

# Good manners: signaling social preferences

Russell Golman<sup>1</sup> 

© Springer Science+Business Media New York 2015

**Abstract** Certain messages, even when not directly payoff relevant, can be a credible form of communication in light of natural social preferences. Social image concerns and other-regarding preferences interact to create incentives to communicate about how one feels about other people. Recognizing the prevalence of the incentive to communicate about one's social preferences suggests that many social and economic phenomena—from norms of etiquette to cooperation to gift exchange—should be seen, in part, as forms of signaling. These behaviors may be surprisingly robust to material costs, yet sensitive to context.

**Keywords** Cheap talk · Credible communication · Etiquette · Signaling · Social preferences

## 1 Introduction

People are social creatures. We care about each other, and we care about how others feel about us. To understand economic behavior, such as public goods contributions, employee relations, consumption of socially responsible products, and more, we must account for the role of social preferences in the choices people make. Economists have long recognized other-regarding preferences, including altruism (Becker 1976; Andreoni 1989), spitefulness, and reciprocity (Levine 1998; Fehr and Gächter 2000b;

---

R. Golman thank Linda Babcock and Sudeep Bhatia for very helpful comments.

---

✉ Russell Golman  
rgolman@andrew.cmu.edu

<sup>1</sup> Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Charness and Rabin 2002; Sobel 2005; Falk and Fischbacher 2006; Ackermann et al. 2014). They have also recognized social image concerns (Holländer 1990; Bernheim 1994; Tadelis 2008; Andreoni and Bernheim 2009; Grossman 2015). It is tempting to try to attribute social behavior (such as giving freely to others) to one of these motives as opposed to the others, but this is a false choice—there is evidence that both kinds of social preferences are at play together (Bowles and Gintis 2005; Della Vigna et al. 2012).<sup>1</sup>

Certainly there is value in modeling either other-regarding preferences or social image concerns in isolation so that we may understand the specific implications of each. The seminal work of Benabou and Tirole (2006) shows the value of modeling them together because having image concerns along with other-regarding preferences opens up the possibility that extrinsic rewards can crowd out prosocial behavior by spoiling its reputational value.<sup>2</sup> This article contends that we get an additional insight from modeling both forms of social preferences together, i.e., from considering the interaction of other-regarding preferences with social image concerns. Together, these social preferences create a new incentive to communicate about one's preferences. Fundamentally, this is because people's beliefs directly affect their utilities (see, e.g., Akerlof and Dickens 1982; Köszegi 2006; Golman and Loewenstein 2015). People care about their beliefs—and about others' beliefs—and communicate accordingly.<sup>3</sup>

Communication about one's preferences can take many forms. Information can be conveyed through one's actions, i.e., via costly signaling, which is generally acknowledged to be credible. Information can also be conveyed through one's words, which are not directly payoff relevant. While such messages are sometimes dismissed as *cheap talk*, in certain cases they may actually be credible. Cheap talk messages can be “self-signaling,” i.e., naturally demonstrating their credibility by the fact that they were sent (see Farrell and Rabin 1996). Crawford and Sobel (1982) show that cheap talk is capable of revealing some truthful information, and Farrell and Rabin (1996) argue that it commonly does. We argue here that messages about how one feels about others (which are not directly payoff relevant and thus seem to be cheap talk) are often naturally credible. We characterize the signaling about social preferences that can emerge through cheap talk in a (partially) separating equilibrium. As in Crawford and Sobel (1982) and Farrell and Rabin (1996), cheap talk achieves credibility despite the opportunity to freely imitate others (i.e., to lie) because it is in the speakers' interests in equilibrium to (partially) reveal their actual types (i.e., to tell the truth). Notably, whereas Crawford and Sobel (1982) and Farrell and Rabin (1996) consider situations in which people want to tell the truth because of the actions the listeners will take in

<sup>1</sup> There is evidence that people also care about social norms like fairness independently of the approval or disapproval they get by conforming to or violating them, but for parsimony our model will not attempt to capture this kind of social preference.

<sup>2</sup> Benabou and Tirole (2006) recognize that self-image can matter as much as social image. The omission of self-image concerns from our model is solely for simplicity.

<sup>3</sup> Social image concerns alone also imply that people care about others' beliefs, but to the extent that people share a common conception of a desirable image, they give everybody the *same* incentives, so credible communication may not frequently occur in equilibrium as individuals cannot necessarily distinguish themselves.

response, in our setting people want to tell the truth because they care intrinsically about the listeners' beliefs.

Andreoni (1989, 1990) distinguishes *warm-glow altruism* from *pure altruism*, recognizing that people typically are not motivated just by concern for the welfare of others but also by the act of contributing to their welfare. Often people are more sensitive to their own role in helping others than to the need for that help. Our model captures warm-glow altruism as the product of pure altruism and the signal sent through the act of giving or helping. A desire to signal altruism or fairness has been recognized as one possible basis for warm-glow motivation (see Andreoni and Bernheim 2009). This signal might be sent to inspire reciprocity (Levine 1998) or simply to earn social approval (Andreoni and Bernheim 2009), or, as we additionally suggest here, to express social approval out of pure altruistic regard for another person. Our schema thus contributes to developing microfoundations for warm-glow motivation. This account of the warm glow can provide insight about the situations in which warm-glow motivation will play a role over and above pure altruism.

Our perspective suggests that many social and economic phenomena should be seen as forms of communication about people's social preferences (see also Fehr and Fischbacher 2002; Sliwka 2007). An illustrative example of this is good manners. Saying "I'm sorry" when one has hurt (or even failed to help) another person conveys that any harm done was not the object of one's actions, but merely incidental to some other purpose. In many (though certainly not all) contexts, a person says "I'm sorry" expressly to make the victim feel better, and doing so credibly signals that the person bears no ill will, or else the person would not want to make the victim feel better. Similarly, saying "thank you" after receiving a favor conveys appreciation for the other person. Positive regard is generally valued for its own sake, and an expression of positive regard can be credible simply because it is given as a form of non-monetary reward.

Of course, proper etiquette is a social norm. Most people are not consciously determining the precise message they would like to convey with their polite remarks. They may think of their behavior as simply conforming to established social norms. But social norms concerning manners arise as equilibria of games of communication. (Other social norms, e.g., concerning fairness, might arise from deep-seated distributional preferences or from a need to reduce conflict among members of a society who frequently interact, but etiquette is inherently about what others will think. See Kahneman et al. 1986; Camerer and Thaler 1995.) These norms make the meaning of arbitrary behaviors easily understood as conveying consideration or disregard for others, as a matter of common knowledge. It requires little or no deliberation to interpret ordinary good manners as polite and violations of norms of etiquette as rude.

Etiquette may seem to be just a cute example of behavior that serves to signal people's social preferences, but its economic relevance should not be dismissed. An organization's efficiency and worker productivity often depend on the teamwork of employees. Organizational citizenship behaviors are essential for highly functioning teams (Organ et al. 2006), and many forms of organizational citizenship should be seen as expressions of proper etiquette. For example, an experienced worker may show a new hire the ropes, offer assistance with a challenging task, or simply introduce herself as part of the team, often as a gesture even if this help is not actually needed. Of course,

it is undeniable that organizational citizenship behaviors (and conformity to norms of etiquette more generally) are sometimes strategically motivated, with one's reputation for a repeated game in mind—e.g., covering for a co-worker in anticipation of future reciprocation. But, it would stretch credulity to believe that people always have such strategic motives in mind (see also [Organ 1988](#)). We do not generally observe workers on the day they retire leaving a mess for their colleagues to clean up, nor do we see all tenured professors resigning from their administrative committees. We could just say that it would feel wrong to do so, but the reason it feels wrong is the signal of disrespect it communicates.<sup>4</sup> We further discuss implications for organizational efficiency and worker productivity in Sect. 4, touching on the robustness of cooperation and gift exchange.

We proceed by presenting a model of people with social preferences, entailing other-regarding preferences and social image concerns, in Sect. 2. Section 3 analyzes credible communication with cheap talk about social preferences, characterizing a (partially) separating equilibrium in the simplest setting. Section 4 discusses signaling of altruism and spite in more general settings and explores broader implications for our understanding of economic phenomena, before concluding.

## 2 Modeling social preferences

We construct a utility function that allows for a wide range of social preferences (as well as pure selfishness), including altruism, spitefulness, reciprocity, and social image concerns. As in standard economic models, individuals in the absence of others have preferences over outcomes that can be represented by an outcome utility function  $\hat{u}(x)$ . In a social interaction (i.e., an  $n$ -person game) with players  $i = 1, \dots, n$ , player  $i$  cares about her overall utility  $v_i$ , which is derived from her own and others' outcome utilities and parameters describing their feelings about each other. We posit a collection of parameters  $\beta_{ij}$ , bounded in  $-1 < \beta_{ij} < 1$ , which describe how player  $i$  feels about player  $j$ . The value of  $\beta_{ij}$  is private information of player  $i$ , whereas player  $j$  just knows the prior distribution of  $\beta_{ij}$ . We give more structure to these parameters below.

Overall utility incorporates two additional terms on top of outcome utility to capture social image concerns and other-regarding preferences, respectively. Taking as fundamental the preference to be liked by others, we introduce a (weakly) concave, strictly increasing function  $S$  to represent utility derived from one's social image. We can define self-centered utility for player  $i$  as

$$u_i = \hat{u}(x_i) + \sum_{j \neq i} S(E_i(\beta_{ji})), \quad (1)$$

where the notation  $E_i(\cdot)$  indicates the expectation given player  $i$ 's information. (We abuse notation by using  $\beta_{ij}$  both as a random variable and as the value taken by this random variable, and it should be clear from context which is meant.) Equation (1) simply

<sup>4</sup> These behaviors might also feel wrong in that they create work for one's colleagues, but while people generally are altruistic (and so they do take this into account), they are not generally so altruistic that they would rather do work themselves than their colleagues could just as well do.

suggests that people want others to like them. We think of the desire to be liked as an intrinsic preference, but this formalism could also capture in reduced form an ongoing interaction in which being liked is instrumental for obtaining some future reward.

Given self-centered utility, which includes outcome utility and social image concerns, we now incorporate other-regarding preferences in the definition of overall utility. Following [Levine \(1998\)](#), other players' (self-centered) utilities enter linearly into an individual's overall utility:

$$v_i = u_i + \sum_{j \neq i} \beta_{ij} u_j. \tag{2}$$

Equation (2) encompasses a broad class of other-regarding preferences. We provide structure to these preferences by assuming, as a slight generalization of [Levine's \(1998\)](#) model of reciprocity, that

$$\beta_{ij} = \frac{a_{ij} + \lambda_i E_i(a_{ji})}{1 + \lambda_i}, \tag{3}$$

where the parameters  $a_{ij}$ , bounded in  $-1 < a_{ij} < 1$ , represent an index of player  $i$ 's intrinsic altruism for player  $j$  and the parameters  $\lambda_i$ , bounded in  $0 \leq \lambda_i < 1$ , represent an index of player  $i$ 's reciprocity. The values of  $a_{ij}$  and  $\lambda_i$  are private information of player  $i$ , and once again, player  $j$  knows only their prior distributions. For simplicity, we assume that these prior distributions are independent, that the distribution of  $a_{ij}$  is absolutely continuous and has full support, and that the support of the distribution of  $\lambda_i$  includes 0.

The specification given by Eq. (3) is flexible enough to capture many kinds of other-regarding preferences. A totally selfish player  $i$  has  $a_{ij} = 0$  and  $\lambda_i = 0$ . If  $a_{ij} > 0$ , then player  $i$  is inclined to act altruistically toward player  $j$ , whereas if  $a_{ij} < 0$ , then player  $i$  is inclined to be spiteful to player  $j$ . When  $\lambda_i > 0$  player  $i$  has a sense of reciprocity, an inclination to give others what they deserve—player  $i$  would be more altruistic to those who she believes would act altruistically towards her and would be more spiteful to those who she believes would act more spitefully towards her. This inclination likely goes beyond strategic reciprocity (as in, say, [Axelrod 1984](#)) and reflects the psychological fact that people generally have warmer feelings toward those who, similarly, like them. We assume that, absent any information about either player  $i$  or player  $j$ , the prior expectation of  $a_{ij}$  is  $\bar{a} \geq 0$ . That is, while individuals may idiosyncratically be spiteful, we do not expect strangers to be spiteful toward each other as a general rule ([Charness and Rabin 2002](#); [Ackermann et al. 2014](#)).

Given heterogeneity in the altruism and reciprocity parameters, along with the assumption that their values are private information, any social interaction presents a game of incomplete information.<sup>5</sup> Players may reveal information about their degrees of altruism (or spitefulness) and reciprocity through actions or through direct com-

<sup>5</sup> Player  $i$ 's self-centered utility, as defined in Eq. (1), depends on other players' types, but because the expectation is taken inside the function  $S$ , this utility is not equal to the expected posterior utility after discovering each other player's type (unless  $S$  is linear). That is, we assume that people care about their

munication. Given others' reciprocity, these signals may affect how others feel about them and in general will affect their own and others' utilities.

Players have two intrinsic motives to shape the signals they send about their own levels of altruism (or spite) and reciprocity. First, because others generally have a sense of reciprocity, the signals players send about their own levels of altruism affect how well others, in turn, like them. Given the fundamental preference to be liked, formalized in Eq. (1), along with the condition that reciprocity is always (weakly) positive, there is a universal motive to appear more nice (i.e., more altruistic or less spiteful) to be more well liked. Second, players are aware that others also care about how they are being viewed. Given the other-regarding preference, formalized in Eq. (2), players internalize to some degree this concern for how others think they are being viewed. Thus, if a player one likes some player two, player one wants to communicate high regard for player two—player two would feel comforted to know he is liked, and player one would feel good about making player two feel better. On the other hand, if player one dislikes player two, player one wants to communicate ill will—player two would feel bad about being disliked, and player one would feel good about making player two feel worse. Both motives shape behavior when a social interaction presents opportunities for signaling. The second motive specifically is necessary for cheap talk to be credible because without it everybody would pretend to be nice. This motive only becomes apparent when we consider social image concerns and other-regarding preferences in conjunction rather than in isolation.

### 3 Credible signaling

The standard *perfect Bayesian equilibrium* solution concept for a dynamic game of incomplete information permits unintuitive equilibria in which a strategy profile is supported by unreasonable beliefs off the equilibrium path. For example, generically in a cheap-talk game there will be pooling equilibria in which players can find no way to distinguish their types even though they would like to. Such pooling equilibria can be supported by a belief that no message reveals any information about a player's type, and since all types are then willing to go along with the same message(s) in equilibrium, indeed no player's type can be revealed. However, if two types would like to distinguish themselves from each other, it would seem strange for them not to try to communicate their differences, and if they do communicate differently, they will reveal themselves as distinct types. To predict such communication, we require a restriction of off-the-equilibrium-path beliefs and (in some situations) an additional assumption about the richness of the players' language.

---

Footnote 5 continued

beliefs about how much others like them (as they may never know for sure how others feel). Technically, we might consider this interaction a psychological game (Geanakoplos et al. 1989; Rabin 1993), but we do not require the machinery of psychological game theory. While psychological game theory allows players' utilities to depend on a full hierarchy of beliefs about other players' strategies, we assume that players' utilities depend only on beliefs about a state of nature. We can still use Harsanyi's (1967) method of positing a type space to model the interaction as a Bayesian game, only we use Eqs. (1) and (2) to define utility, rather than deriving an expected utility from utilities defined when known types interact.

An equilibrium should not be supported by one player's expectation that another player would draw an unreasonable conclusion from an unexpected message. Farrell (1993), Grossman and Perry (1986), and Matthews et al. (1991) offer similar proposals of rules to evaluate whether unexpected messages are credible in the process of forming reasonable beliefs off the equilibrium path. Loosely speaking, they all agree with the following principle: If a player were to observe a move off the equilibrium path, the player should search for a type or, more generally, a set of types for whom that move would be rationalized if that move did cause the player to attribute it to one of these types and to update her beliefs accordingly; if other types would not want to imitate the move, given this attribution, then the player should update beliefs in accordance with this attribution if she ever were to observe this move.<sup>6</sup> In any social interaction with social preferences represented by the utility function  $v$  defined by Eqs. (1) and (2), we should look for an equilibrium satisfying such a restriction of beliefs off the equilibrium path. Such an equilibrium is not guaranteed to exist, but if it does, it is a strong theoretical prediction, grounded on the intuition that types who want to distinguish themselves from each other should be able to do so.

Observable actions allow for signaling as well as for serving their instrumental role. Also, in many social interactions, players can communicate directly. Direct communication often takes the form of cheap talk, i.e., players' messages are not directly payoff relevant.<sup>7</sup> If players have a sufficiently rich language in a game with open-ended cheap talk (i.e., an uncountably infinite message space), then we might think it unreasonable for any type(s) to be able to use all conceivable messages in a mixed strategy, thereby precluding the possibility that anyone could make an unexpected statement. Players should be able to create new, unexpected messages to distinguish themselves when they wish to deviate from a pooling equilibrium. In a game with open-ended cheap talk, Farrell's (1993) *neologism-proof equilibrium* can be defined as a perfect Bayesian equilibrium that does not depend on noisy babbling (i.e., mixing over the entire message space) or on failure to make a reasonable attribution off the equilibrium path were an unexpected message to be sent by precisely those types who would want to send it if they could expect to be accurately recognized. That is, a neologism-proof equilibrium is not broken by an unexpected, self-signaling message.

We cannot make strong, general claims characterizing the set of neologism-proof equilibria for all social interactions allowing for open-ended cheap talk. Instead, we consider the very simplest situations to illustrate the kind of behavior that may emerge. Consider a game in which the only form of possible communication is open-ended cheap talk with a single talker (i.e., only player  $i$  can send a message to player  $j$ , and this message alone constitutes the strategy profile  $s$ ; player's  $i$ 's message space is large; and the material outcome  $x$  is independent of player  $i$ 's message, that is, independent of  $s$ ). This is a simple game indeed. Only one player has a nontrivial strategy set

<sup>6</sup> Farrell (1993) and Grossman and Perry (1986) prescribe attributions of a single unexpected message whereas Matthews et al. (1991) prescribe attributions of a set of unexpected messages. In general, multiple attributions may be possible. Farrell (1993) requires consistency with all possible reasonable attributions, whereas Grossman and Perry (1986) require consistency with some reasonable attribution.

<sup>7</sup> This definition of cheap talk does not assume that it is not credible and allows it to affect utilities if it is believed.

(i.e., standard analysis would not require game theory at all; rational choice theory would suffice), and none of the strategies available affect payoffs directly (i.e., standard analysis would conclude that the player must be indifferent between all strategies). In a world with social preferences, however, it is relevant that players who have no choices to make still observe game play and update their beliefs accordingly. With all other strategic concerns stripped away, this stark setting focuses entirely on communication inspired by social preferences.

We characterize the neologism-proof equilibria<sup>8</sup> of this game as always involving separation into two groups, which we identify as friendly and unfriendly types.<sup>9</sup> There are multiple equilibria that differ only in the particular messages used to reveal this separation, but the outcome is unique.

**Theorem 1** *A game of cheap talk with a single talker has a unique neologism-proof equilibrium outcome: there exists a critical threshold  $\beta^*$  such that friendly types (with  $\beta_{ij} > \beta^*$ ) and unfriendly types (with  $\beta_{ij} < \beta^*$ ) each use distinct partitions of the message space (and the distribution of messages within a partition is independent of the sender's type), and this critical threshold is given by*

$$\begin{aligned} & S\left(E_i\left(\frac{a_{ji} + \lambda_j E_j(a_{ij} \mid \beta_{ij} > \beta^*)}{1 + \lambda_j}\right)\right) + \beta^* S(E_j(\beta_{ij} \mid \beta_{ij} > \beta^*)) \\ & = S\left(E_i\left(\frac{a_{ji} + \lambda_j E_j(a_{ij} \mid \beta_{ij} < \beta^*)}{1 + \lambda_j}\right)\right) + \beta^* S(E_j(\beta_{ij} \mid \beta_{ij} < \beta^*)). \end{aligned} \quad (4)$$

*Proof* We sketch the proof here. Additional details can be found in the appendix. First we show that the fully pooling perfect Bayesian equilibrium is not neologism-proof because friendly types and unfriendly types want to distinguish themselves from each other. Extremely friendly types (with  $\beta_{ij} \approx 1$ ) clearly would prefer to be known to have  $\beta_{ij} > \beta$  (for any  $\beta$ ) rather than to have the unconditional prior distribution because player  $j$ 's updated belief increases his self-centered utility  $u_j$ , which in turn increases player  $i$ 's overall utility  $v_i$ , and also player  $j$  would revise his belief about  $a_{ij}$  upward, making player  $j$  more friendly due to reciprocity, and this too improves player  $i$ 's utility  $v_i$  (via  $u_i$ ). On the other hand, unfriendly types face a tradeoff whereby the first effect of appearing friendly is to decrease utility  $v_i$  (because for this type  $v_i$  moves inversely with  $u_j$ ), but the second effect is still in the positive direction. For an extremely unfriendly type (with  $\beta_{ij} \approx -1$ ), the first effect dominates because this type cares almost as much about the other player's self-centered utility as her own, and player  $j$ 's belief about  $\beta_{ij}$  is revised farther than player  $i$ 's belief about  $\beta_{ji}$ , which is tempered by player  $j$ 's reciprocity  $\lambda_j < 1$ .

<sup>8</sup> Our result would be the same if we considered any of Matthews et al.'s (1991) announcement-proof equilibrium refinements in place of the neologism-proof refinement. We have focused on the neologism-proof equilibrium concept for simplicity of presentation, not to advocate for one solution concept over another. All of these refinements share similar intuition.

<sup>9</sup> Some moderately spiteful individuals may pool with the friendly types.

The preferences of extremely friendly and extremely unfriendly types to distinguish themselves from each other support a partially separating neologism-proof equilibrium. By the intermediate value theorem, there is a critical value of  $\beta$  such that it is precisely those types with  $\beta_{ij} > \beta$  that prefer to be known to have this property. Moreover, because of the concavity of  $S$ , extremely unfriendly types prefer instead to appear to be unfriendly with  $\beta_{ij} < \beta$ . Thus, there is an equilibrium in which both friendly types and unfriendly types are maximizing utility by distinguishing themselves from each other, and the critical value of  $\beta$  that separates these two groups of types is the type that would be indifferent between appearing to be a part of either group.

To complete the proof, we show that no additional separation of types is possible, because within each group, all types have the same incentives. If a subgroup tried to distinguish themselves with a particular message, the rest of the group would have the same incentive to use that message as well. Consider first the friendly types. They all want to appear to be as altruistic as possible. (The subset with  $\beta_{ij}$  satisfying  $\beta^* < \beta_{ij} < 0$  would ideally want to appear simultaneously altruistic (with high  $\alpha_{ij}$ ) and unfriendly (with low  $\beta_{ij}$ ), but this image is impossible to create, and for these types, the motive to appear altruistic (to curry favor) is stronger than the (sadistic) motive to appear unfriendly, so they choose to be friendly.) In order to appear as altruistic as possible, they all share the preference to appear to have low reciprocity as well. The idea is that it is not possible to disentangle a signal about generally friendly regard ( $\beta_{ij}$ ) from intrinsic altruism ( $a_{ij}$ ) or reciprocity ( $\lambda_i$ ), but everybody would prefer their friendliness to be attributed to altruism (so it would be reciprocated), so there is a common incentive to play down reciprocity. Finally, now consider the unfriendly types. They all want to appear to be as spiteful as possible. (They would ideally want to appear altruistic while unfriendly, but this is impossible, and for  $\beta_{ij} < \beta^*$ , the sadistic motive dominates, and the only way to become unfriendly is to be spiteful.) In order to appear as unfriendly as possible, they too prefer to appear to have low reciprocity, simply because higher degrees of reciprocity would make these types friendlier.  $\square$

Theorem 1 identifies a unique equilibrium outcome in which the kind of message a player chooses reveals the player as either friendly or unfriendly, but the particular expression of that message conveys no additional information for more finely distinguishing the player's type. The idea is that friendly types want to be known to be friendly and unfriendly types want to be known as unfriendly. (The assumption that the prior distribution of types has full support is crucial here. If everyone were friendly, then cheap talk would not be meaningful. The friendliest types would be unable to reveal themselves with cheap talk because everyone would imitate them.) Of course, there is trivially a perfect Bayesian equilibrium with full pooling and no communication. If all types pool together, Bayes' rule ascribes messages no meaning, and if messages have no meaning in a game of cheap talk, then there is no incentive to choose any one message rather than another. But, friendly and unfriendly types can attempt to distinguish themselves without any potential adverse consequences, and if they do, their different messages will become meaningful. Thus, only the semi-separating equilibrium, but not the fully pooling equilibrium, is neologism-proof.

We observe that the critical threshold given by Eq. (4) is generally negative, i.e.,  $\beta^* \leq 0$ . Selfish and even some mildly spiteful types pool with the friendly types to ingratiate themselves and protect their image. Only sufficiently hostile types are unfriendly. In the special case of a world without reciprocity (i.e., common knowledge that  $\lambda_j = 0$ ), Eq. (4) admits a simple solution: the critical threshold is then  $\beta^* = 0$ .<sup>10</sup> Without reciprocity the motive to appear friendly to be more well liked vanishes. The remaining motive is simply to confer good news on those one likes and bad news on those one dislikes, and revealing oneself as friendly or unfriendly does exactly that. In this equilibrium all of the spiteful types are happy to distinguish themselves from the altruistic types (and vice versa).

## 4 Discussion

### 4.1 Expressions of altruism and spite

Theorem 1 illustrates that, as Crawford and Sobel (1982) first recognized and Farrell and Rabin (1996) showed intuitively, communication is possible with only cheap talk, even without the availability of costly signals (that Spence 1974, for example, relies on in his analysis of signaling). Friendly and unfriendly types can credibly distinguish themselves with cheap talk. Of course, opportunities for costly signaling allow more extensive communication to take place. Any time one player (say,  $i$ ) can take an action that sacrifices her own material payoff  $x_i$  to improve another player's payoff  $x_j$ , this action could serve as a costly signal of player  $i$ 's friendly regard for player  $j$  (i.e., high  $\beta_{ij}$ ). Although the material cost of the action is fixed, the effective cost depends on player  $i$ 's regard for player  $j$ ,  $\beta_{ij}$ . If the sacrifice is minimal, a player with warm feelings for the other may find it in her interest to take the friendly gesture, even before considering its value as a signaling device. On the other hand, a player with less goodwill for the other (but still friendly, according to the separation of types produced by cheap talk) might be reluctant to actually sacrifice for the other's wellbeing. Depending on the precise cost associated with the sacrifice, such an individual might reluctantly go along with it to avoid signaling a relative lack of altruism or might find the cost of pretending to be so highly altruistic too much to bear. If the sacrifice is sufficiently costly, we might expect to observe separation among friendly types, with a willingness to go out of one's way for someone else indicating a warmer regard for that individual.

Warm-glow altruism emerges in our model from the interplay of pure altruism and social image concerns. Warm glow refers to caring about the act of improving another person's outcome in addition to caring about what the outcome is. The hallmark of warm-glow motivation is incomplete crowding out of aid offered to a recipient when others step in to help the recipient as well. In our model, people care about being recognized for helping another person over and above caring about the person's material wellbeing because the act communicates positive regard for the person. Signaling altruism inspires positive regard in return (due to reciprocity), thus earning social approval, and it also assures the other person as an expression of social approval. In contrast to generosity resulting from pure altruism, in some cases we might actually

<sup>10</sup> In fact, prior knowledge that  $\lambda_j = 0$  is both necessary and sufficient for  $\beta^* = 0$ .

expect greater assistance to be provided to a recipient when others are also doing more for him, i.e., crowding in rather than crowding out (see [Heutel 2014](#)). When others are setting a higher standard, one needs to be more generous to send a positive signal about one's own level of altruism.

Just as generosity can signal altruism, opportunities to punish others can serve as potential signals of a player's ill will. [Fehr and Gächter \(2000a\)](#) observe prevalent voluntary punishment of free-riders in a public goods game even when punishment is costly and cannot provide any material benefits (say, in future interactions). Of course, reciprocity (or spitefulness) alone could be a sufficient motive to punish such a free-rider. Yet people also use cheap talk expressing disapproval as a form of punishment for selfish behavior ([Masclét et al. 2003](#)), suggesting that monetary punishment serves an important signaling function as well. If punishment is sufficiently costly (to the punisher), willingness to punish can indicate strong hostility as part of an equilibrium that involves additional separation among unfriendly types. Even when reward or punishment can be given for free, it, or its absence, can serve as a "costly signal", because with other-regarding preferences, its impact on another player's outcome utility can be seen as a cost or a benefit, depending on one's regard for this other player.

Oddly enough, [Herrmann et al. \(2008\)](#) observe in many societies perverse, antisocial punishment of very generous contributors in a public goods game (see also [Falk et al. 2005](#)). As described above, excessive generosity, when publicly observable, is a form of costly signaling of one's type as particularly altruistic. But this separation of particularly altruistic types imposes a negative externality on other less altruistic types. These others may still want to appear as altruistic as possible, but may not want to pay the cost to send the same signal as the highly altruistic types (who can naturally bear that cost more easily). They are thus revealed to be less altruistic. While the motives behind antisocial punishment are not entirely clear, it could possibly be used to enforce an equilibrium with more pooling of altruistic types.

## 4.2 Implications for economic phenomena

The broader implication here is that economic phenomena often attributed to people's preferences over material outcomes might better be understood as arising from social preferences, and specifically, as manifestations of an incentive to communicate about one's social preferences. Perhaps the most obvious demonstration that social preferences matter involves the dictator game. Giving in the dictator game is often attributed either to altruism (or other-regarding preferences more generally; see, e.g., [Andreoni and Miller 2002](#)) or social image concerns (see, e.g., [Dana et al. 2006](#) or [List 2007](#)). If both of these motives are present, then a third motive emerges as well—the motive to assure the recipient that he is indeed well regarded. Much like [Andreoni and Bernheim \(2009\)](#) and [Grossman \(2015\)](#), we think that dictator game giving is a form of communication about one's social preferences, but the incentive to communicate may include altruistic concerns as well as selfish ones. In this context, cheap talk may, to some extent, be naturally credible. Thus, explicit communication through verbal messages and implicit communication through actions may be substitutes in a dictator game. That would explain the puzzling finding that dictators give less when they can send explicit written messages to recipients ([Andreoni and Rao 2011](#)). Similarly, unfair offers in the

ultimatum game tend to produce negative feelings, and there is a natural incentive for recipients of unfair offers to communicate these negative feelings to proposers. Here, as well, explicit communication appears to be a substitute for costly punishment, as unfair offers are less frequently rejected when responders can also send written messages that let proposers know what they really think of them (Xiao and Houser 2005).

Social preferences also matter in public goods games. Laboratory experiments consistently find substantial contributions in public goods games (Dawes and Thaler 1988).<sup>11</sup> Voluntary contributions have been attributed to warm glow altruism (e.g., Palfrey and Prisbrey 1997), reciprocity (e.g., Fischbacher et al. 2001), or social image concerns (e.g., Lacetera and Macis 2010; Filiz-Ozbay and Ozbay 2013). Our model of social preferences suggests that in addition to these motives, there is also the motive to give to avoid hearing others' disapproval, which can be credibly communicated as cheap talk. (See Masclet et al. 2003 for empirical evidence supporting this hypothesis.) This is an extension of social image concerns. Social image concerns imply that an individual cares what others think of him, but additionally, we believe, an individual cares about hearing these views expressed as cheap talk. This suggests that individuals may voluntarily contribute to the provision of public goods (in field settings) especially in societies in which people care a lot about how they are viewed in their community and have ample opportunities to communicate social approval or disapproval (i.e., in close-knit communities or those with a strong sense of shared cultural identity).

Similarly, in the workplace, workers may cooperate to collectively exert low effort, even when monetary incentives are put in place to encourage relatively higher effort. When relative performance incentive schemes are used to try to overcome the agency problem, a minimum-effort coordination game arises in which no worker wants to fall behind the others in terms of productivity, and there is monetary incentive to have higher productivity than others, but doing so will often generate disapproval from co-workers. A well-designed field study by Bandiera et al. (2005) finds that relative performance incentives may fail because workers cooperate to shirk specifically when they can monitor each other's performance, presumably because expressions of social disapproval are such a strong motivating factor.

Clearly, the implications of this model of social preferences extend to settings of direct economic relevance. A case in point is the phenomena of gift exchange. Gift exchange can be supported by reciprocity, but its foundation becomes more robust when we recognize that people have preferences about other people's beliefs about their preferences rather than only having preferences about final allocations. Thus, as Camerer (1988) argues, gift giving is a form of signaling (see also Ruffle 1999; Prendergast and Stole 2001), and we suggest that specifically what givers are signaling is their regard for the recipients, and this is something they intrinsically care about. Consider, for example, holiday giving. If preferences were exclusively about final allocations (for both giver and receiver, of course), then the most efficient gift

---

<sup>11</sup> Pre-play communication in public goods game actually increases generosity (cf. Sally 1995), in contrast to the dictator game, where it serves as a substitute. An important factor may be that players in the public goods game can both send and receive messages before making their contribution. Andreoni and Rao (2011)'s study of the dictator game finds that sending a message serves as a substitute for giving, but receiving a message actually increases giving. Our model does not explain why receiving a message might systematically stimulate generosity.

often would be cash, which could be used for whatever the recipient desires. However, recognizing gifts as signals of social preferences, the effort to find a uniquely appropriate gift can be understood as conveying information about the giver's regard for the recipient. Similarly, consider the signaling role of tipping service workers. Many people like to use large tips not just to reward good service, but also to let the worker know that good service was appreciated. Even after bad service, many people will leave standard tips to avoid signaling any spitefulness. However, when the quality of service is extremely low, some individuals (with a strong sense of reciprocity) use the tip to signal their ill will. The clearest way to do this is not to leave no tip at all, which could be interpreted as simply forgetting to tip, but rather to leave just a single penny.

Of course, gift exchange has broader economic importance beyond tipping, holiday giving and even charitable giving (Falk 2007). Many labor contracts implicitly involve gift exchange, i.e., paying a living wage in exchange for working hard (Akerlof 1982; Fehr et al. 1998). Gift exchange may be ever so prevalent in labor markets because the incentive to participate arises not only from shared efficiency gains (in terms of material outcomes) but also from the social preferences of the employer and employee (see also Charness and Haruvy 2002). Heterogeneous organizational cultures can emerge as workers self-select into organizations (Kosfeld and von Siemens 2011), and the lines of communication within an organization as well as the norms of etiquette setting expectations there will generally affect whether gift exchange—and cooperation more broadly—can take root.

### 4.3 Conclusion

We have posited that people generally have other-regarding preferences and social image concerns, and we have shown that the interaction of these social preferences creates an incentive to communicate about one's social preferences. Altruistic individuals generally want their goodwill to be recognized, just as (sufficiently) spiteful individuals want to make their feelings clear. We have discussed a few examples of economic phenomena that should be understood in a context with incentives for signaling social preferences. Cooperation and gift exchange, for example, are more robust in this context. Many more examples surely exist. We hope future work acknowledges the role of social preferences, and naturally credible communication about social preferences, in social and economic behavior.

#### Compliance with ethical standards

**Conflict of interest** The author declares that he has no conflict of interest.

## Appendix

### Additional details supporting the proof of Theorem 1

In the pooling equilibrium (which we show is not neologism-proof), player  $i$ 's utility is  $v_i^{\text{pooling}} = S(\bar{\beta}) + \beta_{ij}S(\bar{\beta})$ , where  $\bar{\beta}$  is the prior expectation of  $\beta_{ij}$  and we have nor-

malized the outcome utilities to be 0. Consider an unexpected message (a neologism) that conveys that  $\beta_{ij} > \beta$  for some threshold  $\beta$ . Player  $i$ 's utility from defecting from the pooling equilibrium and sending this message, if it is accepted as credible, is

$$v_i^{\text{neologism}} = S \left( E_i \left( \frac{a_{ji} + \lambda_j E_j(a_{ij} | \beta_{ij} > \beta)}{1 + \lambda_j} \right) \right) + \beta_{ij} S(E_j(\beta_{ij} | \beta_{ij} > \beta)).$$

Clearly, if  $\beta_{ij} \approx 1$ , then  $v_i^{\text{neologism}} > v_i^{\text{pooling}}$  (because both terms are larger). On the other hand, if  $\beta_{ij} \approx -1$ , then  $v_i^{\text{neologism}} < v_i^{\text{pooling}}$  because

$$S \left( E_i \left( \frac{a_{ji} + \lambda_j E_j(a_{ij} | \beta_{ij} > \beta)}{1 + \lambda_j} \right) \right) < S(E_j(\beta_{ij} | \beta_{ij} > \beta)).$$

By the intermediate value theorem, there must exist some  $\beta$  such that  $v_i^{\text{neologism}} > v_i^{\text{pooling}}$  if and only if  $\beta_{ij} > \beta$ . For this value of  $\beta$ , the neologism is indeed credible, so the pooling equilibrium is not neologism-proof.

We obtain the (partially) separating equilibrium similarly by comparing the utilities of appearing friendly or unfriendly relative to the threshold  $\beta^*$ . Player  $i$ 's utility from appearing friendly is

$$v_i^{\text{friendly}} = S \left( E_i \left( \frac{a_{ji} + \lambda_j E_j(a_{ij} | \beta_{ij} > \beta^*)}{1 + \lambda_j} \right) \right) + \beta_{ij} S(E_j(\beta_{ij} | \beta_{ij} > \beta^*)).$$

Player  $i$ 's utility from appearing unfriendly is

$$v_i^{\text{unfriendly}} = S \left( E_i \left( \frac{a_{ji} + \lambda_j E_j(a_{ij} | \beta_{ij} < \beta^*)}{1 + \lambda_j} \right) \right) + \beta_{ij} S(E_j(\beta_{ij} | \beta_{ij} < \beta^*)).$$

To be an equilibrium, we require  $v_i^{\text{friendly}} > v_i^{\text{unfriendly}}$  if and only if  $\beta_{ij} > \beta^*$ . Seeing that  $v_i^{\text{friendly}} - v_i^{\text{unfriendly}}$  is increasing in  $\beta_{ij}$ , an equilibrium is given by Eq. (4), which implies that  $v_i^{\text{friendly}} = v_i^{\text{unfriendly}}$  when  $\beta_{ij} = \beta^*$ . Once again, the intermediate value theorem guarantees that this equation has a solution.

## References

- Ackermann, K., Fleiß, J., Murphy, R. (2014). Reciprocity as an individual difference. *Journal of Conflict Resolution*. doi:10.1177/0022002714541854.
- Akerlof, G. (1982). Labor contracts as partial gift exchange. *Quarterly Journal of Economics*, 97(4), 543–569.
- Akerlof, G., & Dickens, W. (1982). The economic consequences of cognitive dissonance. *American Economic Review*, 72(3), 307–319.
- Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97(6), 1447–1458.
- Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100, 464–477.

- Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5), 1607–1636.
- Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2), 737–753.
- Andreoni, J., & Rao, J. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of Public Economics*, 95, 513–520.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Bandiera, O., Barankay, I., Rasul, I. (2005). Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics*, 120(3), 917–962.
- Becker, G. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of Economic Literature*, 14(3), 817–826.
- Benabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678.
- Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, 102(5), 841–877.
- Bowles, S., & Gintis, H. (2005). Prosocial emotions. In L. Blume & S. Durlauf (Eds.), *The economy as an evolving complex system, III* (pp. 339–366). Oxford: Oxford University Press.
- Camerer, C. (1988). Gifts as economic signals and social symbols. *American Journal of Sociology*, 94–S, S180–S214.
- Camerer, C., & Thaler, R. (1995). Anomalies: Ultimatums, dictators, and manners. *Journal of Economic Perspectives*, 9(2), 209–219.
- Charness, G., & Haruvy, E. (2002). Altruism, equity, and reciprocity in a gift-exchange experiment: An encompassing approach. *Games and Economic Behavior*, 40, 203–231.
- Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117, 817–869.
- Crawford, V., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50, 1431–1451.
- Dana, J., Cain, D., & Dawes, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100, 193–201.
- Dawes, R., & Thaler, R. (1988). Anomalies: Cooperation. *Journal of Economic Perspectives*, 2(3), 187–197.
- DellaVigna, S., List, J., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *Quarterly Journal of Economics*, 127(1), 1–56.
- Falk, A. (2007). Gift exchange in the field. *Econometrica*, 75(5), 1501–1511.
- Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, 73(6), 2017–2030.
- Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54, 293–315.
- Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5, 514–531.
- Farrell, J., & Rabin, M. (1996). Cheap talk. *The Journal of Economic Perspectives*, 10(3), 103–118.
- Fehr, E., & Fischbacher, U. (2002). Why social preferences matter—the impact of non-selfish motives on competition, cooperation and incentives. *Economic Journal*, 112, C1–C33.
- Fehr, E., & Gächter, S. (2000a). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980–994.
- Fehr, E., & Gächter, S. (2000b). Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.
- Fehr, E., Kirchler, E., Weichbold, A., & Gächter, S. (1998). When social norms overpower competition: gift exchange in experimental labor markets. *Journal of Labor Economics*, 16(2), 324–351.
- Filiz-Ozbay, E., Ozbay, E.Y. (2013). Effect of an audience in public goods provision. *Experimental Economics*, forthcoming.
- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3), 397–404.
- Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1, 60–79.
- Golman, R., Loewenstein, G. (2015). An information-gap framework for capturing preferences about uncertainty. In *Proceedings of the fifteenth conference on theoretical aspects of rationality and knowledge*, 141–151.
- Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization*, 117, 26–39.
- Grossman, S., & Perry, M. (1986). Perfect sequential equilibrium. *Journal of Economic Theory*, 39, 97–119.

- Harsanyi, J. (1967). Games with incomplete information played by Bayesian players. Part I: The basic model. *Management Science*, 14, 159–182.
- Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319, 1362–1367.
- Heutel, G. (2014). Crowding out and crowding in of private donations and government grants. *Public Finance Review*, 42(2), 143–175.
- Holländer, H. (1990). A social exchange approach to voluntary cooperation. *American Economic Review*, 80(5), 1157–1167.
- Kahneman, D., Knetsch, J., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *American Economic Review*, 76(4), 728–741.
- Kosfeld, M., & von Siemens, F. A. (2011). Competition, cooperation, and corporate culture. *RAND Journal of Economics*, 42, 23–43.
- Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4), 673–707.
- Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76, 225–237.
- Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1, 593–622.
- List, J. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3), 482–493.
- Maschlet, D., Noussair, C., Tucker, S., & Villeval, M. C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93(1), 366–380.
- Matthews, S., Okuno-Fujiwara, M., & Postlewaite, A. (1991). Refining cheap-talk equilibria. *Journal of Economic Theory*, 55, 247–273.
- Organ, D. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington: Lexington Books.
- Organ, D., Podsakoff, P., & MacKenzie, S. (2006). *Organizational citizenship behavior: Its nature, antecedents, and consequences*. London: Sage Publications.
- Palfrey, T., & Prisbrey, J. (1997). Anomalous behavior in public goods experiments: How much and why? *American Economic Review*, 87(5), 829–846.
- Prendergast, C., & Stole, L. (2001). The non-monetary nature of gifts. *European Economic Review*, 45, 1793–1810.
- Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83, 1281–1302.
- Ruffle, B. (1999). Gift giving with emotions. *Journal of Economic Behavior and Organization*, 39, 399–420.
- Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7, 58–92.
- Sliwka, D. (2007). Trust as a signal of a social norm and the hidden costs of incentive schemes. *American Economic Review*, 97(3), 999–1012.
- Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43, 392–436.
- Spence, A. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Cambridge: Harvard University Press.
- Tadelis, S. (2008). The power of shame and the rationality of trust. Working Paper Series, Center for Responsible Business, UC Berkeley.
- Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Science*, 102(20), 7398–7401.