with a sufficiently small number of partners. This happens in equilibrium, where groups of cooperators endogenously emerge: belonging to a group of limited size ensures that the temptation to cheat when encountering group members is resistible, and that they can be trusted to cooperate because *their* temptation is resistible too. Moreover, each member of the group shows in-group bias, by (mutually) cooperating when playing with ingroup members, while (mutually) defecting when playing against outgroup members. Note that if the moral cost of cheating were instead convex, it would have always been tempting to cheat at least a little bit – i.e. a small fraction of one's partners – and this would have sabotaged within-group cooperation.

If not everyone in society is endowed with a moral cost of cheating we can distinguish between two types – *social types*, who *are* subject to this cost (and for whom the analysis above applies), and *asocial types*, who are not subject to this cost hence always defect in the PD game. The result that social types must form groups in order to maintain cooperation is shown to hold in various informational settings. In particular, after showing the basic result when types are fully observable (Section 2.1), Section 2.2 models incomplete information, so that one's type is one's private information, but individuals know the relative proportions

of types in society. Social types do not mind defecting against (the non-cooperative) asocial types, but, fearing they might mistake a fellow social type for an asocial type, end up being cheated by the latter. This happens in *mixed groups*, where a minority of asocial types free ride at the expense of the social types. In these groups, cooperation and free-riding coexist. Mixed groups will tend to be smaller than the cooperative groups of purely social types that can form when information is complete. More generally, a higher proportion of social types in society will be correlated with larger groups.[1]

Section 3 analyzes a third informational setting, where one's type is still one's private information but signaling is possible, albeit being costly. If the cost of signaling is sufficiently lower for the social types than it is for the asocial types, signaling enables the social types to reveal their type. In this case, *signaling groups*, consisting only of social types who fully cooperate with one another (but not with outsiders), can exist alongside the mixed groups. The social types thus recover (at a cost) the capacity to form larger groups like in the case of complete information, but grouping is still inevitable. Signaling is not enough to facilitate cooperation among *all* the social types because the moral cheating cost still forces them to divide into groups in order to sustain cooperation in equilibrium.

In this third and last informational setting I also perform a welfare analysis. First, I show that the existence of signaling groups decreases the welfare of members of mixed groups. The reason for this is that the availability of the signaling technology allows some social types to separate themselves (at a personal cost) from the rest of society, thus depriving the rest of the benefit of interacting with them. Second, I show that if the proportion of asocial types in society is not too high, the possibility of signaling decreases the welfare of the members of the signaling groups themselves as well, i.e., they would have been better off if they had been denied that possibility. Thus, beyond the private cost to the individual who signals, signaling as a phenomenon imposes a cost on society as a whole. This negative aspect of costly signaling is a significant point that has received little attention in the literature on

[1]If one considers the level of trust in society to be a good proxy for the level of trustworthiness (and hence the proportion of social types), then this correlation is in line with evidence in Porta et al. (1996) and Fukuyama (1995), which indicates a positive correlation between the level of trust in society and the size of firms and other organizations in that society.

in-group bias and social identity so far and should be taken into account when considering the problem of free-riding.[2]

Finally, in Section 4, I develop a simple evolutionary model whose purpose is to demonstrate that in my benchmark model (with complete information) the concave cheating cost is also evolutionarily stable. In this evolutionary model, mutants with a convex cheating cost ("convex types") invade a population composed solely of social types (having a concave cost).[3] In particular, I adopt the stability concept of Dekel et al. (2007) and show that the incumbent population is stable with respect to such invasions. First, I show that if the size of the invasion is small, as is customary to assume in evolutionary models, the incumbents will outperform the invaders hence a convex cheating cost will not spread. Second, I find conditions on the convex cheating cost of the invaders such that even if the number of invaders is as high as that of the incumbents, the invaders will still do worse. A main feature that distinguishes this model from other related models that show stability of cooperation-inducing preferences (such as Herold 2012, Eshel et al. 1998 and García-Martínez and Vega-Redondo 2015) is that in my model interactions take place among all individuals in society (like in my benchmark model) while these other models consider instead local interactions.

2 The basic model

Society consists of a continuum of individuals with a unit measure, who simultaneously interact with each other to play one-shot Prisoner's Dilemma (PD) games. The payoff matrix for the pairwise PD game is as follows.

	C	D
C	$1, 1$	$-\ell, 1+g$
D	$1+g, -\ell$	$0, 0$

The zero payoff for mutual defection has two implications. First, it implies that there is no difference between mutual defection and no interaction at all, hence the model applies also to cases where not every pair of players interacts. Second, it implies that the payoff for

[2]See the discussion of related literature at the end of Section 3.
[3]The asocial types are then a special case of convex types.

mutual cooperation is strictly positive, hence the total return to cooperation increases as the number of one's cooperative partners grows. This captures the idea that larger groups do better than smaller ones (nevertheless, groups will be of bounded size in equilibrium). The payoff for mutual cooperation is normalized to 1 so that cooperating with a mass k of individuals yields a payoff k. The parameter g stands for the *gain* from unilateral defection, and ℓ stands for the *loss* from being the victim of the opponent's unilateral defection. The loss ℓ is assumed to be (strictly) greater than g, implying strategic complementarity.

Let s_{ij} denote the action chosen by player i when interacting with player j. The strategy of player $i \in [0,1]$ is $C_i = \{j \in [0,1] \,|s_{ij} = C\}$ with measure $\mu(C_i)$. With some abuse of notation, I will simply write $C_i = C_i'$ whenever $\mu(C_i \Delta C_i') = 0$.[4]

Society is composed of two types of individuals: *social types* and *asocial types*. The type of player i is denoted by $\tau(i) \in \{s, as\}$. The share of asocial types in each interval is identical and equals $p \in [0,1)$. Asocial types are affected only by the material payoffs of the game, and so for them defection is a dominant strategy. Unlike them, social types are subject to a moral cost of cheating, where cheating means playing D against an opponent who plays C. Let $t(k)$ denote the cost of cheating a mass k of individuals (defined for any $k \in [0,1]$). This moral cost can be thought of as representing the arousal of uncomfortable feelings such as shame or guilt on the side of the defector.[5] The assumptions on $t(k)$ are as follows.

ASSUMPTION 1 *The cheating cost $t(k)$ has the following properties:*

1. $t(0) = 0$.
2. $t(k)$ is weakly increasing: for any $k_1, k_2 \in [0,1]$, $k_1 < k_2 \Rightarrow t(k_1) \leq t(k_2)$.
3. $t(k)$ is weakly concave: for any $k_1, k_2 \in [0,1]$ and any $q \in (0,1)$, $t(qk_1 + (1-q)k_2) \geq qt(k_1) + (1-q)t(k_2)$.
4. $\lim_{k \to 0} t'(k) > g$ (or, if $\lim_{k \to 0} t'(k)$ is not defined, $\lim_{k \to 0^+} t(k) > 0$).
5. $t(1-p) < (1-p)g$.

[4] $C_i \Delta C_i' := (C_i \setminus C_i') \bigcup (C_i' \setminus C_i)$ is the *symmetric difference* between C_i and C_i'.

[5] Adding a psychological cost for being the *victim* of cheating would not affect the qualitative outcomes of the model.

Property 2 implies that the more people are cheated by the individual, the more it costs him, and property 3 (concavity) is the main feature of the moral cost.[6] Properties 4 and 5 are close in spirit to the Inada conditions: a lower bound on the slope at 0 (or otherwise a discrete "jump" at 0) and an upper bound on the cost as k goes to $1-p$ (= the mass of social type in society). The first requirement (property 4) ensures that the moral cost is sufficiently large to allow for at least some cooperation. The second requirement (property 5) ensures that society is sufficiently large to allow for a significant decrease in the marginal cost of cheating as one becomes engaged in sufficiently many bilateral interactions.

Given a strategy profile $\gamma = (C_i)_{i\in[0,1]}$, the payoff $U_i(\gamma)$ of a social type $i \in [0,1]$ is

$$U_i(\gamma) = \mu\left(C_{-i}\right) + g\mu\left(D_i \bigcap C_{-i}\right) - t\left(\mu\left(D_i \bigcap C_{-i}\right)\right) - \ell\mu\left(C_i \bigcap D_{-i}\right), \tag{1}$$

where $C_{-i} = \{j \in [0,1] \,|\, i \in C_j\}$, and D_i and D_{-i} are the complements of C_i and C_{-i} respectively. In equation (1), C_i is the choice variable of player i, who faces C_{-i} which is jointly determined by players $j \neq i$.[7] The payoff of an asocial type is given by setting $t(\cdot) = 0$.

2.1 Grouping and in-group Bias under complete information

2.1.1 In-group Bias

Suppose first that the type of each individual is common knowledge. The result that cooperation with a group of social types can be sustained, but *in-group bias*, i.e., defection when playing against out-group members, is bound to emerge too, is presented in the following proposition.

PROPOSITION 1 *Let $\bar{K} \in (0, 1-p)$ be the unique strictly positive solution of the equation $t(K) = Kg$. A strategy profile $\gamma = (C_i)_{i\in[0,1]}$ is a Nash equilibrium if and only if (i) $C_i =$*

[6] Note that any cost function with a discrete jump at 0 and a weakly increasing and weakly concave continuation afterwards satisfies my condition of concavity as well. In particular, this includes a piece-wise linear cost function that has one fixed component and one component that linearly increases in the number of cheated individuals, and its special case of a step function with a fixed cost of cheating for any $k > 0$, which could represent a binary distinction between one's self image as a cheater and one's self image as someone who does not cheat.

[7] I assume here that C_{-i} is measurable as well.

$C_{-i} = \emptyset$ *when* $\tau(i) = as$, *(ii)* $C_i = C_{-i}$ *and* $\mu(C_i) \leq \bar{K}$ *when* $\tau(i) = s$.[8]

Proposition 1 states that a social type will cooperate only with a subset of the other social types. The intuition is as follows. The dominant strategy of an *asocial type* is $C_i = \emptyset$, and since types are fully observable, part (i) trivially follows. As for the social types, their best response against a defecting partner is to defect as well, but their best response against cooperative partners depends on the mass of such partners. In particular, the concavity of the moral cost (in the number of cooperative partners) and the linearity of the material gain from unilateral defection imply the existence of a threshold on the number of cooperative partners, above which a social type is better off cheating (all of them), whereas below which he is better off cooperating with them all. This threshold is $\bar{K}$ (see Figure 1) and we get part (ii) of the proposition.[9]

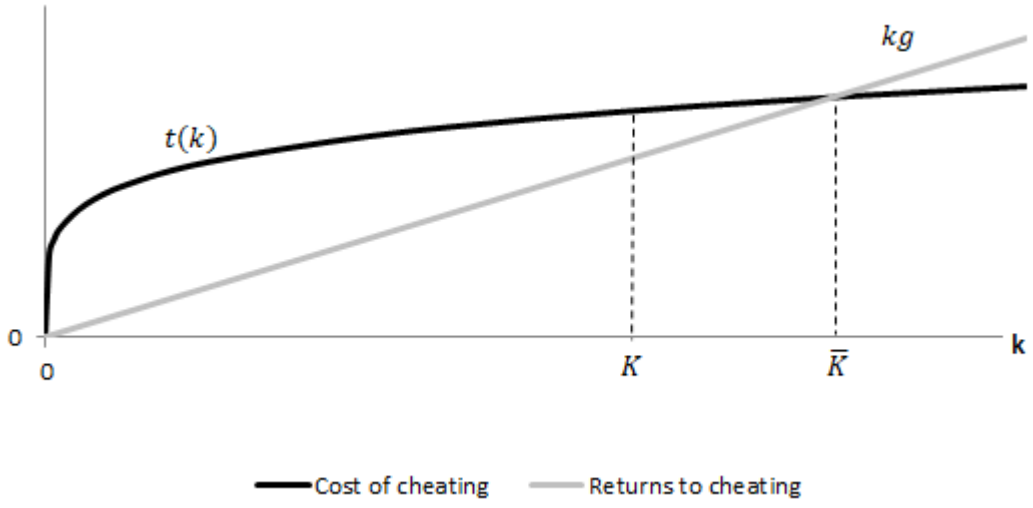

FIGURE 1.— The limit on the number of cooperative partners. For any given mass k of cooperators, the linear gray line depicts the material gain from cheating them, while the curved black line depicts the moral cost of doing so. These lines intersect at $k = \bar{K}$. A social type can cooperate in equilibrium with any set of social types as long as their mass K does not exceed $\bar{K}$, because deviating to defection against any subset of this set of players (of size $k \leq K$) will reduce his utility – the black line is always above the gray line at that range. However, maintaining cooperation with a set whose size is larger than $\bar{K}$ is impossible, because a deviation to defection against all other players will be profitable.

[8] I remind the reader that by writing $C_i = C_{-i}$, I mean that $\mu(C_i \Delta C_{-i}) = 0$.

[9] Since $\bar{K}$ is only an upper limit, also a social type may have $C_i = C_{-i} = \emptyset$ in equilibrium. In particular, the proposition also covers the special case where, as in the standard solution to the PD game, $C_i = C_{-i} = \emptyset$ $\forall i \in [0, 1]$.

Proposition 1 tells us that in-group bias is not contingent on the existence of identity-related preferences. Rather, it is a pure equilibrium phenomenon based on the shape of the moral cost of cheating. The proposition further highlights an important property of in-group bias: in equilibrium, social types show the same level of asociality towards out-group members as asocial types do. This characterization of social types suggests that even people with a clear prosocial orientation are predicted to restrict their cooperation only to interactions within their group (empirical support for this prediction can be found in Ruffle and Sosis 2006 and De Dreu 2010).

It is worth noting that a convex cost of cheating cannot be the explanation for in-group bias. Suppose $t(\cdot)$ is convex with $t'(0) < g$. Then, for a sufficiently large K, we have $t(K) > Kg$, implying that an individual engaged in mutual cooperation with a mass K of partners would rather keep cooperating with them than defecting against them all. Yet cooperation is *not* sustainable in this case. This is because there exists a sufficiently small k for which $t(k) < kg$, implying that this individual i has a profitable deviation from cooperating with all players in the set C_i to defecting against a subset of them of size k or less. In other words, in order to sustain cooperation with K players, it is essential that deviating to defection against *any* subset of this group (of size $k \leq K$) will reduce a player's utility, and this does not hold in the convex case considered here.[10] In Section 4, I further show that a population of individuals endowed with a concave cheating cost is immune to an invasion by types endowed with a convex cost.

2.1.2 Grouping

Define now a binary relation $\sim_c$ on $[0,1]$ by $i \sim_c j$ if $s_{ij} = C$. Proposition 1 implies that $\sim_c$ is symmetric, but it need not be transitive. In other words, society need not necessarily divide into mutually exclusive groups. However, motivated by the abundance of such divisions, and in order to gain further insights on grouping, the rest of the analysis focuses on strategy profiles of the following (transitive) kind: a strategy profile $\gamma = (C_i)_{i\in[0,1]}$

[10] If instead $t'(0) \geq g$ (where $t(\cdot)$ is still assumed to be convex), then $t(K) \geq Kg$ for any $K > 0$, in which case cooperation is sustainable with no bound on the number of partners, so there is no reason for individuals to form groups and to develop in-group bias (hence such a cost cannot account for the observed behaviors this paper aims to explain).

is *partitional* if $C_i \bigcap C_j \neq \emptyset$ implies $C_i = C_j$. I will refer to C_i as the group of player i. Proposition 1 immediately provides the conditions under which a given partitional strategy profile constitutes an equilibrium.

COROLLARY 1 *Any partitional strategy profile in which* $\mu(C_i) = \mu(C_{-i}) = 0$ *if* $\tau(i) = as$ *and* $\mu(C_i) = \mu(C_{-i}) \leq \bar{K}$ *if* $\tau(i) = s$ *forms an equilibrium.*

Corollary 1 sets an upper bound on the size of groups in equilibrium, implying that – in order to survive – groups must limit the number of their members. Note that unlike prevailing models in the literature in which groups try to screen-out only the free riders, the novelty here is in the claim that groups must also restrict the influx of "good guys". This result could potentially explain the limited size of tribes and clans, especially in societies with no central authority (where groups are presumed to form spontaneously). Yet, it is only an upper bound: groups may be of different sizes, as is often the case in reality.

2.2 Grouping and in-group bias under incomplete information

Suppose now that an individual's type is his private information (while the proportion of asocial types p is common knowledge). Can there still be an equilibrium with some cooperation in it?[11]

Since types are private information, the set C_i contains now a proportion of p asocial types. Consider now a social player i who interacts with a set of players of size K such that all social types in the set play C. If player i plays D against $k \in [0, K]$ of them, his expected payoff of is given by

$$U_i = (1-p)K + g(1-p)k - t((1-p)k) - \ell p(K-k), \tag{2}$$

and, in case $k = 0$, equation (2) boils down to

$$U_i = (1-p)K - \ell pK.$$

[11] Formally, this is a model of incomplete information so the equilibrium concept that applies here is Bayesian Nash Equilibrium, but this has no particular significance for the analysis as the only implication of the incomplete information is that payoffs will be computed as expected utility rather than be deterministic.

The following proposition characterizes partitional strategy profiles that constitute an equilibrium under incomplete information.

PROPOSITION 2 *Let* $p_\Lambda \equiv \frac{\Lambda}{\Lambda+\ell}$, *where* $\Lambda \equiv \lim_{k\to 0} t'(k) - g$ (> 0).

(I) If $p > p_\Lambda$, *the unique equilibrium is one where* $C_i = \emptyset$ *for all* $i \in [0,1]$.

(II) If $p < p_\Lambda$, *a partitional strategy profile* $\gamma = (C_i)_{i\in[0,1]}$ *is an equilibrium if and only if (i)* $\tau(i) = as$ *implies* $C_i = \emptyset$; *(ii)* $\{\tau(i) = \tau(j) = s$ *and* $j \in C_i\}$ *implies* $i \in C_j$; *and (iii)* $\mu(C_i) \leq K_p$, *where* $K_p \in \left(0, \bar{K}\right]$ *is the unique strictly positive solution to the equation*

$$t((1-p)\,k) - k[(1-p)g + p\ell] = 0. \tag{3}$$

Furthermore, K_p *is strictly decreasing in* p *in the range* $[0, p_\Lambda)$.[12]

Proposition 2 says that, as long as the proportion of asocial types in society is sufficiently small ($p < p_\Lambda$), social types will still be cooperative toward group members, but such *mixed groups* are bound to be smaller (no bigger than $K_p \leq \bar{K}$, with strict inequality for any $p > 0$). An increase in p, the proportion of asocial types in society, has two effects, both contributing to the decrease of group size. First, the material temptation to defect becomes greater – strategic complementarity implies that the increase in the expected payoff achieved by avoiding the "sucker" payoff ℓ is larger than g, the increase achieved by playing D against a cooperative partner. Second, the moral cost of playing D against a random group member is lower (because some of the "victims" of defection will be asocial). The two effects are illustrated in Figure 2. Consequently, the greater the proportion of asocial types in society is, the more it is tempting for social types to defect. This in turn implies that the groups that *can* sustain cooperation become smaller as p gets larger.

Overall, we get that social types show in-group bias (play C against all group members and D against all outsiders) also under incomplete information, while asocial types play D

[12] If (iii) holds with equality (so that $\mu(C_i) = K_p$), then there can be countably many players j (with total mass of 0) for whom (ii) does not hold – see the proof for details. Furthermore, if $\lim_{k\to 0} t'(k)$ is not defined, implying (by assumption) that $\lim_{k\to 0^+} t(k) > 0$, then equation (3) has a unique strictly positive solution $K_p \in \left(0, \bar{K}\right]$ for any $p \in [0,1)$ and part (II) of the proposition applies.

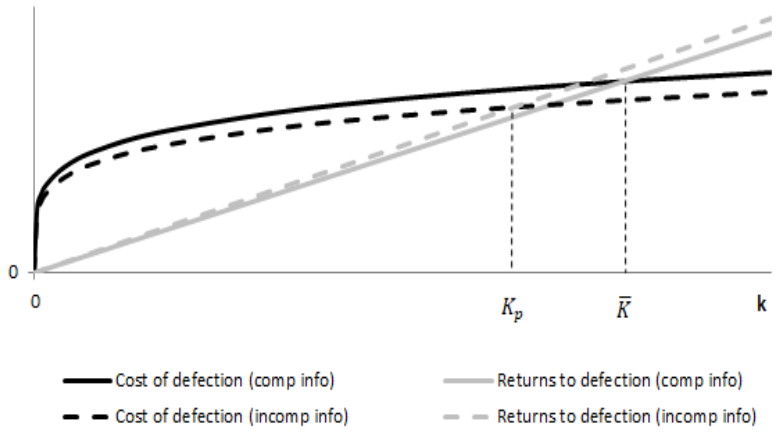

FIGURE 2.— The limit on a mixed group size. For any given mass k of opponents, such that a proportion p of them defect while the others cooperate, the linear dashed gray line depicts the material gain from playing D against all of them, while the curved dashed black line depicts the moral cost of doing so. These lines intersect at $k = K_p$. The linear solid gray and the curved solid black lines that intersect at $k = \bar{K}$ are taken from Figure 1 and are displayed for comparison. The dashed gray line is drawn above the solid gray line because $\ell > g$, and so the returns to defection are larger when facing a mixed group. The dashed black line is drawn below the solid black line because the moral cost applies only to defection against cooperative opponents and there are less of those in a mixed group. It is thus easy to see why $K_p < \bar{K}$ and why K_p is decreasing in p.

against everyone, thus free ride on the social types in their groups. The cooperative behavior of the social types in mixed groups and the sustainability of free-riding in equilibrium are in line with the "weak free-riding hypothesis".[13]

Proposition 2 further says that if the proportion of asocial types in society is high ($p > p_\Lambda$) cooperation is not sustainable. This is where society "breaks up" because too many people cannot be trusted. The threshold for when this is bound to happen is increasing in Λ, which can be thought of as capturing the resilience of social types to small-scale cheating.[14] In particular, if this resilience is sufficiently large ($\Lambda > 1$) and if the proportion of asocial types in society is quite high yet not too high — $p \in \left(\frac{1}{1+\ell}, \frac{\Lambda}{\Lambda+\ell} = p_\Lambda\right)$ — an interesting scenario is revealed. In this case, cooperation is sustainable (because $p < p_\Lambda$), yet the fact that $p > 1/(1+\ell)$ implies that the proportion of asocial types in society is sufficiently high to make the expected payoff of a social type who is a member of a mixed group *negative*

[13]This hypothesis, stating that some people in the group will free ride while others will not, was shown to hold in experimental settings such as the one in Marwell and Ames (1979).
[14]Think of $k = \varepsilon \to 0$, where $t(k) - kg \to t'(0)\varepsilon - \varepsilon g = \Lambda\varepsilon$.

(because his expected payoff is $K[(1-p)-p\ell)] < 0$). This means that, in this case, all the social types would have been better off in a society where everyone else is known to defect (so that they could defect as well with no pangs of conscience and get a zero payoff instead). Yet, mixed groups of size $K \leq K_p$ *are* sustainable in equilibrium, and social types in these groups end up playing C when interacting with other group members in order to avoid hurting other cooperative individuals like themselves, and thus end up getting a negative payoff.

3 Incomplete information with signaling

3.1 Preventing free riding but not in-group bias

If the social types could reliably signal their type, they could screen out the asocial types and restore full cooperation.[15] To study this possibility in my setup, suppose that before the pairwise PD game is played, each individual can send a costly and fully observable signal $x \in X$. Formally, X is an arbitrary finite set of signals that contains also the null signal x_0 whose interpretation is "no signal", hence costs nothing to send. The cost of sending any other signal $x \in X$ is denoted by $x_s > 0$ for the social types and by $x_{as} > 0$ for the asocial types. That is, I assume that – for each type – all non-null signals in X cost the same and differ only in kind. As an illustrative example to keep in mind while reading this section, consider the choice to belong to a religious congregation and participate in its ceremonies and activities. The idea is that attending religious services is potentially a useful form of (costly) signaling, and attending different churches is interpreted as sending different signals that all cost the same. I start the analysis by demonstrating that such differentiation between signals can facilitate a division of society into separate groups that screen out free riders yet show in-group bias towards each other. I then proceed to a welfare analysis.

As opposed to the previous section, here the game is played in two stages (first signaling and then the PD game), hence the equilibrium concept is that of a Perfect Bayesian Equilibrium (PBE). However, beliefs here are confined to the (common) prior about the distribution of types among signalers and non-signalers. The strategy of player i of type τ

[15] Theoretical analyses along these lines can be found in Iannaccone (1992), Akerlof and Kranton (2000), Berman (2000) and Levy and Razin (2012). Aimone et al. (2013) demonstrate it experimentally.

consists of two components: (I) a signaling component $\sigma_i(\tau) \in X$ (as explained above); and (II) a cooperation component $\widetilde{C}_i \in 2^{[0,1]\times X}$. The $\widetilde{C}_i$ component is the signaling-model analog of C_i from the previous section. There, C_i was the set of players with whom i cooperates; hence, analogously, $\widetilde{C}_i$ is the set of (j,x)'s with whom i cooperates. Similarly, I write that $i \in \widetilde{C}_{-j,x}$, which means that i cooperates with j when j uses the signal x, if $(j,x) \in \widetilde{C}_i$.

I am interested in characterizing the conditions under which signaling can solve the free-rider problem of the previous section. For that purpose, I focus on (partitional) strategy profiles $\left(\sigma_i, \widetilde{C}_{\mathrm{i}}\right)_{\mathrm{i}\in\,[0,1]}$ that satisfy the following three properties: (i) if $\sigma_i(\tau) = x \neq x_0$ and $(j,x') \in \widetilde{C}_i$, then $x' = x$; (ii) $\sigma_j(\tau) = \sigma_i(s) = x \neq x_0 \Rightarrow (j,x) \in \widetilde{C}_i$; and (iii) if $\sigma_i(\tau) = x$, then $\sigma_j(\tau') \neq x \Rightarrow (j,x) \notin \widetilde{C}_i$. Properties (i) and (ii) imply that a social type who uses the signal $x \in X$ cooperates with (all) players who – on the equilibrium path – use that same signal x (by ii) and only with them (by i). Condition (iii) further states that player j of type τ' who deviates from his prescribed signaling strategy $\sigma_j(\tau')$ to using another signal x (including the null signal x_0) cannot gain the cooperation of player i who, on the equilibrium path, uses that same signal x.[16]

A strategy profile $\left(\sigma_i, \widetilde{C}_{\mathrm{i}}\right)_{\mathrm{i}\in\,[0,1]}$ that satisfies these three properties and forms an equilibrium (PBE) will be called a *group formation*. The set of group formations is denoted by Γ^*. To indeed solve the free-rider problem, the equilibrium has to be a separating one, as otherwise the set $\sigma^{-1}(x)$ (for any given x) contains a proportion of p asocial types. For that purpose I now define the notion of a *signaling group*.

DEFINITION 1 $\sigma^{-1}(x) \neq \emptyset$ for $x \neq x_0$ is called a signaling group if, for any player $i \in \sigma^{-1}(x)$, $\widetilde{C}_i = \widetilde{C}_{-i,x} = \sigma^{-1}(x)$.

That is, a signal $x \neq x_0$ is associated with a signaling group if *all the players* using the signal cooperate (only) with each other. This implies that there are no free riders (asocial types) in the group. The following proposition provides conditions for the existence

[16] Property (iii) is attainable under various belief systems, in particular one where players who deviate from their prescribed signaling strategy are expected to defect in the PD game. Note that this property does not contradict property (ii) because in property (ii) player i conditions his cooperation with player j not only on player j's use of the signal x but also on x being player j's *prescribed signal* under the signaling profile $(\sigma_i)_{\mathrm{i}\in\,[0,1]}$. Property (iii) thus keeps us close to the basic model where players cannot unilaterally switch groups.

of a separating group formation and characterizes the size of signaling groups in such an equilibrium.

PROPOSITION 3 *Let* $\hat{K} \equiv \min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\}$. *There exists a group formation* $\gamma \in \Gamma^*$ *where* $\sigma^{-1}(x) \neq \emptyset$ *for some* $x \neq x_0$ *and* $\widetilde{C}_i = \widetilde{C}_{-i,x} = \sigma^{-1}(x)\ \forall i \in \sigma^{-1}(x)$ *if and only if the following two conditions hold:*

1. *Individual rationality:* $x_s \leq \bar{K}$
2. *Cost differentiation:* $\frac{x_{as}}{x_s} \geq 1 + g$.

Furthermore, if γ *exists, then under* γ *we have* $\mu(\sigma^{-1}(x)) \in \left[x_s, \hat{K}\right]$.[17]

The proposition provides two conditions for the existence of a group formation with (at least one) signaling group. It also states that this signaling group will have a "medium" size. The two conditions in the proposition – individual rationality and cost differentiation– are necessary for the range $\left[x_s, \min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\}\right]$ to be non empty, and the size of a signaling group must be in this range for it to be sustainable in equilibrium. To see why, denote the size of a signaling group $\sigma^{-1}(x)$ (for $x \neq x_0$) by K. For a signaling group of size K to be sustainable it has to be no smaller than x_s, in order to ensure that a (social) group member, whose payoff is given by $K - x_s$, has no profitable deviation to not signaling and thus getting a zero payoff. It also has to be no larger than $\hat{K} = \min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\}$, because if it was larger than $\bar{K}$, a social player $i \in \sigma^{-1}(x)$ would have a profitable deviation to $(x, \emptyset)$ (i.e. signaling and then cheating the other group members), while if it was larger than $\frac{x_{as}}{1+g}$ the set $\sigma^{-1}(x)$ would contain also asocial types (who never cooperate in equilibrium), violating the requirement that $\widetilde{C}_i = \sigma^{-1}(x)\ \forall i \in \sigma^{-1}(x)$.[18] Finally, the existence of a group formation of this kind when the two necessary conditions hold is guaranteed by construction: the members of $\sigma^{-1}(x)$ whose size is in the non-empty range $\left[x_s, \hat{K}\right]$ have no profitable deviation as just explained, and to individuals outside the group we can assign the strategy $(x_0, \emptyset)$.

Recalling the example of attending a church as a signal, Proposition 3 provides micro-foundations to evidence on religious participation, as reported by Iannaccone (1994). First,

[17] If $\hat{K} = \frac{x_{as}}{1+g}$ then $\mu\left(\sigma^{-1}(x)\right) \leq \hat{K}$ holds with a strict inequality, i.e. $\mu\left(\sigma^{-1}(x)\right) \in \left[x_s, \hat{K}\right)$.

[18] $K(1+g) - x_{as}$ is the payoff that an asocial type who is assigned to a fully cooperative group $\sigma^{-1}(x)$ of size K gains by playing $(x, \emptyset)$. If $K \geq \frac{x_{as}}{1+g}$ this payoff is positive hence sustainable in equilibrium.

Iannaccone reports that, with regard to out-group members, (at least some) religious groups "condemn deviance, shun dissenters, and *repudiate the outside world*" (Iannaccone 1994, p. 1182), i.e., display in-group bias, even toward other sects of the same religion. In the model this stems from the existence of an upper limit on group size, implying that if there are many social types willing to signal, they cannot simply all cooperate with each other but must instead use different signals and divide into separate signaling groups (form different religious sects). In this sense, signaling may solve the free-rider problem but is not a cure for the limit on cooperation: multiple signaling groups, all consisting of social types but using different signals, may coexist and show in-group bias toward each other. Second, Iannaccone reports that, in the US, stricter churches (i.e., those that require a higher cost in terms of members' devotion) tend to be larger. In the model, this stems from the fact that the cost of signaling sets a lower bound on group size, hence the heavier is the cost determined by the group, the larger must it be in order to survive. Third, Iannaccone writes that the data "imply 'optimal' levels of strictness, beyond which strictness discourages most people from joining or remaining within the group" (Iannaccone 1994, p. 1182). This is captured in the model by the fact that if the signaling cost is set too high (above $\hat{K}$, as defined in the proposition), no individual will take part in the group.

The two conditions listed in Proposition 3 do not *guarantee* that a given group formation $\gamma \in \Gamma^*$ contains signaling groups. There is always an equilibrium where everyone plays D, and there are always pooling equilibria in which no one signals yet cooperation among social types is maintained within mixed groups (see Section 2.2). Even more interestingly, there can be "semi-separating" equilibria in which signaling groups coexist alongside mixed groups. This seems to be a sensible characterization of society. A natural interpretation of this situation is that it reflects how different individuals may find different solutions to the problem of cooperation: some choose to endure free-riders in their group, while others choose to engage in costly signaling. The question then arises: does signaling actually raise welfare?

3.2 Signaling: a double-edged sword

I will assume now that *individual rationality* and *cost differentiation* (as defined in Proposition 3) hold, and compare the welfare of individuals under group formations with and without signaling groups. For simplicity, I will use the notation $\sigma^{-1}(x)$ to refer only to signaling groups as defined in Definition 1 (and characterized in Proposition 3), i.e., groups that contain only social types. A first thing worth noting is that the proportion of asocial types among the non-signalers is bound to be higher than p.

COROLLARY 2 *If, under $\gamma \in \Gamma^*$, $\sigma_i(s) = x_0$ and $\mu\left(\widetilde{C}_i\right) \neq 0$, then the expected proportion of asocial types in $\widetilde{C}_i$ is $\frac{p}{1-\mu\left(\bigcup_{x \neq x_0} \sigma^{-1}(x)\right)}$ $(> p)$.*

In order to present the welfare results in a crisp manner, I focus here on the case $\Lambda > 1$ (which implies that mixed groups are sustainable for any $p < \frac{1}{1+\ell}$ – see the last paragraph of Section 2.2). The following result highlights a negative externality of signaling on the rest of society.

PROPOSITION 4 *Let $p \in \left(0, \frac{1}{1+\ell}\right)$ and consider a group formation γ_1 in which there exists $x \neq x_0$ such that $\mu\left(\sigma^{-1}(x)\right) \neq 0$. Then there exists a group formation γ_2 in which $\sigma^{-1}(x) = \emptyset$ for all $x \neq x_0$ and where $U_i(\gamma_1) < U_i(\gamma_2)$ for any social type i for whom, under γ_1, $\sigma_i(s) = x_0$.*

The proposition says that, for any given $p \in \left(0, \frac{1}{1+\ell}\right)$, the expected payoff of *all* the social types who do not signal (any i s.t. $\sigma_i(s) = x_0$), in any group formation that contains a non-zero mass of signalers (any γ_1), can be strictly increased by prohibiting signaling ($\exists \gamma_2$ s.t. $U_i(\gamma_1) < U_i(\gamma_2)$).[19]

The intuition for the proposition is as follows. By Corollary 2, when signaling groups exist in society, the expected proportion of asocial types outside the signaling groups is $\frac{p}{1-\mu\left(\bigcup_{x \in X} \sigma^{-1}(x)\right)}$, i.e. higher than p. This has two negative effects on the expected payoff of members of mixed groups. The first is a greater share of interactions with defecting opponents for any given group size. The second is a decrease in the upper limit on the size of mixed

[19] It can be further shown that, under mild conditions, the expected payoff of all the asocial types can be strictly increased at the same time as well.

groups, which implies a reduction in the maximal expected payoff of group members (of both types in fact). Hence, if signaling is prohibited, mixed groups can be larger and more cooperative, allowing for higher payoffs.

Thus, beyond the individual cost for the signaler, signaling as a social phenomenon imposes a negative externality on the rest of society. This negative externality represents the loss of "good guys", who form their own exclusive clubs, instead of mixing with the other parts of society and lifting the average willingness to cooperate. One may think that at least for the signalers themselves signaling improves welfare. However, the following proposition states that this is the case only for sufficiently large values of p.

PROPOSITION 5 *Suppose that $x_s < \hat{K}$ and let p_c be the unique implicit solution to the equation*

$$\hat{K} - x_s = K_p[1 - p(1+\ell)]. \tag{4}$$

Then $p_c \in \left(0, \frac{1}{1+\ell}\right)$ and, for a social type i,

$$max_{\gamma \in \Gamma^*} \{U_i(\gamma) | \sigma_i(s) \neq x_0\} \geq max_{\gamma \in \Gamma^*} \{EU_i(\gamma) | \sigma_i(s) = x_0\}$$

if and only if $p \geq p_c$.

The proposition identifies a tipping point for a social type – there exists a group formation that maximizes his expected payoff and in which he is signaling if and only if $p \geq p_c$. Equation (4) compares the maximal expected payoff of a social type in a signaling group (LHS) and in a mixed group (RHS). Since the LHS is constant while the RHS decreases in p,[20] social types can be better-off by signaling if and only if the proportion of asocial types in society is sufficiently high, with p_c being the tipping point. Figure 3 illustrates this result. Propositions 4 and 5 together imply that, when $p < p_c$, a group formation in which all groups are mixed and of maximal size (if it exists) Pareto dominates any group formation

[20]The RHS decreases in p because both K_p and $[1 - p(1+\ell)]$ decrease in p.

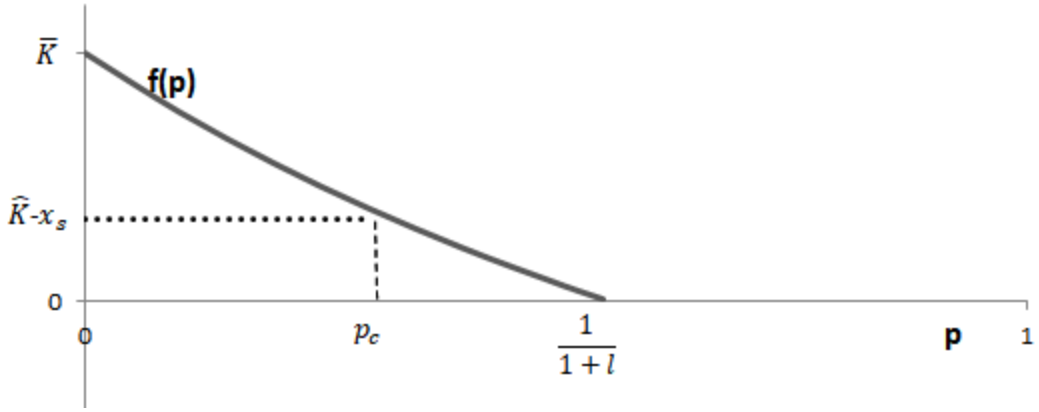

FIGURE 3.— The tipping point for social types. $f(p) \equiv K_p[1 - p(1 + \ell)]$ is the maximum expected payoff of a social type in a mixed group. It is achieved in a pooling equilibrium (i.e., when there is no signaling in society) when the individual's group is of size K_p, which is the maximum possible size given p. $\hat{K} - x_s$ is the expected payoff in a signaling group of the maximum size $\hat{K}$. If p, the proportion of asocial types in society, is smaller than p_c, a pooling equilibrium where all groups are of maximum size Pareto dominates all other equilibria, and so signaling is wasteful. If $p_c > p$, a social type can get a payoff of $\hat{K} - x_s$ in a signaling group of the maximum size, and this payoff is strictly greater than the expected payoff he can achieve in a mixed group.

that includes signaling groups.[21]

The negative aspect of costly signaling is a significant point that has received little attention in the literature on in-group bias and social identity so far and should be taken into account when considering the problem of free-riding. It is also one of the main features differentiating my analysis from similar models that study costly signaling in social settings, most notably Levy and Razin (2012) and Bernard et al. (2016). In Levy and Razin (2012), the aversion of religious types (the equivalent of my social types) to cheating enables them to sustain cooperation while also making them vulnerable to free riding. But the assumption that these religious types believe they will be rewarded for their actions in the afterlife, and that they gain utility from that expected future reward, implies that religion improves rather than deteriorates welfare (Proposition 3 in that paper). In Bernard et al. (2016), signaling is a means for deterring low status individuals from joining a high-status group thereby diluting

[21] Note that this applies also to the asocial types, who gain maximally from free riding in this case (see also Footnote 21). Technically, it may be the case that not everyone can simultaneously be part of a mixed group of the maximal size (if 1 is not divisible by K_p). Similarly, it can be the case that not all social types can be part of signaling groups of maximal size (if $1 - p$ is not divisible by $\hat{K}$). The formal results presented in this section take this into account.

its status. Thus, at least for the "high types", signaling can be welfare improving, insofar as the benefit gained by being perceived as a high type is larger than the signaling cost. In this sense, Bernard et al. (2016) is no different than the canonical job-market-signaling model (Spence 1974), where normally the signalers are better-off *with* the signaling technology. While this is the case also in my own model when the asocial types are numerous ($p > p_c$), the main take away comes from the case where they are few ($p < p_c$), and signaling leads to a Pareto inferior equilibrium.

4 Evolutionary stability of the concave cheating cost

A concave cost of cheating was shown in the paper to be able to explain a variety of group behaviors, but can it survive evolutionary competition against, say, a convex cost? To study this question, I revert back to the benchmark model of Section 2.1 (complete information, no signaling) and analyze the stability of a society composed solely of social types ($p = 0$) to an invasion by other types, in particular types with a convex moral cost of cheating. The focus on a convex cost is without loss of generality. In particular, the analysis holds also for types with a zero cost of cheating (asocial types).

The simple evolutionary model presented in this section adopts the "indirect evolutionary approach", as pioneered by Guth and Yaari (1992) and popularized by Dekel et al. (2007). As opposed to most standard evolutionary models, in this approach the agents are not automatons programmed to take a predefined action but are rather "preference types", in the sense that they are characterized by a payoff function and are assumed to best respond to other individuals' actions based on this payoff function.

Consider a homogeneous population composed of one preference type that is invaded by a small influx of mutants with a difference preference type. An incumbent is denoted by $i \in I$, with $\mu(I) = 1$. A mutant is denoted by $j \in J$. The focus of the analysis is on incumbents who are social types ($\tau = s$), i.e. types who are endowed with a concave cost of cheating $t_s(\cdot)$ (formerly $t(\cdot)$), with its associated $\bar{K}$. The invading mutants are instead endowed with a convex cost of cheating (convex types for short, denoted by $\tau = c$), with their cost function $t_c(\cdot)$ starting below the kg diagonal line ($t_c'(0) < g$) and crossing it once, from

below. I denote by $\underline{K}$ the value of k for which this cost function parallels the kg diagonal line (i.e., $t_c'(\underline{K}) = g$; see Figure 4 for illustration).[22] The key trade-off is easily visible in Figure 4: unlike the social types, convex types are tempted to defect when matched with a small group of cooperators, yet incur high cost when cheating many cooperators. Hence, despite not being able to maintain full cooperation, they might be able to outperform the social types if they can sustain partial cooperation among a sufficiently large group.

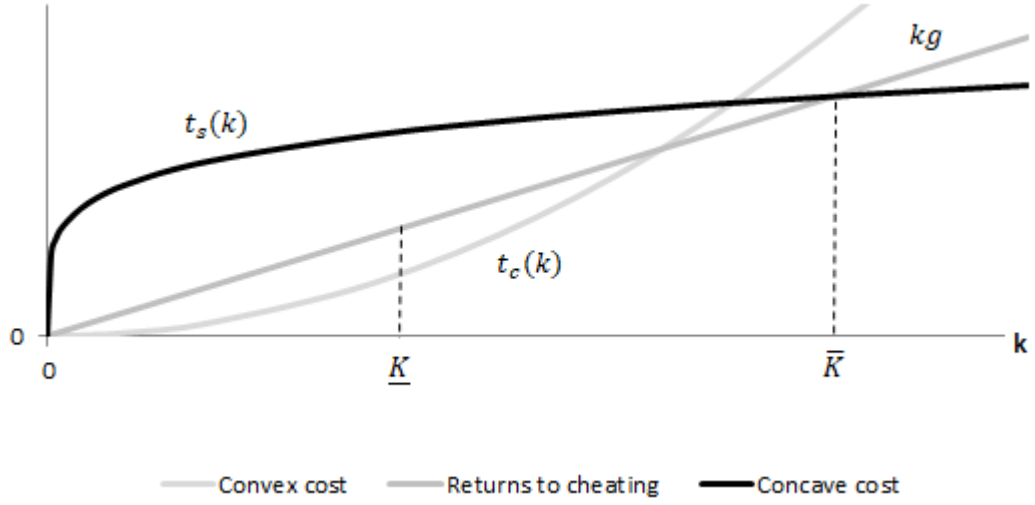

FIGURE 4.— Convex and concave cheating costs and their corresponding associated thresholds $\underline{K}$ and $\bar{K}$. $\underline{K}$ is not necessarily smaller than $\bar{K}$.

The interaction between individuals is modeled like in the rest of the paper: every player plays a one shot PD game against any other player. The payoffs of the game determine the relative fitness, which is then translated to changes in the relative frequencies of the two types in the population. As is standard in the literature on preference types (see e.g. Dekel et al. 2007 and Herold 2012), fitness is determined by the *material* payoffs, V_i and V_j respectively for the incumbents and the mutants, rather than by the experienced utility (U_i and U_j respectively, which also include the moral cost of cheating). That is, the material payoffs of both incumbents and mutants alike, V_i and V_j, are given by setting $t(\cdot) = 0$ in

[22]Note that agents with no cheating cost (asocial types) are behaviorally equivalent to convex types with a sufficiently large $\underline{K}$ (larger than the post-entry population size).

equation (1) of Section 2.1, which yields

$$(5) \qquad V_j(\gamma) = V_i(\gamma) = \mu\left(C_{-i}\right) + g\mu\left(D_i \bigcap C_{-i}\right) - \ell\mu\left(C_i \bigcap D_{-i}\right).$$

4.1 The solution concept

The analysis in this section is based on the stability concept of Dekel et al. (2007), according to which stable populations are those that cannot be invaded by mutants who – in the resulting equilibrium – have higher fitness (larger payoff) than the incumbents. This is a static concept that applies to a *configuration* of society prior to the invasion.[23]

DEFINITION 2 A configuration (of the incumbent population) is a strategy profile $\gamma^0 = (C_i^0)_{i\in I}$ that forms an equilibrium.

Stability is applied to configurations because also fitness is determined by the configuration: fitness depends not only on the composition of society in terms of types but also on the equilibrium behavior of all individuals. For example, in an equilibrium like the one characterized in Proposition 1, the fitness of player $i \in I$ equals the mass of his cooperative partners $\mu\left(C_i^0\right)$.

A configuration is stable if it satisfies two conditions. First, it must be balanced: all individuals must receive the same fitness.

DEFINITION 3 A configuration γ^0 is *balanced* if $V_{i_1} = V_{i_2}\ \forall i_1, i_2 \in I$.

In light of Proposition 1, the requirement of a balanced configuration rules out any configuration where $\mu\left(C_i^0\right)$ differs between individuals. Thus, for a configuration of the incumbent population I to be stable there must exist $K \leq \bar{K}$ such that $V_i = \mu\left(C_i^0\right) = K$ $\forall i \in I$.

Second, a stable configuration must resist entry by mutants in two ways: first, the mutants should not have a higher fitness than the incumbents ($V_j > V_i$); second, the behavior

[23] In Dekel et al. (2007), the definition of configuration includes both the distribution of preferences in the population and an equilibrium strategy profile of the players. In my own definition I do not explicitly mention the distribution of preferences, because it is assumed here to contain (prior to the invasion) only one preference type.

of the mutants should not unravel the original equilibrium behavior causing the distribution of actions to diverge. Dekel et al. (2007) acknowledge that, if there are multiple equilibria (as is the case in my own setup) and all of them are considered, then stability is too hard to satisfy because any entry could destabilize the configuration by arbitrarily triggering a switch to another equilibrium. To cope with that, they focus only on the subset of post-entry equilibria that are *focal* relative to the original configuration. A configuration is focal if incumbents' strategies when facing each other are unchanged, with no restrictions put on the strategies of entrants and on the strategies of incumbents when they interact with entrants. The assumption is that any focal equilibrium can potentially arise following an invasion of mutants, and thus the definition of stability requires that, in all of them, entrants earn no higher fitness than any incumbent. I apply the same methodology and check stability by considering all post-entry focal equilibria, assuming that types (i.e. preferences) are observable.[24] This implies that the solution concept is that of Nash Equilibrium.

Formally, let the size of invasion be arbitrarily small, $\mu(J) = \epsilon$, and denote the post-entry strategies of incumbents and mutants by C_i^ϵ and C_j^ϵ respectively.

DEFINITION 4 Given a pre-invasion configuration γ^0, a focal post-entry equilibrium (relative to γ^0) is a strategy profile $\gamma_\epsilon^0 = (C_i^\epsilon)_{i\in I} \bigcup \left(C_j^\epsilon\right)_{j\in J}$ such that γ_ϵ^0 is an equilibrium and, for any $i \in I$, $C_i^\epsilon \bigcap C_i^0 = C_i^0$ and $D_i^\epsilon \bigcap D_i^0 = D_i^0$.

Note that the set of focal post-entry equilibria is non empty because there always exists an equilibrium in which $C_i^\epsilon = C_i^0 \ \forall i$ and $C_j^\epsilon = \emptyset \ \forall j$. We are now ready to formally define stability of configurations and stability of preference types.

DEFINITION 5 A configuration γ^0 of incumbent population I is *stable* with respect to an invasion by mutant population J if it is balanced and if there exists $\epsilon > 0$ such that, under any focal post-entry equilibrium γ_ϵ^0, $V_i \geq V_j \ \forall i \in I, j \in J$.

DEFINITION 6 A preference type τ_i is said to be stable with respect to (small) invasions by

[24] Dekel et al. (2007) start their analysis by studying observable types and later extend it to include unobservable types too. I stick here to observable types only, like is very common in the literature (see e.g. Bester and Guth 1998 and Heifetz et al. 2007a,b).

another preference type τ_j if an incumbent population of type τ_i has a stable configuration γ^0 with respect to an invasion by mutant population of type τ_j .

The goal of the analysis is to show that the preference type called in this paper "social type" is stable with respect to the preference type "convex type".

4.2 Results

PROPOSITION 6 *Social types are stable with respect to a small invasion of convex types.*

This result follows from the fact that there exists an "efficient" configuration γ^0 in which, upon the invasion, all the incumbent social types are already mutually cooperating with exactly $\bar{K}$ partners hence, under the focality principle, none of them can play C also against new invaders (i.e., $\forall j \in J$, $\mu\left(I \bigcap C_{-j}\right) = 0$).[25] Therefore, in equilibrium, also the convex types defect against the incumbent social types (i.e., $\forall j \in J$, $\mu\left(I \bigcap C_j\right) = 0$). Thus, the payoff of the convex types comes only from "within-group" interactions. Then, given that (i) their size is smaller than $\bar{K}$, (ii) they cannot maintain full cooperation with all the other players in J given their temptation to cheat "a little", and (iii) the underlying PD game is characterized by strategic complementarity ($\ell > g$), the fitness of convex types will be strictly smaller than $\bar{K}$, which is the fitness of the incumbent social types in any post-entry focal equilibrium (focal relative to γ^0). That is, we get that $V_i > V_j$ for any $i \in I$ and any $j \in J$. In fact it is straightforward to show stability of social types with respect to a small invasion by *any* type (because conditions (i) and (iii) above hold and suffice), and this holds even if types are unobservable.[26]

Proposition 6 shows immunity of the social-types' population to a small (and in fact even "medium-sized")[27] invasion of convex types. But can a large invasion be successful?

[25] More precisely, each incumbent social type can play C in a post-entry focal equilibrium only against a zero mass of invaders. Strictly speaking, since individual players have zero mass, one can come up with "crazy" post-entry focal equilibria where some mutants (whose total mass is zero) gain the cooperation of a non-zero mass of incumbents (e.g., a certain mutant $j' \in J$ gains the cooperation of all $i \in I$, which would not raise $\mu\left(C_i^0\right)$ beyond $\bar{K}$). These artificial cases are of no interest and are therefore ignored.

[26] Observability plays no role here because what ensures stability is the existence of a pre-entry configuration in which the incumbents "exhaust their cooperative resources" (the "efficient" configuration γ^0) together with the requirement of focality that implies these incumbents cannot replace their cooperative partners.

[27] Because the proof applies to any invasion of size smaller than $\bar{K}$.

PROPOSITION 7 *If $\underline{K} \geq 1/4$, social types are stable with respect to a same-size (or smaller) invasion of convex types.*

Proposition 7 provides a sufficient condition for immunity of the incumbent population of social types to a large invasion by mutants with a convex cheating cost. This condition is stated in terms of a lower bound on $\underline{K}$ that guarantees stability. The rationale is that, starting from an efficient pre-entry configuration γ^0 of the incumbents as before, the convex types cannot get the cooperation of social types. Thus, for the convex types to have high fitness, they need to be able to establish sufficiently many cooperative relations among themselves without jeopardizing these relations by cheating a lot. To get a sense of what this requires, suppose that there exists $\pi \in (0,1)$ such that $Pr(s_{j_1 j_2} = C) = \pi$ for any $j_1, j_2 \in J$.[28] Then the proportion of within-group interactions in which a mutant j unilaterally defects is $\pi(1-\pi) \leq 1/4$. Given that the total mass of invaders $\mu(J)$ is no larger than $\mu(I) = 1$, it follows that the *mass* of unilateral defections $\pi(1-\pi)\mu(J)$ cannot exceed $1/4$, and the condition $\underline{K} \geq 1/4$ thus implies that this cannot be an equilibrium – the mutants have a profitable deviation to increase their rate of defection.[29] The formal proof takes into account also heterogeneous mixing strategies but the result and intuition stay the same.

The general lesson from this section is that a concave cheating cost is evolutionary stable: a population of types endowed with this cost is stable not only with respect to a *small* invasion by types with a convex cheating cost, but also with respect to "medium-sized" invasions of these types (up to $\bar{K}$ invaders) and even, if $\underline{K}$ is not too small, it is stable with respect to large invasions. This includes also stability with respect to (any size of) invasion by asocial types. Put differently, in order to successfully invade a population of social types, the convex types have to be numerous and their inclination to cheat has to be sufficiently restrained.

[28] Mixing strategies are required because the preferences of convex types imply that they will not cooperate with a cooperative partner unless they have sufficiently many other interactions in which they unilaterally defect, and because unilateral defection in not sustainable in pure strategies.

[29] The borderline case of $\pi(1-\pi) = \underline{K} = 1/4$ is not an equilibrium either, because $t_c'(\underline{K}) = g < \pi g + (1-\pi)\ell$, implying that j would like to increase the mass of unilateral defections ($\pi(1-\pi)$) beyond $\underline{K}$. This is why there is a weak inequality ($\underline{K} \geq 1/4$) in Proposition 7.

5 Conclusion

This paper shows that a simple and quite intuitive assumption about our social conscientiousness, and more specifically, about our moral cost of defecting from cooperation with others who cooperate with us, can explain a plethora of prevailing group behaviors, most prominently the mere existence of groups and the accompanying phenomenon of in-group bias. This assumption about the moral cost is that it increases concavely with the number of unilateral defections, hence maintaining large-scale cooperation becomes increasingly hard as the material incentives for defection keep rising with the size of the population. Inability to distinguish between social types, who are characterized by such a cost, and asocial types, who are not, gives rise either to costly signaling or to sustainable free-riding. The trade-off between the cost of signaling on the one hand, and the cost of having free-riders in the group on the other hand, explains why cooperative groups who engage in costly signaling can co-exist side by side with mixed groups, in which no signaling is practiced but free-riding is likely to happen. If the fraction of asocial types in society is small, the existence of signaling groups is shown to lead to an equilibrium that is Pareto inferior. This is so because those who are not members of the signaling groups have less social types to interact with, while the members of these groups themselves could have avoided paying the cost of signaling at the small price of interacting with few asocial types.

Furthermore, a simple evolutionary extension of the basic model shows that an incumbent population of social types is immune not only to invasion by a small group of types with a convex moral cost – as required by any standard evolutionary stability notion – but also to a "large" invasion of such types, provided that the convex cost triggers "enough" cheating. The intuition for this result is that when the social types are in an efficient equilibrium prior to the invasion, the convex mutants can establish cooperation only within themselves. Then, since the convex types have an unrestrained urge to cheat at least a little bit, this urge must be sufficiently restrained and their numbers must be sufficiently large in order for them to stand a chance to outperform the social (concave) incumbents.

The paper might seem to offer a very gloomy message: from gatherer-hunter societies

in ancient history to contemporary ethnic divisions all around the world, humanity seems to have no hope for unity, only a clash of (micro-)civilizations can endure in our world. However, a more optimistic reading is possible. The paper basically predicts that large-scale cooperation would break due to individuals' temptation to defect, which is in turn a result of large potential benefits from unilateral defection. Thus, if society wishes to establish sustainable large-scale cooperation, the lesson is that society should find a way to make defection less rewarding. One obvious way to do that is to sanction cheaters. If cheating is sanctioned, and in particular if it is sanctioned linearly, so that cheating twice as many people results in a double sanction, then, in principle, large scale cooperation *can* be maintained.

Acknowledgments

I would like to thank Benjamin Bachi, Roland Bénabou, Elchanan Ben-Porat, Ran Eilat, Bård Harstad, Sergiu Hart, Yuval Heller, Andrea Ichino, Rachel Kranton, Amnon Maltz, Andrea Mattozzi, Shiran Rachmilevitch, Yona Rubinstein, Moses Shayo, Paul Slovic, Daniel Spiro, Eyal Winter, Ro'i Zultan, seminar participants at the Hebrew University, University of Oslo, the European University Institute, IDC Herzliya and Ben-Gurion University, as well as participants at the Nordic Conference on Behavioral Economics, the Econometric Society European winter meetings, IMBESS meeting, THEEM conference, the 3rd Psy Games Workshop and LEG2019 conference for their valuable comments. I also wish to thank the anonymous referees for helping me improve this paper. Any remaining error is mine.

A Appendix: proofs

A.1 Proof of Proposition 1

LEMMA 1 *The equation $t(K) = Kg$ has a unique strictly positive solution $\bar{K}$ in $(0, 1-p)$. Moreover, $t(K) > Kg$ for any $K \in \left(0, \bar{K}\right)$ while $t(K) < Kg$ for any $K \in \left(\bar{K}, 1-p\right]$.*

PROOF: First assume by contradiction that there are at least two solutions to the equation $t(K) = Kg$ in the interval $(0, 1-p)$, denoted by K_1 and K_2. Since part 4 of Assumption 1 implies that for $\varepsilon \to 0^+$ we have $t(\varepsilon) > \varepsilon g$, we get that

$$\left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] t(\varepsilon) + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} t(K_2) > \left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] \varepsilon g + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} K_2 g = K_1 g,$$

while the concavity of $t(\cdot)$ implies that

$$\left[1 - \frac{K_1 - \varepsilon}{K_2 - \varepsilon}\right] t(\varepsilon) + \frac{K_1 - \varepsilon}{K_2 - \varepsilon} t(K_2) \leq t(K_1),$$

which contradicts the assumption that K_1 solves the equation $t(K) = Kg$. Thus $t(K) - Kg = 0$ has at most one solution in the range $[\varepsilon, 1-p]$. Next, note that $t(K) - Kg$ is strictly positive at $K = \varepsilon$, strictly negative at $K = 1-p$ (by part 5 of Assumption 1), and any possible discontinuity in between is an increase. Thus $t(K) - Kg = 0$ at least once in the range $[\varepsilon, 1-p]$. Overall, we get that $t(K) - Kg = 0$ *exactly* once in the range $[\varepsilon, 1-p]$, at which point $t(K) - Kg$ changes signs from positive to negative, so that $t(K) > Kg$ for any $K \in \left(0, \bar{K}\right)$ while $t(K) < Kg$ for any $K \in \left(\bar{K}, 1-p\right]$. *Q.E.D.*

PROOF: **(of Proposition 1)** First, it is immediate that $C_i \bigcap D_{-i} = \emptyset$ in any best response. As this holds for any individual i, it follows that $D_i \bigcap C_{-i} = \emptyset$ as well, hence, in equilibrium, $C_i = C_{-i}$ for any i. If i is an asocial type, then $C_i = \emptyset$ in any best response. It thus follows that $C_i = C_{-i} = \emptyset$ if $\tau(i) = as$. Suppose now that i is a social type ($\tau(i) = s$). For C_i to be his best response, it must be that $t(K) - Kg \geq 0$ for any $K \leq \mu(C_i)$, as otherwise, if there exists some $\tilde{K}$ s.t. $\tilde{K} \leq \mu(C_i)$ and $t(\tilde{K}) - \tilde{K}g < 0$, then i would have a profitable deviation to defecting against a set of players $\tilde{D}_i \subseteq C_{-i}$ s.t. $\mu\left(\tilde{D}_i\right) = \tilde{K}$. From Lemma 1 we get that, indeed, $t(K) - Kg \geq 0$ for any $K \leq \mu(C_i)$, if and only if $\mu(C_i) \leq \bar{K}$. *Q.E.D.*

A.2 Proof of Proposition 2

LEMMA 2 *Suppose* $\lim_{k\to 0} t'(k)$ *is well defined and let* $\Delta(k,p) \equiv t((1-p)k) - k[(1-p)g + p\ell]$. *If* $p > p_\Lambda$ *then* $\Delta(k,p) < 0$ *for any* k, *while if* $p < p_\Lambda$ *then* $\Delta(k,p) = 0$ *has a unique strictly positive solution* $K_p \in \left(0, \bar{K}\right]$, *and* $\Delta(k,p) \geq 0$ *if and only if* $k \leq K_p$. *Furthermore,* K_p *is strictly decreasing in* p *in the range* $[0, p_\Lambda)$.

PROOF: The conditions on $t(k)$ and on the payoffs of the game imply that for any given $p \in [0,1)$, we have $\Delta(0,p) = 0$ and $\Delta(k,p) < 0$ for any $k > \bar{K}$ (because $t((1-p)k) \leq t(k)$, $[(1-p)g + p\ell] \geq g$ and for any $k > \bar{K}$ we have $t(k) < kg$). Moreover,

$$\lim_{k\to 0} \frac{\partial \Delta(k,p)}{\partial k} = (1-p) \lim_{k\to 0} t'((1-p)k) - [(1-p)g + p\ell] = (1-p)\Lambda - p\ell$$

and $\Delta(k,p)$ is weakly concave in k. Thus, if $p > p_\Lambda$ then $\lim_{k\to 0} \frac{\partial \Delta(k,p)}{\partial k} < 0$ hence $\Delta(k,p) < 0$ for any k, while if $p < p_\Lambda$ then $\lim_{k\to 0} \frac{\partial \Delta(k,p)}{\partial k} > 0$ and so $\Delta(k,p) = 0$ has a unique strictly positive solution K_p, where $\Delta(k,p) > 0$ for any $k < K_p$, and $\Delta(k,p) < 0$ for any $k > K_p$.[30] Finally $\frac{\partial \Delta(k,p)}{\partial p} = -kt'((1-p)k) - k(\ell - g)[(1-p)g + p\ell] < 0$ (for $k > 0$), hence $\Delta(k,p)$ is strictly decreasing in p, which means that for any $\{p_1, p_2 | p_1 < p_2\}$ we have $\Delta(k,p_2) < 0$ for any $k \geq K_{p_1}$, and so $K_{p_2} < K_{p_1}$, i.e., K_p is strictly decreasing in p. *Q.E.D.*

PROOF: **(of Proposition 2)** Like in the proof of Proposition 1, it is immediate that if $\tau(i) = as$ then $C_i = \emptyset$. Suppose now instead that i is a social type. For $C_i \neq \emptyset$ to be his best response, any deviation to defecting against $\tilde{D}_i \subseteq C_i$ should be non profitable. The expected loss of payoff from this deviation is $\Delta(k,p)$, which by Lemma 2 is (strictly) negative if either $p > p_\Lambda$ or $\mu\left(\tilde{D}_i\right) > K_p$ and (weakly) positive if $p < p_\Lambda$ and $\mu\left(\tilde{D}_i\right) \leq K_p$. Hence $C_i = \emptyset$ if $p > p_\Lambda$ while $\mu(C_i) \leq K_p$ if $p < p_\Lambda$, where K_p is strictly decreasing in p in the range $[0, p_\Lambda)$ by Lemma 2. Suppose now that $p < p_\Lambda$ (hence $\mu(C_i) \leq K_p$), and take $j \in C_i (\neq \emptyset)$ s.t. $\tau(j) = s$. Since Proposition 2 characterizes equilibria that are partitional strategy profiles, it follows that either $C_j \neq \emptyset$, in which case i and j are in the same partition hence $i \in C_j$; or $C_j = \emptyset$, in which case j defects against all players in C_i, whose mass is $\mu(C_i) \leq K_p$, which cannot be true in equilibrium as

[30] Note that if $\Delta(k,p)$ is discontinuous due to discontinuity of $t(k)$, then the same logic of the proof to Lemma 1 applies here too, and so the existence of a solution is guaranteed.

j, being a social type, would have a profitable deviation to playing C against them instead.[31] *Q.E.D.*

A.3 Proof of Proposition 3

PROOF: The IF part: suppose the two conditions – individual rationality and cost differentiation – hold, and consider a strategy profile in which a mass $K \in \left[x_s, \hat{K} = \min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\}\right]$ of players use the signal $x \neq x_0$ (this is feasible because the two aforementioned conditions imply that $\left[x_s, \hat{K}\right]$ is not an empty set). Given that $K \leq \frac{x_{as}}{1+g}$, it is not profitable for an asocial type i to have $\sigma_i(as) = x$, even if $\widetilde{C}_i = \emptyset$, because he cannot gain by this strategy a payoff larger than $(1+g)K - x_{as} \leq 0$, whereas he can secure a weakly positive payoff by choosing the strategy $(x_0, \emptyset)$ instead. It thus follows that all the K players in $\sigma^{-1}(x)$ are social types (which is feasible because $\bar{K} < 1-p \Rightarrow \hat{K} < 1-p$).[32] Now let all these (social) players in $\sigma^{-1}(x)$ play the strategy $\left(x, \sigma^{-1}(x)\right)$ while all players outside $\sigma^{-1}(x)$ play the strategy $(x_0, \emptyset)$. This is a group formation (i.e. an equilibrium) because (1) a player $i \in \sigma^{-1}(x)$ has a positive payoff of $K - x_s$ hence cannot profit by deviating to $\sigma_i(\text{s}) \neq x$ (which would yield a zero payoff – at best – under any $\gamma \in \Gamma^*$); (2) this player also cannot profit by deviating to $\left(x, \widetilde{C}_i\right)$ where $\widetilde{C}_i \neq \sigma^{-1}(x)$ because players outside $\sigma^{-1}(x)$ are not in $\widetilde{C}_{-i,x}$ and because $\widetilde{C}_i \subsetneq \sigma^{-1}(x)$ is not a best response given that $K \leq \hat{K} \leq \bar{K}$ hence $t(k) \geq kg$ for any $k \leq K$;[33] and (3) a player $i \notin \sigma^{-1}(x)$ cannot profit by deviating to $\sigma_i(\tau) = x$ because it will lead with certainty to $\widetilde{C}_{-i,x} \neq \emptyset$ hence to a weakly negative payoff.

The ONLY IF part: A group formation in which $\widetilde{C}_i = \widetilde{C}_{-i,x} = \sigma^{-1}(x) \ \forall i \in \sigma^{-1}(x)$ requires that $\sigma^{-1}(x)$ will contain only social types (because otherwise any asocial type in $\sigma^{-1}(x)$ could profit by deviating to $(x, \emptyset)$, in contradiction to this being an equilibrium). Consider then again a partitional strategy profile where a mass K of social types play the strategy $\left(x, \sigma^{-1}(x)\right)$, but this time $K \notin \left[x_s, \hat{K}\right]$, either because $\left[x_s, \hat{K}\right]$ is an empty set (implying that at least one of the two conditions – individual rationality and cost differentiation – does not hold) or despite this set being non empty. If $K > \frac{x_{as}}{1+g} \geq \hat{K}$ then $\sigma^{-1}(x)$ will contain also asocial types, because an asocial type who plays the strategy $(x, \emptyset)$ gets a strictly positive payoff hence has no profitable deviation. In this case the requirement that $\widetilde{C}_i = \widetilde{C}_{-i,x} = \sigma^{-1}(x) \ \forall i \in \sigma^{-1}(x)$ will not be satisfied. If instead $K \leq \frac{x_{as}}{1+g}$, yet $K \notin \left[x_s, \hat{K}\right]$, the strategy $\left(x, \sigma^{-1}(x)\right)$ for a social player $i \in \sigma^{-1}(x)$ is unsustainable in equilibrium because either $\mu\left(\widetilde{C}_i\right) = K < x_s$, in which case i has a profitable deviation to $(x_0, \emptyset)$; or $\mu\left(\widetilde{C}_i\right) = K > \bar{K}$, in which case i has a profitable deviation to $(x, \emptyset)$. *Q.E.D.*

[31] In the borderline case where $\mu(C_i) = K_p$ there can be countably many players j (with total mass of 0) for whom $C_j = \emptyset$ and it is still an equilibrium.

[32] Strictly speaking, the case of $K = \frac{x_{as}}{1+g}$ is borderline because if we choose a strategy profile in which an asocial player is assigned the strategy $(x, \emptyset)$ while there are $K = \frac{x_{as}}{1+g}$ other (social) players in $\sigma^{-1}(x)$ who are assigned $\widetilde{C}_i = \sigma^{-1}(x)$, this asocial player cannot strictly increase his payoff by deviating from his assigned strategy (as he gains a zero payoff either way). This borderline case is discarded by ensuring, as stated in footnote 18, that if $\min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\} = \frac{x_{as}}{1+g}$, a strict inequality will apply in Proposition 3, i.e. the size of signaling groups will be $\mu\left(\sigma^{-1}(x)\right) \in \left[x_s, \min\left\{\bar{K}, \frac{x_{as}}{1+g}\right\}\right)$.

[33] Here k would be the mass of $\sigma^{-1}(x) \setminus \widetilde{C}_i$.

A.4 Proving the welfare results (Section 3.2)

PROOF: **(of Proposition 4)** Given that, under γ_1, there exists $x \neq x_0$ for which $\mu\left(\sigma^{-1}(x)\right) \neq 0$, it follows that the proportion of asocial types among the non-signalers, $q \equiv \frac{p}{1-\mu\left(\bigcup\limits_{x \in X} \sigma^{-1}(x)\right)}$, is strictly greater than p, their proportion in society. Construct now a partitional strategy profile γ_2 as follows: (i) set $\sigma_i(\tau) = x_0$ for any player i of type τ (implying $\sigma^{-1}(x) = \emptyset$ for any $x \neq x_0$); (ii) if $\tau(i) = s$ and, under γ_1, $\sigma_i(s) = x_0$ and $\mu\left(\widetilde{C}_i\right)|_{\gamma_1} \neq 0$, set $\widetilde{C}_i$ to be of size $\mu\left(\widetilde{C}_i\right)|_{\gamma_2} = \mu\left(\widetilde{C}_i\right)|_{\gamma_1}$;[34] and (iii) divide all remaining players, if they exist, into groups ($\widetilde{C}_i$'s) of non-zero mass.[35] Under this new group formation γ_2, each $\widetilde{C}_i$ has a proportion of p asocial types, and the fact that $p < \frac{1}{1+\ell}$ implies that the payoff of player i of type s with strategy $\left(x_0, \widetilde{C}_i\right)$, which is given by the expression $U_i = (1-p)\,\mu\left(\widetilde{C}_i\right) - \ell p \mu\left(\widetilde{C}_i\right)$, is positive and increases in $\mu\left(\widetilde{C}_i\right)$. Furthermore, U_i decreases in p. We thus get that, for any social type i for whom, under γ_1, $\sigma_i(s) = x_0$,

$$U_i(\gamma_1) = (1-q)\,\mu\left(\widetilde{C}_i\right)|_{\gamma_1} - \ell q \mu\left(\widetilde{C}_i\right)|_{\gamma_1} < (1-p)\,\mu\left(\widetilde{C}_i\right)|_{\gamma_2} - \ell p \mu\left(\widetilde{C}_i\right)|_{\gamma_2} = U_i(\gamma_2).$$

Q.E.D.

PROOF: **(of Proposition 5)** Suppose first that $\sigma^{-1}(x) = \emptyset$ for all $x \neq x_0$, i.e. no player is signaling. Then a social type i who belongs to a mixed group of some size K has $U_i = K[1 - p(1+\ell)]$.[36] This expression is negative if $p > \frac{1}{1+\ell}$, but positive and increasing in the group size K if $p \leq \frac{1}{1+\ell}$, where, given p, it reaches its maximal value $f(p) \equiv K_p[1 - p(1+\ell)]$ when the group is of its maximal feasible size in equilibrium, K_p. Since, for $p \leq \frac{1}{1+\ell}$, both K_p and $[1 - p(1+\ell)]$ are positive and decreasing in p (see Proposition 2), we get that $f(p)$ is also positive and (strictly) decreasing in p at $p \in \left[0, \frac{1}{1+\ell}\right]$. Moreover, $f(0) = \bar{K} > \hat{K} - x_s$, and $f\left(\frac{1}{1+\ell}\right) = 0$. Hence, given that (by assumption) $\hat{K} - x_s > 0$, there is a unique solution to equation (4), denoted by p_c, where $p_c \in \left(0, \frac{1}{1+\ell}\right)$.

Now suppose instead that $\sigma^{-1}(x) \neq \emptyset$ for one or more $x \neq x_0$, i.e. we move to considering group formations with signaling groups. Then it immediately follows that $max_{\gamma \in \Gamma^*}\{U_i(\gamma) | \sigma_i(s) \neq x_0\} = \hat{K} - x_s$.[37]

[34] That is, make sure that all the social-type members of mixed groups under group formation γ_1 will still be members of mixed groups of the same size under γ_2. We know that this is feasible since (1) $K_q < K_p$ (see Proposition 2), hence $\mu\left(\widetilde{C}_i\right)|_{\gamma_2} = \mu\left(\widetilde{C}_i\right)|_{\gamma_1} \leq K_q < K_p$; and (2) the *mass* of social types in these new groups must be higher than under γ_1 (because $p < q \Rightarrow (1-p)\,\mu\left(\widetilde{C}_i\right)|_{\gamma_2} > (1-q)\,\mu\left(\widetilde{C}_i\right)|_{\gamma_1}$), implying that these groups can be constructed by taking the mixed groups that existed under γ_1 (denoted here by $\widetilde{C}_i|_{\gamma_1}$) and replacing some of their asocial-type members by social types who either belonged to signaling groups under γ_1 or were not members of any group.

[35] This is feasible in light of part (II) of Proposition 2 and given that $p < \frac{1}{1+\ell}$ hence $p < p_\Lambda$.

[36] If i belongs to no group we can use the same expression and plug in $K = 0$.

[37] This value of U_i is always achievable in an equilibrium that has exactly one signaling group of size $\hat{K} \leq 1 - p$.

We thus get that, if $p < p_c$, then

$$max_{\gamma\in\Gamma^*}\left\{U_i(\gamma)|\sigma_i\left(s\right)\neq x_0\right\} = \hat{K} - x_s < f\left(p\right) = max_{\gamma\in\Gamma^*}\left\{U_i(\gamma)|\sigma_i\left(s\right)= x_0\right\}.$$

If, on the other hand, $p \geq p_c$, then the converse is true, i.e.

$$max_{\gamma\in\Gamma^*}\left\{U_i(\gamma)|\sigma_i\left(s\right)\neq x_0\right\} \geq max_{\gamma\in\Gamma^*}\left\{U_i(\gamma)|\sigma_i\left(s\right)= x_0\right\}$$

(with strict inequality for $p > p_c$). *Q.E.D.*

A.5 Proofs for Section 4

A.5.1 Proof of Proposition 6

PROOF: Take the pre-entry efficient and balanced equilibrium in which, for any incumbent i, $\mu\left(C_i\right) = \mu\left(C_{-i}\right) = \bar{K}$. Thus we have $V_i = \bar{K}\ \forall i \in I$. In light of definition 4, we know that also in any post-entry equilibrium, and for any $i \in I$, we have $\mu\left(C_i\right) \geq \bar{K}$. Thus, given Proposition 1, there is no post-entry equilibrium in which a player i plays C against a non-zero mass of mutants in J on top of playing C against the $\bar{K}$ players who constitute C_i, i.e. $\mu\left(I \bigcap C_{-j}\right) = 0\ \forall j \in J$.[38] Recalling the strategic complementarity of the PD game's payoffs ($\ell > g$), it immediately follows that, for any small ($< \bar{K}$) invasion of mutants, the invading players cannot get, in expectation, a payoff $V_j \geq \bar{K}$. We thus get that $V_i > V_j\ \forall i \in I, j \in J$ and furthermore, as required by the stability concept, there is no unraveling of the pre-entry equilibrium behavior of the incumbents. *Q.E.D.*

A.5.2 Proof of Proposition 7

In order to show that social types are stable with respect to invasions by convex types, it is enough that there exists a stable configuration with respect to an invasion by convex types. Thus the proof here will apply to the same configuration as in the proof of Proposition 6, i.e. one where, for any incumbent i, $\mu\left(C_i\right) = \mu\left(C_{-i}\right) = \bar{K}$. This configuration will be denoted by γ^0.

For simplicity and to reduce notations, I will write the proof as if individual players have (the same) very small but non-zero mass (so that deviating even against one person has an effect on payoffs). Furthermore, with some abuse of notation, I will say that $j' \in C_j$ if $P\left(s_{j,j'} = C\right) \neq 0$ and that $j' \in D_j$ if $P\left(s_{j,j'} = C\right) \neq 1$. Thus, it may well be that $j' \in C_j$ and also $j' \in D_j$. Similar notations apply to C_{-j} and D_{-j}. In contrast, $\mu\left(C_j\right)$ will denote the expected mass against whom player j *actually* plays C, so that it could be that $\mu\left(C_j\right) < |J'|$, where $|J'|$ is the mass of the set of players $j' \in C_j$. Again, similar notations

[38] There can be a zero mass of mutants for whom this does not hold but I ignore this technicality in my analysis – see footnote 27.

apply to C_{-j} and D_{-j}.[39]

The idea of the proof is to single out the player who gains the least cooperation from his opponents and show that for this player's strategy to be a best response (i.e., to ensure he cheats in at least $\underline{K}$ interactions), the total mass of invaders has to be larger than $4\underline{K}$. This proof is complex because players are free to use different mixing strategies against different opponents. Therefore, to develop the argument rigorously, the proof builds on 7 different lemmas that gradually lead to Proposition 7 and is structured as follows. First, Lemma 3 establishes that in any post-entry equilibrium (w.r.t. the aforementioned configuration γ^0) incumbents do not cooperate with invaders. This implies that invaders cannot gain a positive payoff out of their interactions with incumbents – any positive payoff they gain must originate from within-group interactions. The following lemmas therefore focus only on interactions within the group of invaders J. Lemma 4 establishes that if player j_1 cooperates with player j_2 with some non-zero probability, then also player j_2 cooperates with player j_1 with a non-zero probability. Lemma 5 builds on Lemma 4 to show that if a player uses a mixed strategy against a given set of players, then these players use *an identical* mixing strategy when playing against this player (with probability of cooperation denoted by q_j). This result is later used in Lemma 8 to establish that q_j that characterizes the invader who gains the least cooperation from his opponents (i.e. $min_{j\in J}\{q_j\}$) is strictly lower than 1. But in order to prove this result, two auxiliary lemmas are needed as well: Lemma 6, which shows a monotonicity property of the mixing strategy: the probability with which a player cooperates against different opponents is increasing in the probability with which these opponents cooperate with him; and Lemma 7, which proves that – in a post-entry equilibrium – an invader j who has a strictly positive payoff must be mixing against (at least) a subset of the other invaders (which by Lemma 5 implies the existence of $min_{j\in J}\{q_j\}$). Finally, Lemma 9 is the crucial element of the proof: it shows that the existence of an invader j with a strictly positive payoff requires that the invasion is sufficiently large. In particular, it builds on Lemma 8 ($min_{j\in J}\{q_j\} < 1$) and shows that in order to ensure that the player who gains the least cooperation from his opponents cheats in at least $\underline{K}$ interactions, the invasion has to be of size $|J|$ that is larger than $4\underline{K}$. This ultimately leads to proving Proposition 7, because if $\underline{K} \geq 1/4$ while $|J| \leq 1$, then $|J|$ cannot be larger than $4\underline{K}$, hence by Lemma 9 no invader j can have a strictly positive payoff.

LEMMA 3 *In any post-entry equilibrium (w.r.t. configuration γ^0), $I \bigcap C_{-j} = \emptyset \; \forall j \in J$.*

PROOF: See the proof of Proposition 6. *Q.E.D.*

LEMMA 4 *If, in equilibrium, $P(s_{j_1,j_2} = C) \neq 0$ then $P(s_{j_2,j_1} = C) \neq 0$, $\forall j_1, j_2 \in J$.*

PROOF: Suppose by contradiction that $P(s_{j_1,j_2} = C) \neq 0$ while $P(s_{j_2,j_1} = C) = 0$. Then j_1 has a profitable deviation to $P(s_{j_1,j_2} = C) = 0$, in contradiction to the assumption of this being an equilibrium. *Q.E.D.*

[39]E.g., if all players in J play C against player j with probability $1/2$, then $C_{-j} = J$ while $\mu(C_{-j}) = |J|/2$.

LEMMA 5 *If* $C_{j_1} \bigcap D_{j_1} \neq 0$, *then there exists a value* $q_{j_1} \neq 0$ *s.t.* $P(s_{j,j_1} = C) = q_{j_1}$ *for any* $j \in C_{j_1} \bigcap D_{j_1}$.[40]

PROOF: Take two players $j_2, j_3 \in C_{j_1} \bigcap D_{j_1}$.[41] Then, by Lemma 4, we know that $P(s_{j_2,j_1} = C) \neq 0$ and $P(s_{j_3,j_1} = C) \neq 0$. Suppose by contradiction and w.l.o.g. that $(0 <) P(s_{j_2,j_1} = C) < P(s_{j_3,j_1} = C)$. Then, since (by assumption) $P(s_{j_1,j_2} = C) \neq 0$ and $P(s_{j_1,j_3} = C) \neq 1$, player j_1 has a profitable deviation to strictly decrease $P(s_{j_1,j_2} = C)$ and strictly increase $P(s_{j_1,j_3} = C)$ in a way that keeps the same probability of cheating an opponent but increases the material payoff.[42] Thus $P(s_{j_2,j_1} = C) = P(s_{j_3,j_1} = C) \equiv q_{j_1}$, where $q_{j_1} \neq 0$ by Lemma 4 as mentioned. *Q.E.D.*

LEMMA 6 *If* $\exists j_1, j_2, j_3 \in J$ *s.t.* $P(s_{j_2,j_1} = C) < P(s_{j_3,j_1} = C)$, *then* $P(s_{j_1,j_2} = C) \leq P(s_{j_1,j_3} = C)$.

PROOF: Suppose by contradiction that $P(s_{j_2,j_1} = C) < P(s_{j_3,j_1} = C)$ while $P(s_{j_1,j_2} = C) > P(s_{j_1,j_3} = C)$. Then, as explained in the proof of Lemma 5, player j_1 has a profitable deviation to strictly decrease $P(s_{j_1,j_2} = C)$ and strictly increase $P(s_{j_1,j_3} = C)$ in a way that keeps the same probability of cheating an opponent but increases the material payoff. *Q.E.D.*

LEMMA 7 *If in the post-entry equilibrium* $\exists j \in J$ *for whom* $V_j > 0$, *then* $\exists j' \in J$ *such that* $j' \in C_j \bigcap D_j$.

PROOF: If $V_j > 0$ for some player $j \in J$, it must be that $C_{-j} \neq \emptyset$. Thus also $D_j \bigcap C_{-j} \neq \emptyset$, as otherwise j is not cheating at all and, given that $t'_c(0) < g$, j has a profitable deviation to cheat (i.e. to increase $\mu(D_j \bigcap C_{-j})$). Hence, given Lemma 3, $\exists j' \in D_j \bigcap C_{-j}$, where $j' \in D_j \Rightarrow P(s_{j,j'} = C) \neq 1$ and, given Lemma 4, $j' \in C_{-j} \Rightarrow P(s_{j,j'} = C) \neq 0 \Rightarrow j' \in C_j$, hence $j' \in C_j \bigcap D_j$. *Q.E.D.*

LEMMA 8 *If in the post-entry equilibrium* $\exists j \in J$ *for whom* $V_j > 0$, *then* $min_{j \in J}\{q_j\} < 1$.[43]

PROOF: Let $\widetilde{j} \in J$ be a player for whom $V_{\widetilde{j}} > 0$ in the post-entry equilibrium. Then, by Lemma 7, $\exists j' \in C_{\widetilde{j}} \bigcap D_{\widetilde{j}}$ for whom, by Lemma 5, $P\left(s_{j',\widetilde{j}} = C\right) = q_{\widetilde{j}}$, hence $min_{j \in J}\{q_j\}$ exists and is well defined. Suppose now by contradiction that $min_{j \in J}\{q_j\} = 1$. Then $P\left(s_{j',\widetilde{j}} = C\right) = q_{\widetilde{j}} = 1$. For this to be a best response of j' to $P\left(s_{\widetilde{j},j'} = C\right) \notin \{0,1\}$, there must exist a player $j'' \in D_{j'} \bigcap C_{-j'}$, as otherwise player j' would not be cheating anyone (i.e. $D_{j'} \bigcap C_{-j'} = \emptyset$), hence could profit by decreasing $P\left(s_{j',\widetilde{j}} = C\right)$ given that $\widetilde{j} \in C_{-j'}$ and $t'_c(0) < g$. This implies by Lemma 4 that $P(s_{j',j''} = C) \notin \{0,1\}$. But then either $P(s_{j'',j'} = C) \notin \{0,1\}$, which (by Lemma 5) implies that $q_{j'} < 1$, in contradiction to the assumption that

[40] Recall that $j \in C_{j_1} \bigcap D_{j_1}$ iff $P(s_{j_1,j} = C) \notin \{0,1\}$.

[41] If $C_{j_1} \bigcap D_{j_1}$ contains only one player j_2 then $q_{j_1} = P(s_{j_2,j_1} = C)$ automatically exists and $q_{j_1} \neq 0$ by Lemma 4.

[42] Decreasing $P(s_{j_1,j_2} = C)$ by $\triangle P$ while increasing $P(s_{j_1,j_3} = C)$ by $\frac{P(s_{j_2,j_1}=C)}{P(s_{j_3,j_1}=C)} \triangle P$ $(< \triangle P)$ results in the same probability of playing C against C and D against C, but also in a decrease of the expected material loss by $(\mu(j)\cdot)\ \ell \triangle P \left[1 - \frac{P(s_{j_2,j_1}=C)}{P(s_{j_3,j_1}=C)}\right] > 0$.

[43] q_j is defined in Lemma 5.

$min_{j\in J}\{q_j\}=1$; or $P(s_{j'',j'}=C)=1$, which would contradict Lemma 6 (with player j' in the role of player j_1 in that lemma, and players j and j'' in the roles of players j_2 and j_3 respectively). *Q.E.D.*

LEMMA 9 *If $|J|\leq 4\underline{K}$, then there is no post-entry equilibrium (w.r.t. configuration γ^0) in which $\exists j\in J$ with $V_j>0$.*

PROOF: The existence of a player $j\in J$ for whom $V_j>0$ in the post-entry equilibrium implies, by Lemma 8, the existence of a player j_m such that (i) $j_m=argmin_{j\in J}\{q_j\}$; (ii) $q_{j_m}<1$ and (iii) for any $j'\in D_{j_m}\bigcap C_{-j_m}\neq\emptyset$, $P(s_{j_m,j'}=C)\notin\{0,1\}$ and $P(s_{j',j_m}=C)=q_{j_m}\notin\{0,1\}$.[44] Denote now by J' the set of players j' s.t. $j'\in D_{j_m}\bigcap C_{-j_m}$.[45] We thus get that $\mu(D_{j_m}\bigcap C_{-j_m})=q_{j_m}E_{j'\in J'}[1-q_{j'}]|J'|$ (where the existence of a value $q_{j'}\ \forall j'\in J'$ follows from applying Lemma 5 with j' in the role of j_1 in that lemma). Then, given that $E_{j'\in J'}[1-q_{j'}]\leq 1-q_{j_m}$, we get

$$\mu\left(D_{j_m}\bigcap C_{-j_m}\right)=q_{j_m}E_{j'\in J'}[1-q_{j'}]|J'|\leq q_{j_m}(1-q_{j_m})|J'|\leq\frac{1}{4}|J'|\leq\frac{1}{4}|J|.$$

Next note that for j_m's aforementioned strategy to be a best response, he should have no profitable deviation to decrease $P(s_{j_m,j'}=C)$. Decreasing $P(s_{j_m,j'}=C)$ by a small ΔP would increase his material payoff by $[q_{j_m}g+(1-q_{j_m})\ell]\Delta P|J'|$ while also increasing the cheating cost by $q_{j_m}\Delta P|J'|t'_c(\mu(D_{j_m}\bigcap C_{-j_m}))$. Since this deviation has to be no profitable, the following condition must hold

$$q_{j_m}g+(1-q_{j_m})\ell\leq q_{j_m}t'_c\left(\mu\left(D_{j_m}\bigcap C_{-j_m}\right)\right),$$

which, given that $t'_c(\underline{K})=g$ and $g<\ell$, implies

$$t'_c(\underline{K})=g<q_{j_m}g+(1-q_{j_m})\ell\leq q_{j_m}t'_c\left(\mu\left(D_{j_m}\bigcap C_{-j_m}\right)\right)<t'_c\left(\mu\left(D_{j_m}\bigcap C_{-j_m}\right)\right),$$

and the fact that $t'_c(k)$ is increasing in k then implies that $\underline{K}<\mu(D_{j_m}\bigcap C_{-j_m})\leq\frac{1}{4}|J|$, contradicting the condition $|J|\leq 4\underline{K}$ in the lemma. *Q.E.D.*

PROOF: **(of Proposition 7)** If $\underline{K}\geq 1/4$ and $|J|\leq 1$, then $|J|\leq 4\underline{K}$. Hence, by Lemma 9, $V_j\leq 0\ \forall j\in J$, and so $V_j<V_i\ \forall i\in I, j\in J$. Furthermore, as required by the stability concept, there is no unraveling of the pre-entry equilibrium behavior of the social types because they keep playing the same strategy among themselves. *Q.E.D.*

[44] Item (iii) follows from the existence of player j_m and from the definition and properties of q_{j_m} as they appear in Lemma 5, and $P(s_{j_m,j'}=C)\neq 0$ by the fact that $j'\in C_{-j_m}$ and by Lemma 4.

[45] Recall that $|J'|$ could be larger than $\mu(D_{j_m}\bigcap C_{-j_m})$ because J' contains any player j' for whom $P(s_{j_m,j'}=C)\neq 1$ and $P(s_{j',j_m}=C)\neq 0$, while $\mu(D_{j_m}\bigcap C_{-j_m})$ refers to the expected realizations of cheating (i.e. instances in which j_m plays D while j' plays C).

References

[1] Aimone, J. A., Iannaccone, L. R., Makowsky, M. D., and Rubin, J. (2013), "Endogenous group formation via unproductive costs," *The Review of economic studies*, 80(4), 1215-1236.

[2] Akerlof, G. A. and Kranton, R. E. (2000), "Economics and Identity," *The Quarterly Journal of Economics*, 115(3), 715-753.

[3] Amir, A., Kogut, T., & Bereby-Meyer, Y. (2016), "Careful cheating: people cheat groups rather than individuals," *Frontiers in psychology*, Vol. 7.

[4] Bernard, M., Hett, F., and Mechtel, M. (2016), "Social Identity and Social Free-Riding," *European Economic Review*.

[5] Berman, E. (2000), "Sect, Subsidy, and Sacrifice: An Economist's View of Ultra-Orthodox Jews," *The Quarterly Journal of Economics*, 115(3), 905-953.

[6] Bester, H., and Guth, W. (1998), "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization*, 34(2), 193-209.

[7] Boyd, R. and Richardson, P. J. (1988), "The evolution of reciprocity in sizable groups," *Journal of Theoretical Biology*, 132, 337–356.

[8] Charness, G. and Rabin, M. (2002), "Understanding social preferences with simple tests," *The Quarterly Journal of Economics* 117(3), 817–869.

[9] Chen, D. L., Michaeli, M., & Spiro, D. (2017), "Non-Confrontational Extremists," TSE Working Paper No. 16-694.

[10] Dekel, E., Ely, J. C., & Yilankaya, O. (2007), "Evolution of preferences," *The Review of Economic Studies*, 74(3), 685-704.

[11] de Dreu, C. K. W. (2010), "Social value orientation moderates ingroup love but not outgroup hate in competitive intergroup conflict," *Group Processes Intergroup Relations*, 13(6), 701-713.

[12] Eshel, I., Samuelson, L., & Shaked, A. (1998). Altruists, egoists, and hooligans in a local interaction model. American Economic Review, 157-179.

[13] Fehr, E., and Schmidt, K. M. (1999), "A theory of fairness, competition, and cooperation," *Quarterly journal of Economics*, 114(3), 817-868.

[14] Fukuyama, F. (1995), Trust. New York: Free Press.

[15] García-Martínez, J. A., & Vega-Redondo, F. (2015). Social cohesion and the evolution of altruism. Games and Economic Behavior, 92, 74-105.

[16] Guth, W., and Yaari, M. E. (1992), Explaining reciprocal behavior in simple strategic games: An evolutionary approach. In: Explaining process and change: approaches to evolutionary economics. University of Michigan Press, Ann Arbor, MI, 23-34.

[17] Heifetz, A., Shannon C., and Spiegel, Y. (2007a), "The Dynamic Evolution of Preferences." *Economic*

Theory, 32(2): 251-86.

[18] Heifetz, A., Shannon C., and Spiegel, Y. (2007b), "What to Maximize If You Must." *Journal of Economic Theory*, 133(1): 31-57.

[19] Herold, F. (2012), "Carrot or Stick? The Evolution of Reciprocal Preferences in a Haystack Model," *The American Economic Review*, 102(2), 914- 940.

[20] Iannaccone, L. R. (1992), "Sacrifice and Stigma: Reducing Free-riding in Cults, Communes, and Other Collectives," *Journal of Political Economy*, 100(2), 271-291.

[21] ——— (1994), "Why strict churches are strong," *American Journal of Sociology*, 99(5), 1180-1211.

[22] Kamada, Y., & Kojima, F. (2014), "Voter preferences, polarization, and electoral policies. American Economic Journal: Microeconomics," 6(4), 203-236.

[23] Levy, G., and Razin, R. (2012), "Religious beliefs, religious participation, and cooperation," *American economic journal: microeconomics*, 4(3), 121-151.

[24] Marwell, G., & Ames, R. E. (1979). Experiments on the provision of public goods. I. Resources, interest, group size, and the free-rider problem. American Journal of sociology, 84(6), 1335-1360.

[25] Nowak, M. A. and Sigmund, K. (1998), "Evolution of indirect reciprocity by image scoring," *Nature*, 393, 573-577.

[26] Osborne, M. J. (1995), "Spatial models of political competition under plurality rule: a survey of some explanations of the number of candidates and the positions they take," *The Canadian Journal of Economics* 28 (2), 261–301.

[27] Porta, R. L., Lopez-De-Silane, F., Shleifer, A., and Vishny, R. W. (1996), "Trust in large organizations," National Bureau of Economic Research (No. w5864).

[28] Ruffle, B. J., and Sosis, R. (2006), "Cooperation and the In-Group-Out-Group Bias: A Field Test on Israeli Kibbutz Members and City Residents," *Journal of Economic Behavior and Organization*, 60(2), 147-163.

[29] Schumacher, H., Kesternich, I., Kosfeld, M., & Winter, J. (2017), "One, two, many—Insensitivity to group size in games with concentrated benefits and dispersed costs," *The Review of Economic Studies*, 84(3), 1346-1377.

[30] Slovic, P. (2007), ""If I look at the mass I will never act": Psychic numbing and genocide," *Judgment and Decision Making*, 2(2), 79–95

[31] Spence, A. M. (1974), Market Signalling . Harvard University Press.

[32] Suzuki, S., and Akiyama, E.(2005), "Reputation and the evolution of cooperation in sizable groups," *Proceeding of the Royal Society B*, 272, 1373–1377.

[33] Wilson, E. (1978). 0.(1975) Sociobiology: The New Synthesis.