Group Formation, In-group Bias and the Cost of Cheating*

Moti Michaeli†

Abstract

Group formation and in-group bias – preferential treatment for insiders – are widely observed social phenomena. This paper demonstrates how they arise naturally when people incur a psychological cost as the result of defecting when facing cooperators, when this cost is increasing and concave in the number of such defections. If some group members are asocial, i.e., insusceptible to that cost, then, under incomplete information, free-riding and cooperation can coexist within groups. Signaling of one's type can enable groups to screen out free-riders, but signaling is costly, and its availability may decrease the welfare of *all* the individuals in society.

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

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

1 Introduction

This paper presents a unified theory that explains a number of widely observed social phenomena: (i) people tend to form groups; (ii) group size is limited;

*I would like to thank Roland Bénabou, Elchanan Ben-Porat, Bård Harstad, Sergiu Hart, Rachel Kranton, Yona Rubinstein, Moses Shayo, Paul Slovic, Daniel Spiro, Eyal Winter, Ro'i Zultan, seminar participants at the Hebrew University, the University of Oslo, the European University Institute and IDC Herzliya, as well as participants at the Nordic Conference on Behavioral Economics, the Econometric Society European winter meetings, IMBESS meeting and THEEM conference for their valuable comments.

†The European University Institute, Florence. Contact: moti.michaeli@eui.eu.

(iii) groups tend to show in-group bias; and (iv) free-riders and cooperators can coexist within groups.[1] Many previous theories in the literature have offered an explanation for one or two of these stylized facts, but have rarely aimed to explain all of them together, mainly because these facts are hard to reconcile.[2]

In order to explain the above observations together, I suggest a theory that builds upon one basic assumption about the existence – and the shape – of a *psychological cost of cheating* at the individual level. The term 'cheating' refers here to its common usage in models of the Prisoner's Dilemma game, where it indicates defecting while playing against a cooperative opponent. That is, a person is endowed with a psychological cost of cheating if this person incurs disutility from not reciprocating another's kind actions. The assumption about the shape of this cost is that it rises concavely with the number of cheated individuals, so that cheating has a diminishing marginal cost. This feature of the cost represents *scope neglect* and resonates with recent experimental work on cheating.[3] I show that when people endowed with this psychological cost

[1] For evidence in support of (i) see e.g. Ahn et al (2008,2009), Charness and Yang (2010), Aimone et al. (2013) and Biele et al. (2008). The relation between group size and cooperation (ii) is explored for example in Marwell and Schmitt (1972), Komorita et al. (1992), Isaac et al. (1994), Ledyard (1995) and Holt and Laury (2008). In-group bias (iii) has been demonstrated both in natural environments: Goette et al. (2006), Bernhard et al (2006), Fong and Luttmer (2009) and Shayo and Zussman (2011), and in the lab: Tajfel (1970), Tajfel et al. (1971), Chen and Li (2009), Efferson et al (2008) and de Cremer et al. (2008). For evidence in support of (iv) see e.g. Foster and Rosenzweig (1995), Marwell and Ames (1981), Kim and Walker (1984) and Isaac et al. (1984,1985).

[2] For example, theories that can explain within-group cooperation often fail to explain the limit on group size (e.g., Fehr and Schmidt 1999, Rabin 1993 and Ledyard 1993); others explain the limit on group size but do not account for in-group bias (e.g., Olsen 1965, Bendor and Mookherjee 1987, Boyd and Richerson 1988 and Suzuki and Akiyama 2005; see also Dunbar 1993 for an anthropological account); those explaining in-group bias do not account for group formation, group size, or both (e.g., Becker 1957, Phelps 1972, Arrow 1973, Choi and Bowles 2007 and Fu et al. 2012); theories that may seem to explain stylized facts (i)-(iii), most notably theories of reciprocity, reputation, or kin selection (e.g., Axelrod and Hamilton 1981, Nowak and Sigmund 2005 and Hamilton 1964 respectively), find it hard to explain a coexistence of free-riding and cooperation. Finally, a coexistence of free-riding and cooperation in equilibrium is explained by, e.g., Palfrey and Rosental (1988) and Bramoullé and Kranton (2007), but in these papers groups are exogenously given.

[3] Scope neglect, or scope insensitivity, represents the notion that people tend to be insensitive to the magnitude of outcomes and particularly to the number of victims (or survivors) of a certain intervention. This tendency is widely documented (Kahneman, 1986; Desvousges et al., 1993; McFadden and Leonard, 1993; Nordgren and McDonnell 2010; Västfjäll et al. 2014). For example, by manipulating the expected number of victims of a hypothetical

interact, cooperation is bound to be limited and this stimulates the formation of groups that display in-group favoritism. This in-group bias is not built on any joint agreement between group members to "boycott" out-group members. Rather, it is a pure equilibrium phenomenon based on individual decision making.

To better understand the idea, consider the following thought experiment. Imagine you are a collector of cards who has a pack of cards you wish to exchange in return for other cards you do not possess. On the Internet you may find potential trading partners. Suppose, for the sake of simplicity, that you can exchange only one card with each partner and that there are no card duplicates. Once you find a partner who also wishes to trade cards with you, the exchange is performed in the following way: your partner puts the card you have asked for in an envelope and mails it to your home address, while simultaneously you mail him the card you promised to give in exchange. At the same time, you make similar arrangements with other card collectors, each of whom wishes to exchange one card with you.

Now, what if you do not mail some of the cards you promised? With no duplicates, you attach some (possibly small) value to every single card you are willing to exchange. So any unsent card gives you more or less the same benefit, the benefit of being able to keep that specific card in your collection, regardless of your decision about other cards. However, the disutility you incur by cheating an anonymous partner on the web may depend rather crucially on the total number of partners you cheat. In particular, the assumption of concavity essentially implies that while the total 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 in order to keep one card, it may well be attractive to cheat many partners in order to keep many cards. In fact, cheating is bound to take place once there are so

scenario of chemical leakage, Västfjäll et al. (2014) showed that subjects were sensitive to risking the life of one chemical engineer (compared to zero), while being insensitive to the difference between risking either one, two or three chemical engineers. Recent experimental findings (e.g., Gino et al. 2010 and Gneezy et al. 2013) are also consistent with a concave cost of cheating, as subjects tend to choose a corner solution of either not cheating at all or cheating to the maximum extent.

many partners that the average disutility from cheating a partner falls below the value of keeping a card. But the other side of the coin is that cheating *can* be prevented if the trade in cards is limited to a sufficiently small group of collectors.

The argument of the basic model (Section 2) thus goes as follows. The concavity of the psychological cost of cheating jeopardizes any attempt at large scale cooperation. However, cooperation on a smaller scale is sustainable, thus triggering the formation of groups. Belonging to a group of limited size ensures that the temptation to cheat is resistible, and that others can be trusted to cooperate because *their* temptation is resistible too.[4] Moreover, each member of the group is bound to show in-group bias – an inclination to cooperate only with members of his own group. Otherwise, a person would have "too many" cooperative partners, and the temptation to defect would destroy cooperation both within and between groups. This argument is developed in a model in which individuals interact in a standard pairwise Prisoner's Dilemma game resembling the card-trade example. Each individual decides whether or not to cooperate with any other individual in society. This allows for discriminatory behavior. I show how groups of cooperators can endogenously emerge in such a setting and I characterize the feasible size of such groups (groups can potentially be of different sizes in equilibrium but they are subject to the same limit).[5]

Naturally, not necessarily everyone in society is endowed with a psychological cost of cheating, and the existence of this endowment is generally one's private information (Section 3). To account for this, I model society as consisting of two types – *social types*, who are subject to the psychological cost

[4]It is interesting to note that in a series of Public Goods Game experiments, Isaac et al. (1994) find that in fact cooperation sometimes *increases* with group size. This may seem to contradict my hypothesis. However, this finding holds only in the (arguably unrealistic) case in which the monetary gain of the cheater is independent of his group size. Thus, the finding is in fact in line with the model of this paper: when monetary gains from cheating are independent of the number of cheated individuals, people are indeed predicted to cheat *less* the larger the group is, as cheating more people implies a higher psychological cost.

[5]The endogenous formation of groups of limited size can alternatively be demonstrated using the Public Goods game (with some auxiliary assumptions). However, in the Public Goods game there is no direct modelling of intergroup interaction. Therefore, given that a main goal of the paper is to explain in-group bias, I construct the model in this paper using the pairwise Prisoner's Dilemma game instead.

of cheating, and *asocial types,* who are not subject to this cost. One's type is one's private information, but individuals know the relative proportions of types in society. Social types do not mind cheating asocial types, but, fearing they might mistake a fellow social type for an asocial type, may 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. Moreover, a lower proportion of social individuals in society will be correlated with smaller social structures. If one considers the level of trust in society to be a good proxy for the proportion of social types in it, 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.[6]

Real groups are often capable of screening out free-riders by introducing costly signaling, i.e., demanding that members exhibit some form of payoff-irrelevant self-sacrifice (Section 4). If the cost of signaling is sufficiently low for cooperative group members to bear and at the same time sufficiently high to distance potential free-riders, *signaling groups*, consisting only of social types who fully cooperate with one another (but not with outsiders), can exist. Moreover, multiple such purely cooperative groups can coexist alongside the mixed groups. However, the existence of signaling groups strictly decreases the expected utility of all members of mixed groups, regardless of their type. The reason for this is that the availability of the signaling technology allows social types to separate themselves (at a personal cost) from the rest of society, thus depriving the rest from the benefit of interacting with them. Moreover, 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 too. Thus, beyond

[6] Although I do not aim to establish causality in the empirical sense, my model implies that social structures *reflect* the individual traits of society members. Note that a reversed causality, according to which people who happen to live in small groups tend as a result to be asocial (mainly toward outsiders, maybe due to lack of experience in dealing with strangers), can account only for asociality between groups but cannot explain free-riding within groups. Moreover, it takes the social structure as exogenous, while in this paper it is endogenous.

the private cost to the individual who signals, signaling as a phenomenon imposes a public cost on society. This negative externality 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 accounted for when considering the problem of free-riding.[7]

Sections 2, 3 and 4 respectively present the model under complete information, incomplete information and with signaling and Section 5 concludes. In the appendix I suggest theoretical microfoundations for the psychological cost of cheating and demonstrate how the cost of signaling can be endogenized in the model.

2 The basic model

Society contains a mass 1 of individuals who simultaneously interact with each other to play one-shot Prisoner's Dilemma (PD) games. The payoff matrix for the game is as follows.[8]

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

The payoff from mutual cooperation is normalized to 1 so that cooperating with a mass k of individuals yields a payoff k. g stands for the *gain* from unilateral defection, and ℓ for the *loss* from being the victim of the opponent's unilateral defection. I assume strategic complementarity (i.e., $\ell > g$), which implies that if one's opponent is more prone to defect, one is more prone to

[7] It is also one of the main features differentiating this paper from similar models studying the relation between cooperation and costly signaling, such as Levy and Razin (2012) and Iannaccone (1994).

[8] The model is intentionally simplified. Below I discuss a dynamic (and arguably more realistic) interpretation of the game. It is also possible to develop a repeated-interactions version of the model. However, the repeated PD game is known to be capable of producing cooperation, provided players are sufficiently patient, and so might mask the contribution of the new explanation suggested here. For the payoff matrix of the game I adopt the notations of Kandori (1992) and Ellison (1994). The zero payoff for mutual defection implies that there is no difference between mutual defection and no interaction at all, hence applies also to cases where not everyone interacts. Furthermore, it implies that the payoff for mutual cooperation is strictly positive, hence the total return to cooperation increases as the number of one's cooperative partners increases (nevertheless, groups will be of limited size in equilibrium).

defect too. This assumption is standard in the literature (see e.g. Tabellini 2008 and Levy and Razin 2014). My analysis considers only pure strategies in every pairwise interaction, but individuals can "mix" by discriminating between opponents, i.e., by playing C against some while playing D against others.[9]

2.1 The individual's psychological cost

Beyond the material payoffs of the game, some people are subject to a psychological 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. This psychological cost can be thought of as representing the arousal of uncomfortable feelings such as shame or guilt on the side of the defector.[10] $t(0)$ is set to 0 and $t(k)$ is assumed to (weakly) increase in k – the more people are cheated by the individual, the more it costs him – and to be (weakly) concave. I do not require smooth concavity or even continuity, so any cost function with a discrete jump at 0 and a weakly increasing and weakly concave continuation afterwards satisfies my condition of concavity. In particular, this includes a step function with a fixed cost of cheating t for any $k > 0$, which can capture a binary distinction between one's self image as a cheater and one's self image as someone who does not cheat.[11] Finally, I add two requirements that are close in spirit to the INADA conditions: an infinite slope at 0 (or otherwise a discrete "jump") and an upper bound on the cost as k goes to 1 (= the mass

[9] Mixing in the pairwise interaction level imposes here a modeling ambiguity. As part of the payoff function (the psychological cost) is related to disutility from defecting when playing against a cooperative opponent, it is unclear how to model the disutility of defecting against an opponent who uses a mixed strategy – is it the realization that counts, or maybe the opponent's (impure) intention to cooperate? I prefer to leave these potential controversies aside.

[10] Miettinen and Suetens (2008) indeed show that (most) people feel guilty when defecting in the PD game, but *only* if their partner has not also defected. The assumption that defection is costly only when others cooperate is also in line with Fischbacher et al. (2001) and Frey and Meier (2004), who show that people are generally conditional cooperators, and with Lopez-Perez (2008), with the exception that Lopez-Perez would treat the k cooperators as those who respect the norm and the defector as the norm breaker.

[11] One may also consider a more elaborate model of self image, where, with some probability q, cheating an opponent does not trigger the bad self image (e.g., because the cheater does not pay attention to the cheating or is able to find excuses for it). If this probability q is *iid* across the pairwise encounters, the expected cost of cheating K cooperative opponents would be $\left(1-q^K\right)t$, which is concave in K, hence in line with my psychological assumption.

of individuals in the whole society). The first requirement ensures that the psychological cost is sufficiently large to allow for at least some cooperation. The second requirement 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 interactions.[12] Formally, the assumptions on $t(k)$ beyond positive monotonicity and concavity are:

$$t(0) = 0,\ t(1) < g, \text{ and } \lim_{k \to 0} t'(k) = \infty$$

$$\text{(or, if } \lim_{k \to 0} t'(k) \text{ is not defined, } \lim_{k \to 0^+} t(k) > 0\text{)}.$$

In the appendix I suggest theoretical microfoundations to the psychological cost of cheating, showing that it can be elicited by generally allowing for psychological costs in the PD game, while requiring that these costs are rational and efficient in a particular well-defined sense.

2.2 The society

Society is composed of two types of individuals – *social types* and *asocial types*. 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 prone to the psychological cost of cheating. In the basic model with complete information analyzed in this section, the type of each individual is assumed to be common knowledge. The strategy of each player specifies which action he plays in the PD game against any other player in society. A set of players' strategies forms a Nash equilibrium if, given the strategies of all other individuals, no individual has a profitable deviation from his own strategy.

[12] This second requirement can be replaced by the original INADA requirement, $\lim_{k \to \infty} t'(k) = 0$, if the size of society is modeled as unbounded. As for the first requirement, it is more restrictive than is needed for the results of the basic model to hold. Here, it is enough to have $\lim_{k \to 0} t'(k) > g$. The requirement of an infinite slope at 0 is needed for Proposition 2 (which concerns cooperation under incomplete information). I discuss the effect of relaxing this requirement after presenting that proposition.

2.3 Solving the model

The result that cooperation can be sustained within groups of social types, 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 $\bar{K} \in (0,1)$ be the unique strictly positive solution to the equation $t(K) = Kg$. Then in every Nash equilibrium:*

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

2. Every social type plays C against a mass of individuals of size $\bar{K}$ or less, who play C against him too, and plays D against everyone else.

Proof. See Appendix. ■

The intuition is easy to understand when considering deviations from full cooperation. The concavity of the psychological cost (in the number of cooperative partners) and the linearity of the material gain from unilateral defection imply the existence of a certain threshold on the number of cooperative partners, beyond which a social type is better off by cheating (all of them).[13] This threshold is $\bar{K}$ (see Figure 1). As cheating cannot be sustained in equilibrium – the cheated side always has a profitable deviation to playing D – pairwise interactions in equilibrium are only of two kinds: either C-C or D-D. This implies that all relationships will be mutual. Finally, since asocial types are observable and will always choose to defect, their interactions in equilibrium are only of the D-D kind.[14]

One way to think of the game's strategies is to view them as commitments that each player makes when choosing the action to be played against any other individual in society. In the card game example, for instance, a strategy would

[13] The material gain from unilateral defection is assumed to be linear for simplicity, but this is not a necessary condition for this result to hold. If the material gain is concave, the necessary condition is that the psychological cost will be even more concave.

[14] Note that, as $\bar{K}$ is only an upper limit, some social types may also have only D-D interactions in equilibrium. In particular, the proposition also covers the special case where, as in the standard solution to the PD game, every individual in society plays D against everyone else.

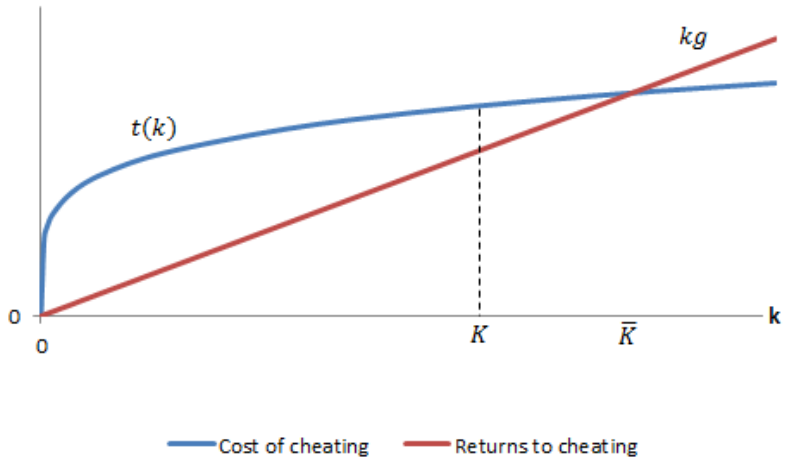

Figure 1: The limit on a cooperative group size. For any given mass k of cooperators, the linear red line depicts the material gain from cheating them, while the curved blue line depicts the psychological cost of doing so. These lines intersect at $k = \bar{K}$. A social type can cooperate in equilibrium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

be a commitment to send cards to a certain subset of the traders' community. These strategies form an equilibrium if no player regrets his committed strategy after all encounters took place and all other players followed their own committed strategies. Under this interpretation, the game does not necessarily have to be simultaneous, i.e., the model as it is can account for sequential PD encounters, as long as there is no flow of information about past behavior of players. The proper equilibrium notion for the sequential version of the game is Sequential Equilibrium (or alternatively, Perfect Bayesian Equilibrium). It can be shown that there is $1:1$ correspondence between the set of *NE* in Proposition 1 and the set of sequential equilibria of the sequential version of the game. In particular, it means that there is no cheating in a sequential equilibrium and that the total number of cooperative partners of each social type does not exceed $\bar{K}$. It also implies that the sequence in which meetings occur does not matter. This follows from the fact that, in equilibrium, not necessarily every pair of social types cooperate, so it may well be the case that the k partners with which a social type cooperates are not the first k social types he encounters.[15]

As my main focus in this paper is on the formation of mutually exclusive groups, I present now the following definition and a relevant corollary of the proposition:

Definition 1 *Let a cohesive group be a collection of individuals who play C with each other and play D against all out-group members.*

Corollary 1 *Any partition of the social types into cohesive groups whose sizes are bounded by $\bar{K}$ can be sustained in equilibrium.*

[15]The claim that there are no sequential equilibria in which a player cooperates with more than $\bar{K}$ players follows from the fact that any system of beliefs that could supposedly support more than $\bar{K}$ cooperative partners for a player implies that the player has a profitable deviation to permanent cheating at the first instance in which he meets a partner he believes will cooperate. The other claim, about no cheating in equilibrium, may seem counter intuitive because there exist values of $k < \bar{K}$ such that a player who has already cheated k times in the past has a dominant strategy to play D from now on (because his marginal cost of cheating at this point is smaller than g). However, cheating does not happen on the equilibrium path. The intuition is that any subgame that contains cheating cannot be reached in the first place since the first instance of cheating will never occur – the player who is supposed to cheat in this instance has a profitable deviation to not cheating.

The result implies that it is easier to sustain cooperation in smaller groups. This 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 result is driven by the concavity of the cost of cheating: as the size of the group increases, it becomes harder to avoid the temptation to defect in order to achieve the ever growing material benefits of unilateral defection. At some point one is bound to surrender to the temptation and cheat. The limit on group size in equilibrium is the threshold above which this is bound to occur. Yet, it is only an upper limit: groups may be of different sizes, as often is the case in reality.

The model is silent about the exact way in which groups are formed. However, in real life, people may use cues to solve the coordination problem. These cues will most often be identity-related, e.g., relating to gender or ethnicity. Thus, players do not necessarily have to be nominally identified for the result to hold. The cues may also be context dependent. For example, if card traders on the Internet are identified solely by their name and geographical location, then gender and nationality are likely to determine the division of traders into trading groups. Moreover, a person may belong to such a group when he engages in card-trading and at the same time belong to a race-based group in his workplace or in the army. Hence, although the setup does not explicitly model multiple identities and group formation, it is applicable in the manner described.

Another aspect of the result is that it generates in-group bias: In equilibrium, social types would show the same level of asociality towards out-group members as asocial types would, while exhibiting sociality only towards in-group members.[16] This suggests that even people who choose a very cooperative life style, e.g. Kibbutz members, would restrict their cooperation to within their group alone. Indeed, in an experiment conducted by Ruffle and Sosis (2006), Kibbutz members exhibited the same level of generosity as that of city residents towards anonymous out-group peers, while showing higher levels of generosity towards anonymous in-group peers. This pattern of in-group bias

[16] If $\bar{K}$ is larger than the mass of social types in society, the social types may be united in one group, showing in-group bias only towards the asocial types.

was also documented in early experimental studies of the Prisoner's Dilemma game in a group context.[17] More recently, de Dreu (2010) found similar patterns using the Intergroup Prisoner's Dilemma–Maximizing Differences Game (IPD-MD). He showed that, compared to individuals with a "chronic pro-self orientation", those with a "chronic prosocial orientation" (i.e., social types) displayed stronger in-group trust and in-group love — they were self-sacrificing to benefit their ingroup — but not more or less outgroup distrust and outgroup hate. As I argue in Section 4, the self-sacrifice practiced by social types is not always intended to benefit the ingroup, but could instead be a means of costly signaling.

3 In-group bias under incomplete information

In this section I relax the somewhat strong assumption that asocial types can be easily distinguished from social types. Instead, here an individual's type is his private information. Let the mass (and proportion) of asocial types in society be p and suppose 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 in the affirmative.

Definition 2 *A mixed group is a collection of individuals of both types, such that:*

- *All social types in the group play C against all other in-group members, and D against all out-group members.*
- *All asocial types in the group play D against both in-group and out-group members.*

Proposition 2 *Given $p \in (0,1)$, $\exists K_p \in (0, \bar{K})$ such that a mixed group of size K is sustainable in equilibrium if and only if $K \leq K_p$. Furthermore, K_p is strictly decreasing in p.*

[17] For example, Wilson and Kayatani (1968) and Dion (1973) found that the competitiveness that characterized inter-group behavior resembled that of individual players, whereas it was the increased proportion of cooperative choices exhibited in intra-group decisions that deviated from typical inter-personal play (see also further analysis in Brewer 1979).

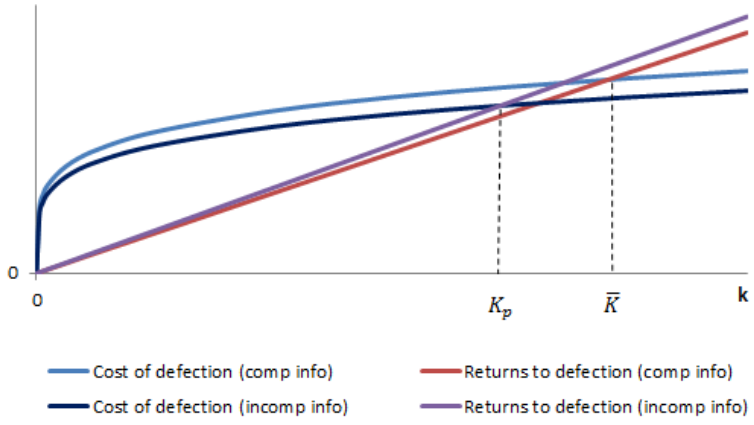

Figure 2: The limit on a mixed group size. For any given mass k of opponents, such that a proportion p of them defect while the others cooperate, the linear purple line depicts the material gain from playing D against all of them, while the curved dark blue line depicts the psychological cost of doing so. These lines intersect at $k = K_p$. The linear red and the curved light blue lines that intersect at $k = \bar{K}$ are taken from Figure 1 and are displayed for comparison. The purple line is drawn above the red line because $\ell > g$, and so the returns to defection are larger when facing a mixed group. The dark blue line is drawn below the light blue line because the psychological 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.

Proof. See Appendix. ■

The intuition for the proposition is similar to that of Proposition 1. From the point of view of a social type, the trade-off is still between cheating some cooperative partners (the social types in his group) and the material gain from doing so. However, here (i) the material temptation to defect is 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 – and (ii) the psychological cost of playing D against a random group member is lower (because some of the "victims" of defection will be asocial). See Figure 2 for illustration.

Corollary 2 *Any partition of society into mixed groups whose sizes are bounded by K_p forms a Bayesian equilibrium.*

The corollary refers to equilibria in which mutually exclusive groups coexist. In these equilibria, social types show in-group bias, by playing C against all group members and D against all outsiders, while asocial types play D against everyone, thus free riding on the social types in their groups.[18] These groups are bound to be smaller than the groups of purely social types in the complete information case (i.e., $K_p \leq \bar{K}$). Moreover, the maximal group size is decreasing in p because the greater the proportion of asocial types in society is, the more it is tempting for social types to defect, hence the smaller the groups are that can sustain cooperation.[19] The cooperative behavior of social types and the sustainability of free-riding in equilibrium are in line with the "weak free-riding hypothesis". 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 (1981).

An interesting scenario is revealed when considering the case of $p > 1/(1+\ell)$. In this case, the proportion of asocial types in society is sufficiently high to make the expected payoff of a social type in a mixed group of size K *negative*, regardless of the value of K (because his expected payoff is $K[(1-p)-p\ell)] < 0$). This means that such a social individual would have been better off in a society where everyone else is known to defect (so that he could defect too with no pangs of conscience and get a zero payoff). However, even in the case of $p > 1/(1+\ell)$, 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

[18] The assignment of the asocial types to different groups thus follows not from their own actions but instead from the identity of the social types who play C against them. Strictly speaking, this setup can also yield Bayesian equilibria in which not all the social types in a mixed-type group cooperate, accompanied by an appropriate system of beliefs. However, in such equilibria, all groups must be of exactly the same size, which is the unique size that would make social types indifferent between cooperation and defection. This restriction makes these equilibria somewhat artificial. Moreover, just as in the basic model, here too I do not characterize all the equilibria that exist under incomplete information. Instead, I focus throughout the paper on the case where separate groups are formed and analyze their characteristics.

[19] However, groups that are sufficiently small *can* sustain cooperation. This is so because in such groups the material payoffs are low, so there is not much to gain by defection, while the psychological cost of cheating kicks-in already on the first occasion of cheating. If we relax the requirement that $\lim_{k\to 0} t'(k)$ is infinite, we get that for sufficiently large values of p cooperation is unsustainable even in small groups.

group members, in order to avoid hurting other cooperative individuals like themselves.[20]

4 Introducing signaling

4.1 Suckers and signalers

In reality, groups often use signaling in order to enhance cooperation and to screen out free-riders. The signaler would usually exhibit some sort of sacrifice, which is meant to reveal his true type.[21] On first impression, signaling may seem incapable of helping the social types in my model – they can never acquire the cooperation of asocial types, and the cooperation of social types in their (mixed) groups is anyway guaranteed. However, if social types condition their own cooperation on the opponent's signaling, an equilibrium with costly signaling, where the cost is compensated for by achieving group cohesiveness, may nevertheless exist.[22]

Suppose therefore that the pairwise PD game is preceded by a signaling stage, in which every individual decides whether on not to signal. A signal is not directed at any specific partner, but is rather a particular payoff-irrelevant sacrifice that is observable by everyone else.[23] The cost of signaling is denoted

[20] One may 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 with so many tax evaders that he would be better off with no tax system and no public service at all.

[21] Analyses along these lines can be found in Camerer (1988), Akerlof and Kranton (2000), Berman (2000), Bacharach and Gambetta (2001), Bénabou and Tirole (2006) and Levy and Razin (2012). In particular, Levy and Razin (2012) study religious participation as a costly signal, and their model and results share some similarities with my own. However, importantly, the driving mechanism in Levy and Razin (2012) is that religious people believe they will be rewarded for their actions in the afterlife, regardless of their opponent's action, and they gain utility from that expected reward. This is particularly apparent in the existence of equilibria in which individuals cooperate in the PD game if and only if they are religious, regardless of their opponent. Thus, it is hard to apply their model to a more general setup like the one discussed here, where beliefs about divine providence play no role.

[22] Camerer (1988) seems to miss this point when he models gift exchange as a signal of willingness to invest later in a relationship. He concludes that in cases where "willing" (social) types anyway invest (cooperate) when facing an unknown type under incomplete information, there is no potential for signaling.

[23] This feature of the signal is quite common in the literature (e.g., Iannaccone 1992, Levy and Razin 2012). Furthermore, it is not unreasonable to have signaling at the level of the group or the society as a means to promote pairwise relationships. See Gintis et al. (2001) for examples of biological signals of this kind and the evidence on Meriam turtle hunters in

by x_s for social types and x_{as} for asocial types. Note that there can be various kinds of different signals such that all impose the same cost on a signaler of a given type. For example, attending religious services may be a useful form of signaling, and the fact that one may choose between different churches to attend implies that each choice can be interpreted as a different signal. The difference between signals that cost the same may then facilitate a division into separate groups that all practice signaling (see below).[24]

I will refer to the signals as "signals of sociality", as everyone, regardless of one's type, would like to be considered a social type and achieve the cooperation of his partners (i.e., even if he intends to cheat them). The signals may then be reliable or unreliable.

Definition 3 *Signaling (of sociality) is said to be reliable if the signaler is guaranteed to be of a social type*

The condition under which signaling is reliable will be stated later. If signaling is indeed reliable, there is a natural interpretation for the group formation process that takes place after the signaling stage: Those who did not signal, be they social or asocial, can form mixed groups as before; those who did signal, and are therefore guaranteed to be social types, can form *signaling groups*.

Definition 4 *A signaling group is a cohesive group whose members signal and condition their cooperation with other group members on them signaling too.*

As the following proposition states, signaling groups are bound to be of a medium size.

Smith et al. (2003).

[24] Another relevant example is the use of different colored handkerchiefs (known as the "hanky code") by males in gay bars in the San Francisco area in order to signal preferences for various kinds of sexual relationships (Gambetta 2009, p. 166-168). In this example, the signal is not required in order to resolve a problem of free riding, but it does help promote cooperation and can facilitate a solution to the coordination problem. This example is somewhat extreme but is useful to demonstrate a case where $x_s << x_{as}$ because, although the monetary cost of a handkerchief of a certain color is the same for everyone, wearing it in a gay bar is much more costly for those who do not wish to take part in the corresponding sexual relationship.

Proposition 3 *A signaling group is sustainable in equilibrium if and only if its size K is in the range* $\left[x_s, \hat{K} \equiv \min\left\{1-p, \bar{K}, \frac{x_{as}}{1+g}\right\}\right]$.

Proof. See Appendix. ■

To see why the conditions in the proposition are required, suppose that a signaling group of size K exists. As cooperation is maintained (only) within the group, the payoff of each member of the group is $K - x_s$. If this payoff is negative, one can just stop signaling and get a zero payoff.[25] This sets the lower bound on group size. The upper bound contains three elements. First, there should be sufficiently many social types in society to form the group ($K \leq 1-p$). Second, K has to be bounded from above by $\bar{K}$, because otherwise every group member has a profitable deviation to cheating even though he truthfully signaled that he is social. Finally, the third element ensures that the payoff of an asocial type who signals and then cheats the other group members, $K(1+g) - x_{as}$, is negative. In fact, this condition is essentially the condition for reliable signaling, as it guarantees that every signaler is social.[26]

The upper limit on group size implies that if there are many signalers in society, they cannot simply all cooperate with each other, but must instead divide into separate signaling groups.[27] In this sense, signaling 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. Proposition 3 further implies that the size of signaling groups is bounded not only from above but also from below, where the cost of signaling for group

[25] Note that the other group members condition their cooperation on him signaling, so he does not have to cheat once he stops signaling.

[26] One can also think of equilibria with signaling that is *unreliable* according to Definition 3. For example, it is possible to construct a pooling equilibrium in which all individuals in society (of both types) pay the cost of signaling and are members of mixed groups. If everyone's expected payoff is positive and if deviation to not signaling implies being treated as someone who is for sure asocial and thus receives a zero payoff, this can constitute an equilibrium. Another example is an equilibrium where asocial types can impersonate being social types (because signaling is unreliable) but do not do so because the equilibrium payoff for each of them is higher without signaling. This can happen if they are members of sufficiently large mixed groups. Such an equilibrium requires strong assumptions about the beliefs of social types who do use the unreliable signaling in equilibrium. These possibilities are left outside the model.

[27] Of course, if for a given proportion q of signalers in society there exists no such feasible division, then no equilibrium with proportion q of signalers exists.

members sets the lower bound. It follows that, if this cost is high, signaling groups should be sufficiently large in order to exist. In reality, groups often determine this cost for themselves. Thus, the heavier is the cost determined by the group, the larger must the group be in order to survive. However, if the cost is set too high (above $\hat{K}$, as defined in the proposition), no individual will take part in such a group. These observations are in line with evidence on religious participation, as reported by Iannaccone (1994). Firstly, Iannaccone provides evidence that, with regard to out-group members, at least some religious groups "condemn deviance, shun dissenters, and *repudiate the outside world*", i.e., they display in-group bias, even toward other sects of the same religion. Secondly, he reports that in the US, stricter churches (i.e., those that require a higher cost in terms of members' devotion) tend to be larger. Thirdly, he writes that the data "imply 'optimal' levels of strictness, beyond which strictness discourages most people from joining or remaining within the group".[28]

Abstracting from the issue of feasible group size, the following corollary lists the conditions on the model parameters that allow for signaling in equilibrium.

Corollary 3 *Signaling groups can exist in equilibrium if and only if the following conditions hold:*

1. Feasibility: $x_s \leq 1 - p$
2. Separability: $\frac{x_{as}}{x_s} \geq 1 + g$
3. Individual rationality: $x_s \leq \bar{K}$

Proof. The corollary follows immediately from a comparison of the lower and the upper limits on the size of signaling groups in Proposition 3. ■

[28] According to Iannaccone (1994), the high cost of adherence to the strict rules of conduct enables the church to screen out potential free-riders and this in turn raises its attractiveness and thus increases its size. In my model, the higher cost of signaling used by stricter churches does not necessarily generates attractiveness, but it implies that they *must* attract many members in order to survive, thus the surviving strict churches are bound to be large. Note also that when analyzing the ability of strict churches to attract followers, it is equally important to analyze what makes imitation by asocial types even more costly, otherwise separability of the types cannot be achieved.

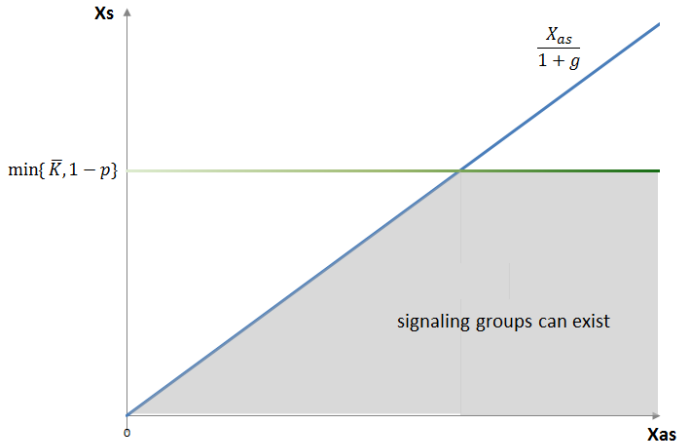

Figure 3: Displaying the conditions that allow signaling groups to exist, as a function of the cost of signaling for asocial types (x_{as}) and for social types (x_s). The blue diagonal line is where the ratio of these type-dependent costs is equal to the ratio of the types' respective payoffs from interacting with a cooperative partner, i.e., where $\frac{x_{as}}{x_s} = 1 + g$. It marks the border between the region where social types can potentially 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 horizontal green line, signaling is either unfeasible, because there are not enough social types in society to form even one signaling group; or it is not individually rational, as the gain from cooperation in a signaling group cannot exceed $\bar{K}$ in equilibrium. The region below the green and the blue lines is where signaling groups of size $K \in \left[x_s, \hat{K} \equiv \min\left\{\bar{K}, \frac{x_{as}}{1+g}, 1-p\right\}\right]$ may exist in equilibrium.

Figure 3 illustrates graphically the conditions of the corollary. The first condition simply guarantees that there are sufficiently many social types for at least one signaling group in society. The second condition compares the cost of signaling for both types. Naturally, in order to achieve separation between the types, the cost of signaling sociality should be lower for social types than for asocial types (a reasonable assumption in itself, reflecting the notion that it should cost more to fake sociality than to signal it when it indeed exists, as argued in Frank 1987). However, as the second condition implies, this is not sufficient, since the gain for an asocial type from being considered as social exceeds that of a truly social type by a ratio of $1 + g$ to 1. Therefore, an asocial type will be willing to pay a higher cost in order to be perceived as social, and so separation requires a ratio of costs that is larger than $1 + g$.[29] Finally, the third condition states that the cost of signaling should not exceed $\bar{K}$. This guarantees that cooperation is sustainable, so that one may benefit from belonging to a signaling group.

The conditions in the corollary can be stated as one condition, $x_s < \hat{K}$. It is important to note that this condition does not *guarantee* that a fully separating equilibrium will indeed emerge. 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. 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. Even more interestingly, there can be semi-separating equilibria in which multiple purely cooperative signaling groups coexist side by side with mixed groups. This seems to be a sensible characterization of society. A natural interpretation of this kind of equilibria is that they reflect how different individuals may find different solutions to the problem of cooperation: some may choose to endure free-riders in their group, while others may choose to engage in wasteful signaling.

[29] In Appendix B, I show how the cost of signaling can be endogenized in the model. There, the cost is determined by the willingness of types to contribute to a public good at the signaling stage, and the separability condition is stated as a requirement on other primitives of the model.

4.2 Signaling: a double-edged sword

The multiplicity of equilibria invites a comparison of them in terms of welfare and stability. As will be shown here, these concepts are tightly related. I start by analyzing the effect of signaling on the welfare of individuals who do not signal and on the signalers themselves, and then introduce a stability concept that reflects the welfare result from a different angle. I keep focusing on partitions of society into mutually exclusive groups, where each group can be either a mixed group or a signaling group.

4.2.1 Welfare analysis

In order to compare the welfare of different partitions of society, I introduce the following definition.

Definition 5 *A coalition formation is a partition of society into mutually exclusive groups, be they mixed groups or signaling groups, such that the individuals' strategies under this partition form an equilibrium.*

The following result highlights the negative externality of signaling on society. It essentially states that the existence of signaling groups strictly decreases the expected utility of *all* members of mixed groups, regardless of their type.

Proposition 4 *For any given $p \in (0,1)$, the expected payoff of all the non-signalers in any coalition formation that contains a non-zero mass of signalers can be strictly increased by prohibiting signaling.*

Proof. See appendix. ■

The intuition for the proposition is as follows. Signaling groups contain only social types. Hence, when signaling groups exist, the actual proportion of asocial types in the rest of the population (i.e., outside the signaling groups) is higher than p. This has two negative effects on the expected payoff of members of mixed groups. The first is a greater expected number of interactions with defecting opponents for any given group size (and a smaller number of interactions with cooperative opponents). The second is a decrease in the upper limit on the size of mixed groups, which implies a reduction in the maximal expected

payoff of group members of both types. Hence, if signaling is prohibited, mixed groups can be larger and more cooperative, allowing for higher payoffs for both types.

Thus, beyond the individual cost for the signaler, signaling as a social 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 willingness to cooperate. One may think that at least for the signalers themselves signaling improves welfare. However, the following lemma states that this is the case only for sufficiently large values of p.

Lemma 1 *Suppose that $x_s < \hat{K}$ and let p_c be the unique implicit solution to the equation*

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

Then there is a tipping point for a social type – in the coalition formations that maximize his expected payoff, he is signaling if and only if $p \geq p_c$.

Proof. See appendix. ■

Equation (1) 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,[30] 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 4 illustrates this result. Together with Proposition 4, Lemma 1 implies that when $p < p_c$, a coalition formation in which all groups are mixed and of maximal size (if it exists) Pareto dominates any coalition formation that includes signaling groups (note that this applies also to the asocial types, who gain maximally from free riding in this case).[31]

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

[31] 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 the paper take this into account.

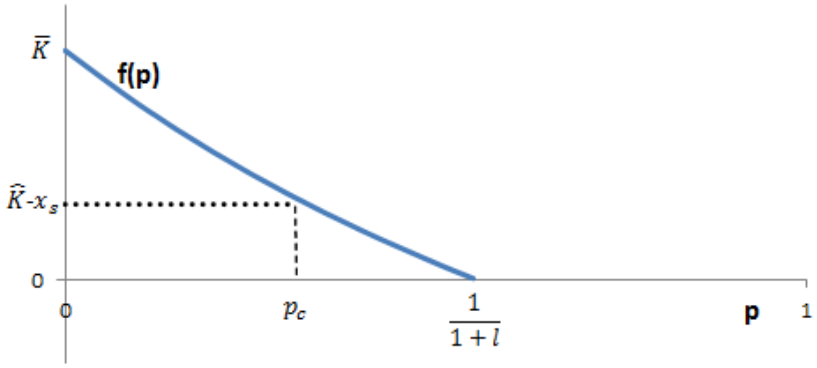

Figure 4: The tipping point for social types. $f(p) \equiv K_p[1 - p(1 + \ell)]$ is the maximum expected payoff of a social type in a mixed group. It is achieved in a pooling equilibrium (i.e., when there is no signaling in society) when the individual's group is of the maximum possible size given p. $\hat{K} - x_s$ is the expected payoff in a signaling group of the maximum size. 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 $p_c > p$, social types can get a payoff of $\hat{K} - x_s$ in signaling groups of maximum size, and this payoff is strictly greater than the expected payoff they can achieve in mixed groups.

4.2.2 Stability of equilibria with signaling

The welfare result in Lemma 1 is related to the concept of *core stability* of coalition formations, which I adopt from Bogomolnaia and Jackson (2002) and Banerjee et al. (2001).

Definition 6 *A coalition formation* Π *is said to be unstable if there exists another coalition formation* Π' *and in it a coalition* $T \notin \Pi$*, such that all members of* T *have strictly higher payoffs under* Π' *than under* Π*.*

In our context, T would be a mixed group or a signaling group that does not exist under the considered coalition formation Π, yet is feasible in equilibrium.

Proposition 5 *If* $p < p_c$*, then any coalition formation with a non-zero mass of signalers is unstable.*

Proof. See appendix. ■

The intuition for this result is rather straightforward in light of the definition. What makes the coalition formation with signaling unstable is that members of both signaling groups and mixed groups can increase their payoffs by forming larger and more prosocial mixed groups, which cannot exist when signaling is used.

This result implies that if one restricts attention only to stable equilibria, then these equilibria can contain signaling only as long as the asocial types constitute a sufficiently large fraction of the population. It is thus very similar in spirit to the conclusion of Iannaccone (1992) about the conditions under which costly signaling will be used by social clubs, cults or communes.[32]

5 Conclusion

This paper shows that a simple and quite intuitive assumption about our social conscientiousness, and more specifically, about our psychological cost of defecting from cooperation with others who wish to cooperate with us, can explain a plethora of prevailing group behaviors. These behaviors range from the mere existence of groups, through in-group bias, to costly signaling of sociality and the positive correlation between the use of such signaling in a particular group and the cohesiveness of that group. Moreover, quite intuitively, an inability to distinguish between social types, who are characterized by such a psychological 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 cohesive groups who engage in costly signaling can coexist 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.

[32] In Iannaccone's (1992) model, 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. Thus, the equivalents to type 1 and type 2 people in my model are asocial types and social types respectively. Proposition 2 in Iannaccone's paper then states that "*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" (the italics are mine).

6 References

References

[1] Ahn, T. K., Isaac, R. M., and Salmon, T. C. (2008), "Endogenous Group Formation," *Journal of Public Economic Theory*, 10(2), 171-194.

[2] ——— (2009), "Coming and going: Experiments on endogenous group sizes for excludable public goods," *Journal of Public Economics*, 93, 336–351.

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

[4] 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.

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

[6] Anderson, S., Goeree, J., Holt, C. (1998), "A theoretical analysis of altruism and decision error in public goods games," *Journal of Public Economics* 70(2), 297–323.

[7] Arrow, K. J. (1973), "The Theory of Discrimination" . In Orley Ashenfelter and Albert Rees, eds. Discrimination in Labor Markets. Princeton, N.J.: Princeton University Press, 3-33.

[8] Axelrod, R., and Hamilton, W. D. (1981), "The evolution of cooperation," *Science*, 211, 1390-1396.

[9] Bacharach, M., and Gambetta, D. (2001), "Trust in Signs". pp. 148-184 in Trust in Society, edited by Karen S. Cook. New York: Russell Sage Foundation.

[10] Banerjee, S., Konishi, H., and Sönmez, T. (2001), "Core in a simple coalition formation game," *Social Choice and Welfare*, 18(1), 135-153.

[11] Becker, G. (1957), "The Economics of Discrimination". Chicago: University of Chicago Press.

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

[13] Bendor, J., and Mookherjee, D. (1987) "Institutional Structure and the

Logic of Ongoing Collective Action," *The American Political Science Review*, 81(1), 129-154.

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

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

[16] Biele, G., Rieskamp, J., and Czienskowski, U. (2008), "Explaining cooperation in groups: Testing models of reciprocity and learning," *Organizational Behavior and Human Decision Processes*, 106(2), 89-105.

[17] Bogomolnaia, A., and Jackson, M. O. (2002), "The stability of hedonic coalition structures," *Games and Economic Behavior*, 38(2), 201-230.

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

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

[20] Bramoullé, Y. and Kranton, R. (2007), "Public goods in networks," *Journal of Economic Theory*, 135, 478 – 494.

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

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

[23] 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.

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

[25] Choi, J., and Bowles, S. (2007), "The Coevolution of Parochial Altruism and War," *Science*, 318, 636-640.

[26] 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.

[27] Desvousges, W. H., Johnson, F. R., Dunford, R. W., Boyle, K. J., Hudson, S. P., and Wilson, K. N. (1993), "Measuring natural resource damages with contingent valuation: Tests of validity and reliability". In J. A. Hausman (Ed.), Contingent valuation: A critical assessment (pp. 91-164). Amsterdam, Netherlands: North Holland.

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

[29] 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.

[30] Dunbar, R. I. M. (1993), "Coevolution of neocortical size, group size and language in humans," *Behavior and Brain Sciences*, 16, 681–735.

[31] Efferson, C., Lalive, R., and Fehr, E. (2008), "The coevolution of cultural groups and ingroup favoritism," *Science*, 321, 1844–1849.

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

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

[34] Fischbacher, U., Gachter, S., and Fehr, E. (2001), "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, 71(3), 397-404.

[35] 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.

[36] Foster, A.D., and Rosenzweig, M.R. (1995), "Learning by doing and learning from others: human capital and technical change in agriculture," *J. Polit. Economy*, 103(6), 1176–1209.

[37] Frank, R. H. (1987), "If Homo Economicus Could Choose His Own Utility Function Would He Want One with a Conscience?," *The American Economic Review*, 77(4), 593-604.

[38] Frey, B. S., and Meier, S. (2004), "Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment," *Ameri-*

can Economic Review, 1717-1722.

[39] Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., and Nowak, M. A. (2012), "Evolution of in-group favoritism," *Scientific reports*, 2.

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

[41] Gambetta, D. (2009), Codes of the underworld. Princeton University Press.

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

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

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

[45] 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.

[46] Hamilton, W. D. (1964), "The genetical evolution of social behaviour," *Journal of Theoretical Biology*, 7(1), 1–16 and 17–32.

[47] 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.

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

[49] ——— (1994), "Why strict churches are strong," *American Journal of Sociology*, 99(5), 1180-1211.

[50] Isaac, R. M., McCue, K. F., and Plott C. (1985), "Public Goods Provision in an Experimental Environment," *Journal of Public Economics*, 26, 51-74.

[51] Isaac, R. M., Walker, J. M., and Thomas, S. H. (1984), "Divergent Evidence on Free Riding:An Experimental Examination of Possible Explanations," *Public Choice*, 43, 113-149.

[52] 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.

[53] Kahneman, D. (1986), "Comments on the contingent valuation method". In R. G. Cummings, D. S. Brookshire, & W. D. Schulze (Eds.), Valuing environmental goods: An assessment of the contingent valuation method (pp. 185-193). Totowa, NJ: Rowan and Allanheld.

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

[55] Kim, O. and Walker, M. (1984), "The Free Rider Problem: Experimental Evidence," *Public Choice*, 43, 3-24.

[56] Komorita, S. S., Parks, C. D., and Hulbert, L. G. (1992), "Reciprocity and the induction of cooperation in social dilemmas," *Journal of Personality and Social Psychology*, 62(4), 607-617.

[57] Laury, S. (1996), "Experimental studies on the voluntary provision of public goods". Unpublished doctoral dissertation, Indiana University.

[58] Ledyard, J. (1993), "Is there a problem with public goods provision?". In: J. Kagel and A. Roth, eds., The handbook of experimental economics, Princeton University Press.

[59] ——— (1995), "Public Goods: A Survey of Experimental Research". In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.

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

[61] ——— (2014), "Rituals or Good Works: Social Signalling in Religious Organizations," *The Journal of European Economic Association*, 12(5), 1317–1360.

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

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

[64] Marwell, G. and Ames, R. (1981), "Economists Free Ride, Does Anyone Else?" *Journal of Public Economics*, 15, 295-310.

[65] Marwell, G., and Schmitt, D. R. (1972), "Cooperation in a three-person Prisoner's Dilemma," *Journal of Personality and Social Psychology*, 21(3), 376-383.

[66] McFadden, D., and Leonard, G. (1993), "Assessing use value losses caused by natural resource injury". In J. A. Hausman (Ed.), Contingent valuation: A critical assessment (pp. 341-363). Amsterdam, Netherlands: North Holland.

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

[68] Nordgren, L. F., and McDonnell, M. H. M. (2011), "The Scope-Severity Paradox: Why Doing More Harm Is Judged to Be Less Harmful," *Social Psychological and Personality Science*, 2(1), 97-102.

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

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

[71] 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.

[72] Phelps, E. S. (1972), "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62, 659-61.

[73] 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).

[74] Rabin, M. (1993), "Incorporating fairness into game theory and economics," *The American economic review*, 83(5), 1281-1302.

[75] Rapoport, H., and Weiss, A. (2003), "The optimal size for a minority," *Journal of economic behavior & organization*, 52(1), 27-45.

[76] 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.

[77] 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.

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

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

[80] Tabellini, G. (2008), "The Scope of Cooperation: Values and Incentives," *The Quarterly Journal of Economics*, 123(3), 905-950.

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

[82] 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.

[83] Västfjäll, D., Slovic, P., Mayorga, M., and Peters, E. (2014), "Compassion fade: Affect and charity are greatest for a single child in need," *PloS one*, 9(6), e100115.

[84] 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.

A Appendix: Theoretical microfoundations to the psychological cost of cheating

This section endogenizes the psychological cost of cheating by suggesting theoretical microfoundations that support my assumption that cheating (and only cheating) is what triggers a psychological cost, and that this cost is plausibly described by an increasing and concave function. The argument is based on showing that a cost of cheating is both *rational* and *efficient*.

Consider a PD game and suppose that beyond the material payoffs of the game, people are subject to psychological costs that are related to the possible realizations of the game. These costs can capture feelings such as shame, anger, pity, envy, pride, frustration, guilt, etc., that are aroused if a certain outcome

of the game is realized. However, a psychological cost is required to be *rational* in the following sense.

Definition 7 *A psychological cost is said to be rational if it has the potential to strictly increase the material payoff of the individual incurring it.*

Claim 1 *A psychological cost is rational if and only if it is associated with the realization [D,C]*

Proof. In order to increase the material payoff of the individual, the psychological cost must induce cooperation by his opponent. As the opponent may cooperate only if he believes that the individual will cooperate too, the psychological cost must facilitate the cooperation of the individual himself, and so can only be attached to playing D. However, this requirement is insufficient, as having a psychological cost attached to the realization $[D,D]$ cannot induce cooperation. To see why, suppose this cost is so high that the individual prefers $[C,D]$ to $[D,D]$. Then, the opponent's best strategy is to play D and get the maximal payoff, $1+g$. Hence, having a psychological cost attached to the realization $[D,D]$ can only decrease the material payoff of the individual incurring this cost. This concludes the "only if" part of the claim. In order to see that a psychological cost that is associated with the realization $[D,C]$ has the potential to strictly increase the individual's payoff, suppose that both the individual and his opponent share such a psychological cost and this cost is sufficiently high to make them prefer $[C,C]$ to $[D,C]$, and this is common knowledge. Then $[C,C]$ becomes an equilibrium of the game played between them, and both players' payoffs are strictly greater than without the psychological cost. ■

Let $t(k)$ denote the psychological (and rational) cost of playing D against a mass k of individuals who play C. I now show that positive monotonicity and concavity are sufficient conditions for *efficiency* of this cost in the following sense.

Definition 8 *A rational psychological cost $t(k)$ is said to be inefficient if $\exists K_1, K_2 \in (0,1]$ such that $K_1 < K_2$, $t(K_1) < K_1 g$, and $t(K_2) \geq K_2 g$.*

That is, what makes the cost inefficient is that it makes the individual prefer playing C against K_2 cooperative opponents to playing D against them, yet cooperation with all these K_2 individuals is not sustainable, because, at the same time, he prefers to cheat against a subset of them of size K_1 (and he can do so as interaction is pairwise). From this definition immediately follows the next corollary.

Corollary 4 *A rational psychological cost $t(k)$ is efficient if and only if $\exists K \in [0,1]$ such that $\forall k \leq K$, $t(k) \geq kg$, and $\forall k > K$, $t(k) < kg$.*

The corollary states that once $t(k)$ falls below the linear line kg, it should also stay below it. This does not guarantee that $t(k)$ is increasing and concave. However, note that the main result of the basic model, Proposition 1, holds for *any* efficient psychological cost $t(k)$.[33] As positive monotonicity and concavity seem to be plausible assumptions with behavioral foundations and they help in keeping the model tractable, I limited the analysis in the paper to the special case of an increasing and concave $t(k)$.

B Appendix: Endogenizing the cost of signaling

Section 4 investigated the prospects for signaling in equilibrium when the cost of signaling for each type was exogenously given. It showed that, in order to enable separation between the two types, this cost should be sufficiently higher for asocial types ($\frac{x_{as}}{x_s} \geq 1+g$). Here I suggest a way to endogenize the cost of signaling by correlating it to the one characteristic that distinguishes the two types in my model, namely the existence (or lack of) a psychological cost of cheating.

Suppose that we model social interaction in two stages. In the first stage, groups are endogenously formed and each group plays a public good game. In the second stage, everyone plays one-shot PD against everyone else in society. If the same psychological cost of cheating that characterizes social types in the second stage applies to the public good game of the first stage as well, this cost

[33] Proposition 1 holds exactly as it is as long as there is a unique crossing point $\bar{K} \in (0,1)$ at which $t(k) = kg$ (see the proof of that proposition). Otherwise the proposition needs to be slightly changed.

can turn the contribution to the public good into a signal of sociality. Beyond endogenizing the cost of signaling, this setup differs from the one in the text in two additional ways. First, here I implicitly assume that the signal is visible only within the group. Thus, by signaling, the signaler may only expect to gain the cooperation of social types who belong to his original group. Second, since now groups are formed *before* signaling takes place, he may expect to have a proportion of $1-p$ social types in his group, regardless of his decision whether to signal or not.

Formally, let c denote the material cost of contributing to the public good (for both types). Suppose that a group of size K is formed in the first stage and that there exists a separating equilibrium in which all the social types in the group, and only them, contribute to the public good in the first stage and then form a signaling group of their own in the second. For the contribution to be a separating signal, it should be unprofitable for asocial types to mimic. This happens if and only if the cost c exceeds the expected benefit from being perceived by the social types in the group as a social type too (in the second stage), i.e., if and only if c exceeds the expected return to cheating them, $K(1-p)(1+g)$. As for the social types, since, compared to the asocial types, they rip off smaller material benefits from cooperative partners (1 instead of $1+g$ from each partner), they will contribute to the public good only if the psychological cost of shirking from contribution is sufficiently high. Let this psychological cost be similar to that of cheating in the PD game, so that it costs $t(k)$ to shirk from contribution in the presence of a mass k of other contributors. Then a social type has no profitable deviation from contributing if and only if the return to cooperating in the second stage with the other social types (who do signal), minus the cost of contribution, exceeds the payoff of not contributing, i.e., if and only if $(1-p)K - c \geq 0 - t((1-p)K)$. Combining these two conditions and denoting the size of the emergent signaling group by $S \equiv (1-p)K$, we get that contribution to a public good as a preliminary signaling stage can lead to type separation and the formation of signaling groups only if

$$Sg \leq c - S \leq t(S). \tag{2}$$

Recall now that as long as $S \leq \bar{K}$, Sg is (weakly) smaller than $t(S)$, and so there exist values of c that satisfy this inequality. This implies the potential existence of equilibria in which social types screen out free-riders and form signaling groups. Moreover, as the condition $Sg \leq t(S)$ implies that $\bar{K}$ is an upper limit on S, we get that, just as with the signaling groups of Section 4, here too the limit on cooperation *among* social types restricts the size of groups.

Furthermore, signaling groups will be bounded in size both from above and from below. Just as when the cost of signaling was exogenous, the risk of mimicry by the asocial types limits the signaling group size (S) from above while individual rationality limits it from below. As follows from the first inequality in equation (2), the upper limit is $\frac{c}{(1+g)}$, because otherwise $K > \frac{c}{(1-p)(1+g)}$, in which case, as explained above, asocial types find it profitable to contribute to the public good and then cheat in the second stage. As for the lower limit, the second inequality in equation (2) implies that $S + t(S)$ (which is monotonically increasing in S and equals 0 when $S = 0$) must be greater than c. However, there is another sense in which the act of contribution should be individually rational: As social types can refrain from taking part in any group in the first stage and thus secure a zero payoff, individual rationality implies also that their total payoff in a signaling group must be positive. This adds a restriction on the benefit created by the public good. Let $k \cdot b$ denote the benefit to each member of the group from the public good when a mass k of individuals contribute. Then, individual rationality requires $Sb - c + S \geq 0$, where $Sb - c$ is the payoff in the first stage and S is the payoff in the second. Thus, S is limited from below also by $\frac{c}{(1+b)}$. The size of a signaling group must therefore be at the range $\left[\frac{c}{(1+b)}, \frac{c}{(1+g)}\right]$. This further implies that contribution to the public good can serve as a separating signal only if $b \geq g$, i.e., only if the public benefit it creates exceeds the private benefit from cheating.

C Appendix: proofs

C.1 Proof of Proposition 1

Lemma 2 *The equation $t(K) = Kg$ has a unique strictly positive solution $\bar{K}$ in $]0,1[$. Moreover, $t(K) > Kg$ for every $K \in \left]0, \bar{K}\right[$ while $t(K) < Kg$ for every $K \in \left]\bar{K}, 1\right]$.*

Proof. *First assume by negation that there are at least two solutions to equation $t(K) = Kg$ in the interval $]0,1[$, denoted by K_1 and K_2. Since the conditions on $\lim_{k\to 0} t'(k)$ and $\lim_{k\to 0^+} t(k)$ imply 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 $t(K) = Kg$. Next, note that $t(K) - Kg$ is strictly positive at $K = \varepsilon$, strictly negative at $K = 1$ (by assumption), and any possible discontinuity in between is an increase. Thus $t(K) - Kg = 0$ at least once in the range $[\varepsilon, 1]$. Therefore we get that $t(K) - Kg = 0$ exactly once in the range $[\varepsilon, 1]$, at which point $t(K) - Kg$ changes signs from positive to negative, so that $t(K) > Kg$ for every $K \in \left]0, \bar{K}\right[$ while $t(K) < Kg$ for every $K \in \left]\bar{K}, 1\right]$. ■

Proof of Proposition 1. Denote the action played by player i in the PD game against player j by s_{ij}. 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 statement (1). Next, it follows from Lemma 2 that $t(K) > Kg$ for every $K < \bar{K}$ while $t(K) < Kg$ for every $K > \bar{K}$. If a mass K of individuals play C against a social type, and $K \leq \bar{K}$, then his best response is to play C against all of them, as deviating to defection against any

subset of them (of size $k \leq K$) would impose on him a net cost of $t(k)-kg \geq 0$. 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 $Kg - t(K) > 0$. This completes the proof of statement (2). ■

C.2 Proof of Proposition 2

Lemma 3 *Let $\Delta(k,p) \equiv t((1-p)k) - k[(1-p)g + p\ell]$. Then for any $p \in (0,1)$, $\exists K_p \in (0,\bar{K})$ s.t. $\Delta(k,p) \geq 0$ if and only if $k \leq K_p$. Furthermore, K_p is strictly decreasing in p.*

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 every $k > \bar{K}$ (because $t((1-p)k) \leq t(k)$, $[(1-p)g + p\ell] > g$ and for every $k > \bar{K}$ we have $t(k) < kg$). Moreover, $\lim_{k\to 0} \frac{\partial \Delta(k,p)}{\partial k} = +\infty$ (or, if $\lim_{k\to 0} t'(k)$ is not defined, $\lim_{k\to 0^+} \Delta(k,p) = \lim_{k\to 0^+} t(k) > 0$) and $\Delta(k,p)$ is weakly concave in k. Thus, $\exists K_p \in (0,\bar{K})$ such that $\Delta(K_p,p) = 0$, $\Delta(k,p) > 0$ for every $k < K_p$, and $\Delta(k,p) < 0$ for every $k > K_p$.*[34] *Finally, $\Delta(k,p)$ is strictly decreasing in p, which means that for any $\{p,q|p<q\}$ we have $\Delta(k,q) < 0$ for every $k \geq K_p$, and so $K_q < K_p$, i.e., K_p is strictly decreasing in p.* ■

Proof of Proposition 2. First, it is straightforward that asocial types have no profitable deviation, since as members of mixed groups they already play their dominant strategy D against everyone else in society. As for the social types, consider a social type who is a member of a mixed group of size K. By the definition of a mixed group, all the social types in the group play C against all other group members, including him. Moreover, because of the incomplete information, the individual cannot choose to defect only against the asocial types in the group. Defecting against any subset of the group, of mass $k \leq K$, of which only a fraction of $(1-p)$ are social,[35] would result in an increase in the

[34] If $\Delta(k,p)$ is discontinuous due to discontinuity of $t(k)$, then the same logic of the proof to Lemma 2 applies here too.

[35] Strictly speaking, the distribution of realizations (in terms of the exact proportion of social types) over an interval of size k is not defined. However, it is common to assume that the measure p applies to any subinterval of the original range $[0,1]$. One way to explicitly model this is to represent society by $[0,1]^2$, where the choice of partners is applied only in one dimension (represented by choosing a subinterval of $[0,1]$), while the other dimension

expected material payoff of $k[(1-p)g+p\ell]$, but the expected total payoff would also decrease by $t((1-p)\,k)$ due to the cost of cheating. Thus, the individual would have no profitable deviation if and only if $t((1-p)\,k) \geq k[(1-p)g+p\ell]$ for every $k \leq K$, i.e., iff $\Delta\,(k,p) = t((1-p)\,k) - k[(1-p)g+p\ell] \geq 0$ for every $k \leq K$. By Lemma 3, this holds if and only if $K \leq K_p$, where K_p is the unique value of k for which $\Delta\,(k,p) = 0$. It thus follows that mixed groups of size K are sustainable if and only if $K \leq K_p$. That K_p is decreasing in p is proved in Lemma 3. ■

C.3 Proof of Proposition 3

Proof. Consider a signaling group of size K. Then: (i) There are sufficiently many social types to form the group if and only if $K \leq 1-p$. (ii) Group members are social types hence have no profitable deviation to not signaling if and only if $x_s \leq K$. (iii) Group members have no profitable deviation to cheating if and only if $K \leq \bar{K}$. (iv) The asocial types who are not part of any group in equilibrium have no profitable deviation to signaling and then cheating if and only if $K(1+g) - x_{as} \leq 0$. ■

C.4 Proving results on welfare and stability (Sect. 4.2)

Proof of Proposition 4. Take any coalition formation Π that contains a non-zero mass of signalers. Then q, the proportion of asocial types among the non-signalers, is strictly greater than p, their proportion in society. Take now a different partition Π' with no signaling, such that all the members of mixed groups under partition Π are still members of mixed groups of the same size under Π'. This partition can be sustained in equilibrium since $K_q < K_p$ (see Proposition 2). Moreover, under partition Π', each of these individuals gains a strictly higher expected payoff than under partition Π. This is so because the fraction of asocial types in the groups decreases from q under Π to p under Π', and both types gain from this decrease. ■

Proof of Lemma 1. First recall that the expected payoff of a social type in a mixed group of size K is $K[1-p(1+\ell)]$, which is negative if $p > \frac{1}{1+\ell}$, but positive and increasing in the group size if $p \leq \frac{1}{1+\ell}$, where, given p, it reaches

guarantees that the proportion of asocial types is p for every chosen set of partners.

its maximal value $f(p) \equiv K_p[1-p(1+\ell)]$ when the group is of maximal size, 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 $x_s < \hat{K}$, there is a unique solution to equation (1), denoted by p_c, and $p_c \in \left(0, \frac{1}{1+\ell}\right)$. Next, since $\hat{K} - x_s$ is the maximal equilibrium payoff in signaling groups, we get that if $p < p_c$ this payoff is strictly smaller than the payoff achievable by social types in mixed groups. If on the other hand $p \geq p_c$, then the converse is true – the maximal payoff achievable by social types in mixed groups is smaller than $\hat{K} - x_s$, the maximal payoff achievable in signaling groups. ■

Proof of Proposition 5. Let Π be a coalition formation with a non-zero mass of signalers, and let q denote the proportion of asocial types among the non-signalers under this coalition formation. Consider now a different coalition formation Π' in which there is no signaling, and which contains a mixed group T of size K_p. If $p < p_c$, it follows that $p < \frac{1}{1+\ell}$, in which case the expected payoff of every member of T is strictly higher than the maximal expected payoff he can obtain in a mixed group under partition Π (because $p < q \Rightarrow f(p) > \max\{f(q), 0\}$ – see Figure 4 and the proof to Lemma 1). Furthermore, since T is of maximal size, the fact that $p < p_c$ implies (by Lemma 1) that the expected payoff of social types in T is higher than the expected payoff of any member of a signaling group under partition Π. Thus, the expected payoffs of all the members of T are strictly higher than their expected payoffs under coalition formation Π, and so Π is unstable. ■