The Origins of In-Group Bias and the Cost of Signaling Sociality*

Moti Michaeli†

Abstract

All around us we see that people form groups, that these groups are often indifferent to other groups in the best case, or hostile to other groups in the worst case, and that many cohesive groups push their members to signal their belonging to the group by performing actions that involve some self-sacrifice. In this paper we show that the tendency of people to form groups of limited size and to show in-group favoritism can be traced back to a fundamental characteristic of our mentality – the psychological cost we pay for not reciprocating the kind actions of others. Moreover, a difficultly in spotting asocial individuals, who are not subject to this cost, may lead to the emergence of costly signaling of sociality. Groups that adopt such practices are characterized by a high level of cooperation among group members, and can coexist alongside groups with no signaling and a lower level of cooperation. When such coexistence is sustained by an envy-free equilibrium, the welfare of all individuals is strictly lower than would have been if signaling was impossible. Thus, the cost of signaling is twofold: the individual cost of producing it, and the social cost of ending up in an inferior equilibrium.

*I would like to thank Moshe Shayo and Eyal Winter for their guidance and precious advices. I wish also to thank Elchanan Ben-Porat, Bård Harstad, Andrea Ichino, Chloe Le Coq, Yosef Rinott, Assaf Romm, Elyashiv Wiedman, and participants at the Hebrew University, the University of Oslo, and the 8th Nordic Conference on Behavioral and Experimental Economics in Stockholm, for their valuable comments.

†Department of Economics and the Center for the Study of Rationality, the Hebrew University, motimich@gmail.com.

Keywords: In-Group Bias, Costly Signaling, Group Formation, Evolution of Cooperation, Prisoner's Dilemma Game.

JEL Classification: D82, Z13, D03, D7, C72.

1 Introduction

In the literature on human group size and on the development of sociality in Homo Sapiens, one prominent hypothesis suggests that the size of a "natural" human group is bounded by our cognitive skills - the need to memorize all the interactions and relationships between all members of the group consumes a lot of memory space and thus limits the group size. This hypothesis is based on research of various animals, which showed a positive correlation between animal group size and the relative size of the neocortex in the animal's brain, a correlation that led the researches to hypothesize that the bound is actually on the number of relationships that an individual animal can successfully monitor (Sawaguchi and Kudo 1990, Dunbar 1992).[1] These results were extended to anatomically modern humans, for whom a maximal group size of 148, commonly known as "Dunbar's number", was predicted (Dunbar 1993).[2]

Although there is some evidence in support of "Dunbar's number", larger groups are also known to exist, even in hunter-gatherer societies (Stewart 1955, Service 1962, and Birdsell 1970). Moreover, this theory can explain the limit on the cooperative group size, but cannot explain cooperation itself. We suggest here an alternative theory. We believe that at least when it comes to modern human beings, the bound on group size is not due to cognitive limitations, but rather due to the nature of human social conscientiousness. In particular, most human beings are endowed with a "psychological cost of cheating", i.e., they have disutility from cheating or betraying another person by not reciprocating the other person's kind actions.[3] But one should be careful not to automatically

[1] For example, an increase in group size from 40 to 50, entails an increase from 780 to 1225 in the number of pairwise interactions to memorize, suggesting that brain complexity crucially limits group size (Aiello & Dunbar 1992).

[2] According to this theory, when groups significantly exceed this size, they can no longer be egalitarian in their organization but must increasingly develop stratification involving specialized roles relating to social control (Naroll 1956, Forge 1972).

[3] Note that cheating here is not lying: one's actions are what counts, and not the consistency between one's statements and one's actions. Lopez-Perez (2012) demonstrates that

assume that this cost rises linearly with the number of betrayed individuals. In fact, though it is quite reasonable that this cost of cheating increases in the number of cheated individuals, the most salient difference is probably between cheating no one and cheating someone. Moreover, the marginal cost is bound to decrease in the number of cheated individuals. Thus, a plausible assumption would be that the "psychological cost of cheating" is concave.[4] Since the gain from cheating increases more or less linearly in the number of cheated individuals, one is inclined to be tempted to cheat if the number of cooperators exceeds a certain threshold.[5] Thus, belonging to a group of limited size ensures that the temptation is resistible, and that others can be trusted to cooperate because their temptation is resistible too.

Note that as opposed to the hypothesis about cognitive limitations as the source of restriction on group size, our hypothesis does not imply that the human brain imposes a hard-wired constraint on group size. Therefore, we do not predict a fixed limit on group size, but rather a flexible bound that is sensitive to the material returns to cheating. In particular, groups of larger size can be sustained if they find reliable ways to reduce the material gains from unilateral defection of a group member.

The assumption that the cost of cheating is increasing in the number of cheated individuals, yet it does so in a concave manner, generates two distinct refutable predictions. In natural situations, where the material benefit from unilateral defection is (more or less linearly) *increasing* in group size (i.e., more "suckers" to exploit if one defects from cooperation), we expect the tendency to cheat on the group to increase with group size. However, if for some reason

indeed lying aversion is not enough to induce cooperation in PD.

[4] One may think of this concavity as depicting a state where the more social connections one has, the weaker is one's empathy to one's weakest connection, and as a consequence the psychological cost of breaking the weakest connection decreases with the total number of connections. However, we will not assume the existence of groups, so there is no reason to presuppose that some people are inherently closer than others. Moreover, this depiction seems to suggest that groups are formed because people end up cheating only those who are detached enough from them, and that the cheated people will be considered the out-group members. However, we do not assume such discrimination exists, and therefore we show a different mechanism that leads to group formation, where in fact there is no cheating at all in equilibrium.

[5] Unless one assumes that the utility from monetary gains is even more concave than the cost of cheating.

the material benefit from unilateral defection is *constant* in group size, the individual should be less prone to cheat when the group is larger, because cheating more people would inflict a higher cost on him, with no increase in benefit. These two distinct predictions were neatly demonstrated in experiments of the public good game conducted by Isaac et al (1994) and surveyed in Ledyard (1995) and Holt & Laury (2008). In these experiments, subjects divide their allocation of tokens between a private account and a group account. In order to create an incentive to free-ride, the experimenters set the *marginal per capita return* from the group account (MPCR, defined as the ratio of benefits to costs for moving a single token from the individual to the group account) to be in the range of $(0, 1)$. Moreover, the design of the experiments was such that the monetary return to unilateral defection was inversely related to the MPCR. Isaac et al. showed that in treatments in which MPCR was independent of group size (i.e., the material benefit from unilateral defection was *constant* across group sizes), the rates of defection in groups of size 40 and 100 were lower than in groups of size 4 and 10, in line with our theory (and contrary to most economists' expectations to find more free riding in larger groups). On the other hand, when they compared the rates of defection in groups of different size in situations where the MPCR was decreasing in group size (i.e., monetary return to unilateral defection *increasing* in group size), they found higher defection rates in the larger groups, again, in line with our prediction.[6] Although this is not a validation of our hypothesis, these experiments demonstrate its potential to explain some prominent group behaviors.

The limit on group size has another important implication. As we show in the paper, cooperation *within* groups (of limited size) emerges side by side with defection *between* groups. That is, the cost of cheating leads at the same time to the formation of groups and to the development of *in-group bias* - an inclination to cooperate only with members of one's own group. Otherwise,

[6] The most striking evidence was probably the comparison of groups of sizes 4 and 10, where in both kinds of groups a token contributed to the group account was multiplied by the same multiplier, 3 (corresponding to MPCR's of 0.75 and 0.3 respectively). The experimenters found a significantly higher rate of defection in groups of size 10. This result was not replicated in a different experiment that compared groups of sizes 40 and 100, but this experiment with larger size groups was not conducted with monetary incentives, which may possibly affect the results.

a person would have "too many" cooperative partners, and the temptation to defect would destroy cooperation both within and between groups. This is true in particular to *social types*, i.e., people whose social conscientiousness makes them subject to the aforementioned psychological cost of cheating. In our model we do not presuppose any initial difference in their empathy or commitment towards different individuals, yet we show that in equilibrium they are all non-cooperative toward out-group members. This result is in line with the experimental findings of Tajfel (1970), Tajfel et al. (1971), and more recently Chen and Li (2009) and de Cremer et al (2008), who show that the effect of in-group bias can be easily triggered by even the most trivial and arbitrary group categorization.

We distinguish the social types from *asocial types,* i.e., people who are *not* subject to the cost of cheating. When these people are easily spotted, they cannot form any social connections at all. However, when it is hard to spot the asocial types, the *social types* cannot form cooperative groups without having to lose something. In particular, if one's type is one's private information, then only two distinct kinds of groups can emerge in society. The first kind, which we call a *mixed-type group*, contains individuals of both types, where a minority of asocial types free ride at the expense of the social types. In a sense, in any modern state where most people pay taxes but some do not, yet everyone enjoys the social benefits provided by the state, a similar situation prevails. We show that this kind of group can always be sustained in equilibrium, but the limit on the group size is stricter than before. So in the absence of enforceable contracts and central authority, higher proportion of selfish or asocial individuals will be correlated with smaller social structures (e.g., families instead of tribes).[7]

Groups of the second kind, which we call *cohesive groups*, consist only of social types, and their members fully cooperate with one another. Yet they need to screen out potential free riders. They do so by enforcing a practice of *costly signaling*, which means that members of the group obtain the trust and cooperation of the other group members only by exhibiting some self-sacrifice.

[7]Note that the explanation that goes in the other direction, saying that people are asocial *because* they live in small families and not in big tribes, takes the social structure as exogenous, while we believe it should be treated as endogenous.

We further show that the two kinds of groups can coexist in equilibrium. However, if such equilibrium is *envy-free*, in the sense that no one wishes to be in the shoes of someone else in society, then the possibility to signal strictly decreases the expected utility of *everyone* in society, regardless of their type and the group they belong to. Thus, beyond the private cost for the individual who signals, signaling as a phenomenon imposes a public cost on society. This public cost represents society's loss of "good guys", who form their own exclusive clubs instead of mixing with the other parts of society and lifting the average willing to cooperate.

One common practice of costly signaling is self-mutilation. Akerlof and Kranton (2000) list the various facets of this practice: "tattooing, body-piercing (ear, nose, navel, etc.), hair conking, self-starvation, steroid abuse, plastic surgery, and male and female circumcision". In this paper we choose to demonstrate the costly signaling aspect of three other phenomena, or practices, that are not often presented as such. The choice of the phenomena, which are discussed thoroughly towards the end of the paper, is motivated by the ability to demonstrate through them how the two kinds of groups, the mixed-type and the cohesive, can coexist, and how signaling inflicts a cost on society as a whole. The first practice is related to the term "acting white". This term is mostly used to describe the pressure that is imposed on Black people who invest in particular behaviors (especially acquiring higher education) by their social peer group (Fordham and Ogbu 1986; Austen-Smith and Fryer 2005). We suggest here to interpret the personal sacrifice of a Black individual who concedes to the pressure and *refrains* from these behaviors as a form of costly signaling. That is, by not acquiring higher education, the individual signals that he can be trusted not to forsake the Black brotherhood in pursuit of selfish goals at the expense of others. The second practice we discuss is religious rituals. We distinguish between mere believers and active participants in religious rituals (such as Sunday prayers), and show that religious rituals enable the practitioners to signal their social value to the community and to screen-out potential free-riders, but the exclusion of the practitioners from the whole society comes at a cost for everyone.[8] The third practice we discuss is social activism. Like in

[8] Levy and Razin (2012) develop a model where religious organizations play a significant

the case of religious rituals, we distinguish between the supporters of an agenda and those who actively pursue it, and show that at least to some extent, activism serves as a signaling device, which enables the activists to screen-out potential free-riders and to achieve internal cohesiveness.

The paper relates mostly to four literatures. The first is the literature on cheating and deception, the second is the literature on the link between cooperation and group size, the third is the literature on in-group bias, and the fourth is the literature on costly signaling.

Cheating, deception, lying and dishonesty, have all been recently in the spotlight of experimental study in behavioral economics (e.g., Gino, Norton & Ariely 2010, Hurkens & Kartik 2009, Gneezy et al. 2013, and Lundquist et al. 2009). The assumption of the current paper that the psychological cost of cheating is concave in nature is related to a concept called the "what the hell effect", which is generally used to describe behaviors that, once triggered, burst into full-fledge expression instead of developing gradually. Gino et al (2010) documented the "what the hell effect" of cheating in the dimension of time. They found that a person may be unwilling to cheat for a long period of time, but once he cheats for the first time, he often succumbs to full-blown cheating afterwards. Another dimension of the "what the hell effect" of cheating, the dimension of the size of lie, was reported by Gneezy et al (2013), who showed that when monetary payoffs were positively correlated to the size of lying, most subjects who decided to cheat a fellow participant chose the maximum size of lie. Moreover, Hurkens & Kartik (2009) found that their subjects could be divided into two distinctive types - those who lie whenever they can monetarily gain from lying (our asocial types), and those who never lie (our social types, assuming that the monetary temptation used in the experiment was not big enough).

---

role in enhancing cooperation through establishing belief in reward and punishment and through the possibility to signal membership in these organizations. However, a belief in punishment for bad deeds is conceptually different than our "cost of cheating", as the punishment is conditioned only upon one's own actions, while ignoring the potential effect of the expectation about the opponent's actions. Moreover, in their model, the existence of religion can be beneficial to everyone in society, and is never bad for the secular types, whereas in our model abolishing religion would have a Pareto-improving effect. In Section 5.2 we compare the assumptions of their model and ours, and the implications thereof.

The second related literature is the one about the link between cooperation and limited group size. The problem of sustaining cooperation in sizable groups was raised already by Olsen (1965). Bonacich et al. (1976), Bendor & Mookherjee (1987), Boyd & Richerson (1988), and Suzuki & Akiyama (2005), have all used the N-person Prisoner's Dilemma game in order to analyze this problem under various assumptions. However, these works do not try to explain the tendency to cooperate only with in-group members. Choi & Bowles (2007) do provide an evolutionary model that explains at the same time altruism within the group and parochialism between groups, but do not account for group size. Their work can be seen as a link to the third related literature, which is the one that documents in-group bias.

We already mentioned some lab experiments that demonstrated the minimal group effect, i.e., that in-group bias can be triggered by arbitrary group categorization. Goette et al. (2006) showed a similar effect in a field experiment, where the arbitrary group categorization was the division of soldiers into platoons in the Swiss army. When it comes to naturally formed groups, such as ethnic or racial groups, Bernhard et al (2006) showed in-group bias among ethnic groups in Papua New Guinea, and Fong and Luttmer (2009) showed racial in-group bias among contributors to Hurricane Katrina victims.[9] All these works, whether in the lab or in the field, whether with randomly assigned groups or with natural ones, involved subjects playing canonical experimental games, such as the dictator game and the Prisoner's Dilemma. But recently, in-group bias was verified also using naturally occurring data. Shayo and Zussman (2011) were able to expose in-group bias in real-life decisions by professionals, where the decisions had significant implications to the parties involved. They analyzed judicial decisions in Israeli courts, where strong nondiscriminatory norm applies, and demonstrated empirically the existence of in-group bias in

[9] It is interesting to note that experiments that use the Trust Game instead of allocation games tend to show much more variation in behavior towards out-group members. Hennig-Schmidt et al. (2009) find no in-group bias when letting Germans, Israelis and Palestinians play the Trust Game with in-group and with out-group members. Similarly, Bornhorst et al. (2010) find no regional discrimination in an experiment involving students of different European nationalities who are matched to play this game in mix-nationality groups. Even more strikingly, Fershtman and Gneezy (2001) reveal out-group favoritism among Israeli Jews of eastern decent, who show more trust towards Israeli Jews of western decent.

the decisions of judges.

The fourth related literature is the one on costly signaling. The canonical works in this literature are Spence's (1974) model of education as a signal in the labor market, and the models of reputation signaling in firm competition by Kreps & Wilson (1982) and Milgrom & Roberts (1982). When it comes to signaling as a means to acquire cooperation and social connections, Gintis et al (2001) develop an evolutionary model of costly signaling as a promoter of cooperation in the group level, and Camerer (1988) analyzes gift exchange as signaling intentions for future investments in pairwise relations. Even more closely related to our paper are the work of Akerlof and Kranton (2000), who discuss costly signaling of one's identity, lannaccone's (1992) work on social clubs, where signaling is used by individuals as a means to be accepted to desired groups, and the work of Levy and Razin (2012), where participation in religious rituals signals a greater inclination to cooperate. Finally, Benabou and Jean Tirole (2006) suggest a model where pro-social behavior (charity in their case) is used as a means for signaling quality, and not as an indicator of it's independent existence.

The structure of the paper is derived mostly from the stylized facts that we wish to explain. Our benchmark model with complete information (Section 2) captures the tendency of people to form groups that exhibit in-group bias, and the tendency of social individuals to be "kind" (cooperative) only to in-group members.[10] Our model with incomplete information (Section 3) captures the connection between the cohesiveness of a group and the use of costly signaling, and analyzes the prospects for having a society in which groups with different levels of cooperation coexist. In Section 4 we discuss the conditions under which such coexistence is envy-free, and show that under them costly signaling has a negative effect on the welfare of all individuals. This result is further shown to be restricted to cases where the proportion of asocial types is not large enough to make envy-free coexistence impossible, in which case signaling has a positive effect on the welfare of social types. Section 5 demonstrates the main assertions by analyzing three examples as special cases of costly signaling.

[10]For experimental findings that support this assertion about social individuals, see de Dreu (2010) and the discussion of these findings in Section 2.

Section 6 concludes.

2 The formation of cooperative groups and in-group bias

We model society as containing $N+1$ individuals[11] who simultaneously interact with each other to play one-shot Prisoner's Dilemma (PD) games. We follow the notations of Kandori (1992) and Ellison (1994) and use the following payoff matrix for the game:

	C	D
C	$1,1$	$-l, 1+g$
D	$1+g, -l$	$0,0$

The zero payoff for mutual defection suggests that there is no difference between mutual defection and no interaction at all, thus relaxing the somewhat unrealistic assumption that each individual is practically engaged in simultaneous plays against all members of society (an assumption that aims to keep the model as parsimonious as possible). Furthermore, it implies that the payoff for mutual cooperation is strictly positive, hence the total return to cooperation increases in group size (nevertheless, groups will be of limited size in equilibrium). g stands for the *gain* from unilateral defection, and l for the *loss* from being the victim of the opponent's unilateral defection. We assume strategic complementarity, i.e., $l \geq g$, which implies that if one's opponent is more prone to defect, one is more prone to defect too. Our analysis considers only pure strategies at the pairwise level, but individuals can discriminate between opponents, i.e., cooperate with some while defecting against others.[12]

Society is composed of two types of individuals, $\tau \in \{s, as\}$, where s stands for *social type* and as stands for *asocial type*. Asocial types are affected only by the material payoffs of the game, and so for them defection is a dominant strategy against any opponent. Unlike them, social types may lose utility by

[11] For most applications it is helpful to assume that N is very big.

[12] Mixed strategies pose here a modeling ambiguity. Since part of the payoff is going to be related to disutility from defecting against a cooperative opponent, it is not clear how one should feel when defecting against an opponent who uses a mixed strategy - is it the realization that counts, or maybe the (impure) intention to cooperate? We prefer to leave these potential controversies aside.

cheating, where cheating means playing D against an opponent who plays C.[13]

Let $t(k)$ denote the cost of cheating against k individuals. This can be thought of as a psychological cost caused by the arousal of uncomfortable feelings such as shame or guilt on the side of the defector.[14] We naturally assume that $t(0) = 0$, and that $t(k)$ is weakly increasing in k - the more people are cheated by the individual, the (weakly) more it costs him. Additionally, we put some restriction on the form of this increase. In particular, we assume that the "what the hell effect" of cheating, as discussed in the introduction, applies. With regard to modeling, this effect can be modeled as a cost function $t(k)$ that is concave in k. We do not require smooth concavity or even continuity, so that any cost function with a discrete jump at 0 and a weakly concave continuation afterwards satisfies our concavity condition, and in particular this includes one with a fixed cost of cheating for any $k > 0$.[15] The other requirements are a "flat enough" slope as k goes to infinity, and a "steep enough" slope at 0 (if $t(k)$ is continuous at 0). Formally, the assumptions on $t(k)$ beyond positive monotonicity and concavity are:

$$t(0) = 0, \quad \lim_{k\to\infty} t'(k) < g,$$
$$\text{and } \lim_{k\to 0} t'(k) > g \text{ (or } \lim_{k\to 0} t(k) > 0 \text{ if } \lim_{k\to 0} t'(k) \text{ is not defined).}$$

In the benchmark model with complete information that we analyze in this section, we assume that the type of each individual is common knowledge. The strategy of player i is the N-tuple whose j's element is the action played in the PD encounter with player j. We denote this element by s_{ij}. We say that society is in (Nash) equilibrium if, given the strategies of all other individuals, no individual has a profitable deviation from his strategy. We further say that

[13] Note the difference between defecting, i.e., playing D, and cheating, i.e., playing D against an opponent who plays C. Miettinen and Suetens (2008) indeed show that people feel guilty when defecting in the PD game only if the partner has not defected as well.

[14] This interpretation is in line with that of Lopez-Perez (2008), with the exception that he would treat the k cooperators as those who respect the norm, and the defector as the norm breaker. In Lopez-Perez (2008) $t(k)$ is linear in k and the groups are of fixed size.

[15] In terms of social identity theory, a discrete jump captures the change in one's perceived self image from a self image of someone who never cheats, to a self image of someone who potentially cheats (the border between these two distinct characters is nicely captured in the recent experiments on lying aversion of Hurkens & Kartik 2009 and Gneezy et al. 2013).

a *cooperative group* exists if all members of the group cooperate with each other, and that the group members show *in-group bias* if they defect against all out-group members. The following result shows why cooperative groups can be formed and sustained in equilibrium as long as they are limited in size, and why in-group bias is bound to emerge too.

Proposition 1 *Let $\bar{K} > 0$ be the unique strictly positive solution to the equation $t(K) = Kg$. Then in equilibrium:*

1. Every asocial type plays D against everyone else, and everyone else plays D against him.

2. Every social type plays C against mostly $\bar{K}$ individuals, who play C against him too, and plays D against everyone else.

Proof. Since, for both types, defection is a best response against an opponent playing D himself, we get that in equilibrium, if $s_{ij} = D$ then $s_{ji} = D$. Hence, since D is a dominant strategy for asocial types, and types are common knowledge, we get (1). As for social types, if K players play C against a social type, and $K \leq \bar{K}$, his best response to all of them is C, since deviating to defection against any subset of them (of size $k \leq K$) would impose on him a net cost of $t(k) - kg > 0$. Otherwise, if $K > \bar{K}$, then playing C against all of them cannot be his best response, because deviating to playing D against all of them would increase his total payoff by $Kg - t(K) > 0$. ■

Corollary 2 *If $\bar{K} \geq 1$, then any division of the social types into cooperative groups of size $\bar{K} + 1$ at most, whose members show in-group bias, can be sustained in equilibrium.*

This result implies that it is easier to sustain cooperation in smaller groups. It sounds plausible when considering the limited size of tribes and clans, especially in societies with no central authority, where groups are presumed to form spontaneously. The driving force behind this result is the "what the hell effect" of cheating - as the size of the group increases, it becomes harder to avoid the temptation to defect and achieve the ever growing material benefits

of unilateral defection. At some point this effect is going to burst out, leading to cheating across the board. The limit on group size in equilibrium is the threshold above which such across the board defection is bound to occur.

Another aspect of the result is its built-in in-group bias. It turns out that social types would show the same level of asociality towards out-group members as would asocial types, while exhibiting sociality only towards in-group members. This result is in line with experimental studies of the Prisoner's Dilemma. For example, Wilson & Kayatani (1968) and Dion (1973) find that the competitiveness which characterizes inter-group behavior resembles that of individual players, whereas it is the increased proportion of cooperative choices exhibited in intra-group decisions that deviates from typical inter-personal play (see further analysis in Brewer 1979). More recently, de Dreu (2010) uses the Intergroup Prisoner's Dilemma to show that compared to individuals with a "chronic pro-self orientation", those with a "chronic prosocial orientation" (these would be the social types in the jargon of the current paper) display stronger ingroup trust and ingroup love — they self-sacrifice to benefit their ingroup — but not more or less outgroup distrust and outgroup hate. As we show in the next section, the self-sacrifice practiced by social types is not always intended to benefit the ingroup, but can rather be a means of signaling membership in the group.

3 Cohesive groups and membership costs

3.1 The effect of incomplete information

The basic model with complete information implicitly assumed that a social type can consider the cooperation of other group members as guaranteed. This assumption is a bit unrealistic when considering pairwise PD game. Moreover, the assumption that asocial types can be easily distinguished from social types is quite strong. We therefore turn now to consider the case where the individual's type is his private information. We assume that each individual is randomly assigned a type $\tau \in \{s, as\}$, with probability p to be assigned $\tau = as$, and that this is common knowledge. Can there still be an equilibrium with some cooperation in it? The following proposition, preceded by a

definition, shows that the answer is affirmative.

Definition 3 *A mixed-type group is a collection of individuals of both types, such that all social types in the group play C against all other group members, while all asocial types in the group play D against all other group members.*

Proposition 4 *If $p \leq \frac{t(1)-g}{t(1)+l-g}$, then there exists a unique integer $K_p \in [1, \bar{K}]$ such that a mixed-type group of size $K+1$ is sustainable in equilibrium if and only if $K \leq K_p$. Furthermore, K_p is decreasing in p.*

The proof of the proposition follows the next lemma.

Lemma 5 *Let $h(x)$ be an increasing and concave function defined for $x \geq 0$ with $h(0) = 0$. If $x \sim Bin(n, p)$, then:*

1. *Given a fixed $p \in [0, 1]$, $E_n h(x)$ is increasing and concave in n.*
2. *Given a fixed $n > 0$, $E_n h(x)$ is increasing in p.*

Proof. *(1) That $E_n h(x)$ is increasing in n is clear from the fact that*

$$E_{n+1}h(x) = pE_n h(x+1) + (1-p)E_n h(x),$$

and $h(x+1) \geq h(x)$. For proving concavity, we can write

$$E_{n+2}h(x) = (1-p)^2 E_n h(x) + 2p(1-p)E_n h(x+1) + p^2 E_n h(x+2).$$

Then we need to show that $E_{n+2}h(x) + E_n h(x) \leq 2E_{n+1}h(x)$. Substituting the above expressions in this inequality, it boils down to showing that $p^2 E_n h(x) + p^2 E_n h(x+2) \leq 2p^2 E_n h(x+1)$, which indeed holds by the concavity of $h(x)$ and the linearity of the expectation operator. (2) We will prove by induction. For $n = 1$ the inequality holds: if $x \sim Bin(n, p)$ and $y \sim Bin(n, q)$ with $q > p$, then $E_1 h(y) = qh(1) \geq ph(1) = E_1 h(x)$. Assume now that the inequality holds also for some n, so that $E_n h(y) \geq E_n h(x)$. Then

$$\begin{aligned} E_{n+1}h(y) &= qE_n h(y+1) + (1-q)E_n h(y) \\ &\geq pE_n h(x+1) + (1-p)E_n h(x) = E_{n+1}h(x), \end{aligned}$$

which completes the proof by induction. ■

Proof of Proposition 4

Consider an individual of type s who plays C against exactly K other individuals. Defecting against any (randomly chosen) $k \leq K$ of them, of which $X \in [0,k]$ are of type s, would result in an increase in expected material payoff of $Xg + (k-X)\,l$, but the expected total payoff would also decrease by $t(X)$ due to the cost of cheating. Since $X \sim Bin(k, 1-p)$, the individual would have no profitable deviation if $E_k\left[t(X)\right] \geq E\left[Xg + (k-X)\,l\right] = k[(1-p)g + pl]$ for every $k \leq K$. Let $\Delta(k) \equiv E_k\left[t(X)\right] - k[(1-p)g + pl]$. The conditions on $t(k)$ and on the payoffs of the game imply that $\Delta(0) = 0$ and $\lim_{k\to\infty} \Delta(k) < 0$. From Lemma 5 part (1) we know that $E_k\left[t(X)\right]$ is concave in k, and therefore so is $\Delta(k)$. It can be verified that if If $p \leq \frac{t(1)-g}{t(1)+l-g}$ then $\Delta(1) \geq 0$, in which case $K_p \geq 1$ is the floor of K_p^*, the unique strictly positive solution to the equation $\Delta(k) = 0$. Moreover, $K_p \leq \bar{K}$ because $l \geq g$ and $E_k\left[t(X)\right] \leq t(k)$, and so $\Delta\left(\bar{K}\right) \leq t(\bar{K}) - \bar{K}g = 0$. Next, from Lemma 5 part (2), we get that $E_k\left[t(X)\right]$ is decreasing in p for a fixed k, and so $\Delta(k)$ is also decreasing in p for any $k > 0$ (remembering that $l \geq g$ and so $[(1-p)g + pl]$ is increasing in p). That is, since $\Delta\left(K_p^*\right) = 0$, we have $E_{k,q}\left[t(X)\right] - k[(1-q)g + ql] \leq 0$ for any $q > p$, which in turn implies that $K_q^* \leq K_p^*$ and so $K_q \leq K_p$.■

Corollary 6 *Any division of society into mixed-type groups whose sizes are bounded by $K_p + 1$ forms a Bayesian equilibrium.*

In this equilibrium, social types would show in-group bias, by playing C against all group members and D against all outsiders, while asocial types would play D against everyone, thus "free-riding" on the social types in their group. Such groups are bound to be smaller than the groups of purely social types in the complete information case (i.e., $K_p \leq \bar{K}$), because here the temptation to defect is larger (avoiding the sucker payoff l is assumed to increase expected payoff at least as much as gaining g by defecting against a cooperative opponent) and the cost of cheating is lower (because defecting against an asocial type who defects himself is not psychologically costly). It can be shown that naturally the maximal group size is decreasing in p, i.e., the greater is the

proportion of asocial types in society, the more it is tempting for social types to defect, thus the smaller are the groups that can sustain cooperation.[16] The behavior of social types in this equilibrium bears some similarities to the behavior of "conditional altruists" in Palfrey & Rosental (1988). However, they assume that the payoffs of a contributor (= cooperator) are unaffected by the opponent's strategy, and so even "conditional altruists", who condition their strategy on their expectations from the opponent, contribute only because they fear from a mutual defection and not because they feel obliged to contribute when others do so. The main differences between our results and theirs are that in ours group size plays a significant role in determining the players' strategies, and the threshold for cooperation of social types is affected by believes that are derived from the actual proportion of social types in society.[17]

An interesting scenario is revealed when considering the case of $p(1+l) > 1$. In this case, the proportion of asocial types in society is high enough to make the expected payoff of a social type in a cooperative group of $K+1$ members *negative*, regardless of the exact group size ($K[1-p(1+l)] < 0$, $\forall K$). This means that the social types would have been better off in a society with full-blown defection (where the payoff of everyone is zero), yet, if $K \leq K_p$, they end up playing C against their group members for a negative expected payoff. One can think of this situation as resembling the frustrating state of someone who pays taxes in order not to free ride other people like him, in a country that is so corrupt that he would be better off with no tax system and no public service at all. This state of affairs raises the question of the possible introduction of costly signaling - can such signaling be efficiently used by social types to distinguish themselves from asocial types? We consider this option in the next subsection.[18]

[16] However, groups that are small enough can still sustain cooperation, because in such groups the material payoffs are low so there is not much to gain by defection, yet the psychological cost of cheating kicks in already with the first potential cheating.

[17] Strictly speaking, we can also get Bayesian equilibria in which not all social types in mixed-type groups cooperate, accompanied by an appropriate system of beliefs, but in such equilibria all mixed-type groups must be of the same size, which is the unique size that would make social types indifferent between cooperation and defection. We believe that such restrictions make these equilibria less interesting.

[18] In a similar model, Camerer (1988) models gift exchange as a system of costly signalling intentions for a long term investment in relationship. For some values of the parameters in

3.2 Costly signaling

Costly signaling is a well-known means to achieve a separating equilibrium (e.g., Spence 1974). In our model, it is signaling "sociality" that has the potential of achieving a separating equilibrium. This follows from the fact that anyone, regardless of one's type, would like to be regarded as a social type and be trusted by his opponent, because an opponent playing C gives one a potential for a higher expected payoff.[19] So a necessary condition for a separating equilibrium is that the cost of signaling sociality will be lower to social types, compared to asocial types. However, this condition is not sufficient, since the gain of an asocial type from being considered social exceeds that of a truly social type, so an asocial type will be willing to pay a higher cost in order to be perceived as social.

Let x_s and x_{as} be the cost of signaling sociality for types s and as respectively. The signal is not directed at any specific opponent, but is rather a singular payoff-irrelevant sacrifice that is observable by everyone else, just like

his model, he gets that the "Willing" types, who prefer investing if and only if their partner invests too (similar to our social types), would choose to invest when playing against an unknown type when there is incomplete information. This is equivalent to the case of social types playing C against an unknown group member in our model. However, Camerer jumps then to the conclusion that in this case there is no potential for signaling, because the purpose of signaling is to elicit investment by a "Willing" opponent, who anyway invests. We claim that not only is it possible to prove that separating equilibria with costly signaling *can* exist in such a case, but also that the emergence of signaling is plausible and almost even self evident in some cases, e.g., when $p(1+l) > 1$ in our model.

[19] In the case of $p(1+l) > 1$ discussed above, social types get a negative payoff as members of cooperative groups, thus do worse than they could do by being perceived as asocial types. In this case, it seems that they have incentive to signal *asociality*. If believed, their opponents would defect, and they would then be able to defect too without any pangs of conscience, while improving their expected payoff. But should they be believed indeed? Since truly asocial types would not like to be revealed as such (they have a strictly positive expected payoff as members of a mixed-type group), they themselves would not want to signal asociality. Clearly, in such a case, signaling asociality would in fact reveal one as *social*. So it seems at first glance that we can get an equilibrium where only social types signal asociality, and paradoxically by that distinct themselves from the crowd, which in turn would enable them to form cooperative groups. However, as opposed to signaling sociality, signaling asociality should not be more expensive to asocial types than to social types (if anything, it would be cheaper to them). And since asocial types can only profit from being perceived as social types, they are bound to imitate the social types in signaling asociality, thus bringing the situation to its original state of affairs. Foreseeing this, the social types would not signal asociality in the first place.

in the examples described in the introduction (self-mutilation, avoiding education, etc.).[20] We assume that $x_s < x_{as}$, i.e., it costs more to fake sociality than to signal it when it indeed exists (see e.g. Frank 1987). A person would signal sociality if this signal made others treat him as social, and if his increase in total expected payoff from being treated as social exceeded the cost of signaling. Being treated as social means getting the cooperation of potential group members. Recall now that the return to both cooperation and defection is increasing in the number of cooperative partners. So what is the lower bound on the number of cooperative individuals that makes signaling sociality profitable to each type?

Assume that a separating equilibrium exists, i.e., types are believed to be social if and only if they signal, and these beliefs are consistent with reality. Then asocial types would not be trusted by anyone and thus have a zero payoff, while social types would be able to form cooperative groups of size $K+1$ and get a payoff of $K - x_s$. Of course, the upper limit on K from the benchmark model with complete information still applies here, i.e., $K \leq \bar{K}$. Furthermore, to be indeed consistent with reality, no type should have a profitable deviation from this equilibrium. If a social type deviates to not signaling, he can expect to be treated as asocial and thus meet only defecting opponents and get a zero payoff, but to save the cost x_s. This deviation would not be profitable if $K > x_s$. So $K_s \equiv x_s$ is a lower bound on the number of cooperative individuals that makes signaling sociality profitable to social types. If an asocial type deviates to signaling, he can make $1+g$ against every social type that plays C against him, but signaling would cost him x_{as}. So this deviation would be profitable if he can expect to meet at least $K_{as} \equiv \frac{x_{as}}{1+g}$ such cooperative opponents. Therefore, a social type can be sure that his K signalling group mates are truly social only if $K < K_{as}$. This means that $K_s < K_{as}$ is a necessary condition for a separating equilibrium, or, written as a condition on the ratio

[20] It is not unreasonable to have signaling in the level of the group or the society as a means to promote pairwise relationships. As Gintis et al (2001) write in their paper titled 'costly signaling and cooperation', "it is often the case that biological signals in other domains such as mate choice, resource competition, and even predator-prey interactions are not private to an intended receiver, but are emitted without the signaler knowing exactly with which among a population of possible observers it might influence". Evidence from Meriam turtle hunters is consistent with this claim (Smith et al 2003).

of the costs of signaling, $\frac{x_{as}}{x_s} > 1 + g$. That is, to get separation it is not sufficient that signaling sociality would be cheaper to social types, but the ratio of costs should also exceed the ratio of marginal gains from a cooperative opponent. If this condition of separability holds, and if the condition of individual rationality, $x_s \leq \bar{K}$, holds too, then in a separating equilibrium we can get cooperative groups of purely social types, where the size of each such group will be $\left\{K+1|\ K_s < K < \min\left\{K_{as}, \bar{K}\right\}\right\}$. Otherwise, if $\hat{K} \equiv \min\left\{K_{as}, \bar{K}\right\} < K_s$, then social types cannot distinct themselves from the asocial types by signaling, either because they would have incentive to cheat after signaling and getting the cooperation of other social types (if $\bar{K} < K_s$), and thus will not be trusted to cooperate in the first place, or because they can be imitated by asocial types (if $K_{as} < K_s$).

It is important to note that $K_s < \hat{K}$ is only a necessary condition for a separating equilibrium, i.e., it does not *guarantee* that a separating equilibrium will indeed emerge. There is always an equilibrium where everyone plays D, and there are always pooling equilibria in which society is divided into mixed-type groups of sizes $K \leq K_p + 1$ and no one signals. In these cases, a social type cannot hope to gain from a unilateral deviation to signaling his type, even if the signal is known to be truthful.

The fact that the condition $K_s < \hat{K}$ does not guarantee separation supports a stylized fact we want to explain here - the coexistence of cooperative groups with costly signaling (which will be referred to as *cohesive groups*) side by side with groups that are less cooperative, and whose members do not engage in costly signaling. This situation can emerge if a fraction $1-\lambda$ of the social types form cohesive groups of purely social types (of sizes at the range $[K_s+1, \hat{K}+1]$), while the other social types are members of mixed-type groups in which asocial types "free-ride" on the social ones. The splitting of the social types into two kinds of groups increases the proportion of asocial types in the mixed-type groups, thus further constrains the size of this kind of groups. That is, let $q \equiv \frac{p}{p+\lambda(1-p)}$ be the proportion of asocial types in the mixed-type groups. Following the same analysis as before, K_q ($\leq K_p$) will be the new upper bound on the size of mixed-type groups. We call such an equilibrium with both kinds of groups a *hybrid equilibrium*. Figure **??** demonstrates the effect of the cost of

signaling for each type on the potential for separating and hybrid equilibria.

The cooperative groups of purely social types who signal sociality are the *cohesive groups* the section title refers to. We use this term to denote a group of people who cooperate with each other, defect against all others, and pay the cost of signaling, a "membership cost", in order to do so. These cohesive groups can be thought of as social clubs, cults or communes, as in the work of Iannaccone (1992). It is interesting to note that Proposition 2 in Iannaccone's paper says that if society consists of two types of people, type 1 and type 2, such that type 1 people participate in group activities and value group quality less than type 2 people, "then, *as long as people of type 1 constitute a sufficiently large fraction of the population*, there will exist a signaling equilibrium in which type 2 people end up in groups that require their members to sacrifice a valued resource or opportunity and type 1 people end up in groups that require no such sacrifice". The equivalent to type 1 and type 2 people in our model are asocial types and social types respectively. And although in our model the proportion of asocial types is not a binding condition on the existence of a separating equilibrium, a "sufficiently large" proportion of them makes separation more plausible than pooling or hybrid equilibria. This is so, because if p is large enough, hybrid equilibrium are not *envy-free*, as will be explained in detail in the next section.

4 "Bad" signaling

The multiplicity of equilibria invites a comparison of them in terms of stability and welfare. In particular, we will consider *envy-free* equilibria, and show that such equilibria are inefficient if they contain groups of both kinds coexisting. The idea of an envy-free equilibrium is related to stability: if an individual i envies an otherwise identical individual j who is a member of a different group where the expected payoff is larger, than this inequitable allocation may be considered unstable, as individual i will try to replace individual j once it is possible. This is one reason to focus on this sort of equilibria. Another reason is related to a natural way to think of the setup developed above. So before formally presenting the definition of *envy-free* equilibria and its welfare implication, let us loosely describe the setting we refer to.

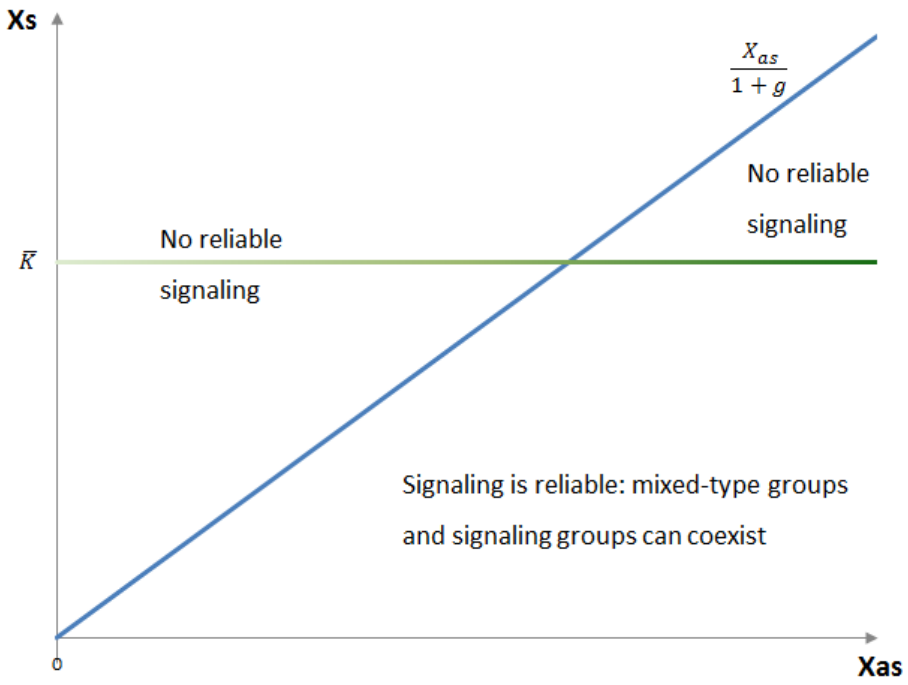

Figure 1: Displaying the necessary conditions for separating and hybrid equilibria. The blue diagonal line is where the ratio of the costs of signaling for each type is such that $\frac{x_{as}}{x_s} = 1 + g$. It marks the border between the region where social types can distinguish themselves from the asocial types by signaling (below it to the right) and the region where they cannot (above it to the left). Moreover, if x_s, the cost of signaling for the social types, is above the green line, then cooperative groups with signaling of purely social types cannot be sustained in equilibrium, because if a social type is able to achieve so many cooperative partners to make signaling worthwhile, he'd better cheat them and play D.

Suppose that there are many institutions called "clubs", some of whom require signaling in order to join them, and some of whom do not. Suppose also that each individual can choose which club to join, and that individuals can move freely between clubs. What can we expect to happen? A plausible implication is that the expected payoff of social types will tend in the long run to be the same in all clubs, and in particular, in both kinds of clubs – with signaling and without signaling. Otherwise there would remain an incentive to move to a club where expected payoff is larger once an opportunity shows up.[21] This makes the case of envy-free equilibria more plausible. We next introduce a formal definition of envy-free equilibria, and a proposition about the social undesirability of signaling.

Definition 7 *An equilibrium is envy-free if no individual wishes to switch groups with a different individual in society.*

Proposition 8 *Every envy-free hybrid equilibrium is strictly Pareto dominated by its corresponding envy-free pooling equilibrium (i.e., the one with the same proportion of asocial types in society).*

Proof. First note that in an envy-free equilibrium, the expected payoff of social types in all groups must be equal. Let this expected payoff be U. Moreover, all groups of the same kind must be of equal size, as expected payoffs depend on group size. Let there be an envy-free hybrid equilibrium in which the size of mixed-type groups is K. Then we have

$$U = K[(1-q) * 1 + q(-l)] = K[1 - q(1+l)].$$

Now compare U to the expected payoff of social types in a corresponding pooling equilibrium with no signaling, where mixed-type groups are of the same size K (these groups are sustainable in equilibrium because the proportion of asocial types in the pooling equilibrium is $p < q$, and we know from Section 3.2 that $K_q \leq K_p$). The expected payoff of social types in the pooling equilibrium will

[21] Another possible implication is that each club will reach its maximal capacity (corresponding to the bound on group size for that kind of club), but this effect is marginal in importance for our analysis.

be $K[1-p(1+l)] > K[1-q(1+l)] = U$, i.e., the pooling equilibrium is strictly better for them. As for the asocial types, in both envy-free equilibria they are all members of the mixed-type groups of size K (otherwise a "group-less" asocial type will envy an asocial type who is a member of a mixed-type group). It follows that their expected payoff in the hybrid equilibrium, $K\left(1-q\right)\left(1+g\right)$, is strictly smaller than $K\left(1-p\right)\left(1+g\right)$, their expected payoff in the pooling one. ■

Note that each envy-free hybrid equilibrium can also be compared to a corresponding envy-free separating equilibrium, where only cohesive groups exist (and they are of the same size as those in the hybrid equilibrium). Since social types get the same expected payoff in both kinds of groups in the envy-free hybrid equilibrium, and this payoff stays the same in cohesive groups under the separating equilibrium, while asocial types are clearly better-off in the hybrid one, where they can free-ride social types, this separating equilibrium is Pareto dominated by the hybrid one, which was shown to be Pareto dominated by the pooling one. So the proposition essentially tells us that except for the case where the proportion of asocial types is high to begin with (such that $p(1+l) > 1$, and so social types cannot have the same expected payoff in both kinds of groups; but see also below), social types would probably, and quite paradoxically, be better-off if there were no cohesive groups at all. In other words, the signal is not only costly but wasteful. It is not new that social customs that may be interpreted as signaling are self-harming – the annual *Ashura* ritual of the *Shia* Muslims, where the participants beat themselves with iron chains and swords until blood sheds is a very visual manifestation of the idea.[22] However, the context here is novel, as the existence of mixed-type groups with social types in them alongside the cohesive groups suggests, on the face of it, that signaling is sustainable because it is in the self interest of the signalers. We devote the next section to discuss some real-life scenarios that seem to be in line with this result.

[22] Boyer (2001) provides more examples of seemingly wasteful religious practices.

5 Examples of wasteful signaling

5.1 On Acting White

One salient case of costly (and probably wasteful) signaling in cohesive communities is the one related to the "acting White" accusation in the Black American society. When thinking about "acting White", many tend to focus on those who do try to acquire education, and the social cost they have to bear by doing so, but we believe that the focus should be instead on those who *do not* try to acquire education. That is, the cost is in fact for "remaining Black", not for "acting White".

In order to see why "acting White" can be explained with our model, let action D in the PD game be interpreted as pursuing individual goals, and let action C be interpreted as contributing to the Black community one comes from. Using the PD game to model this situation implies that from a selfish perspective, pursuing individual goals is always better, but everyone in the Black community would be better off if all contributed to it than if all pursued individual goals.[23] The social types are the Black individuals who are willing to sacrifice some self profits for the benefit of their community if others do it too (unless the number of contributors is big enough to make them free-ride). In the case of incomplete information, people in the community cannot know who will eventually comeback to the community to contribute and who will shirk from contribution. Then, the costly signal is naturally the self-sacrifice of a Black person who refrains from the pursuit of individual goals such as education or career opportunities in order to avoid being perceived as 'acting like White people do'.

Consider now the case of a hybrid equilibrium, where the costly signal is the personal cost of giving up education and staying in the Black neighborhood. In such equilibrium, some people will give up education and form cohesive groups in their communities, and some will acquire education. Those giving up education will enjoy the cooperation and support of their group mates, at the cost of staying uneducated. Those acquiring education will consist of social

[23] This should not necessarily apply to the White community too for various reasons, such as differences in socioeconomic status or in community structure.

types who go back to the community to contribute, and asocial types who leave their communities in pursuit of their individual goals. Note that in our model, these are only the social types in the mixed-type groups, i.e., those who acquire education and comeback to contribute, that suffer from the defection of the asocial types (the members of the cohesive groups are only affected indirectly through the need to costly signal in order to distinguish themselves). A plausible explanation for that would be that the departure of the asocial educated Blacks imposes a higher burden on the educated Blacks who return to the community (because they share this burden with less people), while from the point of view of those who stay in the community, the total contribution acquired is the same.

The payoff structure captures correctly the fact that the asocial types are clearly better-off by acquiring education, and are much better-off if others (the educated social types) pay back to the community on their behalf too. As for the social types, the lesson from the previous section is that unless the proportion of asocial types is so large that even if all social types acquire education still the burden of coming back to serve the community afterwards is high ($p > p_c$), social types would have been better-off if all of them acquired education and absorbed the absence of the asocial types together as a group. By splitting into cohesive groups of non-educated people on the one hand, and a fraction who become educated on the other hand, the social types are all worse-off: the non-educated could have had higher utility by acquiring education, and the educated could have gained from sharing their burden with all the other social types. In this sense, "acting white" is a shameful waste of human capital. When it comes to high education it is not reasonable to apply policies that eliminate signaling by making this education mandatory. However, if the gains from education will continue to increase, the cost of signaling is bound to increase too, and the model predicts that eventually signaling would stop being individually rational, and consequentially would cease to exist.

5.2 Religious practices[24]

That stronger social ties (i.e., higher levels of cooperation) and religious practice are positively correlated is not new and was empirically demonstrated by Ellison and George (1994). In particular, Levy and Razin (2012), henceforth LR, develop a model where this correlation is facilitated by the belief of religious people in reward and punishment (fostering cooperation), and by the positive correlation between belief and religious practice. Moreover, religious practice in their model serves as a (costly) signaling device. We believe that by incorporating these ideas of LR about the nature of belief and the nature of religious practice into our model, we can account for the main results we wish to highlight here, i.e., the coexistence of the two kinds of groups and the wasteful aspect of signaling.

The main departure between our model and LR is that in LR belief is conditioned on paying the costly signal of joining a religious service, while in our story one can be a believer and not go regularly to church,[25] with the cost of participating in the religious rituals being smaller for a believer. By conditioning the belief on religious practice, LR cannot get a Pareto improvement through abolishing religion, because without an institutionalized religion they have no belief, and without belief they have no cooperation in equilibrium. Moreover, in their story believers expect reward or punishment regardless of their opponents' actions, and so some of their believers are better called "saints", as they unconditionally cooperate, i.e., even when they know for sure that the opponent is going to defect. As opposed to that, the story we have in mind is a story about people with high morals, call them believers, who use religious practice to segregate themselves from society, and so society loses good cooperative people, and these people lose by paying a costly signal that can be spared.

More formally, let s denote a believer type, and let as denote the nonbelievers. Next, let participation in religious services be the costly signal. As

[24] See Berman (2000) for a more economically oriented analysis of signaling in the Ultra-orthodox Jewish communities.

[25] The separation between belief and actual religious practice is well demonstrated by Huber (2005), who finds a large variability in the degree to which religious beliefs are associated with decisions to participate in religious services.

in LR, believers believe in reward and punishment, and so they are (the only ones) endowed with the "psychological cost of cheating".[26] Any individual may decide whether or not to join a religious order, at some cost, but the difference in costs for the two types is large enough to make it a reliable signal of belief. A cohesive group will consist purely of believers, who share religious rituals (such as Sunday prayers), and fully cooperate with each other, where the participation in rituals serves them to signal that they can be trusted. A mixed-type group will consist of both believers and nonbelievers, where the believers unconditionally cooperate with everyone in the group (they are not "saints" – they do so because the group is small enough and there are enough people like them in it), while the nonbelievers free-ride. Indeed, in most religions one can find believers who do not actively participate in religious services (at least not in the public ones, which are the ones used for signaling).

Note that our model implies that the religious congregations are limited in size, and are segregated from society in terms of social ties (and that there can be more than one such group in society). We believe these features capture important aspects of religion groups in real-life. By refraining from segregation and joining the other parts of society, the members of these groups can "dilute" the proportion of unbelievers in society and raise the average level of cooperation, and so raise welfare for everyone.

5.3 The Occupy Movement

In the case of social activism – participation in protests, demonstrations, clashes with police, etc. – one may argue that activists are purely concerned with the issues at hand and are not trying to signal anything. However, at least to some extent, activism serves as a clear signaling device, enabling participants to acquire higher levels of trust and cooperation among themselves, and to screen out potential free-riders. This in turn benefits the activists and at least partly compensates them for the cost they pay in participating. However, as opposed to the case of "acting white", successful activism gives additional gains to the activists (in the sense of the achieved goals), and so is not necessarily a

[26] As mentioned above, we assume that they do not believe in getting a punishment for defecting against an opponent who defects too.

wasteful act.

The Occupy Movement was a recent international protest movement against social and economic inequality, its primary goal being to make the economic and political relations in all societies less vertically hierarchical and more flatly distributed. One of the objectives of the movement was to replace the supposedly estrangement of the market system, with its alienated financial sector, with social networks of exchange, e.g., through the use of barter and free exchange of goods. Accordingly, the movement has put a great emphasis on solidarity between people.[27]

Solidarity can take many forms, but it is plausible to assume that, like with cooperation in the PD game, one benefits from the solidarity of others, while showing solidarity to others entails some sacrifice, and that people are better-off when everyone shows solidarity then when no one does. In the context of the model, a social type would be someone who feels obliged to show solidarity in return to the solidarity of others, at least as long as free riding is not too tempting, while asocial types do not feel such obligation. However, one's type is naturally a private information. If it wasn't, social types could spontaneously form groups that practice the kind of market-bypassing free exchanges that the Occupy Movement advocated, thus manifesting the solidarity they feel to one another without fear of intrusion by asocial types. In reality, although such groups did emerge at the time of the Occupy Movement (especially in the form of spontaneous and local arrangements), they were naturally of mixed-type.[28]

[27] The prominent role of solidarity is not unique to this movement. Sidney Tarrow (1994) defines a social movement as "collective challenges [to elites, authorities, other groups or cultural codes] by people with common purposes and *solidarity*".

[28] Some of the activities were inherently prone to exploitation by free riders. For example, at Occupy LSX (London Stock Exchange) the protesters created a Tent City University, where they offered free lectures and workshops, organized alternative city guides, and every night screened documentaries. At Occupy Wall Street (OWS), the OWS kitchen served free food around the clock to thousands of people every day without turning anyone away. If one adopts a broader interpretation of "asocial types" in the context of the Occupy Movement, such that it describes all those who do not share the values of movement, then there is ample evidence of such mixing of types. For example, one commentator, writing for The Cambridge Citizen, noted during the peak of the movement that "citizens from a variety of backgrounds, *whether supportive of the Occupy movement or not*, have realized that local trading of goods and services are worthwhile endeavours to become a part of" (link:http://cambridgecitizen.ca/the-occupy-movement-community-groups/).

At the same time, those who were willing to engage in costly actions such as living in a tent in the park, or participating in clashes with police forces, were able to take advantage of higher levels of cooperation and solidarity. That is, these acts were perceived as truthfully signalling solidarity, and thus enabled the formation of cohesive groups of protesters. The cohesiveness of these groups was apparent in the way they made collective decisions in the general assemblies they formed. Although theoretically accessible by anyone, the right to vote in these assemblies was practically given only to those actively participating in the protests, as you had to be present in the venues in which the assemblies took place in order to participate in the voting. As Andreas Bieler, a professor of Political Economy and an active participant in Occupy London Stock Exchange noted, "everything happened so fast in the occupation that it was impossible to keep abreast if you weren't there most of the time". With regard to the cost one had to pay in order to be part of the group, and the exclusion of non-devotees that this cost implied, Bieler wrote: "An occupation is extremely exhausting for its participants. For those permanently on site it was a full time job, often with a severe lack of sleep. [Thus] the very nature of an occupation excludes many from participation".[29] Naturally, the high level of solidarity achievable by the groups of protesters organizing and participating in the general assemblies could not be achieved by *passive* supporters of the agenda, who could only form ad-hoc groups of collaborators that were not able to screen out potential "asocial types". And so, once more we see the power of public signaling as a means of achieving cooperation and cohesion, but at a certain non negligible cost to the signalers.

6 Conclusion

The main conclusion of the paper is that a simple and quite intuitive assumption on our social conscientiousness, and more specifically – on the psychological cost of defecting from cooperation with others who wish to cooperate with us, can explain a plethora of prevailing group behaviors. These range from the mere existence of groups, through in-group bias, to costly signaling of sociality and

[29] See http://andreasbieler.blogspot.co.il/2013/02/the-occupy-movement-lasting-legacy.html.

the positive relation between the use of such signaling in a particular group and the cohesiveness of that group, a relation that was demonstrated recently in the lab by Ahn et al (2009).[30] Moreover, quite intuitively, inability to distinguish between social types, who are characterized by such social conscientiousness, and asocial types, who are not, gives rise either to costly signaling or to 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 cohesive groups who engage in costly signaling can coexist side by side with mixed-type groups where no signaling is practiced, but free riding is likely to happen. Of these two costs, it is the signaling cost that is bound to be more harmful from the point of view of society, unless the proportion of asocial types is so large that the mere existence of mixed-type groups in equilibrium is questionable. Finally, it would be interesting to directly investigate the exact shape of the psychological cost of cheating as a function of the number of cheated partners (e.g., is it fixed, smoothly concave, or is characterized by a "jump"?), possibly in experiments.

7 References

References

[1] Ahn, T. K., Isaac, R. M., and Salmon, T. C. (2009), "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

[2] Aiello, L. C. and Dunbar, R. I. M. (1992), "Neocortex Size, Group Size, and the Evolution of Language," *Current Anthropology*, 34(2), 184-193.

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

[4] Austen-Smith, D. and Fryer, R. G. (2005), "An Economic Analysis of "Acting White"," *The Quarterly Journal of Economics*, 120(2), 551-583.

[30] The key result in Ahn et al (2009) was that subjects in a restricted entry treatment (where one was incentivized to signal "sociality" or "cooperativeness" in order to be accepted as a group member) achieved substantially higher earnings, due to higher levels of cooperation, than subjects in the other treatments. Note that the total earning of a signaling member could have been lower, depending on the amount he spent on signaling before entering the group. Charness and Yang (2008) report similar evidence using a different mechanism.

[5] Benabou, R. J. M., and Tirole, J. (2006), "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678.

[6] Bendor, J., and Mookherjee, D. (1987) "Institutional Structure and the Logic of Ongoing Collective Action," *The American Political Science Review*, 81(1), 129-154.

[7] Bernhard, H., Fischbacher, U., and Fehr, E. (2006), "Parochial Altruism in Humans," *Nature*, 442, 912-915.

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

[9] Birdsell, J. B. (1970), "Local group composition among the Australian Aborigines: A critique of the evidence from fieldwork conducted since 1930," *Current Anthropology*, 11, 115-142.

[10] Bonacich, P., Shure, G. H., Kahan, J. P., and Meeker, R. J. (1976), "Cooperation and Group Size in the N-Person Prisoners' Dilemma," *The Journal of Conflict Resolution*, 20(4), 687-706.

[11] Bornhorst, F., Ichino, A., Kirchkamp, O., Schlag, K. H. and Winter, E. (2010), "Similarities and differences when building trust: the role of cultures," *Experimental Economics*, 13(3), 260-283.

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

[13] Boyer, P. (2001), Religion Explained. Basic Books, New York, NY.

[14] Brewer, M. B. (1979), "In-Group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis," *Psychological Bulletin*, 86(2), 307-324.

[15] Camerer, C. (1988), "Gifts as Economic Signals and Social Symbols," *American Journal of Sociology*, 94, Supplement: S180-S214.

[16] Charness, G. B. and C. Yang (2010), "Endogenous Group Formation and Public Goods Provision: Exclusion, Exit, Mergers, and Redemption," Department of Economics, UC Santa Barbara, Working Paper.

[17] Chen, Y., and Li, Sh. X. (2009), "Group Identity and Social Preferences," *American Economic Review*, 99, 431-457.

[18] Choi, J., and Bowles, S. (2007), "The Coevolution of Parochial Altruism

and War," *Science*, 318, 636-640.

[19] de Cremer, D., van Knippenberg, D. L., van Dijk, E., and van Leeuwen, E. (2008), "Cooperating if one's goals are collective-based: Social identification effects in social dilemmas as a function of goal-transformation," *Journal of Applied Social Psychology*, 38(6), 1562–1579.

[20] Dion, K. L. (1973), "Cohesiveness as a determinant of in-group-outgroup bias," *Journal of Personality and Social Psychology*, 28, 163-171.

[21] 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.

[22] Dunbar, R. I. M. (1992), "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, 22, 469-93.

[23] ——— (1993), "Coevolution of neocortical size, group size and language in humans," *Behavior and Brain Sciences*, 16, 681–735.

[24] Ellison, C. G., and George, L. K. (1994), "Religious Involvement, Social Ties, and Social Support in a Southeastern Community," *Journal for the Scientific Study of Religion*, 33, 46-61.

[25] Ellison, G. (1994), "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching," *The Review of Economic Studies*, 61(3), 567-588.

[26] Fershtman, C., and Gneezy, U. (2001), "Discrimination in a Segmented Society: An Experimental Approach," *The Quarterly Journal of Economics*, 116(1), 351-377.

[27] Fong, C. M., and Luttmer E. F. P.(2009) "What Determines Giving to Hurricane Katrina Victims? Experimental Evidence on Racial Group Loyalty," *American Economic Journal: Applied Economics*, 1(2), 64-87.

[28] Fordham, S., and Ogbu, J. (1986), "Black Students' School Successes: Coping with the Burden of 'Acting White'," *The Urban Review*, 18 (3), 176-206.

[29] Forge, A. (1972), "Normative factors in the settlement size of Neolithic cultivators (New Guinea)," in Man, settlement, and urbanism. Edited by P. Ucko, R. Tringham, and G. Dimbleby, 363-376. London: Duckworth.

[30] Frank, R. H. (1987), "If Homo Economicus Could Choose His Own Util-

ity Function Would He Want One with a Conscience?," *The American Economic Review*, 77(4), 593-604.

[31] Gino, F., Norton, M. I., and Ariely, D. (2010), "The Counterfeit Self : The Deceptive Costs of Faking It," *Psychological Science*, 21(5), 712-720.

[32] Gintis, H., Smith, E. A., and Bowles, S. (2001), "Costly signaling and cooperation," *Journal of Theoretical Biology*, 213, 103-119.

[33] Gneezy, U., Rockenbach, R., and Serra-Garcia, M. (2013), "Measuring lying aversion," *Journal of Economic Behavior & Organization*, 93, 293–300.

[34] Goette L., Huffman, D. and Meier, S. (2006), "The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups," *The American Economic Review*, 96(2), 212-216.

[35] Hennig-Schmidt, H., Selten, R., Walkowitz G., and Winter, E. (2009), "Cultural Biases in Bilateral Trust Among Israelis, Palestinians, and Germans," *mimeo*.

[36] Holt, C. A. and Laury, S. K. (2008), "Theoretical Explanations of Treatment Effects in Voluntary Contributions Experiments," Handbook of Experimental Economics Results, Volume 1, Ch. 90, Elsevier B.V.

[37] Huber, J.D. (2005), "Religious belief, religious participation, and social policy attitudes across countries," working paper, Columbia University.

[38] Hurkens, S. and Kartik, N. (2009), "Would I lie to you? On social preferences and lying aversion," *Experimental Economics*, 12(2), 180-192.

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

[40] Isaac, R. M., Walker, J. M., and Williams, A. W. (1994) "Group size and the voluntary provision of public goods," *Journal of Public Economics*, 54, 1-36.

[41] Kandori, M. (1992), "Social Norms and Community Enforcement," *The Review of Economic Studies*, 59(1), 63-80.

[42] Kreps, D. M., and Wilson, R. (1982), "Reputation and Imperfect Information," *Journal of Economic Theory*, 27, 253-279.

[43] Ledyard, J. O. (1995), "Public Goods: A Survey of Experimental Research". In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.

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

[45] Lopez-Perez, R., (2008), "Aversion to norm-breaking: A model," *Games and Economic Behavior*, 64, 237–267.

[46] Lopez-Perez, R., (2012), "The power of words: A model of honesty and fairness," *Journal of Economic Psychology*, 33, 642–658.

[47] Lundquist, T., Ellingsen, T., Gribbe, E. and Johannesson, M. (2009), "The Aversion to Lying," *Journal of Economic Behavior and Organization*, 70(1), 81-92.

[48] Miettinen, T., and Suetens, S., (2008), "Communication and guilt in a Prisoner's dilemma," *Journal of Conflict Resolution*, 52, 945–960.

[49] Milgrom, P., and Roberts, J. (1982), "Predation, Reputation and Entry Deterrence," *Journal of Economic Theory*, 27, 280-312.

[50] Naroll, R. (1956), "A preliminary index of social development," *American Anthropologist*, 58, 687-715.

[51] Olson, M.(1965), "The Logic of Collective Action". Cambridge: Harvard University Press.

[52] Palfrey, T. R., and Rosenthal, H. (1988), "Private incentives in social dilemmas: The effects of incomplete information and altruism," *Journal of Public Economics*, 35(3), 309-332.

[53] Sawaguchi, T., and Kudo, H. (1990), "Neocortical development and social structure in primates," *Primates*, 31, 283-290.

[54] Service, E. R. (1962), "Primitive social organization: An evolutionary perspective." New York: Random House.

[55] Smith, E. A., Bliege Bird, R. L., and Bird, D. W. (2003), "The benefits of costly signalling: Meriam turtle hunters," *Behavioral Ecology*, 14, 116-126.

[56] Shayo, M, and Zussman, A. (2011), "Judicial Ingroup Bias in the Shadow of Terrorism," *The Quarterly Journal of Economics*, Vol.126(3), 1447-1484.

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

[58] Steward, J. H. (1955), "Theory of culture change: The methodology of multilinear evolution". Urbana: University of Illinois Press.

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

[60] Tajfel, H. (1970), "Experiments in intergroup discrimination," *Scientific American*, 223, pp.96-102.

[61] Tajfel, H., Billig, M. G., Bundy, R. P., and Flament, C. (1971), "Social categorization and intergroup behavior," *European Journal of Social Psychology*, 1, 149-178.

[62] Tarrow, S.(1994), "Power in Movement: Collective Action, Social Movements and Politics." Cambridge University Press.

[63] Wilson, W., and Kayatani, M. (1968), "Intergroup attitudes and strategies in games between opponents of the same or of a different race," Journal of Personality and Social Psychology, 9, 24-30.