



Munich Personal RePEc Archive

## **Intentions, Insincerity, and Prosocial Behavior**

Amegashie, J. Atsu

University of Guelph

12 January 2006

Online at <https://mpra.ub.uni-muenchen.de/3223/>

MPRA Paper No. 3223, posted 14 May 2007 UTC

## **Intentions, Insincerity, and Prosocial Behavior\***

J. Atsu Amegashie  
Department of Economics  
University of Guelph  
Guelph, Ontario  
Canada N1G 2W1

May 13, 2007

Email: [jamegash@uoguelph.ca](mailto:jamegash@uoguelph.ca)  
Phone: 519-824-4120 ext. 58945  
Fax: 519-763-8497

\*My thanks are due to Charles Becker, Claire Hill, Qiang Fu, Mike Hoy, Navin Kartik, Preston McAfee, Zane Spindler, Henry Thille and seminar participants at Dalhousie, Minnesota (School of Law) and Ryerson universities for very helpful comments and encouragement on earlier version of this paper. I thank Roland Benabou for drawing my attention to related work and SSHRC for financial support. This paper is a *significantly* revised version of CESifo Working paper No. 1757.

## Abstract

Consider a world with two people, 1 and 2, where person 1 (the proposer) may offer to help person 2 (the responder). The proposer may be altruistic towards the responder either out of a genuine desire to make her happy or out of guilt. The responder derives disutility from apparent acts of altruism motivated by guilt because she considers them to be insincere. She rejects some offers, depending on her beliefs about the proposer's type. I model this social interaction as a game with interdependent preference types under incomplete information where the responder cares about the *intentions* behind the proposer's prosocial behavior. I consider two recent formulations of *endogenous* guilt: *simple guilt* and *guilt from blame*. These formulations make the social interaction a psychological game. I find that the beliefs held by the players can lead to an equilibrium in which all offers are sincere and so no mutually beneficial trades are rejected, although the responder has incomplete information about the proposer's type. Equilibria with insincere offers are possible under *simple guilt* but are impossible under *guilt from blame*. I discuss intrinsic and instrumental motivations for sincerity. I also discuss the implications of insincerity aversion for co-operation, altruism, political correctness, choice of identity, and trust.

Keywords: guilt, intentions, insincerity, interdependent preference types, psychological game, social interaction.

JEL Classification: C73, J16, Z13.

*“Sincerity makes the very least person to be of more value than the most talented hypocrite.” - Charles Spurgeon*

*“To give real service you must add something which cannot be bought or measured with money, and that is sincerity and integrity.” - Douglas Adams*

## **1. Introduction**

In standard economics and game theory, only actions affect payoffs. Intentions are irrelevant. It is the final outcome that matters not the process. But there are clearly situations where *intentions* affect payoffs. The same action might induce different payoffs depending on the intentions of the parties or players. Indeed, intentions matter in important ways. It is the basis for the legal distinction between murder and manslaughter and partly explains the attitudes of certain groups towards racial profiling. In the former case, the same action (i.e., taking a person’s life) may attract a different punishment depending on whether the action is believed to be premeditated or not. In the latter case, a traveler at an airport or a motorist who is searched by the police may react differently depending on whether he believes that the search was random or was motivated by his race or religion.

The purpose of this paper is to analyze the following class of social interactions or prosocial behavior. Suppose someone offers to help you but you thought the offer might be motivated by guilt than a genuine desire to help, will you accept the offer? Suppose you are tolerated as opposed to being genuinely accepted by your peers and “friends”. In particular, suppose you are invited to a party, movie, dinner, etc not because your company is desired but because the inviter would feel guilty if she did not invite you, or you got a job at an elite institution but you wouldn’t

have been offered the job if you were not a minority, or someone gives you a present because they felt obliged to do so not because they really wanted to give you a present? Or a friend is expected to give you a phone call because you need her emotional support. If you have a caller ID and you think she is making the call reluctantly, will you answer the phone? If your boss, supervisor, or professor tells you to feel free to come talk to her anytime you encounter problems in your work, will you take her up on that offer, if you thought she was making the offer grudgingly? Does one's enjoyment from sex depend on whether her partner's intention is a long-term relationship or casual relationship? Will the answer affect the decision to accept or reject an offer into a sexual relationship? In all of these cases, it is conceivable that the *intention* behind the action will matter and hence will affect your payoffs. The intention will matter if the target of the offer is averse to insincerity.

The average reader may be able to relate to some of these situations from personal experience. These examples are common and interesting social interactions worthy of study. They are the basis of friendships and relationships at work, school, church, and in our daily lives. They determine who we choose to go to lunch with, play with, and in general socialize with. They determine the frequency and enjoyment of our social interactions.

One may assume that there is already some kind of superficial, implicit, or lower-level relationship between the two parties. For example, they may work at the same place or they may be neighbours. The question is: will the parties necessarily engage in mutually beneficial trades in a world where the sincerity of actions matter?

In what follows, I refer to the player who offers to help or extends an invitation to a social event (e.g., dinner) as the *proposer*, and the other player as the *responder*.

Gul and Pesendorfer (2005) show that intentions can also be modeled as stemming from interdependent type preferences. In the social interaction in the present paper, the proposer has a one dimensional type space: a social type which captures whether he likes the responder's company or not or whether he is helping the responder out of guilt or out of a genuine desire to help. The responder has interdependent type preferences because she has preferences over the proposer's social types.<sup>1</sup> Knowledge of the proposer's social type will help the responder determine the sincerity of the proposer's offer and hence determine the intention behind his offer. This requires that the responder forms beliefs about the proposer's social type.

Psychological game theory pioneered by Geanakoplos, Pearce, and Stacchetti (1989) models intentions as beliefs about beliefs, where players have belief dependent preferences. In an important extension, Rabin (1993) applied this framework to study intentions-based reciprocity or kindness. Falk, Fehr, and Fischbacher (2000, 2006), Brandts and Sola (2001), Falk et al. (2003), McCabe et al. (2003), and Offerman (2002) present experimental evidence which support the idea that intentions matter in reciprocal relationships.

This paper has both elements of psychological games and interdependent type preferences. Both approaches to modeling intentions in games are complementary.

---

<sup>1</sup> Gul and Pesendorfer (2005) consider more general cases where all players could have interdependent type preferences. This makes their model much more difficult to analyze because of the potential circularity of the formulation of types. We do not have this problem because the proposer does not have interdependent preferences. He does not care about the responder's type. His social preferences are independent of any characteristics of the responder. This makes our model much simpler. Similarly, in Kartik and McAfee (2006) only the voters have interdependent type preferences.

Indeed, Searle (1969) argues, in a very influential philosophical work, that sincerity is linked to a person's state of mind (i.e., his beliefs).<sup>2</sup> So insofar as this paper is concerned with the sincerity of a person's behavior or altruism, psychological game theory and/or a model with interdependent type preferences gives an appropriate analytical framework.

I consider *endogenous* guilt by allowing the proposer's cost of guilt to depend on the responder's expectation of an offer. In particular, I consider two formulations of such endogenous guilt due to Battigalli and Dufwenberg's (2006): *simple guilt* and *guilt from blame*. These formulations make the social interaction a dynamic psychological game with interdependent preference types. I characterize the equilibria of this social interaction.

I find that the beliefs of both parties play a key role in generating an equilibrium with sincere or insincere offers. In particular, the beliefs held by the players can lead to an equilibrium in which the responder does not reject mutually beneficial trades (i.e., sincere offers), although she has incomplete information about the responder's type. Equilibria with insincere offers are possible under *simple guilt* but are impossible under *guilt from blame*. I also discuss the implications of insincerity aversion for altruism, political correctness, choice of identity, and trust.

The rest of the paper is organized as follows. In the next section, I briefly discuss the role of guilt and insincerity in social interactions. I also discuss how my model differs from Benbou and Tirole (2006a). Section 3 presents a simple model of

---

<sup>2</sup> Walker (1978) discusses the limitations of this concept. My formal model has a precise definition of sincerity or insincerity and therefore is not subject to philosophical critiques of the definition of sincerity. In addition to Walker (1978), see Ridge (2006) for a definition, critical analysis, and extension of Searle's concept of sincerity.

social interaction under incomplete information in a psychological game and characterizes its equilibria. I discuss applications in section 4. Section 5 concludes the paper.

## **2. Guilt aversion and insincerity aversion**

In the social interaction studied, guilt plays an important role. As Baumeister, Stillwell, and Heartheron (1994, p. 243) note “... guilt is something that happens between people rather than just inside them. That is, guilt is an interpersonal phenomenon that is functionally and causally linked to communal relationships between people. The origins, functions, and processes of guilt all have important interpersonal aspects.” They continue “[T]his is not to deny that some experiences of guilt can take place in the privacy of one’s individual psyche, in social isolation. Still, many of those instances may be derivative of interpersonal processes and may reflect highly socialized individuals...”

Building on a well-known idea in psychology (e.g., Baumeister et al., 1994), Charness and Dufwenberg (2006) and Battigalli and Dufwenberg (2006) introduce the term *guilt aversion* to describe the behavior of people who suffer guilt if they believe that they have hurt another person because they did not meet that person’s expectation. It refers to the disutility felt from disappointing others or letting them down. They show how guilt aversion can sustain good or co-operative behavior.

In a related contribution, Huang (2003) examines how guilt can motivate securities professionals' behavior in their fiduciary relationships with their clients and the legal implications of guilt for the regulation of securities professionals. In a

different but related context, Huang and Wu (1994) examine show how remorse can lead to better social order. In both papers, the basic notion is that guilt provides an internal mechanism beyond the external mechanisms for legal compliance provided by private litigation, public enforcement, and formal sanctions. As noted by Huang (2003), the ability of guilt to regulate behavior is applicable to any situation that involves trust.

A very important difference in the present paper is that insincere offers, motivated by guilt aversion, impose a cost (disutility) on the responder. This insincerity-induced disutility or insincerity aversion produces an effect that is absent in previous works on guilt aversion (e.g., Huang and Wu, 1994; Dufwenberg, 2002; Huang, 2003; and Charness and Dufwenberg, 2006). In particular, while guilt aversion in these papers can sustain cooperation or good behavior, the likelihood of such cooperation may fall because guilt-induced cooperation may be *perceived* by one party as insincere and hence may be rejected because this party dislikes insincere or forced cooperation.

A person may be a sincerity pragmatist where in certain contexts, she may care about sincerity but in others, she may not. She may have an *instrumental* value for sincerity in certain situations but may have an *intrinsic* value for sincerity in others. This kind of cost-benefit calculus by such sincerity pragmatists is alluded to by the Nobel laureate, Albert Camus when he opined that “[H]ow can sincerity be a condition of friendship? A taste for truth at any cost is a passion which spares nothing.”<sup>3</sup> In the same vein, Kang (2003a, 2003b) makes a case for insincerity in a

---

<sup>3</sup> This quote and the two others at the beginning of this article are taken from <http://www.brainyquote.com>.

democracy. He argues that insincerity in public discourse is necessary for tolerance and mutual co-existence in liberal democracies.<sup>4</sup> I return to this issue when I discuss political correctness and other applications in section 4.

Still on the preceding point, it is conceivable that in financial matters of the kind analyzed in Huang (2003), a person may have an instrumental value for sincerity but in non-financial matters, the same person may more likely have an intrinsic value for sincerity. So in financial matters, guilt-aversion is likely to sustain co-operation because insincerity aversion is less likely to matter.

The preceding point calls for reasons why a person may be averse to insincerity or have a preference for sincerity, and why such a preference may be driven by instrumental or intrinsic motivations. Notice that in the above papers (e.g., Dufwenberg, 2002; Huang, 2003; and Charness and Dufwenberg, 2006), only one player is guilt-averse or only one player's guilt aversion is relevant for the analysis. My model could be seen as one in which both players are guilt-averse but for different reasons. Under this interpretation, the proposer extends insincere offers to assuage his guilt while the responder dislikes them because she feels guilty if she believes that she is forcing someone to accept her or be nice to her out of guilt. While the proposer feels guilty for disappointing others, the responder feels guilty if she believes that she is manipulating the proposer's guilt for her personal gain. The responder does not feel guilty if she rejects an offer.

People may be insincerity averse if they believe that the intention behind an offer or an apparent prosocial behavior is to make them feel morally obliged to

---

<sup>4</sup> Markovits (2006) presents a related argument.

reciprocate in the future or requires them to stroke their benefactor's ego by being held to an emotional ransom of a perpetual demonstration of gratitude.

Insincerity-aversion may also stem from the belief that those who act out of guilt are ultimately not trust-worthy.<sup>5</sup> They can fake their behavior for only a short while but eventually their true feelings and behavior will come out. So the responder may be insincerity-averse because she wants to interact with people that she can trust. To avoid the cost of being unpleasantly surprised, insincerity-averse people will terminate cooperation sooner than later.

Related to the previous point is the observation that the desire to know the sincerity of others in socio-economic relationships may stem from the fact that knowledge of such sincerity or the degree thereof may determine the effort that an insincerity-averse person puts into the relationship.<sup>6</sup> The cost of insincerity is then the cost of over-investing in the relationship based on the erroneous information or presumption that the person being dealt with was sincere. Hill and O'Hara (2007) examine how the law should intervene to either promote more accurate trust levels or to mitigate the costs of mistaken assessments in contractual or non-contractual relationships.

Ayres and Klass (2005, 2004) present a lucid and interesting examination of the legal implications of insincere promises and misrepresented intent. A promise is insincere if the promisor never intended to fulfill the promise. According to Ayres and Klass (2005), a promisee cares about the sincerity of the promisor because "... breach-of-contract damages are not fully compensatory." If such damages were fully

---

<sup>5</sup> There is now a growing literature on trust in economics (see, for example, Laibson et al., 2000; Charness and Dufwenberg, 2006; and the references cited therein)

<sup>6</sup> I thank Claire Hill for this point.

compensatory, a promisee will not care about the sincerity of the promisor. This is consistent with our earlier point that a person may be insincerity-averse because dealing with an insincere person is costly.<sup>7</sup> Hill and O'Hara (2007) observe that full compensation for breach-of-contract damages may lead to excessive levels of trust in contracting relationships.

Benabou and Tirole (2006a) study a model in which prosocial behavior (e.g., contribution to a public good) can yield different payoffs depending on whether the such behavior is perceived to have been motivated by altruism (e.g., would like to remain anonymous) or by a desire for public show or good public image.<sup>8</sup> Anonymous donors may then be perceived as being more sincere than donors who have a strong preference to have their contributions publicized. In their model, a person's *reputational* payoff is increasing in his intrinsic motivation as perceived by others but is decreasing in perceived extrinsic (monetary) motivation. Insincerity-aversion by others (e.g., beneficiaries) may cause benefactors (donors) to contribute less because their altruism (intrinsic motivation) is called into question leading to a dampening effect on their

---

<sup>7</sup> To be sure, any moral hazard behavior in a principal-agent relationship could be considered as insincere behavior. However, in a standard principal-agent model, the principal would not derive any disutility from an agent who exerts a high effort or desists from moral hazard behavior out of guilt. The principal only cares about actions not intentions. And if the principal cares about intentions, it is only in an instrumental sense insofar intentions affect actions. In contrast, my model is also applicable to situations where intentions have intrinsic value for people and therefore the same action will yield different payoffs depending on the intention behind it.

<sup>8</sup> On the latter motivation, see also Glazer and Konrad (1996). For an interesting and extensive survey of the literature on altruism and philanthropy, see Andreoni (2006).

incentive to engage in prosocial behavior.<sup>9,10</sup>

Kartik and McAfee (2006) study a game of electoral competition in which voters have preferences over the character and campaign promises of politicians. If two politicians propose the same policy, voters get a higher utility from the candidate who is perceived to have character. This higher utility does not stem from a belief that the politician with character will honor his campaign promise while the politician without character will renege on his promise. It is simply due to the fact that voters directly value character *per se*. Hence the voters have an intrinsic value, as opposed to an instrumental value, for the sincerity of politicians.

As a reason for why voters may have this preference, Kartik and McAfee (2006) argue that voters may not like politicians who are willing to pander or say anything in order to get votes. They like a politician who will run on a platform that he truly believes in and thinks is best for the country, even if that will not get him votes. In this sense, voters' concern about character is similar to a concern for

---

<sup>9</sup> Psychologists have long argued that monetary incentives can crowd out intrinsic motivation for desired behavior. Economists have recently begun to pay attention to this possible effect (see Frey, 1997; Benabou and Tirole, 2006a, and the references cited therein).

<sup>10</sup> There are differences between my model and Benabou and Tirole (2006a). First, while donors in their model care about the opinion of others, this concern does not depend on the endogenous expected payoff of others. Agents engaged in prosocial behavior do not have belief-dependent preferences, so their model is not a psychological game. Second, in their model, an agent derives disutility from being perceived as motivated by extrinsic incentives even if this is not known with certainty. In my model, the proposer who extends an insincere offer feels guilty *if and only if* the responder can tell with certainty that his offer was insincere. Third, in my model, the responder rejects offers with positive probability. This is a punishment to the proposer since some sincere offers are rejected with positive probability. In contrast, all offers in Benabou and Tirole (2006a) are accepted but donors are punished through the expected loss of *reputational* payoff. Fourth, in my model, the disutility from insincere offers is captured in the payoff of the intended beneficiary, while in Benabou and Tirole (2006a), it enters the payoff of the donor. I return to this difference in section 4. Fifth, Benabou and Tirole (2006a) do not focus on guilt or different formulations of guilt. And sixth, the choice of prosocial behavior is binary in my model, while it is continuous or binary in Benabou and Tirole (2006a).

sincerity.<sup>11</sup>

### **3. A Game of Social Interaction with Guilt**

In this section, I consider a very simple model to examine the several examples of social interaction mentioned in section 1, where the sincerity of actions matters. While the model is applicable to those examples, I use one or two specific examples in this section for the sake of exposition. In particular, I focus on situations where the proposer has the option of helping the responder in an activity or inviting her to some social event like a party, dinner, movie, etc. In section 4, I demonstrate how this simple model can be adapted to the issue of political correctness.

Consider two people, 1 and 2. I use male pronouns for player 1 and female pronouns for player 2. Player 1 has the option of proposing to help player 2 in some activity or to invite her to some social event (e.g., a movie or dinner). If player 2 believes that player 1 genuinely wants her company or wants to help her, she gets a utility,  $v > 0$ , given that she accepted player 1's offer. If she believes that player 1's offer is insincere (i.e., his help is out of guilt or does not want her company), she incurs a psychic cost of  $\theta > 0$ , given that she accepted player 1's offer.

---

<sup>11</sup> The voters in Kartik and McAfee (2006) have interdependent type preferences as in Gul and Pesendorfer (2005).

Player 1 need not show that his offer is out of guilt or dislikes player 2's company. It is sufficient for player 2 to *believe* that player 1's offer is insincere. It is player 2's inference about player 1's intentions that matters. Therefore, the *same* action (i.e., offer) by player 1 could give player 2 *different* payoffs depending on her beliefs about player 1's intentions. Given that given  $v > 0$ , player 2 would accept any offer from player 1 if she did not care about player 1's intentions.

Player 2 may simply have an intrinsic value for sincerity or may have an instrumental value for sincerity. Under the latter interpretation, one may imagine that accepting player 2's offer requires an exogenous investment of time and effort into an activity or relationship, which yields a net benefit of  $v$  if player 1 is sincere but a cost of  $\theta$ , if player 1 is insincere.

Let  $v$  be a random variable that is commonly known to be continuously distributed on  $[\underline{v}, \tilde{v}]$  with density  $f(v)$  and corresponding distribution function,  $F(v)$ ,  $\underline{v} > 0$ . I assume that  $F(v)$  is a strictly increasing function.

Suppose that nature gives player 1 a social type. If person 1 is of social type  $w_H > 0$ , then he derives a psychic *benefit* (joy) of  $w_H$  from helping player 2. If he is of social type  $w_L$ , then he incurs a *cost* of  $w_L > 0$  of helping player 2. Let the probability distribution of these types be such that  $\Pr(w_H) = p$  and  $\Pr(w_L) = 1-p$ ,  $p \in (0,1)$ . Furthermore, player 1 feels guilty, if he does not offer to help player 2. I assume that player suffers a guilt cost denoted by  $G$ .

An offer is insincere if it is extended by player 1 of type  $w_L$  and it is sincere if it is extended by player 1 of type  $w_H$ . These social types are similar to

Benabou and Tirole's (2006a) intrinsic and extrinsic valuations for prosocial behavior.

As in Battigalli and Dufwenberg (2005, 2006), player 1's guilt depends on the extent to which he believes that he has disappointed player 2. In particular, I assume that  $G = \alpha D_2$ , where  $D_2$  is the disappointment felt by player 2 when player 1 does not offer to help her. I shall endogenize  $D_2$  but it is easier to do so when part of the solution to the game has been discussed. This is because  $D_2$  depends on endogenous second-order beliefs making the game a dynamic psychological game (Battigalli and Dufwenberg, 2005, 2006).

After observing his social type, player 1 has two actions: offer to help (I) or do not offer to help (N). Player 2 has two actions: accept (A) or reject (R) an offer from player 1. The game is sequential. Player 1 is the first-mover and player 2 is the second-mover.

It is important to note that player 1 does *not* feel guilty so long as he offers to help player 2, even if he does not want player 2 to accept his offer. If his social type is  $w_L$ , he might offer to help player 2 and if player 2 rejects it, then he suffers no guilt. While the motivation for this behavior may be straightforward, it may be helpful to elaborate further. One explanation is that player 1 does not feel guilty because he can justify his behavior on the grounds that he, after all, took the risk of offering to help player 2.<sup>12</sup> This is what Baumeister et al. (1994) refer to as the *deconstruction* of guilt.

---

<sup>12</sup> If player 1's social type is  $w_L$ , then offering to help player 2 is risky because she might accept his offer. In Benabou and Tirole (2006a) such an agent suffers some disutility even if it is not known with certainty that his offer was insincere.

Of course, if player 2 could definitely tell that player 1 extended an insincere offer with the goal of getting his offer rejected, then a rejection of an insincere offer from player 2 could make player 1 feel guilty.<sup>13</sup> However, due to incomplete information, player 2 cannot, *in general*, be certain of the insincerity or otherwise of player 1's offer. So due to incomplete information, the rejection of an insincere offer does not make player 1 feel guilty.<sup>14</sup>

Player 1's payoff is

- (a)  $u_1 = w_H$ , if he plays I, his social type is  $w_H$ , and player 2 plays A
- (b)  $u_1 = -w_L$ , if he plays I, his social type is  $w_L$ , and player 2 plays A
- (c)  $u_1 = -G$ , if he plays N
- (d)  $u_1 = 0$ , if he plays I and player 2 plays R

Player 2's payoff, assuming that she knows player 1's social type, is

- (i)  $u_2 = -\theta$ , if she plays A, given that player 1 of type  $w_L$  played I
- (ii)  $u_2 = v$ , if she plays A, given that player 1 of type  $w_H$  played I
- (iii)  $u_2 = 0$ , if she plays R

The players have common priors. All this information is common knowledge. However, player 1's social type is his private information.

---

<sup>13</sup> Of course, player 1's guilt need not depend on player 2's words or actions. This is at the heart of the distinction between *simple guilt* and *guilt from blame* considered in this paper. For an interesting discussion of the social importance of guilt, see Baumeister et al. (1994).

<sup>14</sup> It must clearly be emphasized that player 1's guilt need not be dependent on what player 2 knows. It may sometimes appear that what we refer to as *guilt* should actually be called *shame*. The distinction does not affect our analysis or results. However, our use of the term guilt is consistent with Baumeister et al. (1994, p. 243) who argue that "... guilt is something that happens between people ... and is causally linked to communal relationships between people". Tangney (1992) also found that guilt had strong interpersonal causation.

### 3.1 Equilibrium analysis with simple guilt

I look for psychological perfect Bayesian equilibria (PPBE) of this game.<sup>15</sup> Note that if player 1 plays N, then player 2 does not have to respond. So player 2's behavior is restricted to her response when player 1 plays I. For player 1, I consider both decisions (i.e., offer to help or not offer to help).

Let  $\sigma \in [0,1]$  be the probability that player 2 rejects an offer from player 1 and let  $\lambda \in [0,1]$  be the probability that player 1 will offer to help player 2 when his social type is  $w_L$ . Notice that when player 1's social type is  $w_H$ , it trivially follows that he will offer to help player 2 with certainty.

Given player 1's strategy,<sup>16</sup> player 2 computes the posterior probabilities

$$\rho_L \equiv \rho(w_L|I) = \frac{\rho(I|w_L)\Pr(w_L)}{\sum_{i=L,H} \rho(I|w_i)\Pr(w_i)} = \frac{\lambda(1-p)}{\lambda(1-p) + p} \quad (1)$$

and

$$\rho_H \equiv \rho(w_H|I) = \frac{p}{\lambda(1-p) + p}. \quad (2)$$

Note that  $\rho(w_H|I) > p$  for  $\lambda \in [0,1]$ .

Player 2's expected equilibrium payoff if she *accepts* an offer from player 1 could be written as

---

<sup>15</sup>A psychological perfect Bayesian equilibrium (PPBE) is a PBE with the additional requirement that players' expectations are correct in equilibrium (i.e., endogenous first-order and higher-order beliefs are correct in equilibrium).

<sup>16</sup> Of course, we shall find player 1's *optimal* offer probabilities.

$$U_2(\lambda) = \rho(w_H|I)v - \rho(w_L|I)\theta \quad (3)^{17}$$

Player 2 rejects an offer, if  $U_2(\lambda) < 0$ . It follows that player 2 of type

$$\hat{v}(\lambda) = \frac{\rho(w_L|I)\theta}{\rho(w_H|I)} = \frac{\lambda(1-p)\theta}{p}$$

is indifferent between accepting or rejecting an

offer. Therefore,

$$\sigma = \int_{\underline{v}}^{\hat{v}(\lambda)} f(v)dv = F(\hat{v}(\lambda)), \quad (4)$$

Then it immediately follows that  $\partial\sigma/\partial\lambda > 0$ . Hence, player 2 increases her rejection probability, if she believes that the probability of insincere offers is higher.

I now endogenize player 2's level of disappointment. As in Battigalli and Dufwenberg (2006), player 2's disappointment is a function of the difference between his expected payoff and his actual payoff.

---

<sup>17</sup> Clearly, player 2 makes her decision based on her expected payoff. *Ex post*, the interaction between player 1 and player 2, for example at a social event like a movie or party, can be described in the following three different ways: (a) player 1 is lukewarm but civil towards player 2 when his social type is  $w_L$ , given that player 2 accepted his offer. Player 1 sees nothing wrong with a lukewarm and civil interaction, so he feels no guilt. However, player 2 suffers a disutility of  $\theta$  from player 1's lukewarm attitude towards her, when his social type is  $w_L$  and enjoys a payoff of  $v$  from player 1's warm attitude when his social type is  $w_H$ . Then consistent with our analysis player 2's expected payoff, when she accepts an offer, is the expression in equation (3); (b) having extended an insincere offer (i.e., his social type is  $w_L$ ), player 1 pretends that he enjoys helping player 2 given that player 2 has accepted his offer. Player 2, having accepted an offer, is aware of this potential pretence, but enjoys the interaction based on the fact that the expected benefits are not smaller than the expected costs. Obviously, this is the case when the expected payoff in equation (3) is non-negative and this is indeed her expected payoff, *ex post*; and (c) having accepted an offer, player 2 is able to convince herself that player 1 genuinely enjoys helping her company, even if player 1 is actually pretending. However, this does not change her *ex ante* decision rule by causing her to necessarily accept every offer. This is because her *present self* (pre-acceptance self) is not able to anticipate the behavior of her *future self* (post-acceptance self). This is similar to the behavior of naïve types (naifs) in Rabin and O'Donoghue (1999).

Let  $\lambda_1$  be player 2's first-order belief of  $\lambda$  and player 1's belief (second-order) of  $\lambda_1$  be  $\lambda_2$ .<sup>18</sup>

If player 2 does not get an offer or assistance from player 1, her actual payoff is zero. If she gets an offer and she accepts it, then she expects a payoff of  $U_2(\lambda) > 0$ . So if she plans on accepting an offer, then her disappointment, given that she did not get an offer, is  $U_2(\lambda) - 0 > 0$ . On the other hand, if she rejects an offer, she must believe that her payoff if she had accepted it would have been  $U_2(\lambda) < 0$ . So in this case, her disappointment from a non-offer is  $U_2(\lambda) - 0 < 0$ . So she actually suffers no disappointment from not getting an offer.

Based on the above discussion, we may write player 2's disappointment as

$$D_2(\lambda_1, v, \theta) = \max [U_2(\lambda_1) - 0, 0] \quad (5)$$

Player 1 needs to determine his optimal offer probabilities,  $\lambda$ . But to do so he has to form beliefs about  $\lambda_1$  since  $D_2(\lambda_1, v, \theta)$  is a function of  $\lambda_1$ . Hence player 1's optimal choice of  $\lambda$  depends on his second-order beliefs (i.e.,  $\lambda_2$ ) of  $\lambda$  and thus on player 2's expectation of the equilibrium play of the game. Player 1's payoff does not only depend on player 2's actions but also depends on his endogenous beliefs of player 2's beliefs. Indeed, since  $\rho(w_H|I)$  is a function of  $\lambda$ , it follows that player 1's payoff depends on his beliefs of player 2's updated beliefs of his social type.

---

<sup>18</sup> As in Battigalli and Dufwenberg (2006), we consider beliefs, at most, of the fourth order.

Therefore, we may write player 1's cost of guilt as

$$G = \alpha D_2(\lambda_2, v, \theta), \quad (6)$$

where  $\alpha$  is common knowledge and measures player 1's sensitivity to guilt. The formulation in (6) makes this game a psychological game in the sense of Battigalli and Dufwenberg (2005, 2006), where player 1 has belief-dependent preferences about player 2's belief about his social type. Battigalli and Dufwenberg (2006) refer to this formulation of guilt as *simple guilt*.

It is important to reiterate that player 1 might feel guilty if and only if he does not offer to help player 2, and does not feel guilty if player 2 rejects his offer. Indeed, as argued above, player 2 feels no disappointment if she rejects player 1's offer. Therefore, the formulation of guilt based on equation (6) is consistent with Battigalli and Dufwenberg's (2006, note 3) argument that a player "... cannot be guilty for others' behavior." So if player 2 rejects player 1's offer, player 1 cannot assume any responsibility for player 2's reluctance to accept his offer.

Player 1 must also know  $v$  and  $\theta$  in order to determine his optimal  $\lambda$ . In what follows, we assume that player 1 knows  $\theta$  and knows the distribution of  $v$ . So he uses the expected value of  $v$  (i.e.,  $\bar{v}$ ) in his decision making. That is, he assumes that player 2's disappointment from a non-offer is

$$D_2 = \max[\rho_H \bar{v} - (1 - \rho_H)\theta, 0], \text{ where } \rho_H \bar{v} - (1 - \rho_H)\theta = \int_{\underline{v}}^{\bar{v}} U_2(\cdot) dF(v).$$

I characterize the equilibria of this game under the following three exhaustive cases: (a)  $\sigma < 1 - G/w_L$ , (b)  $\sigma = 1 - G/w_L$ , and (c)  $\sigma > 1 - G/w_L$ .

Case (a):  $\sigma < 1 - G/w_L$

Suppose  $\lambda_1 = \lambda_2 = 0$ . Then player 2 believes that player 1 will not offer to help her when his social type is  $w_L$ . And also player 1 believes that player 2 believes that player 1 will not offer to help her when his social type is  $w_L$ .

Suppose also that  $(1-\sigma)w_L > \alpha \{\max[\rho_H \bar{v} - (1-\rho_H)\theta, 0]\} = \alpha(\rho_H \bar{v} - (1-\rho_H)\theta) > 0$ . Then player 1's optimal response is  $\lambda = 0$ . Note that

$(1-\sigma)w_L > \alpha(\rho_H \bar{v} - (1-\rho_H)\theta) = G > 0$  holds if  $\bar{v}$  is sufficiently low and/or  $\theta$  is sufficiently high and/or  $\alpha$  is sufficiently low. This gives the following proposition:

**Proposition 1:** *If player 1 suffers from simple guilt, and (a) player 2 does not expect any insincere offers, (b) player 1 believes that player 2 does not expect any insincere offers, (c) player 2 has a sufficiently low valuation for sincere offers, and/or a sufficiently high disutility for insincere offers and/or player 1 has a sufficiently low sensitivity to guilt, then there exists a PPBE in the social interaction game where all offers are sincere and player 2 accepts all offers.*

Case (b):  $\sigma = 1 - G/w_L$

Now consider some  $\lambda_2 \in (0, 1)$ . Then player 1 believes that player 2 believes that player 1 will offer to help her with probability  $\lambda_2 \in (0, 1)$  when his social type is  $w_L$ . Suppose also that  $\rho_H \bar{v} - (1-\rho_H)\theta > 0$ , given  $\lambda_2$ . Then player 1 believes that player 2 will suffer a disappointment of

$D_2 = \max[\rho_H \bar{v} - (1-\rho_H)\theta - 0, 0] = \rho_H \bar{v} - (1-\rho_H)\theta > 0$ , if he does not offer to

help her. Then  $G(\lambda) = \alpha(\rho_H \bar{v} - (1-\rho_H)\theta) > 0$ . Therefore, given that player 1 is of social type  $w_L$ , he is indifferent between extending an offer and not doing so if

$-(1 - \sigma)w_L = -\alpha(\rho_H \bar{v} - (1 - \rho_H)\theta)$ . This gives  $\sigma = 1 - \alpha(\rho_H \bar{v} - (1 - \rho_H)\theta)/w_L$ .

It follows from (4) that that  $F(\hat{v}(\lambda)) = \sigma = 1 - G(\lambda)/w_L$ . The solution to this equation gives  $\lambda^*$ . Assume that  $\lambda^* \in (0, 1)$ . Imposing consistency of beliefs gives  $\lambda_1 = \lambda_2 = \lambda^*$ . This gives the following proposition:

**Proposition 2:** *If player 1 suffers from simple guilt, then there exists a PPBE in which player 1 offers to help player 2 if his social type is  $w_H$ . If his social type is  $w_L$  and  $\sigma^* = 1 - g/w_L$ , then he offers to help player 2 with probability  $\lambda^* \in (0, 1)$ . Player 2 rejects player 1's offer with probability  $\sigma^*$ . In this equilibrium, player 2 of type  $v \leq \hat{v}(\lambda^*)$  rejects offers, and player 2 of type  $v > \hat{v}(\lambda^*)$  accepts offers.*

Case (c):  $\sigma > 1 - G/w_L$

Now suppose that  $\lambda_1 = \lambda_2 = \lambda^* = 1$ . Also, assume that  $-(1 - \sigma)w_L > -\alpha(\rho_H \bar{v} - (1 - \rho_H)\theta)$ , where  $\rho_H = p$  and  $\rho_H \bar{v} - (1 - \rho_H)\theta > 0$ . Then the following proposition holds:

**Proposition 3:** *If player 1 suffers from simple guilt, then there exists a PPBE in which player 1 always offers to help player 2 and player 2 rejects player 1's offer with probability  $\sigma^{**} = F(\hat{v}(1)) > 1 - g/w_L > 0$ , where  $\hat{v}(1) = (1 - p)\theta / p$ . In this equilibrium, player 2 of type  $v \leq \hat{v}(1)$  rejects offers, and player 2 of type  $v > \hat{v}(1)$  accepts offers.*

Note from (2) that  $\rho_H$  is decreasing in  $\lambda$ . So comparing proposition 1 where  $\lambda_1 = \lambda_2 = \lambda^* = 0$  with propositions 2 and 3, where both players believe that

an insincere offer will be extended with some *positive* probability, it follows that  $\tilde{\rho}_H > \hat{\rho}_H$ , where  $\tilde{\rho}_H$  is player 2's belief that player 1 is of social type  $w_H$  in proposition 1 and is also player 1's belief of player 2's belief.  $\hat{\rho}_H$  is similarly defined for propositions 2 and 3.

For proposition 1 to hold, we required  $(1-\sigma)w_L > \alpha(\tilde{\rho}_H\bar{v} - (1-\tilde{\rho}_H)\theta) > 0$ . For propositions 2 and 3, we required  $(1-\sigma)w_L \leq \alpha(\hat{\rho}_H\bar{v} - (1-\hat{\rho}_H)\theta)$  to obtain equilibria with insincere offers. Note that  $\tilde{\rho}_H\bar{v} - (1-\tilde{\rho}_H)\theta > \hat{\rho}_H\bar{v} - (1-\hat{\rho}_H)\theta > 0$  since  $\tilde{\rho}_H > \hat{\rho}_H$ . It follows that if  $(1-\sigma)w_L > \alpha(\tilde{\rho}_H\bar{v} - (1-\tilde{\rho}_H)\theta)$  holds, then  $(1-\sigma)w_L \leq \alpha(\hat{\rho}_H\bar{v} - (1-\hat{\rho}_H)\theta)$  cannot hold and vice versa. An implication is that the equilibria above are unique given the specified conditions.<sup>19</sup> It also gives the following corollary:

**Corollary 1:** *If players 1 and 2 hold beliefs that generate the equilibrium with only sincere offers, then there exists no beliefs that can generate an equilibrium with insincere offers. Conversely, if they hold beliefs that generate an equilibrium with insincere offers, then there exists no beliefs that can generate an equilibrium with only sincere offers.*

Note that we do not have to worry about out-of-equilibrium beliefs in any of the equilibria above. Suppose that in proposition 3, player 2 observed an out-of-equilibrium action of N by player 1. Then the game ends, so player 2's beliefs

---

<sup>19</sup> Note that if  $\sigma > 1 - G/w_L$ , then  $\lambda = 1$  is player 1's unique response. If  $\sigma < 1 - G/w_L$ , then  $\lambda = 0$  is his unique response. Since  $F(\hat{v}(\lambda))$  is monotonic in  $\lambda$ , it follows that  $\sigma = F(\hat{v}(\lambda))$  is unique given that  $\lambda$  is unique. Suppose that when  $\sigma = 1 - G/w_L$ , player 1's optimal  $\lambda$  in the mixed strategy is unique. Then  $\sigma$  is also unique.

are irrelevant. In propositions 1 and 2 player 1 plays either N or I with positive probability, so player 2 will continue to update her beliefs using Bayes' rule.

### 3.2 *Guilt from blame*

In the *simple guilt* formulation of Battigalli and Dufwenberg (2006), a player feels guilty as a result of the disappointment felt by others, even if the affected people do not blame him for his actions. Battigalli and Dufwenberg (2006) also consider another formulation of guilt, where a player who has disappointed another player feels guilty depending on the extent to which the affected player blames him for his actions. They refer to this as *guilt from blame*.

One can interpret *simple guilt* as the guilt felt from blaming one's own self and *guilt from blame* as the guilt felt from being blamed by others. In the analysis below and in Battigalli and Dufwenberg (2006), it is implicitly assumed that for the same level of disappointment incurred by player 2, player 1's guilt sensitivity (i.e.,  $\alpha$ ) is the same whether he blames himself or he is blamed by player 2. This may not be the case in practice, although it seems to be the correct methodological assumption to make. In that way, differences in equilibrium behavior from these formulations of guilt will only be attributed to the differences in the strategic incentives that they induce as opposed to differences in a player's distaste for feelings of guilt.

Note that there can be no *guilt from blame* when player 1 offers to help player 2 or when player 2 rejects player 1's offer. So *guilt from blame* is only possible when player 1 does not offer to help player 2.

In what follows, odd-numbered-order beliefs apply to player 2 and even-numbered-order beliefs apply to player 1. Following Battigalli and Dufwenberg's (2006), player 2's blame of player 1 depends on her inference of the extent to which player 1 is willing to disappoint her. But this inference must depend on her beliefs of player 1's second-order beliefs,  $\lambda_2$ . This is player 2's third-order belief,  $\lambda_3$ . This is because her inference of the extent to which player 1 is willing to disappoint her depends on her beliefs of player 1's beliefs of her expected payoff from accepting an offer. But for player 1 to know the blame that player 2 will apportion to him, he must have beliefs about  $\lambda_3$ . This is his fourth-order belief,  $\lambda_4$ .

However, in our model, the case of *guilt from blame* turns out to be much easier to analyze. To this see this, observe that player 1 feels no guilt if he does not offer to help player 2 because player 2 will not blame him for doing so. Player 2 understands that if player 1 does not offer to help her, then it must be the case that his social type is  $w_L$ . And since player 2 dislikes insincere offers, she does not get disappointed and so does not blame player 1.<sup>20</sup> So under *guilt from blame*, player 1 will not extend insincere offers. Consistency of beliefs requires  $\lambda_j = 0$  for  $j = 1, 2, 3, 4$ . Clearly, propositions 2 and 3 are not possible under *guilt from blame*. The proposition below then follows:

**Proposition 4:** *If player 1 suffers from guilt from blame, then there is a unique equilibrium in which all offers are sincere and no offers are rejected.*

Unlike *guilt from blame*, player 1 blames himself under *simple guilt* for not offering to help player 2, even if his social type is  $w_L$ . This makes sense

---

<sup>20</sup> She does not infer that player 1 wants to disappoint her. Her inference is that player 1 is not extending her an offer because player 1 knows that she does not like insincere offers.

because it is not uncommon for people to feel guilty or blame themselves for having certain antisocial preferences or for being of an antisocial type.

#### **4. Discussion and Applications**

Proposition 1 is interesting because it shows that even if player 2 is suspicious of player 1's intentions, there are beliefs which can sustain an equilibrium where mutually beneficial trades are never rejected. When player 1's social type is  $w_H$ , he derives satisfaction from helping player 2 and player 2 also derives a benefit from receiving this help. If player 2 rejects his offers in this case due to incomplete information, then there are clearly gains from trade that are not realized. Proposition 1 shows that this inefficient outcome can be precluded given the appropriate beliefs, even though player 2 has incomplete information. This clearly has implications for the ability of people to genuinely communicate their altruistic intentions to others.

But proposition 4 relative to the propositions under *simple guilt* is even more interesting. Unlike *simple guilt*, proposition 4 shows that under *guilt from blame*, there cannot be equilibria with insincere offers. This accords with intuition because if player 2 is averse to insincerity and if player 1 is sensitive to blame from player 2, then player 2 will place the *minimal* blame possible on player 1 mindful of the fact that it is player 1's guilt aversion which causes him to extend insincere offers. With such minimal blame, player 1 has no incentive to extend insincere offers in order to assuage his guilt. On the other hand, if player 2's

blame has no effect on player 1's guilt (i.e., *simple guilt*), then player 2 cannot guarantee sincerity.

Notice that player 1 offers to help player 2 if he believes that player 2 expects an offer and will be sufficiently disappointed otherwise. That is, player 1 offers to help player 2 if he believes that player 2 expects an offer and  $\rho_H \bar{v} - (1 - \rho_H)\theta > 0$  is sufficiently high. This accords very well with casual empiricism. The emotional cost (i.e., guilt) of disappointing someone *coupled with* that person's expectations could force us to be kind to them, although we would have preferred to act otherwise.

The preceding observation applies generally to the way we tolerate others who we would otherwise not have tolerated. In some cases, we do so only because such people *expect* to be treated with respect.

Except for the equilibria in propositions 1 and 4, all other equilibria involve some insincere offers due to player 1's guilt aversion. One may then conclude that guilt breeds insincerity. While this is sometimes true, proposition 1, for example, suggests that this is not always the case. In addition to guilt aversion, the players' expectations or beliefs play a crucial role in generating an equilibrium with insincere or sincere offers. If player 2 expects insincere offers and player 1 believes that player 2 expects an insincere offer, then these beliefs coupled with a high guilt sensitivity (i.e., high  $\alpha$ ) may indeed lead to an equilibrium with insincere offers. On the other hand, if player 2 expects sincere offers and player 1 believes that player 2 expects sincere offers, then these beliefs coupled low guilt

sensitivity (i.e., low  $\alpha$ ) yield an equilibrium with only sincere offers. This is the message behind corollary 1.

But even if guilt aversion breeds insincerity, is that necessarily a bad thing? Not really. As Charness and Dufwenberg (2006) demonstrate, guilt aversion and verbal promises can create commitment power which may foster trust and cooperation. A similar point is made in Huang (2003). However, in our model, guilt aversion need not sustain cooperation or good behavior because player 2 may perceive player 1 as cooperating reluctantly or cooperating out of guilt. Therefore, the issue may not be whether guilt aversion leads to insincerity but whether the insincerity *per se* has an adverse effect on the utility of other relevant players.

As argued in the section 2, insincerity-induced disutility is less likely in financial matters of the kind analyzed in Huang (2003). However, it may still matter in different ways. To see this, note that in my model, the disutility from insincere offers is captured in the payoff of the intended beneficiary, while in Benabou and Tirole (2006a), it enters the payoff of the donor (benefactor). All donations in Benabou and Tirole (2006a) are accepted. Hence one can think of the model in Benabou and Tirole (2006a) as one in which insincerity-averse beneficiaries do not have the option of rejecting donations or choose not to reject donations but treat insincere benefactors with contempt. So the fact that insincere offers may not be rejected does not mean that it does not matter in the sense of either reducing the benefactor's payoff as in Benabou and Tirole (2006a) or reducing the intended beneficiaries' payoffs as in the present model.

Still on the issue of insincerity, Loury (1994, p. 435) defines a regime of political correctness as "... an equilibrium pattern of expression and inference within a given community where receivers impute undesirable qualities to senders who express themselves in an "incorrect" way and, as a result, senders avoid such expressions."<sup>21</sup>

One can adapt this simple model to political correctness as follows: When player 1's social type is  $w_H$ , he gets a benefit of  $w_H$  from using politically-correct language (e.g., affirmative action is a good policy). When his social type is  $w_L$ , he prefers to use politically-incorrect language and therefore using politically-correct language imposes a cost of  $w_L$  on him. This cost may stem from the mental and emotional effort required to restrain his language or suppress his true opinion. However, there is a cost of using politically-incorrect language which depends on social norms of appropriate language or the expectations of one's peers. This is the cost of guilt in the model. Player 2 derives a benefit of  $v$  when player 1's use of politically correct language is sincere, and a cost of  $\theta$ , if it is insincere. When player 1 uses politically-correct language, player 2's options are to either treat him with admiration (accept) or treat him with contempt (reject). If player 1 uses politically-incorrect language, then player 2's payoff is zero. She derives no disutility from politically-incorrect language, so long as it is sincere. Examples of such politically-incorrect language may be a member of a majority group who argues that most minorities at elite institutions would not have been there in the

---

<sup>21</sup> As noted earlier, the sincerity of language has occupied the interest of philosophers beginning with the influential work of Searle (1969) and recently by Ridge (2006).

absence of affirmative action or that women are incapable of excelling in math and science subjects.<sup>22</sup>

In a politically-correct equilibrium, some  $w_L$  types mimick (pool with)  $w_H$  types as in propositions 2 and 3. But there will still be some people who will deviate from the politically-correct equilibrium (i.e., not offer to help) if  $w_L$  is sufficiently high as in proposition 1 (i.e.,  $(1-\sigma)w_L > \alpha G$ ).<sup>23</sup>

Political correctness may have the disadvantage that people are more likely to be suspicious of each other's intentions and hence a decrease in social interactions akin to the positive probability of rejections as in the equilibrium in propositions 2 and 3.

Again an insincere behavior such as political correctness need not be a bad thing even if it causes people to be suspicious of the intentions of others. One thing missing from the model is that person 2 does not derive any disutility from not receiving an offer or not being invited to a social event (i.e., a disutility from being rejected). If she did, then we could argue that she derives utility from the mere act of being invited or from an offer to be helped even if she intends to reject the offer. Therefore, political correctness need not be a bad thing if people derive disutility from politically correct language *per se*. For example, people may derive utility from others restraining their use of racial slurs or derogatory

---

<sup>22</sup> Morris (2001) presents a more elaborate but different model of political correctness. Morris' (2001) model is a cheap talk game because the actions of the proposer (advisor in his model) does not directly affect any player's payoffs. It only indirectly affects payoffs through its effect on the responder's (decision-maker) beliefs. In the present model, an action by the proposer could impose a direct cost of  $w_L$  or  $G$  on him.

<sup>23</sup>Bernheim (1991) finds that people with sufficiently extreme preferences will deviate from social norms.

language, even if they know that these people harbor such thoughts. If they do not value political correctness (i.e., insincerity) *per se*, then it could be welfare reducing as in the present model.<sup>24</sup>

To be sure, there are certain situations in which people prefer insincerity:  $\theta < 0$  (e.g., they want their peers to not use racial slurs and instead use politically correct language). However, these same people may dislike insincerity in other situations:  $\theta > 0$  (e.g., don't support affirmative action or don't offer to help me if your offer/support is insincere). As noted in section 2, such people may be called sincerity pragmatists.

Indeed, as noted in the introduction and subsequently in the conclusion, Kang (2003a, 2003b) and Shklar (1984) forcefully argue that insincerity is necessary for mutually peaceful co-existence in a democracy. Markovitz (2006) takes the view that because people can have multiple intentions, listeners may interpret sincere statements differently, and people may say things different from what they intended to say, the quest for sincerity in a democracy has to be tampered with caution.

Propositions 2 and 3 imply that player 2 rejects offers when  $\theta$  is sufficiently high. A very high  $\theta$  may be the characteristic of a person with a very

---

<sup>24</sup>In a related but different context, Morris (2001) finds that political correctness could lead to the suppression of socially valuable information. For example, a policy advisor who does not want to be perceived as racist may recommend an affirmative action policy when in fact he believes that affirmative action is a bad policy. This is similar to Kuran (1993) who argues that sincere political discourse leads to the exchange of valuable information and thus better social decisions. This argument is correct insofar as we limit ourselves to the kind of language or communication that results in socially valuable information. Certainly, politically incorrect language like racial slurs and sexist language do not achieve this goal and the argument by Kuran (1993) and Morris (2001) is different from the argument in this paper.

high sense of identity or self-image,<sup>25</sup> which is consistent with why she may derive a high disutility from associating with people who really don't like her. Associating with people who really do not like her imposes a cost on her similar to the cost stemming from a loss of identity in Akerlof and Kranton (2000).<sup>26</sup> If so, the rejection of player 1's offer when  $\theta$  is very high may be player 2's way of choosing her identity by choosing who to associate with, in the sense of Akerlof and Kranton (2000). Consistent with Akerlof and Kranton (2000), my model will predict that women may reject attempts to entice them to traditionally male professions, if they believe that they will only be tolerated but not truly accepted.<sup>27</sup> A difference between my explanation and Akerlof and Kranton (2000) is that identity is an observable characteristic while intention is not. Intention can be inferred but not necessarily observed.

The analysis may also explain why a *high-ability* person may decide to reject a job offer, if she believes that affirmative action influenced the decision. This effect is more likely to kick in, if the person has equally-attractive or better outside options, which she believes were not influenced by affirmative action.<sup>28</sup> If high-ability people are very proud people with a high sense of identity and self

---

<sup>25</sup> Conditional on knowing the insincerity of an offer, a person who rejects such offers may have a very high self esteem. But without knowing for sure whether an offer is insincere, a high rejection rate may be the characteristic of a person with low self esteem who is paranoid about insincerity and therefore may think that most offers are insincere when this is not actually the case. For a discussion of the self see Baumeister (1998).

<sup>26</sup> For other recent economic models of identity, see Darity, Mason, and Stewart (2006) and Benabou and Tirole (2006b).

<sup>27</sup> On a related point, see Case (2003).

<sup>28</sup> Notice also that if the person had an instrumental value for sincerity, as opposed to an intrinsic value for sincerity, then it will be optimal to accept a job from a high-guilt employer in order to exploit their guilt for material gain (e.g., future higher salary).

confidence then, as argued in the preceding paragraph, they will tend to have a high  $\theta$  and would not like to associate with people who do not really like them.

In relationships which require short-term investments by both parties, guilt aversion is more likely to support co-operation because an insincerity-averse person might believe that a guilt-averse person could be of good behavior for a short period. However, if the relationship requires long-term investment, then an insincerity-averse person would not believe that a guilt-averse can sustain his good behavior, so guilt-aversion is less likely to sustain co-operation. In this case, the insincerity-averse person has an instrumental value for sincerity.

On the preceding point, whether a person accepts a potentially insincere offer depends on the costs of insincerity (i.e., the value of  $\theta$ ). However, there are some people who will accept insincere offers because forcing people to be nice to them out of guilt gives them a sense of power. In a different but related context, imagine an affirmative action law that requires certain minorities to be employed at a public institution. A member of a minority group may feel empowered by working at this place, even if her superiors hired her reluctantly and therefore do not want her there. But whether such a minority decides to work in such an environment depends on her belief in the legal system to protect her from unfair treatment while there. Hence the costs of insincerity will influence her choice. This is related to Ayres and Klass (2005) point that a promisee will not care about the sincerity of a promisor, if legal damages are fully compensatory in the event of a breach of contract.

The analysis has been based on the assumption that player 1 incurs no cost if his offer is rejected. It is conceivable that if and only if his social type is  $w_H$ , he might find a rejection embarrassing.<sup>29</sup> The absence of this cost explains why if player 1's social type is  $w_H$ , he always invites or offers to help player 2. However, we sometimes do not invite certain people into closer relationships not because we do not like them. On the contrary, we like them but we are not sure if it is appropriate to invite them or offer to help them. By keeping the relationship at the original lower level, we do not rock the boat. Indeed, a rejection can even push the relationship to a much lower level. For example, imagine how telling a friend that you are romantically interested in them could damage a hitherto platonic and exciting friendship if your proposal is rejected.

Including this cost of rejection or embarrassment to player 1 will not alter our results. Note that including the cost of rejection to player 1 will induce player 2 to moderate her rejection rate in order to encourage player 1 to invite her if player 1's social type is  $w_H$  but it will not affect the qualitative results in the paper. Suppose  $k > 0$  is the cost of rejection or embarrassment to player 1 when his social type is  $w_H$ . Then he will not invite player 2, if  $\sigma k - (1-\sigma)w_H > G$ . This holds if  $k$  is sufficiently high. Indeed, in propositions 1 and 4 there is no rejection of offers in equilibrium, so the cost of rejection will have no effect.

---

<sup>29</sup> Of course, he does not suffer this cost if his social type is  $w_L$ , since he wants his offer to be rejected in this case anyway.

## 5. Conclusion

I have presented an analysis of a common social phenomenon. Using a very simple model, I depart from previous analysis of guilt aversion by taking into account insincerity-induced disutility stemming from guilt aversion. Insincerity-aversion affects trust in relationships, cooperative behavior, and leads to deadweight losses (i.e., mutually beneficial trades may not be realized).

However, due to incomplete information, guilt aversion still results in some cooperation, even if people are averse to insincerity. Clearly, the responder's acceptance probability is not zero, if the cost of insincerity,  $\theta$ , is sufficiently small. But more importantly, the beliefs held by the players can lead to an equilibrium in which no mutually beneficial trades are rejected. And whether this result occurs with certainty depends on the nature of guilt (i.e., *simple guilt* versus *guilt from blame*).

To quote Shklar (1984, p. 77), “[T]he democracy of everyday life, which is rightly admired by egalitarian visitors to America, does not arise from sincerity.... Not all of us are even convinced that all men are entitled to a certain minimum respect. Only some of us think so. But most of us always act as if we really did believe it, and that is what counts.” However, as the analysis in this paper points out people, driven by guilt, may choose to be insincere when sincerity need not disturb mutually peaceful co-existence. On the other hand, sincerity pragmatists may be insincerity-averse in certain situations but not in others. The ‘truth’ hurts but not always.

## References

Akerlof, G., and Kranton, R. (2000). Economics and Identity. *Quarterly Journal of Economics* CXV: 715-753.

Andreoni, J. (2006). Philanthropy. In *Handbook on the Economics of Giving, Reciprocity and Altruism*. Serge-Christophe Kolm and Jean Mercier Ythier (Eds).

Ayres, I., and Klass, G. (2004). Promissory Fraud without Breach. *Wisconsin Law Review* 507: 507-534.

Ayres, I., and Klass, G. (2005). *Insincere Promises: the Law of Misrepresented Intent*. Yale University Press.

Battigalli, P., and Dufwenberg, M. (2005). Dynamic Psychological Games. <http://www.u.arizona.edu/~martind1/Papers-Documents/dpg.pdf>

Battigalli, P., and Dufwenberg, M. (2006). Guilt in Games. *American Economic Review, papers and proceedings*, forthcoming. <http://www.u.arizona.edu/~martind1/Papers-Documents/gig.pdf>

Baumeister, R.F., Stillwell, A.M., and Heatherton, T.F. (1994). Guilt: an interpersonal approach. *Psychological Bulletin* 115: 243-267.

Baumeister, R.F. (1998). The self. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed.; pp. 680-740). New York: McGraw-Hill.

Baumeister, R.F. (1999). *The Self in Social Psychology*. Philadelphia, PA: Psychology Press (Taylor & Francis).

Benabou, R., and Tirole, J. (2006a). Incentives and prosocial behavior. *American Economic Review* 96: 1652-1678.

Benabou, R., and Tirole, J. (2006b). A cognitive theory of identity, dignity, and taboos. <http://www.wws.princeton.edu/rbenabou/identity%20october%20b.pdf>

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

Brandts, J., and Sola, C. (2001). Reference Points and Negative Reciprocity in Simple Sequential Games, *Games and Economic Behavior* 36: 138-157.

Brunnermeier, M.K., and Parker, J.A. (2005). Optimal expectations. *American Economic Review* 95: 1092-1118.

- Case, M.A. (2003). Developing a Taste for not Being Discriminated Against. *Stanford Law Review* 55: 2273-2291.
- Charness, G., and Dufwenberg, M. (2006). Promises and partnerships. *Econometrica* 74: 1579-1601.
- Darity, W.A., Mason, P.L., and Stewart, J.B. (2006). The Economics of Identity: The Origin and Persistence of Racial Identity Norms. *Journal of Economic Behavior and Organization* 60: 283-305.
- Dufwenberg, M. (2002). Marital Investments, Time Consistency, and Emotions. *Journal of Economic Behavior and Organization* 48: 57-69.
- Falk, A., Fehr, E., and Fischbacher, U. (2000). Testing Theories of Fairness - Intentions Matter. Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 63. <http://www.iew.unizh.ch/wp/iewwp063.pdf>
- Falk, A., Fehr, E., and Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry* 41: 20-26
- Falk, A., and Fischbacher, U. (2006). A Theory of Reciprocity. *Games and Economic Behavior* 54: 293-315.
- Fehr, E., and Schmidt, K. (2006). The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories. In *Handbook on the Economics of Giving, Reciprocity and Altruism*. Serge-Christophe Kolm and Jean Mercier Ythier (Eds).
- Frey, B. (1997). *Not Just for the Money: An Economic Theory of Personal Motivation*. Cheltenham: Edward Elgar.
- Geanakoplos, J., Pearce, D., and Stchetti, E. (1989). Psychological Games and Sequential Rationality. *Games and Economic Behavior* 1: 60-79.
- Glazer, A., and Konrad, K.A. (1996). A Signaling Explanation for Charity. *American Economic Review* 86: 1019-1028.
- Gul, F., and Pesendorfer (2005). The canonical type space for interdependent preferences. <http://www.princeton.edu/~pesendor/interdependent.pdf>
- Hill, C., and O'Hara, E.A. (2007). A Cognitive Theory of Trust. *Washington University Law Quarterly*, forthcoming. [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=869423](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=869423)

- Huang, P.H., and Wu, H-M. (1994). More Order without More Law: A Theory of Social Norms and Organizational Cultures. *Journal of Law, Economics, and Organization* 10: 390-406.
- Huang, P.H. (2003). Trust, Guilt and Securities Regulation. *University of Pennsylvania Law Review* 151: 1059-1095.
- Kang, J.M. (2003a). The uses of insincerity: Thomas Hobbes's theory of law and society. *Law & Literature* 15: 371-393.
- Kang, J.M. (2003b). The case for insincerity. *Studies in Law, Politics and Society* 29: 143-164.
- Kartik, N., and McAfee, R.P. (2006). Signaling character in electoral competition. *American Economic Review*, forthcoming.
- Kuran, T. (1993). Mitigating the tyranny of public opinion: Anonymous discourse and the ethic of sincerity. *Constitutional Political Economy* 3: 41-74.
- Laibson, D., Glaeser, E.L., Scheinkman, J., and Soutter, C. (2000). Measuring trust. *Quarterly Journal of Economics* 115: 811-846.
- Levine, D.K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1: 593-622.
- Loury, G.C. (1994). Self-Censorship in Public Discourse: A Theory of "Political Correctness" and Related Phenomena. *Rationality and Society* 6:428-461.
- Markovits, E. (2006) . The Trouble with Being Earnest: Deliberative Democracy and the Sincerity Norm. *Journal of Political Philosophy* 14: 249-269.
- McCabe, K., Rigdon, M., and Smith, V. (2003). Positive Reciprocity and Intentions in Trust Games, *Journal of Economic Behavior and Organization*. 52: 267-275.
- Morris, S. (2001). Political correctness. *Journal of Political Economy* 109: 231-265.
- Offerman, T. (2002). Hurting Hurts More than Helping Helps: The Role of the Self-Serving Bias. *European Economic Review* 46: 1423-1437.
- Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83: 1281-1302.
- Rabin, M., and O'Donoghue, T. (1999). Doing it now or later. *American Economic Review* 89: 103-124.

Ridge, M. (2006). Sincerity and Expressivism. *Philosophical Studies* 131: 487-510.

Searle, J. (1969). *Speech Acts: an Essay into the Philosophy of Language*. Cambridge: Cambridge University Press.

Shklar, J.N. (1984). *Ordinary vices*. Cambridge, Mass: Harvard University Press.

Tangney, J.P. (1992). Situational determinants of shame and guilt in young adulthood. *Personality and Social Psychology Bulletin* 18: 199-206.

Walker, A.D.M. (1978). The Ideal of Sincerity. *Mind*, New Series, 87: 481-497.