Pride and Prejudice: The Human Side of Incentive Theory

Tore Ellingsen* Magnus Johannesson**

November 29, 2006
(Under revision)

Abstract

Many people are sensitive to social esteem, and their pride is a source of pro–social behavior. We present a game-theoretic model in which sensitivity to esteem varies across players and may depend on context as well players' beliefs about their opponents. For example, the pride associated with a generous image is greater when the player holding the image is in fact generous and believes the observers to be generous as well. The model can account both for the fact that players' behavior sometimes depends on the opponents' unchosen options and for the prevalence of small symbolic gifts. Perhaps most importantly, the model offers an explanation for motivational crowding out: Control systems and pecuniary incentives may erode morale by signaling to the agent that the principal is not worth impressing.

JEL CLASSIFICATION: D01, D23, D82, Z13
KEYWORDS: Motivational crowding out, Esteem, Incentives, Framing, Social preferences.

---

*Address: Department of Economics, Stockholm School of Economics, Box 6501, S—113 83 Stockholm, Sweden. Email: gte@hhs.se.

**Address: Department of Economics, Stockholm School of Economics, Box 6501, S—113 83 Stockholm, Sweden. Email: hemj@hhs.se.

We are grateful to the Torsten and Ragnar Söderberg Foundation (Ellingsen) and the Swedish Research Council (Johannesson) for financial support. Thanks to George Baker, Kjell-Arne Brekke, Florian Englmaier, Ernst Fehr, Martin Flodén, Oliver Hart, Bengt Holmström, John Moore, Anna Sjögren, Jean-Robert Tyran, Robert Östling, and especially Michael Kosfeld for helpful discussions. The paper has also benefited from comments by many other seminar and conference participants. Errors are ours.

1 Introduction

> Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard. She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive.
>
> Adam Smith (1790, Part III, Section I, Paragraph 13)

Few controversies in the social sciences are more heated than the debate over incentive theory. For example, in the field of organizational behavior, microeconomic incentive theory is frequently regarded as outright dangerous. McGregor's (1960) celebrated management book *The Human Side of Enterprise* argued that managers who subscribe to the conventional view that employees dislike work – McGregor labelled it Theory X – may create workers who are "resistant, antagonistic, uncooperative" (page 38). That is, managerial control and material incentives may trigger the very behaviors that they are designed to avert. Conversely, managers who subscribe to the more optimistic view that employees see their work as a source of self-realization and social esteem – Theory Y – may create workers who voluntarily seek to fulfill the organization's goals.[1]

Any theory of incentives must be based on assumptions about human nature, and the theorist must balance the desire for realism against the desire for parsimony. Over the last decade, microeconomists working on incentive theory have become increasingly inclined to discard the common simplification – the cornerstone of Theory X – that people are guided solely by material self–interest.[2] To a considerable extent, the change in attitude is due to empirical studies that document the prediction failures of Theory X. Two prominent observations are the *wage level puzzle* that higher wages sometimes induce better performance (Fehr, Kirchsteiger and Riedl, 1993, Bewley, 1999), and the *incentive intensity puzzle* that stronger monetary incentives sometimes induce worse performance (Frey and Oberholzer-Gee, 1997, Gneezy and Rusticini,

[1]For an updated version of the arguments against Theory X, see Chapter 4 of Pfeffer (1994), aptly entitled "Wrong Heroes, Wrong Theories, and Wrong Language."

[2]Kreps (1997) and Baron and Kreps (1999) are watershed contributions. Others are mentioned below. In a recent survey that seems representative of the current mainstream view, Sobel (2005, page 432) concludes that the assumption of narrow selfishness should not be taken for granted: "A philosophical refusal to consider extended preferences leads to awkward explanations of some phenomena. It limits the questions that can be asked and restricts the answers. It is a handicap."

2000a, 2000b, Bohnet, Frey and Huck, 2001, Fehr and Gächter, 2002, Fehr and Rockenbach, 2003, and Fehr and List, 2004). Both observations violate the standard principal–agent model, which predicts that the agent's effort should be unaffected by the level of pay, and that stronger incentives should always entail higher effort.[3] The incentive intensity puzzle aptly illustrates McGregor's critique: A principal who believes that agents are opportunistic has reason to utilize relatively strong material incentives, and when these incentives create more opportunistic behavior, the principal's belief is self-fulfilling.

While theories of fairness and reciprocity can account for the wage level puzzle, they generally fail to explain the incentive intensity puzzle. Thus, theorists have recently turned to other explanations, closer to the territory of McGregor's Theory Y. In two recent papers, Bénabou and Tirole have argued that private information could be the culprit. The first paper, Bénabou and Tirole (2003), focuses on self-realization and shows that material incentives might backfire if offered by a principal who is more knowledgeable than the agent, because the agent may interpret the incentive as bad news about his talent or about the difficulty of the task. Bénabou and Tirole (2006) instead focuses on social esteem and shows that material incentives might likewise backfire when the agent has private information about multiple personal characteristics, such as materialism and altruism (see Seabright, 2004, for a related argument). For example, an altruistic agent may donate less blood after the introduction of an incentive, because the incentive will be attractive to materialistic types and hence dilute the signaling value of blood donation.[4] Although Bénabou and Tirole's two models have considerable explanatory power, they both fail to explain a striking regularity uncovered in recent experimental studies by Fehr and Rockenbach (2003), Fehr and List (2004), and Falk and Kosfeld (2005), namely that material incentives have a negative effect on agents' performance even when the principal lacks private information about the agent's characteristics, but only when (the agent knows that) the principal can choose whether or not to impose the incentive. Thus, the question is: Why does the principal's choice set matter even though the principal lacks private information about the agent?

In this paper, we propose a model that can explain why the principal's choice set matters even when the principal lacks information about the agent's type. According to our model, the principal's distrust has a negative effect on

[3]We here abstract from the effect of wealth changes on labor supply. In most of the empirical studies, wealth effects can safely be assumed to be negligible due to low stakes and short horizons.

[4]Titmuss (1970) famously originated the idea that material incentives crowd out voluntary blood donation. For a supportive field experiment, see Mellström and Johannesson (2006).

agents' effort because low expectations are demoralizing. The argument builds on two key assumptions. The first assumption is that people care about social esteem and therefore want to signal favorable traits; in this respect, our model is closely related to Bénabou and Tirole (2006).[5] The second assumption is that the value of esteem depends on the audience. We want to be thought well of by all people, but it also matters *who* thinks well of us. Roughly put, the agent wants the principal's respect, but more so if the principal is respectable.[6]

We think that both our assumptions are uncontroversial. While most people certainly do seek material rewards, immaterial rewards like esteem or social status matter too. Thus, a person may work hard not only to earn a larger material reward and to contribute to society, but also in order to make a favorable impression. Chester Barnard (1938, page 145) put it succinctly: "The opportunities for distinction, prestige, personal power, and the attainment of dominating positions are much more important than material rewards in the development of all sorts of organizations, including commercial organizations." Psychologists from Maslow (1943) to Baumeister and Leary (1995) agree that esteem is a fundamental source of motivation, as did the classical thinkers, from the Greeks to Adam Smith; see Brennan and Pettit (2004, Chapter 1). In fact, it seems to us that almost everyone realizes that making a good impression is rewarded through the respect, attention, and tribute paid by principals, peers, or other observers. Likewise, it is widely agreed that the value of respect depends on its source. As David Hume (1739, Book II, Part I, Sect. XI) expresses it in his account of humans' fundamental love of fame: "tho fame in general be agreeable, yet we receive a much greater satisfaction from the approbation of those, whom we ourselves esteem and approve of, than those, whom we hate and despise." Hume's student, Adam Smith (1790, Part II, Section III, Paragraph 10), articulates the same idea: "What most of all charms us in our benefactor, is the concord between his sentiments and our own, with regard to what interests us so nearly as the worth of our own character, and the esteem that is due to us."

At a general level, our model offers a new approach to the modelling of reciprocity. In the reciprocity literature, a major question has been to explain the important role that people appearently ascribe to others' intentions. Two

[5]Although human concern for social esteem has been well understood for a long time, the first satisfactory formal model of signalling for esteem purposes is probably due to Bernheim (1994).

[6]This is not to say that ruthless principals can exploit their agents by pretending to be respectable: In equilibrium, respectable principals can only convey their respectability by engaging in costly signalling. That is, in order to credibly signal their faith in Theory Y the principal must be either more optimistic or more concerned with social esteem than are the principals who subscribe to Theory X.

previous explanations have been formalized. Levine (1998) suggests that people's altruism or spite depend on their beliefs about their opponents, and that this may explain reciprocity. For example, if a player is more altruistic toward other altruists, a generous action by an opponent may be rewarded, and a less generous action may be punished - because it is a signal that altruism is low. If the situation is the same, except the opponent does not have the opportunity to be generous, lack of altruism can no longer be inferred, and the same action may go unpunished. Our model is closely related to Levine's inasmuch as we too focus on signaling and the impact of the opponents' type on a players' utility. However, in Levine's model players do not mind what opponents think about them, as long as it does not affect their behavior. One argument in favor of our approach over Levine's is that we can potentially explain why behavior in experiments like the Dictator game does not always end up at corners of the feasible set when stakes are small.[7] According to Levine's model, either altruism, spite, or selfishness ought to be the dominant motive in this case - so subjects ought to give all or nothing. (If fairminded subjects are allowed, there might also be a spike at the equal split.) When people seek social esteem, behavior may be interior as players engage in exactly the amount of altruism or spitefulness that is necessary to signal their type to others. Giving, say, twenty percent of one's endowment in the Dictator game then makes sense as a strategy to stand out from the group of selfish and spiteful subjects.

A second way to accommodate intention–based reciprocity is to allow agents to have preferences defined directly over others' actions, not only over outcomes. This approach, which builds on concepts introduced by Geanakopolos, Pearce, and Stacchetti (1989), has been pioneered by Rabin (1993), Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2005), and Falk and Fischbacher (2006). While this model is fascinating and potentially useful for understanding reciprocity, it is rather complicated and it takes a longer step away from conventional game theoretic models than we do. More importantly, it fails to explain the incentive intensity puzzle identified by Falk and Kosfeld (2005).

Before presenting our model, let us therefore fix ideas by considering Falk and Kosfeld's experimental evidence on the hidden costs of control. Their setting is maximally simple. An agent has an endowment of 120 money units and can make transfers to the principal. For every unit that the agent gives up, the principal receives two units. Hence, the principal can receive an amount of at least 0 and at most 240. Before the agent decides how much to transfer voluntarily, the principal has the opportunity to impose a compulsory transfer of

[7]Rotemberg (2006) discusses some other shortcomings of Levine's model and proposes a solution to them.

10 (receiving 20). Note that all conventional theories, and even the intention–based reciprocity model of Falk and Fischbacher (2006), predict that the principal should control the agent. The reason is that only a relatively selfish agent would ever give less than 10, and if the agent is selfish trusting her makes no sense. In stark contrast, Falk and Kosfeld find that the majority of principals trust their agent, abstaining from the compulsory transfer, and also that such trust on average pays significantly better than distrust. One achievement of our model is to rationalize Falk and Kosfeld's findings.[8]

2 A model of pride and prejudice

There are already many models in which players are assumed to care about the beliefs of others, i.e., to be proud. Our model will add the seemingly minor twist that not everyone is equally proud, and that not every opponent's belief is equally important.

In addition to differential pride, we will admit heterogeneous beliefs. This assumption, which generates a role for prejudice, is less common. Although most of our qualitative results can be derived with heterogeneous pride alone, we argue that heterogeneous beliefs could be quantitatively important for understanding much of the experimental evidence.

2.1 Players and actions

There are two players.[9] We restrict attentions to games in which each player moves at most once. The set of (pure) actions for player i is denoted $\mathcal{A}_i$ and a generic action is denoted a_i. Mixed actions are probability distributions over $\mathcal{A}_i$. Let $\mathcal{X}$ denote the set of mixed actions.

[8]Sliwka (2003) and Ellingsen and Johannesson (2005) have explored the closely related idea that incentives may backfire by conveying pessimistic expectations, thereby reducing agents' feeling of guilt in case of shirking. (The first paper uses psychological game theory; the second, like us, uses a conventional signaling model.) In other words, material incentives backfire because pessimistic expectations legitimize opportunism; a similar mechanism is hinted at in Charness and Dufwenberg, 2005. If this explanation were correct, the less opportunistic agents would feel relieved whenever the principal imposed non-binding constraints on them. Falk and Kosfeld (2005) document that agents are frustrated – not relieved – when subjected to control, whether it is binding or not.

[9]The analysis is straightforwardly extended to the case of more players, but at some notational cost.

2.2 Characteristics

Players are heterogeneous. A player's type is given by a vector of characteristics, or personality traits, $\theta_i \in \Theta = (\Theta^1, ..., \Theta^n) \subset \mathbb{R}^n_+$.

Each agent's type is drawn independently from the same joint distribution $f_t(\theta)$, with the associated joint cumulative distribution function $F_t(\theta)$. (At the cost of cluttering the notation, we could instead let the two agents' types be drawn from separate distributions.)

2.3 Beliefs

In order to accommodate the heterogeneity in beliefs that is observed in expriments, we assume that players beliefs about the opponent's type are correlated with their own types. Instead of just assuming the correlation, we provide primitive assumptions that generate it.[10] To this end, we assume that the true distribution is not known to the agents; they only know that the distribution is drawn from a set of joint cumulative distribution functions $\mathcal{F}$.

The set $\mathcal{F}$ is common knowledge between the players. We also assume that the distribution from which F_t is drawn is common knowledge. However, since players know their own types, and these types may differ, Bayesian updating implies that players may hold different beliefs about each other. We say that the players have a common *metaprior*, but that they have different – and privately known – *priors*.

Let $p_i^0 = p(F|\theta_i)$ denote player i's prior on the joint distribution (the probability density over the set of possible cumulative distributions), and let p_i^{0k} denote the associated prior about the marginal distribution of trait k. In other words, p_i^0 is player i's initial beliefs about player j. As the game progresses, players will update their beliefs. Let h_i denote the history of actions observed by player i when it is i's turn to move, and let $p_i(\theta_j|h_i, \theta_i)$ denote i's conditional belief about j's type. Finally, let $\hat{p}_j$ denote i's belief about (the function) p_j.

If the game is sequential, we assume that players update their beliefs using Bayes' rule following any action by the opponent.

[10] It is well known that people tend to think that others are like them. Psychologists initially concluded that people therefore systematically *overestimate* the degree of similarity – creating a "false–consensus" effect; see Ross, Green and House (1977). However, as noted by Dawes (1989), it is rational to use information about one's own inclinations to infer the likely inclinations of others. Our model will capture the rational consensus effect alluded to by Dawes. (We take no stand on the issue of whether empirically observed consensus effects are primarily rational or not.)

2.4 Strategies and solution concepts

The model centers around the effect that an agent's action a_i has on the opponent's posterior belief, and especially how actions are affected by the strategic motive to influence these beliefs. The solution concept is perfect Bayesian equilibrium (PBE) and refinements thereof. Since we confine attention to games in which each player moves at most once, the set of possible histories for player i when moving is $\mathcal{H}_i = \{\mathcal{A}_j \cup \emptyset\}$.

A strategy for player i is a mapping $\sigma_i : \Theta \times \mathcal{F} \times \mathcal{H}_i \to \mathcal{X}$. In words, player i's (mixed) action depends on the own type, the belief about the opponent's type, and any observed prior actions by the opponent. We seek pairs of strategies (σ_1^*, σ_2^*) and beliefs (p_1^*, p_2^*) such that σ_i^* is a best response to σ_j^* given the beliefs p_i^* and we require that p_1^* and p_2^* satisfy Bayes' rule whenever it applies. In addition, we insist that the beliefs are "reasonable" off the equilibrium path in the sense that they satisfy either the Never a Weak Best Response (NWBR) property of Cho and Kreps (1987) or the (stronger) D1 criterion of Cho and Sobel (1990); for formal definitions and comparisons of these solution concepts, see for example Fudenberg and Tirole (1991).[11] Finally, we insist that beliefs about the opponent's conditional beliefs are correct on the equilibrium path, i.e., $(\hat{p}_1, \hat{p}_2) = (p_1^*, p_2^*)$. The latter assumption is probably the strongest of all our assumptions; to us it is not even clear that players should hold correct beliefs about the opponent's prior after learning the opponent's type.

2.5 Preferences

Players care both about the material consequences of their actions and about what the opponent will think about them.[12] More precisely, they care about the opponent's assessment of their type.

For simplicity, we assume that players only care about first moments. Let

$$r_j^k = E[\theta_i^k | \theta_j, h_j]$$

denote player j's assessment, or *rating*, of player i's characteristic k. The rating potentially depends both on player j's type – since the type affects the prior belief – and on any move that i has taken.

The key elements of the model concern the impact of others' rating on players' utility. We allow player i's sensitivity to player j's rating to depend on

[11] As usual in signaling games, the main results would survive, subject to suitable restrictions on the players' priors, if we instead were to apply the Undefeated Equilibrium concept of Mailath, Okuno-Fujiwara and Postlewaite (1993).

[12] In principle, players may also care about what other spectators think. We return to this issue below.

what i thinks about j's personality. For example, appreciation for skilfulness is sweeter when it comes from others who are skilled. On the other hand, revenge is sweet primarily when observed by the perpetrator. Thus, we probably want our gifts to be observed by altruists and our punishments to be observed by egoists. The impact of others' beliefs on a player's utility is also allowed to depend on which traits are *salient* in the situation. Some situations call for generosity, others for courage, and the distinction between situations can be subtle. For example, in an otherwise identical situation, the personality traits that are salient for someone dressed as a soldier may be less so for someone dressed as a vicar.

To capture the dependence of players utility on opponents' characteristics and the nature of the situation, let $s_{ij} = (s_{ij}^1, ..., s_{ij}^n) : \Theta \times \Theta \to \mathbb{R}_+^n$ be the weights that player i's assign to player j's rating; we refer to s_{ij} as player i's *salience weights*. In the applications that we consider in this paper, we shall make the simplification that salience weights are independent of i's own type. Thus, we shall let $s_j = s(\theta_j)$ be the weight that player i puts on the assessment of player j of type θ_j.

For concreteness, we assume that player i's concern for esteem can be expressed by the function

$$\begin{aligned} V(\hat{p}_j, \theta_i) &= g(\theta_i) \sum_{k=1}^{n} E_{\theta_j}[s(\theta_j) r_j^k | \theta_i, h_i] \\ &= g(\theta_i) \sum_{k=1}^{n} E_{\theta_j}[s(\theta_j) E[\theta_i^k | \theta_j, h_j] | \theta_i, h_i] \end{aligned} \tag{1}$$

Here, $g(\theta_i)$ mesures how sensitive player i is to social esteem in general. Note that if $g(\theta_i)$ were constant, player i's type matters only through i's beliefs. In many contexts it seems more plausible that the own characteristics affect how deeply one cares about the opinion of others.

Let $\pi_i(a_i, a_j, \theta_i)$ denote the material payoff to player i. If players are entirely selfish, they only care about their own material payoff; otherwise, they also care about the material payoff of the opponent. Player i's total utility can thus be written

$$U_i(\theta_i) = M(\pi_i(a_i, a_j, \theta_i), \pi_j(a_i, a_j, \theta_j), \theta_i, p_i) + V(\hat{p}_j, \theta_i), \tag{2}$$

where $i \neq j$. We refer to M as the player's *material well-being* and to V as the player's *pride*. The formulation assumes, for simplicity, that players do not care about others' pride. Players are assumed to be risk-neutral, so when payoffs are uncertain, player i maximizes the expectation of U_i.

Note that by including p_i as an argument of M we admit Levine's (1998) point that players may care directly about their opponent's type. In fact,

our model nests both Levine (1998) and Bénabou and Tirole (2006) as special cases. Compared to Levine's model, the main innovation is the function V. Compared to Bénabou and Tirole's model, the main innovation is the symmetric treatment of the two players; in their setting, only one player cares about pride, and only one player holds private information. In order to focus sharply on our contribution, we shall henceforth dispense with the key ingredients of both Levine and of Bénabou and Tirole: We assume that M is independent of p_i and, for most of the time, that players' characteristics are uni-dimensional.

2.6 The two-type case

Consider the simplest possible case, with unidimensional characteristics and only two types of player. That is, $\theta \in \{\theta_L, \theta_H\}$ with $\theta_H > \theta_L$.

The two players know whether their own type is H(igh) or L(ow), but not the type of the opponent. Suppose players are drawn from one of two distributions, G(ood) and B(ad). (I.e., the set $\mathcal{J}$ has two elements.) Let $P(\theta_H|G) = h$ and $P(\theta_H|B) = l < h$ be the probability of drawing type H in state G and state B respectively. We say that player i's personal signal "S" is "H" if $\theta_i = \theta_H$ and "L" if $\theta_i = \theta_L$. Let both states be equally likely a priori, i.e., $P(G) = P(B) = 1/2$. The players' metaprior can then be expressed as

$$p^0 = P(\theta_H|\emptyset) = P(G)h + P(B)l = \frac{h+l}{2}.$$

The personal priors concerning the true state, $P(G|\text{“}H\text{”})$ and $P(G|\text{“}L\text{”})$, and concerning the opponent's type, p_H^0 and p_L^0, can now be computed using Bayes' rule.[13] Comparing these priors to the metaprior p_0, we see that $p_H^0 > p^0 >$

[13]A player of type H makes the inference

$$\begin{aligned} P(G|\text{“}H\text{”}) &= \frac{P(\text{“}H\text{”}|G)P(G)}{P(\text{“}H\text{”}|G)P(G) + P(\text{“}H\text{”}|B)P(B)} \\ &= \frac{h(1/2)}{h(1/2) + l(1/2)} = \frac{h}{h+l} > 1/2, \end{aligned}$$

whereas a player of type L makes the inference

$$\begin{aligned} P(G|\text{“}L\text{”}) &= \frac{P(\text{“}L\text{”}|G)P(G)}{P(\text{“}L\text{”}|G)P(G) + P(\text{“}L\text{”}|B)P(B)} \\ &= \frac{(1-h)(1/2)}{(1-h)(1/2) + (1-l)(1/2)} = \frac{1-h}{2-h-l} < 1/2. \end{aligned}$$

Having updated her beliefs concerning the true distribution, a type H player forms a prior concerning the opponent's type, call it

$$p_H^0 = \text{Prob}(\theta_j = \theta_H|\theta_i = \theta_H) = P(\text{“}H\text{”}|G)P(G|\text{“}H\text{”}) + P(\text{“}H\text{”}|B)P(B|\text{“}H\text{”})$$

p_L^0. Despite starting out with a common view of the world, type H is more optimistic about the opponent than is type L.

In the two–type case, $p_i(\theta_i)$ denotes player i's conditional belief that $\theta_j = \theta_H$, and $\hat{p}_j(\theta_j)$ denotes player i's belief that player j believes that $\theta_i = \theta_H$. Let s_I be the weight that player i assigns to the assessment of player j of type I, where $I \in \{L, H\}$. Then, the level of pride for player i of type I is

$$\begin{aligned} V_i(\hat{p}_j, \theta_I) &= g(s_I)[p_i(\theta_I)s_H[\hat{p}_j(\theta_H)\theta_H + (1 - \hat{p}_j(\theta_H))\theta_L] \\ &\quad +(1 - p_i(\theta_I))s_L[\hat{p}_j(\theta_L)\theta_H + (1 - \hat{p}_j(\theta_L))\theta_L]]. \end{aligned} \tag{3}$$

Note how player i likes to impress player j; i's pride is increasing in $\hat{p}_j$. Importantly, j's type matters to player i's pride. If $s_H > s_L$, it is more valuable for player i to impress player j the more likely it is that player j is of type H. As we shall see, this may create an additional incentive, over and above j's own pride, for player j to convey the impression of being type H.

The model is ready for its first test.

3 The trust game

By now it is well known that people care not only about ultimate outcomes, but also the process leading up to the outcomes. A player who had a chance to be trusting or kind, but failed to be so, induces another behavior in the opponent(s) than a player who handed the same choice set to the opponents, but who had no choice. More precisely, McCabe, Rigdon and Smith (2003) show that subjects' behavior in the "voluntary trust game" depicted in Figure 1a is radically different from their behavior in the "involuntary trust game" depicted in Figure 1b. Although player 2 has exactly the same choice set in both situations, the frequency of strategy R is much greater in the voluntary trust game than in the involuntary trust game.

$$= h\frac{h}{h+l} + l\left(1 - \frac{h}{h+l}\right) = \frac{h^2 + l^2}{h+l}.$$

Analogously, type L's prior is

$$\begin{aligned} p_L^0 &= \text{Prob}(\theta_j = \theta_H | \theta_i = \theta_L) = P(\text{“}H\text{”}|G)P(G|\text{“}L\text{”}) + P(\text{“}H\text{”}|B)P(B|\text{“}L\text{”}) \\ &= \frac{h - h^2 + l - l^2}{2 - h - l}. \end{aligned}$$

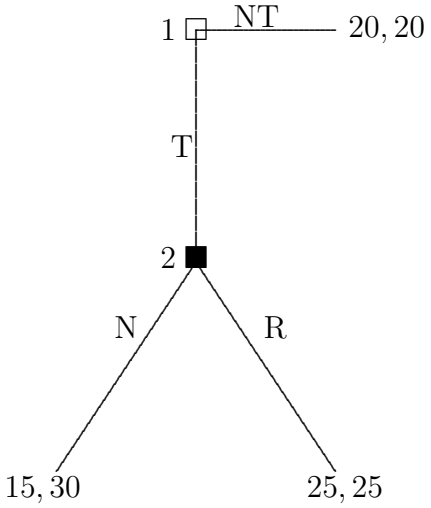

Figure 1a: Voluntary trust

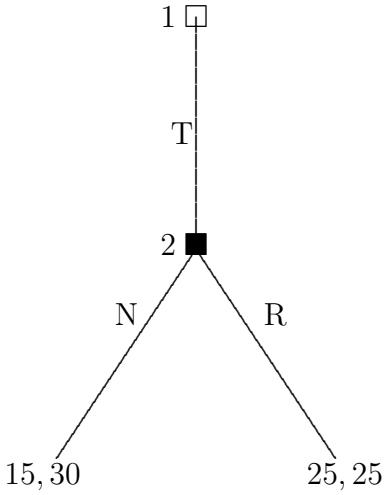

Figure 1b: Involuntary trust

To see how our model rationalizes this experimental finding, we must first ask what is the source of heterogeneity among players, i.e., what is the salient characteristic of a player when in this situation? Since monetary payoffs are given exogenously, it must be some social preference. Let us posit that the relevant characteristic is inequality aversion of the kind proposed by Fehr and Schmidt (1999), and that for simplicity the preferences of player i can be written

$$U_i = \pi_i - |\pi_i - \pi_j|\theta_i + V(\hat{p}_j, \theta_i). \tag{4}$$

The formulation implies that a player's "superiority aversion" is as strong as the player's "inferiority aversion"; it is straightforward to relax this assumption. Another plausible generalization would be to introduce internal salience weights to account for the fact that even the purely personal aversion to inequality may depend on the situation.

Let us start by analyzing the problem of player 2 in the involuntary trust game. We want to derive a condition under which player 2 chooses N (not reward) regardless of his type. Since type H players have most to gain by playing R (reward), we impose the out–of–equilibrium belief restriction that play of R will be interpreted as a sure sign that player 2 is type H. A type H player then chooses to play N if

$$\begin{aligned} & 30 - (30 - 25)\theta_H \\ + \; & [p_H^0 s_H(p_H^0\theta_H + (1 - p_H^0)\theta_L) + (1 - p_H^0)s_L(p_L^0\theta_H + (1 - p_L^0)\theta_L)]g_H \\ > \; & 25 + [p_H^0 s_H\theta_H + (1 - p_H^0)s_L)\theta_H]g_H, \end{aligned}$$

or equivalently

$$g_H < \frac{5(1-\theta_H)}{(\theta_H-\theta_L)(1-p_H^0)(p_H^0 s_H + (1-p_L^0)s_L)}. \tag{5}$$

We see that a necessary condition is $\theta_H < 1$, which is all right as Fehr and Schmidt usually sets $\theta_H = 0.6$. The corresponding condition for a type L player is obviously weaker, so we can neglect it.

Turning to the voluntary trust game, player 2 of type H is supposed to play R. Given the proposed equilibrium expectations, the type H player plays R if

$$25 + g_H\theta_H s_H > 30 - \theta_H(30-15) + g_H\theta_L s_H,$$

or equivalently

$$g_H > \frac{5-15\theta_H}{(\theta_H-\theta_L)s_H}. \tag{6}$$

Player 2 of type L is supposed to play N and will do so if

$$25 + g_L\theta_H s_H < 30 - \theta_L(30-15) + g_L\theta_L s_H,$$

or equivalently

$$g_L < \frac{5-15\theta_L}{(\theta_H-\theta_L)s_H}. \tag{7}$$

Finally, player 1 of type H is willing to play T if

$$\begin{aligned} & p_H^0(25 + g_H\theta_H s_H) + (1-p_H^0)(15 - \theta_H(30-15) + g_H\theta_H s_L) \\ > \;& 20 + p_H^0 g_H\theta_L s_H + (1-p_H^0)g_H\theta_L s_L, \end{aligned}$$

or equivalently

$$p_H^0 > \frac{5 + 15\theta_H - (\theta_H-\theta_L)g_H s_L}{10 + 20\theta_H + (\theta_H-\theta_L)(s_H-s_L)g_H}, \tag{8}$$

whereas player 1 of type L is willing to play NT if

$$\begin{aligned} & p_L^0(25 + g_L\theta_H s_H) + (1-p_L^0)(15 + g_L\theta_H s_L) \\ < \;& 20 + p_L^0 g_L\theta_L s_H + (1-p_L^0)g_L\theta_L s_L, \end{aligned}$$

or equivalently

$$p_L^0 < \frac{5 - (\theta_H-\theta_L)g_L s_L}{10 + (\theta_H-\theta_L)(s_H-s_L)g_L}. \tag{9}$$

It is immediate from (5), (6), and (7) that there is an open set of parameter vectors $(g_L, g_H, p_L^0, p_H^0, s_L, s_H, \theta_L, \theta_H)$ such that all these three conditions hold.

Moreover, this is true for any priors (p_L^0, p_H^0). Since we can always find priors such that both (8) and (9) hold, we have proved that there is an open set of parameters such that type H trusts in the role of player 1 and rewards voluntary trust but not involuntary trust i the role of player 2, whereas type L does not trust in the role of player 1, and fails to reward trust irrespective of whether trust is voluntary or not. Call the relevant parameter set $\mathcal{S}$. It is tedious but straightforward to show that, for these parameters, the depicted equilibrium is the only equilibrium to satisfy standard equilibrium refinement criteria.[14]

Proposition 1 *There exists an open set of parameters such that in the unique perfect Bayesian equilibrium satisfying the NWBR criterion, player 1 trusts voluntarily if and only if she is of type H, and player 2 rewards trust if and only if he is of type H and trust is voluntary.*

In Figure 2, we depict $\mathcal{S}$ for a numerical example with homogeneous expectations, $p_L^0 = p_H^0 = 2/5$. Moreover, $s_L = \underline{s} = 0, \theta_L = 0, \theta_H = 1/5$.[15]

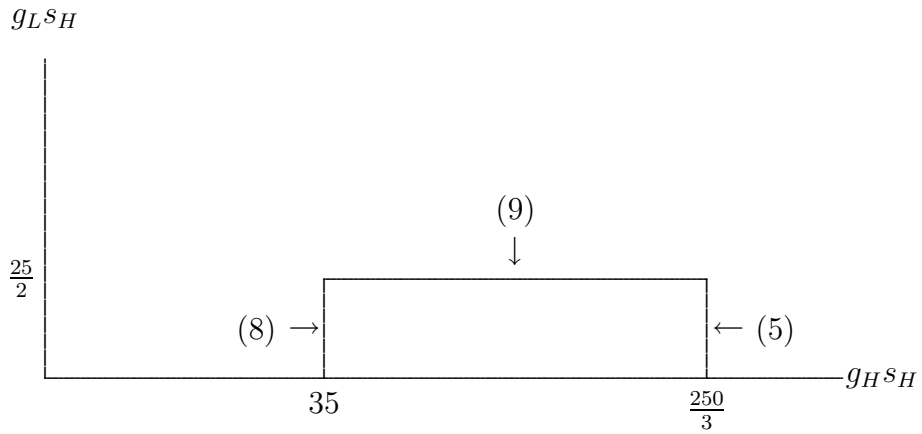

Figure 2: The set S in the trust game example

A necessary condition is that $s_H > 0$. To set $s_H = \bar{s}$ makes good sense in the

[14]The proof proceeds by showing that there exists plausible out–of–equilibrium beliefs that support the proposed outcome while there does not exist plausible beliefs that support any other outcome. Since full details are given for Proposition 2, where each player has larger strategy spaces, we skip details here.

[15]Our value for θ_H is low in comparison to the values proposed by Fehr and Schmidt (1999). However, they needed the high values in order to account for ultimatum game evidence in the absence of esteem considerations; when players are driven partly by pride, the parameter must necessarily be smaller.

trust game, because the source of esteem is generosity. As one would expect, the key to generating the desired equilibrium is then that g_L is sufficiently small, which guarantees that type L always plays N, while g_H takes an intermediate value which is large enough to make type H players trust voluntarily in the role of player 1, but not so large that they reward involuntary trust as player 2. Note also that the parameter box in Figure 2 can be made to expand by allowing heterogeneous priors; in particular, a smaller p_L expands upwards, and a larger p_H expands both sides.

In Appendix 2, we illustrate how the model can similarly be used to explain behavior in the (mini-) ultimatum game. Here, we proceed to consider our main application.

4 The principal-agent relationship

In a principal-agent setting, one player, the principal, first chooses a contract and the other player, the agent, then chooses an effort level. Denote the effort level a and let the contract $t : \mathcal{A} \to \mathbb{R}$ specify the agent's remuneration. The agent's effort yields a benefit $B(a)$ to the principal and a cost $C(a, \theta)$ to the agent. The cost of the contract to the principal is denoted $T(t(a))$.

We assume that $T(a) \geq t(a)$. When the two are equal, t can be thought of as a pure transfer; otherwise the principal pays more than the agent receives or receives less than the agent pays. In the case that the agent is penalized ($t(a) < 0$) without the principal being rewarded ($T(a) \geq 0$) we can think of the principal engaging in control of the agent's actions. When nothing else is said, we assume $B(a)$ is non–negative, increasing and concave, that $C(a)$ is everywhere twice differentiable and weakly concave in a. We confine attention to incentive schemes that have at most one "jump"; more precisely, we assume that t is differentiable everywhere, except possibly for an upward jump at some action a^u. In particular, we allow fixed wages with or without firing threats, bonus contracts, and linear incentive schemes.

The key feature of our model, compared to the standard principal–agent analysis, is that both the agent and the principal engage in signaling. The agent's action a is potentially informative of θ_A and the principal's contract proposal $t(a)$ is potentially informative of the principal's characteristics, θ_P.

For most of the paper, and all of this section, we confine attention to the two–type example specified in Section 2.6.

4.1 Selfish players

When players are entirely selfish, any private information concerns monetary costs and benefits. We shall consider the case in which only the agent's pro-

ductivity matters for material payoffs. More precisely, we write the principal's utility as

$$U_P = B(a) - T(t(a)) + V(\hat{p}_A(t), \theta_P), \tag{10}$$

and the agent's utility as

$$U_A = t(a) - C(a, \theta_A) + V(\hat{p}_P(a, t), \theta_A). \tag{11}$$

Note that (10) implies that the principal cares about how her productivity will be rated by the agent, even if the principal does not actually work. We think that this is realistic in many cases, and it preserves the model's symmetry, but as will become clear our main results would go through even if we were to neglect the principal's pride by setting $V(\hat{p}_A(t), \theta_P) = 0$.

4.1.1 The agent's problem

In this version of the model, the two agent types can be called productive (θ_H) and unproductive (θ_L).

The principal's prior describes a subjective probability p_P^0 that the agent is productive. The probability p_P^0 is drawn from the set $\{p_L^0, p_H^0\}$.

The contract offer t may carry information about the principal's type. Thus, $p_A(t, \theta_I)$ denotes the posterior belief about the principal's type held by an agent of type I. Likewise, $\hat{p}_P(\emptyset, t, \theta_I)$ denotes the posterior belief about the principal's prior held by an agent of type I, and $\hat{p}_P(a, t, \theta_I)$ denotes the agent's posterior belief about the principal's posterior. Recall that it is the own belief about the opponent's posterior beliefs that enters the pride function.

In order for the agent's signaling problem to be interesting, we assume that concern for esteem is sufficiently strong to make the unproductive agent contemplate mimicking the productive agent's action. (For the moment, we neglect the issue of whether this assumption is consistent with optimal incentives $t(a)$.) More precisely, assume that there are unique maximizers

$$a_i^* = \arg\max_a [t(a) - C(a, \theta_i)],$$

and assume that

$$t(a_L^*) - C(a_L^*, \theta_L) + V_A(0, \theta_L) < t(a_H^*) - C(a_H^*, \theta_L) + V_A(1, \theta_L), \tag{12}$$

where

$$V_A(0, \theta_I) = \theta_L[p_A(\theta_I)s_H + (1 - p_A(\theta_I))s_L]$$

is the type I agent's pride when (he believes that) his type is believed by the principal to be L, and

$$V_A(1, \theta_I) = \theta_H[p_A(\theta_I)s_H + (1 - p_A(\theta_I))s_L]$$

is the pride when (he believes that) his type is believed by the principal to be H. Obviously, we can always make (12) hold by increasing the difference $\theta_H - \theta_L$.

When (12) holds, the productive agent cannot distinguish himself completely through the action a_H^* because if this action were sufficient to come across as productive, then even the unproductive agent would prefer a_H^* to a_L^*. We are now ready to state our first result.

Proposition 2 *Assume that (i)* $\theta \in \{\theta_L, \theta_H\}$; *(ii)* $\mathcal{A}_\mathcal{A} = \mathbb{R}_+$; *(iii)* $t(a) - C(a, \theta_i)$ *have unique maximizers,* $a^*(\theta_i)$*, and are decreasing in* a *for* $a > a^*(\theta)$*; (iv)* $\partial^2 C(a,\theta)/\partial a \partial \theta < 0$ *for all* a *and* θ*; (v) the inequality (12) is satisfied. Then, the unique PBE outcome to satisfy NWBR entails actions by the agent,* $a_L = a_L^*$ *and* $a_H = a_H^S$*, such that* a_H^S *is the solution to*

$$t(a_L^*) - C(a_L^*, \theta_L) + V_A(0, \theta_L) = t(a_H^S) - C(a_H^S, \theta_L) + V_A(1, \theta_L). \tag{13}$$

Proof: See Appendix.

Essentially, Proposition 2 recapitulates the job market signaling argument of Spence (1973). There, productive workers acquire education in order to impress the prospective future employers; here productive workers put in high effort so as to impress their current employers. The context is slightly different, but the signaling problem is the same.

Observe that exogenous changes in the material incentives t do not affect immaterial incentives V at all. Thus, if material incentives are strengthened (for example, through an increase in the slope of $t(a)$), it follows directly from (13) that a_H^S must increase. Proposition 2 thus confirms that material incentives promote effort in the one–sided incomplete information model if agents differ along one dimension only. A similar observation is made by Benabou and Tirole (2006), who go on to investigate what happens when agents differ along two dimensions.

The first truly novel feature of our model is that the agent's beliefs about the principal's type affect the agent's action. Differentiation of (13) yields

$$\frac{da_H^S}{dp_A(\theta_L)} = \frac{(\theta_H - \theta_L)(s_H - s_L)}{C'(a_H^S, \theta_L) - t'(a_H^S)}. \tag{14}$$

The denominator is positive under weak assumptions; notably, it is sufficient that $t(a)$ is concave and $C(a, \theta_L)$ is convex, and at least one of them strictly so. Therefore, the agent's action is generally increasing in the agent's optimism regarding the principal's type. Intuitively, the agent is more keen to make a favorable impression on the type H principal, so the more likely it is that the principal is of type H, the more intensively does the type H agent need to work

in order to credibly signal his type. Our major point, to be elaborated in the next subsection, is that the principal can use the incentive scheme t not only to directly affect the agent's action (the material incentive effect) but also to affect it indirectly through the agent's beliefs (the immaterial incentive effect).

Before turning to the principal's problem, let us investigate what happens when the agent's actions are bounded above. It is immediately obvious that full separation may then no longer be sustainable; when the unproductive agent is willing to take the highest possible action in return for the highest possible esteem, separation breaks. In a related model, Denrell (1998) suggested that in this case there would be no signaling; however that conclusion only holds if we confine attention to pure strategies. There does not seem to be any justification for doing so here. When mixed strategies are allowed, and the upper bound $\bar{a}$ is in the interval (a_L^*, a_H^S), the equilibrium will typically be semi-separating, with type H always choosing the highest feasible action and type L randomizing between this action and a_L^*.

The argument runs as follows. Suppose the type H agent takes an action a with probability 1. Let $x(a) \in (0,1)$ be the probability with which (the principal believes that) the type L agent takes the same action a. Then, the principal's posterior upon seeing a is

$$p^X(a, \theta_I) = \frac{p^0(\theta_I)}{p^0(\theta_I) + (1 - p^0(\theta_I))x(a)}. \tag{15}$$

The equilibrium pride of a type L agent following the action a_L^* is $V_A(0, \theta_L)$ as before. The equilibrium pride of a type L agent following the action a is

$$\begin{aligned} V_A(p^X, \theta_L) &= p_A(\theta_L) s_H[p^X(a, \theta_H)\theta_H + (1 - p^X(a, \theta_H))\theta_L] \\ &\quad + (1 - p_A(\theta_L)) s_L[p^X(a, \theta_L)\theta_H + (1 - p^X(a, \theta_L))\theta_L]. \end{aligned}$$

Since $V_A(p^X, \theta_L)$ is monotonically decreasing in x, there can be only one such semi–separating equilibrium for a given a. The job of equilibrium refinements is to rule out equilibria with other supports than $(a_L^*, \bar{a})$.

Proposition 3 *Retain the assumptions of Proposition 2, except replace (ii) by the assumption* $\sup \mathcal{A}_A = \bar{a} < a_H^S$. *Then, the unique PBE outcome to satisfy D1 is for type H to play* $a_H = \bar{a}$ *and for type L to play* $\bar{a}$ *with probability* x^* *and* a_L^* *with probability* $(1 - x^*)$, *where* x^* *is the solution to*

$$t(a_L^*) - C(a_L^*, \theta_L) + V_A(0, \theta_L) = t(\bar{a}) - C(\bar{a}, \theta_L) + V_A(p^X, \theta_L). \tag{16}$$

Proof: See Appendix.

Because of the upper bound, type H is unable to separate completely from type L, and thus chooses the action that yields the maximum amount of separation,

i.e., $\bar{a}$. Strengthening the monetary incentive serves to increase the equilibrium fraction x of type L that chooses the high action. Hence, there are no negative effects of incentives that are imposed for exogenous reasons.

4.1.2 The principal's problem

Everything else equal, the principal would like the agent to believe that she is of type H. There are two reasons: The principal's own pride and the agent being more concerned with the esteem of the type H principal.

If the principal merely claims to be type H, there is no reason why the agent should believe the claim; it can be credible only if the type H principal offers a contract t that the type L principal prefers not to mimic.

Intuitively, the way for the principal to convince the agent of her high type is to be generous in case the agent is of type L. The type H principal has a comparative advantage in being generous to type L agents, because of the belief that type L agents are relatively unlikely. Being generous in case of poor performance can be funded by less generosity in case of good performance, because the productive agents are kept productive by their fear of being classified as unproductive.

An ambitious objective would be to characterize the principal's optimal choice from a large set of incentive schemes. However, this exercise is meaningful only if we impose participation or limited liability constraints on the agent; otherwise the optimal incentive scheme is degenerate. Moreover, the general problem is quite complicated and involves many cases in which immaterial incentives play a minor role. Therefore, we here only present an example where the set of available material incentive scemes is limited and immaterial incentives come to the forefront.

A high wage as a signal of high expectations

A leading example of trustful contracts is the employment contract that specifies a high fixed wage no matter what the worker does. The prevailing explanation for this finding is that people want to reciprocate; if the principal behaves kindly, the agent wants to return the favor. (We shall discuss the merits of that explanation below.) An alternative explanation emanates from our model with selfish preferences: The best agents work hard when the principal sets a high wage, because the high wage is a credible signal that the principal is productive *and therefore worth impressing.* The high wage is a credible signal, because productive principals are more optimistic about their agents.

Here is the formal argument. Suppose the agent's set of actions $\mathcal{A}$ is unbounded.[16] Suppose that the principal can only choose the wage level. That is,

[16]Our argument goes through with a bounded action set as well, but the formulas are

$t \in \mathbb{R}_+$ and $T = t$. Under our assumption that $C'(a)$ is always increasing, i.e., the marginal cost of effort is always positive, the unproductive agent will not work at all due to the lack of pay incentive. That is, $a_L^S = 0$. Using equation (13), we see that the productive agent's effort level $a_H^S(p_A(t, \theta_L))$ is given by

$$C(a_H^S, \theta_L) = C(0, \theta_L) + (\theta_H - \theta_L)[p_A(\theta_L, t)s_H + (1 - p_A(\theta_L, t))s_L]. \quad (17)$$

Observe that the payment t has no direct effect on the effort; t only serves to signal the principal's type, i.e., to affect p_A. As we have already shown, the effort level a_H^S is increasing in p_A. In order for the type H principal to want to separate from the type L principal by increasing t, she must have more to gain from the expected increase in effort by the type H agent. A sufficient condition is

$$p_P^0(\theta_H)B'(a) > p_P^0(\theta_L)B'(a), \quad (18)$$

which holds because the productive principal is more optimistic.

Proposition 4 *Suppose the conditions of Propositions 2 hold and determine the agent's behavior for fixed agent beliefs. Suppose in addition that: (vi) $t(a)$ can be any constant function, (vii) $T = t$. Then the unique PBE outcome is for the type L principal to set $t = t_L = 0$ and for the type H principal to set $t = t_H$ as the solution to*

$$t_H = p_P^0(\theta_L)B(a_H^S(1)) + V_P(1, \theta_L) - V_P(0, \theta_L). \quad (19)$$

PROOF: See Appendix.

Even if the principal would not care about her own pride, we see from (19) that the more optimistic principal would signal her high expectations through a high wage. In expectation, she would still earn a positive material surplus $[p_P^0(\theta_H) - p_P^0(\theta_L)]B(a_H^S(1))$.

Our model produces a new twist on the efficiency wage argument. It resembles the gift–exchange model of Akerlof (1982). However, Akerlof invokes reciprocity among employers and workers – workers put in more effort when the wage is high because they want to return the employer's gift. In Akerlof's words, workers acquire a sentiment for the firm. Our argument, at least so far, is based on strict selfishness. Productive workers are willing to work harder when employers have higher expectations (more optimistic priors), because additional esteem from such "demanding" employers give them more pride.

While our argument may seem far–fetched to some economists, it will be familiar to many sociologists, psychologists, and business practitioners. The model gives a formalization of the self–fulfilling prophecy argument by, among

somewhat different.

others, Livingston (1969), Archibald (1974) and Eden (1984). All of these argue that management ought to motivate workers by conveying high, but realistic, expectations. What we add to their story is that expectations cannot always be conveyed through words alone. The employer may have to put her money where her mouth is and pay high wages as a credible signal of optimism.

4.2 Social preferences

Many of the results with selfish preferences and heterogeneous productivity have direct counterparts in the case of heterogeneous social preferences. In particular, low effort costs are isomorphic to a concern for efficiency.

Following the seminal paper of Fehr, Kirchsteiger and Riedl (1993), several experiments have shown that high wages induce high effort. The experimental evidence typically concern situations in which productivity is known, so the relevant heterogeneity must concern tastes, like altruism, or fairness. The best current interpretation of the evidence is that the worker wants to reciprocate the employer's favor; see Charness (2004). The social preference version of our model offers an alternative interpretation: The employer who offers a high wage signals high expectations about the worker's concern for efficiency. Workers value esteem from such employers more than esteem from employers with low expectations, and hence generous workers put in more effort following a high wage offer than following a low wage offer. The formal argument is isomorphic to that Propositions 2 and 4, so we do not repeat it.

Relabelling the parameter θ as altruism, Proposition 2 offers an explanation for observable acts of generosity. Indeed our analysis even explains why some people give positive and "interior" amounts in the dictatorship game.[17] To see this, let a be the amount transferred from the agent to the principal and suppose the principal's only feasible incentive scheme is $t(a) = T(a) = 0$. Furthermore, for illustration let $B(a, \theta_P) = a/\theta_P$ and $C(a) = a/\theta_A$. With these assumptions, $a_L^* = a_H^* = 0$, so neither agent type would give anything absent a concern for esteem. Since $V_A(1, \theta_L) > V_A(0, \theta_L)$, (12) holds. From Proposition 2 it then follows that the relative altruists, the type H agents, will give a strictly positive amount a_H^S merely to signal their altruism, whereas the relative egoists give nothing.

If instead the parameter θ denotes inequality aversion, the single–crossing condition in Proposition 2 will no longer hold. Since we believe that concerns

[17] In the Dictatorship game, one player - the Dictator - has an amount of money and is free to divide it in any way between herself and a passive recipient. If people are entirely selfish they should not give anything; if they are generous enough to give something, it is puzzling that they don't always give everything (in case of altruism) or exactly half (in case of inequality aversion), but quite often settle on intermediate amounts.

for equality are important in many experiments, let us give a general result for this case.

Proposition 5 *Retain the assumptions of Proposition 2, except replace (iv) by the assumption* $\partial^2 C(a,\theta)/\partial a \partial \theta < 0$ *for all* $a < a^*$ *and* $\partial^2 C(a,\theta)/\partial a \partial \theta > 0$ *for all* $a > a^*$*, where* $a^* \in [a_L^*, a_H^S]$*. Then, the unique PBE outcome to satisfy D1 is for type H to play* $a_H = a^*$ *and for type L to play* a^* *with probability* x^{**} *and* a_L^* *with probability* $(1 - x^{**})$*, where* x^{**} *is the solution to*

$$t(a_L^*) - C(a_L^*, \theta_L) + V_A(0, \theta_L) = t(a^*) - C(a^*, \theta_L) + V_A(p^X, \theta_L). \tag{20}$$

Proof: See Appendix.

Apart from any difference between $\bar{a}$ and a^* Proposition 5 is essentially identical to Proposition 3. At first sight, it perhaps seems strange to assume that type H's marginal cost is first lower and then higher than type L's marginal cost. However, inequality aversion gives preferences precisely this structure. Inequality averse individuals are relatively willing to increase their effort until rewards are equal, but relatively unwilling to increase their effort beyond this point.

Again, it is straightforward to investigate the effect of the agent's beliefs about the principal on the agent's behavior. Differentiation of (16) yields

$$\frac{dx^*}{dp_A(\theta_L)} = \frac{s_H p^X(\theta_H) - s_L p^X(\theta_L)}{-\left[p_A(\theta_L) s_H \frac{\partial p^X(\theta_H)}{\partial x} + (1 - p_A(\theta_L)) s_L \frac{\partial p^X(\theta_L)}{\partial x}\right]}, \tag{21}$$

and the result for x^{**} is analogous. Since $p^X(\theta_I)$ is decreasing in x, the expression is positive. Thus, the probability that the type L agent emulates the type H agent is increasing in the agent's optimism about the principal.

We are now ready to discuss evidence from a recent experiment that previous theories fail to explain, namely Falk and Kosfeld (2005).

4.3 Control

We gave a brief description of Falk and Kosfeld's study in Section 2. Their receiver corresponds to our principal and their donor to our agent. The principal can rule out some small donations by the agent, but not affect the choice between the remaining actions. In the language of our model, the principal can impose $t(a) = -\infty$ for the subset of donations $a < 10$, which suffices to keep donations at or above this level.[18] For all other donations, $t(a) = 0$. Finally,

[18] Falk and Kosfeld also studied the effect of varying the control option; instead of 10 they substituted 5 and 20 respectively.

$T(a) = 0$ for all a, since the imposition of control is costless. The agent's set of actions is $A = [0, 120]$.[19] The monetary benefit of the principal is $2a$ and the monetary cost of the agent is $a - 120$.

Falk and Kosfeld's evidence can be summarized as follows:

1. The agents' average donation was 17.5 with control and 23 without. There were few donations above 40. About half of the agents choose to donate exactly 10 if controlled.

2. If control is exogenously imposed, in the sense that the principal must leave at least 10 to the agent, the negative effect of control vanishes.

3. About 30 percent of the principals choose to control. (The remaining 70 percent trust.)

4. Principals make roughly correct predictions about agent's actions following their own control choice. Controlling principals underestimate what the average donation would have been had they trusted.

We first attempt to fit all aspects of the evidence. Afterwards, we select the findings that are most likely to be robust as players gather experience, and discuss the model's implications for these.

Note that the full evidence cannot be replicated by an equilibrium of our model if we insist on correct expectations. With correct expectations, all principals would have to trust – since material payoffs are higher than under control (and esteem can hardly be lower). Heterogeneous priors is therefore necessary to fit all aspects of the evidence. Since many agents choose donations of 40 and few choose higher donations, we further infer that equal payoffs are salient. Thus, we assume that inequality aversion is the relevant social preference. Utilities are thus

$$U_P = 2a - |2a - (120 - a)|\theta_P + V(\hat{p}_A(t), \theta_P)$$

and

$$U_A = t(a) - a - |120 - t(a) - a - 2a|\theta_A + V(\hat{p}_P(a, t), \theta_A).$$

Since the principal does not have the opportunity to impose an unequal distribution, the agent does not have any reason to be spiteful according to Fehr and Schmidt. As negative reciprocity is not an issue, according to our model the agent will always care more about esteem from generous principals. To

[19] In the experiment, the set of actions has an upper bound, but the data does not suggest that this upper bound ever played a role.

simplify expressions, we make the normalization $\theta_L = 0$. As the private information parameter takes only two values, we use equation (3) to specify $V(\cdot)$.

In order to fit the heterogeneous behaviors, our equilibrium needs to separate, at least partially, generous agents from selfish agents and generous principals from selfish principals. Moreover, the most generous donations should be greater when agents are trusted. For example, can we construct an equilibrium with the feature that type L agents always donate the smallest admitted amount and type H agents donate 40 if they are trusted and 20 if they are not trusted? (The average payoff in such an equilibrium will then be 15 for controlling principals and 20 for trusting principals, roughly as in the data.)

Let us list the equilibrium conditions. Start with the type L agent's problem. Since principals fully separate in the equilibrium under consideration, a "controlled" agent of type L prefers donating the minimum amount of 10 rather than mimicking type H's donation of 20 if

$$-10 = -20 + g_L s_L \theta_H.$$

or equivalently,

$$g_L s_L \theta_H = 10. \tag{22}$$

A trusted agent of type L is indifferent between donating 0 and 40 if

$$0 = -40 + g_L s_H \theta_H,$$

or equivalently,

$$g_L s_H \theta_H = 40. \tag{23}$$

The controlled type H agent is content to give 20 rather than any higher amount if

$$\theta_H \leq 1/3. \tag{24}$$

The type L principal prefers control to trust if

$$p_L^0 40 + (1 - p_L^0)20 \geq p_L^0(80 + g_L s_H \theta_H) + (1 - p_L^0)(0 + g_L s_L \theta_H)$$

or equivalently

$$p_L^0 \leq \frac{20 - g_L s_L \theta_H}{60 + g_L \theta_H (s_H - s_L)}. \tag{25}$$

Finally, the type H principal prefers trust to control if

$$\begin{aligned} & p_H^0(40 - (100 - 40)\theta_H) + (1 - p_H^0)(20 - (110 - 20)\theta_H) \\ \leq \; & p_H^0(80 + g_H s_H \theta_H) + (1 - p_H^0)(0 - 120\theta_H + g_H s_L \theta_H) \end{aligned}$$

or equivalently

$$p_H^0 > \frac{20 + 120\theta_H - g_H s_L \theta_H}{60 + 90\theta_H + g_H \theta_H (s_H - s_L)} \tag{26}$$

Observe that the right hand side is decreasing in g_H. Since $g_H \geq g_L$, we can replace g_H by g_L to get an upper bound. If we set $\theta_H = 1/5$ as before, the last two conditions then become $p_L^0 \leq 1/9$ and $p_H^0 \geq 2/99$. Only the first condition is restrictive. Our example thus mimicks the main features of Falk and Kosfeld's evidence if we are willing to allow type L agents to be sufficiently (over)pessimistic.[20] Moreover, under the assumed parameters it is straightforward to check that the proposed equilibrium uniquely survive the refinement criteria that we have imposed.

When asked about their reaction to the principal's decision to control them, subjects who reduce their donations express feelings of being restricted and distrusted (Falk and Kosfeld, Figure 2). Our model captures this phenomenon inasmuch as both types of agents would prefer to be trusted. This is most obvious for the type L agent, who loses 10 without any compensation in terms of pride when controlled. The type H agent keeps more money when controlled, but the monetary gain is more than offset by a reduced sense of pride.

Finally, let us consider what happens when control is exogenously imposed rather than chosen by the principal. The exogenous control treatment implemented by Falk and Kosfeld essentially limits the action set of the agent to $[10, 120]$. (They do not use the word "control" in the instructions to the subjects. As we shall see below, wording could affect the results.) Recall that the donation of type H agents is given by the need to separate from type L agents. Given that the agent has no information about the principal, the separating equilibrium condition pinning down the type H agent's donation becomes

$$-10 = -a_H + g_L \theta_H (p_L^0 s_H + (1 - p_L^0) s_L).$$

Recall that under endogenous control, due to full separation of principal types, the condition was

$$-10 = -a_H + g_L \theta_H s_L.$$

The solution (a_H) must be larger when control is exogenous than when it is endogenous, because $s_H > s_L$. When control is endogenous, the agent

[20] The example does not quite mimick the fact that controlling principals in the experiment have correct expectations about controlled agents' average donations; in the example, the prediction is $(8/9) \cdot 10 + (1/9) \cdot 20$ which is smaller than the average donation of 15. Also, the fraction of principals who trust in the example (one half) is smaller than the fraction who trust in the experiment (three quarters). On both these counts we can improve the model's fit by considering a semi-separating equilibrium, in which type L principals are indifferent between trusting and controlling.

understands that the principal is of type L, and is thus not so concerned with getting the principal's esteem. When control is exogenous, the principal might be of type H and thus worth impressing. If we plug in the numbers from the numerical example, the difference in donation is relatively small; if $p_L^0 = 1/10$, a_H increases from 20 to 23. This is substantially less than Falk and Kosfeld finds; to mimick their numbers, type H's donation with exogenous control should be around 40 - the same as for trusted high types. One reason for the discrepancy is that our assumptions of common meta-priors and Bayesian updating force the type H agent to correctly assess the low type's prior, p_L^0. If we instead assume that subjects believe others to share their own beliefs, the prediction becomes radically different. If type H's belief is $p_H^0 = 5/8$ (which is still considerably less distorted than type L's assumed belief), and type H believes type L to share it, type H's donation becomes 38.75, which is almost the same as when the principal trusts. The reason why trust and exogenous control can have so similar effects on the high type's action is that two forces cancel: On one hand, exogenous control increases the action of the low type, forcing the high type to increase the donation in order to separate, but on the other hand the agent does not know whether the principal's type is high or low, depressing the low type's incentive to mimick generosity.

The experimental evidence shows that subjects had erroneous expectations. Since more experienced subjects might be able to make better estimates, we end the discussion by predicting the behavior under correct beliefs – all other model parameters held constant. It is straightforward to check that the only equilibrium to satisfy D1 under correct beliefs is for all principals to trust, for type L agents to donate 0, and for type H agents to donate 25.[21] Observe that the type H agent's donation is smaller than before, reflecting the fact that the agent's feeling of pride when trusted is smaller now that both types of principals are understood to trust. Note also that principals could earn more money by not trusting; the expected revenue would then be $(1/2) \cdot 20 + (1/2) \cdot 40 = 30$, whereas in this all–trust equilibrium, the principal's expected monetary payoff is only $(1/2) \cdot 50 = 25$. However, in the example this difference is more than compensated by the additional esteem that the type L principal obtains from pooling with the type H principal.

[21] The key conditions to check are the type L agent's indifference condition

$$0 = -a_H^S + g_L \theta_H (s_H/2 + s_L/2)$$

and the type L principal's preference condition

$$(1/2)\dot{2}0 + (1/2) \cdot 40 < (1/2) \cdot 2a_H^S + g_L \theta_H (s_H/2 + s_L/2)/2.$$

4.4 Penalties and trust

Gneezy and Rustichini (2000a) conduct a field experiment in which several daycare centers are induced to impose a fine on parents who collect their children too late in the evening. The effect of the fine is to increase the prevalence of late collection. When the penalty is removed, parents continue to collect their children later than before. Thus, a good theory needs to explain both why the fine backfires to begin with and why behavior does not revert when the fine is removed.

Our model offers explanations for both findings. Without the fine, the daycare manager displayed trust, and parents attempted to collect children in time in order to gain the manager's approval. The imposition of fines signal a reduction in trust, which implies that the value of approval declines. If the fine is relatively small, parents will behave less diligently. If the fine is seen to be removed because it did not work, rather than because the manager suddenly trusts the parents again, the parents will continue their less diligent ways – indeed, they should become even less diligent than before, because the esteem incentive and the pecuniary incentive are both gone.

Two recent studies by Fehr and Rockenbach (2003) and Fehr and List (2004) consider a version of the trust game, but with two twists: The trustor always states a desired back-transfer, and the trustor can choose upfront whether to punish a smaller-than-desired back-transfer by imposing a fine. Concretely, the trustor has 10 money units. For every unit given, the trustee receives 3 units. (Thus, the best egalitarian outcome is for the trustor to give all 10 and for the trustee to make a back-transfer of 20.) The fine is 4 money units, and if the fine kicks in these units are wasted - not transfered to the trustor. A main finding is that the decision by the trustor to impose a fine leads to lower back-transfers, so trustors are better off by foregoing the punishment option. It is straightforward to demonstrate that our model rationalizes the behavior: The imposition of the fine signals low expectations, thereby reducing the value of esteem. As a result of the fine, the smallest donations go up, but the largest donations go down.

4.5 Framing

Framing has a significant effect on behavior in many games. As shown by Liberman, Samuels and Ross (2004), people's actions in the prisoners' dilemma game depend heavily on whether experimenters call it the Community Game or the Wall Street Game.[22] A natural explanation is that concern for cooperation

[22]For a seminal contribution along the same lines, see Eiser and Bhavnani (1974). See also Pillutla and Chen (1999) and Rege and Telle (2003) for closely related work.

is more salient in the "community" than on Wall Street. Our model is able to capture variation in salience through the salience weights s. All we need to do is to specify a mapping from the set of decision frames to the set of salience weights. In doing so, we may borrow heavily from research in social psychology, which has devoted considerable effort to the classification of social frames; see for example Levin et al.(1998).

Recently, several experiments have studied the effect of framing on incentives. Fehr and Gächter (2002) investigate the role of fines in a gift exchange experiment. The principal commits to a wage and states a desired effort level, but can also choose a wage combined with a fine – to be imposed if the agent's effort is discovered to be lower than desired (the principal can observe the effort with some probability smaller than 1). The contract with a fine tend to entail lower effort for a given wage level, much as we might expect by now. However, when the authors rephrase the experiment, so that the principal chooses either a plain wage or a plain wage combined with a bonus in case the agent exerts the desired effort (or, more precisely, is not discovered to be shirking), the incentive works considerably better. Note that the two treatments give the principal exactly the same set of options: A wage of w coupled with a fine f is equivalent to a wage of $w - f$ coupled with a bonus of f. So why does the latter contract work better? Since the only difference is in the words "bonus" and "fine", we posit that these words must trigger different associations.

Houser et al.(2005) replicate the experiment of Fehr and Rockenbach (2003), but include a treatment in which the punishment threat is random and beyond the control of the principal. They find that the imposition of (weak) punishment threats have a negative effect on the agent regardless of whether the principal imposes it voluntarily or it is imposed randomly. The authors suggest that the punishment threat induces a cognitive shift away from cooperative behavior. In the terms of our model, cooperation becomes less salient.

Irlenbusch and Sliwka (2005) document a similar effect. The mere opportunity of paying a piece rate in a gift exchange experiment lowers the agent's effort. As in the daycare experiment of Gneezy and Rustichini, the detrimental effect of incentives prevail after the incentive is removed; when the trust game is played again without availability of piece rates, agents display less reciprocity than when piece rates were never available.

4.6 Other issues

Let us here briefly mention some issues that we have neglected so far.

4.6.1 Incentives affect the principal's desires

Up until now, we assumed that principals always appreciate "good" characteristics. However, credit card companies do not ordinarily want customers to pay their debts on time. As long as customers pay eventually, the high interest rate make the companies better off when the customers pay late. Because of the incentive scheme, the principal wants to encourage sloppy payment practices rather than diligence. As a result, there is no reason for the customers to feel appreciated by the credit card company when paying on time.

The daycare experiment can perhaps be interpreted in this way too. To the extent that the imposition of a fine indicates that the daycare center is now indifferent towards parents' behavior (or even prefers late collection), why should parents take pride in punctuality? It seems to us that the salience weight on the punctuality parameter ought to be zero in this case.

4.6.2 Multiple tasks: incentives as communication

We have assumed that the agent's effort is unidimensional and the principal's objective is publicly known. In many realistic cases effort is multi-dimensional, and the principal's objective is privately known.

Suppose for a moment that agents do not know what behaviors are in the principal's interest, but do know the principal's expectations about agent characteristics. It is now possible that agents' behavior reacts strongly to small but informative clues, with little or no material payoff relevance to the agents. For example, small bonuses - and even gold stars and other materially worthless distinctions - can be used to indicate which behaviors are deemed to be desirable. A bonus based on group performance indicates that concern for others' output is valued; bonuses based on individual performance indicate that concern for others is less important. Small rewards play the essentially same role as verbal communication. If rewards are increased, a reasonable interpretation is now that high effort is more valuable to the principal than before. In this case, there is a positive incentive multiplier: The direct effect through higher material reward is complemented by an indirect effect through increased salience of high effort for esteem purposes.

4.6.3 Multiple observers

In our principal–agent examples, we have so far assumed that the principal plays three roles: She designs the contract, is the beneficiary of agent cooperation, and is the sole observer whom the agent may impress. One way to distinguish the predictions of our model from the predictions of reciprocity models is to separate these three roles. For example, reciprocity models pre-

dict that the agent will be kind in order to reward kindness, whereas our model predicts that the agent is kind in order to impress an audience.

5 Final remarks

The analysis can be extended in several directions. We think that it might explain why people dislike being monitored, and why this dislike is stronger when monitoring is linked to future rewards, as demonstrated by Enzle and Anderson (1993). It might also be able to explain why careful ex ante planning by one contracting party "indicates a lack of trust and blunts the demands of friendship" as suggested by Macaulay (1963, p64). Finally, it is natural to explore the nature of optimal incentive schemes under the assumption that people care about esteem. Our preliminary investigations suggest that a fixed wage can be optimal under quite plausible assumptions.

All this is not to say that material incentives never work, or even that the case for material incentives is weaker than the traditional principal–agent model suggests. In order to explain a set of puzzles, we have been led to consider cases in which esteem incentives are naturally aligned with the principal's interest. Recognizing that agents may seek esteem from others than the principal, the presence of esteem incentives could well *strengthen* the case for material incentives. For example, asking a CEO to engage in heavy cost cutting could be extremely difficult if the CEO cares more about esteem from the firm's employees than from the firm's owners. We believe that this is one reason why corporate restructuring either has to wait until employees are sufficiently "crisis conscious" or is carried out by a new manager on strong material incentives, who often leaves when the job is done.

Finally, our model of human nature can be applied to analyze many other social settings than those we have considered here.

Appendix 1: Proofs

All the proofs are standard applications of the NWBR and D1 criteria. In fact, since senders care directly about receivers' beliefs rather than about the receiver's rational actions following these beliefs, the notation is somewhat simpler than usual.

Proof of Proposition 2

From condition (iv) it follows that $a_L^* \leq a_H^*$, with equality only if $a_H^* = a^u$ (type H's favorite outcome is at a point where the material incentives jump

upwards). From (iv) and (12) it follows that $a_H^S > a_H^*$. Conditions (ii) and (iii) together with differentiability of C and t (except for possibly one upward jump) ensure that a_H^S exists and is unique.

Next, we check that there exist "reasonable" out–of–equilibrium beliefs that sustain the proposed equilibrium outcome: Let the principal believe that any action $a < a_H^S$ is taken by a type L agent. Then, by definition, a_L^* dominates all actions $a \in (a_L^*, a_H^S)$ for type L. Since a_H^S dominates all actions $a > a_H^S$ for type L, it follows from (13) that L has no incentive to deviate from a_L^*. To check that type H has nothing to gain from deviating from a_H^S, observe that any higher action entails lower utility, and that the best lower action is a_H^*. Type H prefers a_H^S to a_H^* if

$$t(a_H^*) - C(a_H^*, \theta_H) + V_A(0, \theta_H) < t(a_H^S) - C(a_H^S, \theta_H) + V_A(1, \theta_H). \qquad (27)$$

Observe that $V_A(1, \theta_I) - V_A(0, \theta_I) = [p_A(\theta_I)s_H + (1 - p_A(\theta_I))s_L][\theta_H - \theta_L]$ for $I \in \{L, H\}$. Since $p_A(\theta_H) \geq p_A(\theta_L)$ it follows that $V_A(1, \theta_H) - V_A(0, \theta_H) \geq V_A(1, \theta_L) - V_A(0, \theta_L)$. It is thus sufficient to show that (27) is satisfied when $V_A(1, \theta_H) - V_A(0, \theta_H) = V_A(1, \theta_L) - V_A(0, \theta_L)$. Substituting for $t(a_H^S)$ from (13) and rearranging, the condition then becomes

$$C(a_H^S, \theta_H) - C(a_H^*, \theta_H) < C(a_H^S, \theta_L) - C(a_L^*, \theta_L) + t(a_L^*) - t(a_H^*). \qquad (28)$$

Since a_L^* is the unique solution to type L's problem, $t(a_L^*) - C(a_L^*, \theta_L) > t(a_H^*) - C(a_H^*, \theta_L)$, or equivalently, $t(a_L^*) - t(a_H^*) \geq C(a_L^*, \theta_L) - C(a_H^*, \theta_L)$. Thus, the right hand side of (28) cannot be smaller than $C(a_H^S, \theta_L) - C(a_L^*, \theta_L)$. In other words, a_H^S is a best response for type H if

$$C(a_H^S, \theta_H) - C(a_H^*, \theta_H) < C(a_H^S, \theta_L) - C(a_L^*, \theta_L),$$

and this inequality is fulfilled due to the single–crossing condition (iv). Thus, we have proved that the beliefs sustain the proposed equilibrium strategies. To see that the principal's beliefs are reasonable out of equilibrium, note that a deviation by type L to an action in the interval (a_L^*, a_H^*) would be justified if the principal were to believe that the action was taken by a type H agent. Thus, the belief that any such deviation is due to a type L player satisfies the NWBR-property.

The final step is to show that all other perfect Bayesian equilibria fail the NWBR property. Consider first other fully separating equilibria. We know that these entail $a_L = a_L^*$ and $a_H > a_H^S$. In order for type H not to deviate to a lower action (which, by condition (iii) gives higher utility if beliefs are constant), the principal must believe that there is some probability that this lower action is taken by a type L agent. But by condition (iii) and equation (13) type L is always worse off taking an action $a > a_H^S$ than with a_L^*. Thus,

the principal's belief fails NWBR. Consider next any pooling equilibrium $a_L = a_H = a^*$. To destroy it, show that there exists an action $a^d > a^*$ which is such that type H is willing to deviate to a^d if the principal becomes convinced that the agent's type is H, but type L is not willing to make the deviation regardless of the principal's beliefs. (Finding a^d is analogous to finding a_H^S, again single–crossing is the key, so we do not repeat the step.) Since type H can only be deterred from deviating to a^d if the principal ascribes a positive probability to the deviation being caused by a type L agent, the pooling equilibrium belief must fail NWBR. Finally, consider semi–separating equilibria, where there is at least one action a^s that both types play with positive probability. Again, this equilibrium can be destroyed by identifying a deviation that is potentially attractive only to type H. □

Proof of Proposition 3

Most of the proof is analogous to the proof of Proposition 1, so we are brief: First show that there are out–of–equilibrium beliefs that sustain the proposed outcome as a PBE. In particular, suppose that the principal believes that any action $a \in (a_L^*, \bar{a})$ is taken by a type L agent. Clearly, all actions $a \in (a_L^*, \bar{a})$ are then dominated by a_L^*. Given that the Principal updates the prior using Bayes' rule, the probability assigned to type H given action $\bar{a}$ is p^X. Thus, if (16) holds, type L is indifferent between actions a_L^* and $\bar{a}$. Since type H has lower marginal cost of effort than type L, and at least as much to gain by raising V_A from $V_A(1, \cdot)$ to $V_A(p^X, \cdot)$, it follows that type H strictly prefers $\bar{a}$ to all other available actions given the assumed beliefs. Thus, we have indeed described a PBE. The equilibrium satisfies D1, because there is a larger set of (alternative) beliefs that would justify any action $a \in (a_L^*, \bar{a})$ by the type L agent than by the type H agent. It remains to check that no other PBE survives D1. Consider any equilibrium in which there is some partial pooling at an action $a^* < \bar{a}$. This equilibrium must be sustained by "sufficiently pessimistic" beliefs following a deviation to all higher actions (otherwise, type H would deviate). But due to the single–crossing conditions on C and V_A, the set of beliefs justifying a deviation to $\bar{a}$ by type H is again larger than the corresponding set for type L. Thus, by D1, the principal must ascribe the deviation to a type H player, contradicting the original pessimistic beliefs. □

Proof of Proposition 5

Most of the proof is analogous to the proof of Proposition 3. The main difference is that we also need to consider equilibria with partial pooling at some action $\hat{a} > a^*$. These equilibria fail D1 because they must assume relatively

pessimistic beliefs following actions in the half-open interval $[a^*, \hat{a})$, despite the set of beliefs justifying a deviation to an action in this interval being greater for type H. □

Appendix 2: The mini ultimatum game

The mini ultimatum games depicted in Figures 3a and 3b can be analyzed in much the same way as the trust games if Figures 1a and 1b.

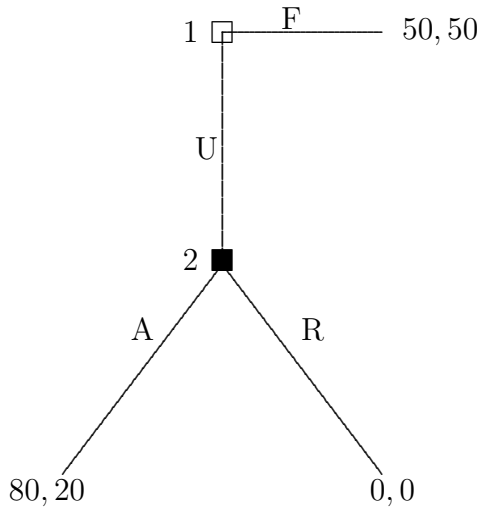

Figure 3a: Voluntary ultimatum

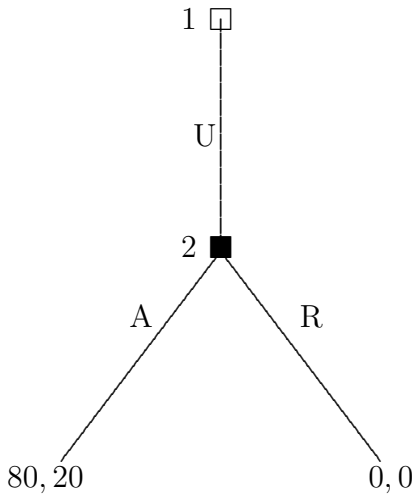

Figure 3b: Involuntary ultimatum

In experiments, subjects in the role of player 2 frequently plays R (reject) in the voluntary ultimatum game, while playing A (accept) in the involuntary ultimatum game.[23] Moreover, in the voluntary ultimatum game, both F (fair) and U (unfair) are quite common choices among subjects in the role of player 1.

In order to explain the behavior with the assumed preferences, let us look for an equilibrium in which type H players play F in the role of player 1 and play R as player 2 if and only if player 1's move is voluntary. Type L players should play U in the role of player 1 and always play A in the role of player 2.

To save on notation, we make the normalization $\underline{s} = 0$ from now on.

Let us start by analyzing the problem of player 2 in the involuntary ultimatum game. We want to derive a condition under which player 2 chooses A

[23] See Falk, Fehr and Fischbacher (2003) and references therein. In the experiments, player 2 is usually allowed to accept or reject even when player 1 chooses the 50:50 option. As expected, no players reject this fair offer in favor of a 0:0 outcome. Also, our analysis is virtually unaffected if the additional player 2 choice is admitted.

regardless of his type. The first question is whether there is any esteem benefit to be had from rejecting 20 in favor of 0, given that player 1 had no choice? We think not. Envy is not a socially approved characteristic. If anything it is a source of derision. Setting $s_L = s_H = \underline{s} = 0$, player 2 thus accepts the 80:20 outcome if

$$20 - (80 - 20)\theta_I > 0,$$

or $\theta_H < 1/3$.

Rejecting an intentionally unfair offer, on the other hand, is socially approved. Indeed, such vengefulness is sometimes considered an essential part of maintaining one's honor.[24] To capture this intuition, we assume that $s_L = \overline{s} > s_H = \underline{s}$ for player 2. At the same time, we maintain the assumption that $s_L = \underline{s} < s_H = \overline{s}$ for player 1. These assumptions starkly illustrate our point that salience depends on the situation. Player 1 has the opportunity to signal generosity, which might impress a generous player 2. Player 2, if called on to play, has the opportunity to signal vengefulness, which is meant to make an impression on a selfish player 1.[25]

In the voluntary ultimatum game, according to the proposed equilibrium, player 2 of type H rejects the 80:20 outcome if

$$g_H\theta_H\overline{s} > 20 - (80 - 20)\theta_H + g_H\theta_L\overline{s}$$

or equivalently

$$g_H\overline{s} > \frac{20 - 60\theta_H}{\theta_H - \theta_L}. \tag{29}$$

Analogously, player 2 of type L accepts the 80:20 outcome if

$$g_L\overline{s} < \frac{20 - 60\theta_L}{\theta_H - \theta_L}. \tag{30}$$

Player 1 of type H plays F if

$$\begin{aligned} & 50 + p_H^0 g_H\theta_H\overline{s} + (1 - p_H^0)g_H\theta_H\underline{s} \\ > \; & p_H^0 g_H\theta_L\overline{s} + (1 - p_H^0)(80 - (80 - 20)\theta_H + g_H\theta_L\underline{s}), \end{aligned}$$

[24] In the Viking age, a male Viking who did not take revenge on a thief was socially despised. Likewise there is a culture of honor in the southern states of the U.S., see Cohen et al. (1996).

[25] In comparison, reciprocity models based on psychological game theory would impose a change in the parameter θ_i depending on whether the player j behaves kindly or unkindly. We think that both models are realistic. We may wish misery on someone who hurts us, and we usually prefer the perpetrator know that the misery is due to retribution and not a bout of bad luck.

or equivalently (using $\underline{s} = 0$) if

$$p_H > \frac{30 - 60\theta_H}{g_L(\theta_H - \theta_L)\overline{s} + 80 - 60\theta_H} \tag{31}$$

Analogously, player 1 of type L plays U if

$$p_L < \frac{30 - 60\theta_L}{g_L(\theta_H - \theta_L)\overline{s} + 80 - 60\theta_L}. \tag{32}$$

Again, it is straightforward to check that there exist parameter vectors such that the proposed equilibrium exists. More interestingly, there is considerable overlap with the parameters that worked for the trust game. Suppose for example that $\theta_H = 1/5, \theta_L = 0, p_H = 2/5$ as above, and that $p_L = 1/3$. Then, the conditions on $(g_L, g_H, \bar{s})$ are weaker in the example depicted above. Note, however, that the slight reduction in p_L was necessary. If $p_L > 3/8$ even the type L player is better off with 50 than with the gamble on 80. Indeed, we conjecture that most of the heterogeneity in player 1 behavior in the mini-ultimatum game is due to heterogeneous expectations.

References

Archibald, W.P. (1974). Alternative explanations for self-fulfilling prophecies, *Psychological Bulletin* 81, 74-84.

Baron, J.N. and Kreps, D.M. (1999): *Strategic Human Resources: Frameworks for General Managers,* New York: John Wiley & Sons.

Barnard, C. (1938). *The Functions of the Executive.* Cambridge: Harvard University Press.

Baumeister, R.F. and Leary, M.R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin* 117, 497-529.

Bewley, T.F. (1999). *Why wages don't fall during a recession.* Cambridge: Harvard University Press.

Bénabou, R. and Tirole, J. (2003). Intrinsic and extrinsic motivation. *Review of Economic Studies* 70, 489-520.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review,* forthcoming.

Bernheim, D.B. (1994): A theory of conformity. *Journal of Political Economy* 102, 841-877.

Bohnet, I., Frey, B.S. and Huck, S. (2001). More order with less law. *American Political Science Review* 95, 131-144.

Brennan, G. and P. Pettit (2004). *The Economy of Esteem.* Oxford: Oxford University Press.

Camerer, C.F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton: Princeton University Press.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117, 817-869.

Charness, G, and Dufwenberg, M (2005). Promises and partnership, *Econometrica,* forthcoming.

Cho, I.-k. and Kreps, D.M. (1987): Signaling games and stable equilibria. *Quarterly Journal of Economics* 50, 179–221.

Coletti, A.L., Sedatole, K.L., and Towry, K.L (2005). The effect of control systems on trust and cooperation in collaborative environments. *Accounting Review* 80, 477-500.

Dawes, R.M. (1989). Statistical criteria for establishing a truly false consensus effect, *Journal of Experimental Social Psychology* 25, 1-17.

Deci, E.L. (1971). The effects of externally mediated rewards on intrinsic motivation, *Journal of Personality and Social Psychology* 18, 105-115.

Deci, E.L., Koestner, R.M. and Ryan, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation *Psychological Bulletin* 125, 627-668.

Denrell, J. (1998). Incentives and Hypocrisy. Chapter 2 in the PhD dissertation *Essays on the economics of vanity and career concerns.* Stockholm: Stockholm School of Economics.

Eden, D. (1984). Self-fulfilling prophecy as a management tool: Harnessing Pygmalion, *Academy of Management Review* 9, 64-73.

Eiser, J.R. and K. Bhavnani (1974). The effect of situational meaning on the behavior of subjects in the Prisoners' Dilemma game, *European Journal of Social Psychology* 4, 93-97.

Ellingsen, T. and Johannesson, M. (2005). Trust as an incentive, mimeo., Stockholm School of Economics

Engelmann, D. and Strobel, M. (2004): The false consensus effect: Deconstruction and reconstruction of an anomaly, mimeo.

Enzle, M.E. and S.C. Anderson (1993). Surveillant intentions and intrinsic motivation. *Journal of Personality and Social Psychology* 64, 257-266.

Falk, A. and Kosfeld, M. (2005). Distrust - the hidden cost of control. *American Economic Review,* forthcoming. (Originally IEW Working Paper No. 193, University of Zürich.)

Fehr, E. and Rockenback, B. (2003). Detrimental effects of sanctions on human altruism,*Nature* 422, 137-140.

Fehr, E. and Gächter, S. (2002). Do incentive contracts undermine voluntary cooperation? Working Paper No. 34, Institute for Empirical Research in Economics, University of Zürich.

Fehr, E., Kirchsteiger, G. and Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108, 437-459.

Fehr, E. and List, J. (2004). The hidden costs and returns of incentives - trust and trustworthiness among CEOs, *Journal of the European Economic Association* 2, 743-771.

Fehr, E. and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114, 817-868.

14,

Frey, B.S. (1993): Does monitoring increase work effort? The rivalry with trust and loyalty. *Economic Inquiry* 31, 663-670.

Frey, B.S. (1997). *Not Just for the Money. An Economic Theory of Personal Motivation.* Cheltenham: Edward Elgar.

Frey, B.S. and Oberholzer-Gee, F. (1997). The cost of price incentives: an empirical analysis of motivation crowding out. *American Economic Review* 87, 746-755.

Frey, B.S., Oberholzer-Gee, F. and Eichenberger (1996). The old lady visits your backyard: a tale of morals and markets. *Journal of Political Economy* 104, 1297-1313.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*, Cambridge: MIT Press.

Geanakopolos, J, Pearce, D, and Stacchetti, E. (1989). Psychological games and sequential rationality, *Games and Economic Behavior* 1, 60-79.

Gneezy, U. (2003). Do high wages lead to high profits? An experimental study of reciprocity using real effort. Mimeo, University of Chicago Graduate School of Business.

Gneezy, U. and Rustichini, A. (2000a). A fine is a price. *Journal of Legal Studies* 29, 1-17.

Gneezy, U. and Rustichini, A. (2000b). Pay enough or don't pay at all. *Quarterly Journal of Economics* 115, 791-810.

Houser, D., Xiao, E., McCabe, K., and Smith, V. (2005): When punishment fails: Research on sanctions, intentions and non-cooperation, mimeo., George Mason University.

Hume, D. *A Treatise of Human Nature,* London: John Noon.

Irlenbusch, B., and Sliwka, D. (2005): Incentives, decision frames, and motivation crowding out - an experimental investigation, IZA Discussion Paper No. 1758.

Kreps, D. (1997). Intrinsic motivation and extrinsic incentives. *American Economic Review, Papers and Proceedings* 87, 359-364.

Levin, I.P. et al. (1998). All frames are not created equal: A typology and critical analysis of framing effects, *Organizational Behavior and Human Decision Processes* 76, 149-188.

Levine, D.K. (1998). Modeling altruism and spitefulness in experiments, *Review of Economic Dynamics* 1, 593–622.

Liberman, V., Samuels, S.M. and Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin* 30, 1175-1185.

Livingston, J.S. (1969). Pygmalion in management. *Harvard Business Review* 47 (July–August), 81–89.

Macaulay, S. (1963): Non-contractual relations in business: A preliminary study. *Americal Sociological Review* 28, 55-70.

McCabe, K.A., Rigdon, M.L. and Smith, V.L. (2003): Positive reciprocity and intentions in trust games, *Journal of Economic Behavior and Organization* 52, 267-275.

McGregor, D. (1960): *The Human Side of Enterprise,* New York: McGraw-Hill.

Maslow, A.H. (1943). A theory of human motivation. *Psychological Review* 50, 370-396.

Mellström, C. and Johannesson, M. (2006). Crowding out in blood donation: Was Titmuss right? Mimeo. Stockholm School of Economics.

Pfeffer, J. (1994). *Competitive Advantage through People: Unleashing the Power of the Work Force* Boston: Harvard Business School Press.

Pillutla, M.M. and Chen, X. (1999): Social norms and cooperation in social dilemmas: The effects of context and feedback, *Organizational Behavior and Human Decision Processes* 78 (2), 81-103.

Rege, M. and Telle, K. (2004): The impact of social approval and framing on cooperation in public good situations, *Journal of Public Economics* 88, 1625-1644.

Rotemberg, J.J. (2006): Minimally acceptable altruism and the ultimatum game, mimeo. Harvard Business School.

Ross, L., Greene, D. and House, P. (1977): The "false-consensus" effect: An egocentric bias in social perception and attribution processes, *Journal of Experimental Social Psychology* 13, 279–301.

Seabright, P. (2004). Continuous preferences can cause discontinuous choices: An application to the impact of incentives on altruism. CEPR discussion paper 4322.

Sliwka, D. (2003): On the hidden costs of incentive schemes. IZA Discussion paper No. 844.

Smith, A. (1790). *The Theory of Moral Sentiments* 6th edition (first edition 1759). London: A. Millar.

Sobel, J. (2005). Interdependent preferences and reciprocity, *Journal of Economic Literature* 43, 392–436.

Titmuss, R. (1970). *The Gift Relationship: From Human Blood to Social Policy.* London: George Allen and Unwin.