Intentions, Insincerity, and Prosocial Behavior*

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

May 23, 2007

Email: 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, Ronnie Schöb, Zane Spindler, Henry Thille and seminar participants at Dalhousie, Minnesota (School of Law), and Ryerson universities and a 2007 CESifo Area conference for very helpful comments and encouragement. 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 might reject 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*. These results are applicable to both 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, prosocial behavior, psychological game.
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.[1] 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.[2] Battigalli and Dufwenberg (2005) extend Geanakoplos et al. (1989) to dynamic psychological games where players' utility depend on the beliefs of others and on updated beliefs. Falk, Fehr, and Fischbacher (2000, 2006), Brandts and Sola (2001), Falk et al. (2003), McCabe et al. (2003), and

[1] 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.

[2] See Levine (1998) for a model which studies intentions-based reciprocity by using interdependent preferences.

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.[3] Both approaches to modeling intentions in games are complementary. 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).[4] 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 (2006): *simple guilt* and *guilt from blame*. These formulation 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

[3] The model is a psychological game due to the specific formulations of endogenous guilt. If guilt were exogenous, the model would have been a standard game with interdependent preference types. An earlier version of this paper had exogenous guilt. Such a model is not capable of producing results such as propositions 1 and 4 and is unable to examine various forms of guilt. More importantly, it does not model the source of guilt.

[4] 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.

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 Benabou and Tirole (2006a). Section 3 presents a simple model of social interaction under incomplete information in a dynamic 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…"[5]

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

[5] See also Tangney (1992).

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."[6] In the same vein, Kang (2003a, 2003b) makes a case for insincerity in a democracy. He argues that insincerity in public discourse is necessary for tolerance and mutual co-existence in liberal democracies.[7] I return to this issue when I discuss political correctness and other applications in section 4.

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

[6] This quote and the two others at the beginning of this article are taken from http://www.brainyquote.com.
[7] 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.[8] 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.[9] 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. In this regard, 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 and 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

[8] 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)

[9] 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.[10] However, Hill and O'Hara (2007) observe that full compensation for breach-of-contract damages may lead to excessive levels of trust in contracting relationships.

It is conceivable that in formal and financial matters of the kind analyzed in Huang (2003), a person may have an instrumental value for sincerity but in non-financial and informal matters, the same person may be more likely to have an intrinsic value for sincerity. Indeed, in formal relationships protected by the law, guilt-aversion is more likely to sustain cooperation because the law reduces the cost of insincerity, even if it does not eliminate it. Hence insincerity aversion will matter less in such relationships than in informal 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., the donor would like to remain anonymous) or by a desire for public show or good public image.[11] 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

[10] 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.

[11] 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).

intrinsic motivation as perceived by others but is decreasing in perceived extrinsic 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 incentive to engage in prosocial behavior.[12]

Benabou and Tirole (2006a) focus on the effect of extrinsic incentives (i.e., material rewards) on prosocial behavior in world where the sincerity (i.e., intrinsic motivation) of such behavior matters. I focus on the ability of guilt-aversion to achieve the kind of prosocial behavior that will be accepted by the intended beneficiaries in a world where the sincerity of such behavior matters. My model applies to social interactions that involve a very small number of people (e.g., two people). It is in such interactions that benefactors can explicitly make offers to intended beneficiaries that might lead to an acceptance or rejection. On the other hand, the model in Benabou and Tirole (2006a) is more applicable to situations where insincerity-averse beneficiaries cannot reject donations (i.e., offers) because the donations are targeted at a large number of anonymous

[12] 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).

beneficiaries (e.g., donating blood).[13]

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 sincerity.[14]

[13] There are other 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).

[14] The voters in Kartik and McAfee (2006) have interdependent type preferences as in Gul and Pesendorder (2005).

Sobel (1985) studies a signaling game where a receiver has an instrumental value for the sincerity of the sender. The sender might be a friend or an enemy. His model differs from the present model in the following respects: (1) his model is only applicable to situations in which the receiver has an instrumental value for sincerity, so she derives a positive utility from the sender's good behavior even if such behavior is out of guilt, (2) the sender is not driven by guilt nor by various forms of endogenous guilt (i.e., *simple guilt* and *guilt from blame*), (3) the sender's second-order beliefs do not enter his utility function, and (4) he discusses different applications from those in the present paper.

3. A Game of Social Interaction with Guilt

In this section, I consider a very simple model to examine the several examples of prosocial behavior mentioned in section 1, where the sincerity of actions matters. While the model is applicable to those examples, I use one specific example 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. 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. Suppose that nature gives player 1 a social type which is his private information. If person 1 is of social type $w_H > 0$, then he derives a psychic *benefit* (joy) of w_H from helping player 2.[15] 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 $\Pr(w_L) = 1\text{-}p$, $p \in (0,1)$. 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.

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).

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.

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, she incurs a psychic cost of $\theta > 0$, given that she accepted player 1's offer.

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. I assume that v is player 2's private information but θ is common knowledge.

[15] One could think of this as a "warm glow" of giving (Andreoni, 1990).

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.

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 for a moment 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

Player 1 need not show that his offer is out of guilt when his social type is w_L. 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.

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.[16] This is what Baumeister et al. (1994) refer to as the *deconstruction* of guilt (see example on *Seinfeld* episode in section 4).

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. [17] 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.[18]

The players have common priors. All this information is common knowledge. In what follows, I assume that player 2 has intrinsic value for sincerity. Later, I shall show that the model is applicable even if player 2 has instrumental value for sincerity.

[16] 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.

[17] 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).

[18]It may sometimes appear that what we refer to as *guilt* should actually be called *shame*. Accordingly, one may argue that what Battigalli and Dufwenberg (2006) refer to as *guilt from blame* should be called *shame*. In order not to get bogged down by semantics, we stick to Battigalli and Dufwenberg's (2006) terminology.

3.1 Equilibrium analysis

I look for psychological perfect Bayesian equilibria (PPBE) of this game. A psychological perfect Bayesian equilibrium (PPBE) is a PBE with the additional requirement that players' endogenous first-order and higher-order beliefs are correct in equilibrium. Battigalli and Dufwenberg (2006) adapt the *sequential equilibrium* solution concept of Kreps and Wilson (1982) to their game. However, a psychological PBE is equivalent to a psychological *sequential equilibrium* in this simple social interaction game because, as shown below, out-of-equilibrium beliefs are either irrelevant or player 1 plays either strategy with positive probability in equilibrium.

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,[19] player 2 computes the posterior probabilities

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

and

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

[19] Of course, we shall find player 1's *optimal* offer probabilities.

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

$$U_2(\lambda) = \rho(w_H|I)v - \rho(w_L|I)\theta \qquad (3)$$[20]

Player 2 rejects an offer, if $U_2(\lambda) < 0$. It follows that player 2 of type $\hat{v}(\lambda) = \dfrac{\rho(w_L|I)\theta}{\rho(w_H|I)} = \dfrac{\lambda(1-p)\theta}{p}$ is indifferent between accepting or rejecting an offer. Therefore,

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

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.

[20] Clearly, player 2 makes her decision based on her expected payoff. *Ex post*, the interaction between the parties, given that player 1 of type w_L is helping player 2, 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 θ 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).

3.1.1 Simple guilt

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 her expected payoff and her actual payoff.

Let λ_1 be player 2's first-order belief of λ and player 1's belief (second-order) of λ_1 be λ_2.[21]

If player 2 does not get an offer from player 1, her *actual* payoff is zero. If she gets an offer and she accepts it, then she *expects* a payoff of $U_2(\lambda_1) > 0$. So if she plans on accepting an offer, then her disappointment, given that she did not get an offer, is $U_2(\lambda_1) – 0 > 0$. On the other hand, if she rejects an offer, she must believe that her payoff if she had accepted it would have been $U_2(\lambda_1) < 0$. So in this case, her disappointment from a non-offer is $U_2(\lambda_1) – 0 < 0$. So she actually suffers no disappointment from not getting an offer.

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

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

Player 1 needs to determine his optimal offer probability, λ. But to do so he has to form beliefs about λ_1 since $D_2(\lambda_1, v, \theta)$ is a function of λ_1. Hence player 1's optimal choice of λ depends on his second-order beliefs (i.e., λ_2) of λ 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 $\rho_H \equiv \rho(w_H|I)$ is a function

[21] As in Battigalli and Dufwenberg (2006), we consider beliefs, at most, of the fourth order.

of λ, it follows that player 1's payoff depends on his beliefs of player 2's updated beliefs of his social type. Therefore, we may write player 1's cost of guilt as $G = \alpha D_2(\rho_H(\lambda_2), v, \theta)$. I abuse notation by rewriting player 1's cost of guilt as

$$G = \alpha D_2(\lambda_2, v, \theta), \tag{6}$$

where α 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 the formulation of guilt in (6) 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.

Since player 1 knows θ and knows the distribution of v, he uses the expected value of v (i.e., $\overline{v}$) in his decision making. That is, he assumes that player 2's disappointment from a non-offer is $D_2 = \max[\rho_H\overline{v} - (1-\rho_H)\theta, 0]$, where $\rho_H\overline{v} - (1-\rho_H)\theta = \int_{\underline{v}}^{\tilde{v}} U_2(\cdot)dF(v)$.

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

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 $(1-\sigma)w_L > \alpha\{\max[\rho_H\overline{v} - (1-\rho_H)\theta, 0]\} = \alpha(\rho_H\overline{v} - (1-\rho_H)\theta) > 0$. Then player 1's optimal response is $\lambda = 0$. Note that $(1-\sigma)w_L > \alpha(\rho_H\overline{v} - (1-\rho_H)\theta) = G > 0$ holds if $\overline{v}$ is sufficiently low and/or θ is sufficiently high and/or α 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 psychological PBE in the social interaction game where all offers are sincere and player 2 accepts all offers.*[22]

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

[22] In proving proposition, I assumed that $\rho_H\overline{v} - (1-\rho_H)\theta > 0$. This proposition also holds if $\rho_H\overline{v} - (1-\rho_H)\theta \leq 0$ which gives $G = \alpha D_2 = 0$.

social type is w_L. Suppose also that $\rho_H\bar{v} - (1-\rho_H)\theta > 0$, given λ_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 λ^*. 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 psychological PBE 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* σ^*, *where 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 psychological PBE 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 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 ρ_H is decreasing in λ. 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 $(1-\sigma)w_L > \alpha(\tilde{\rho}_H\overline{v} - (1-\tilde{\rho}_H)\theta))$. For propositions 2 and 3, we required $(1-\sigma)w_L \leq \alpha(\hat{\rho}_H\overline{v} - (1-\hat{\rho}_H)\theta)$. Note that $\tilde{\rho}_H\overline{v} - (1-\tilde{\rho}_H)\theta > \hat{\rho}_H\overline{v} - (1-\hat{\rho}_H)\theta$ since $\tilde{\rho}_H > \hat{\rho}_H$. It follows that if $(1-\sigma)w_L > \alpha(\tilde{\rho}_H\overline{v} - (1-\tilde{\rho}_H)\theta)$ holds, then $(1-\sigma)w_L \leq \alpha(\hat{\rho}_H\overline{v} - (1-\hat{\rho}_H)\theta)$ cannot hold and vice versa. An implication is that the equilibria above are unique given the specified conditions.[23] 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.*

[23] 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 λ, it follows that $\sigma = F(\hat{v}(\lambda))$ is unique given that λ is unique. Suppose that when $\sigma = 1 - G/w_L$, player 1's optimal λ in the mixed strategy is unique. Then σ is also unique.

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 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.

It is straightforward to apply the model to a situation in which player 2 has an instrumental value for sincerity. Imagine that accepting player 1's offer requires an investment, e, by player 2 into an activity, which yields a net benefit of ve if player 1 is sincere and cost of θe if player 1 is insincere. This cost may be incurred because player 1 of type w_L does not put enough effort into the activity. However, player 2 will not suffer this cost if player 1 exerts the required effort, regardless of whether he did so out of guilt or wholeheartedly. The cost of effort to player 2 is C(e) which is an increasing and strictly convex function. Then player 2's expected payoff is $U_2(e, \lambda_1) = [\rho_H v - (1- \rho_H)\theta]e - C(e)$. It immediately follows that $e^*(\lambda_1) = \text{argmax } U_2(e, \lambda_1) = C'^{-1}(\rho_H v - (1 - \rho_H)\theta)$. The inverse of $C'(e)$ exists because it is a monotonic function. Clearly, e* is decreasing in λ_1 and by the envelope theorem, $U_2(e^*(\lambda_1), \lambda_1)$ is also decreasing in λ_1. Then $D_2 = \max[U_2(e^*(\lambda_1), \lambda_1) - 0, 0]$ is also decreasing in λ_1 for $U_2(e^*(\lambda_1), \lambda_1) > 0$.[24]

Player 1 of type w_L extends an offer hoping that player 2 will reject it. In doing so, he compares $(1 - \sigma)w_L$ to $G = \alpha D_2$. But if player 2 accepts the offer, then *ex post*, player 1 of type w_L will not invest in the activity (i.e., renege on his offer to help) if $w_L > G$. This latter condition is consistent with $(1 - \sigma)w_L > G$ and

[24] Of course, player 2 will set $e^* > 0$ if and only if $U_2(e^*, \lambda_1) > 0$. Otherwise, she will set $e^* = 0$.

$(1 - \sigma)w_L \leq G$. By imposing the restriction $w_L > G$, player 1 of social type w_L will always be insincere *ex post* (i.e., after his offer has been accepted). Hence whether player 2 has an intrinsic or instrumental value for sincerity makes no difference to the analysis. Therefore all the above propositions continue to hold.

3.1.2 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., α) 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 (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, λ_2. This is player 2's third-order belief, λ_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 λ_3. This is his fourth-order belief, λ_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.[25] 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:

[25] 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.

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

Suppose blaming a person for being antisocial can over time change their preferences to prosocial preferences. This is how people are sometimes socialized to imbibe good values or norms. Then one may argue that player 2 may still blame player 1 for not extending an offer even if she knows that player 1's social type is w_L. Player 2 may do this with the goal of changing player 1's social type from w_L to w_H. In a multi-period model, this could result in an insincere offer in the current period, but sincere offers in all subsequent periods because player 1 would have been permanently socialized as a prosocial type. However, if player 2 has a sufficiently small discount factor, then the cost of an insincere offer in the current period will outweigh the discounted value of future sincere offers. Hence, she may not blame player 1 of type w_L if he does not extend an offer. Therefore, proposition 4 will remain unchanged. Furthermore, every society has a positive proportion of antisocial types regardless of how hard it tries to inculcate prosocial values in its members. Besides, to argue that blaming player 1 can change his social type is tantamount to endogenizing preferences which is clearly beyond the scope of the present paper.[26]

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 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, even if they do not change your preferences.

[26] For a recent attempt to endogenize preferences, see Akerlof and Kranton (2005).

4. Discussion and Applications

Let me begin this section by noting that the story could be told differently but with similar results. In particular, the timing of actions could be reversed where player 2 is the first mover and player 1 is the second mover. In this case, player 2 initiates a request by asking player 1 for help (H) or makes no such request (NH). Player 1's response to a request for help is yes (Y) or no (N). Player 1 does not feel guilty if player 2 does not ask for help and he does not offer to help. Then λ is now the probability that player 1 of type w_L plays Y and σ is the probability that player 2 will play NH. This formulation is identical to the original formulation except that player 1's conjecture about the probability that player 1 will accept his offer is 1, given that player 2 will only ask for help if she intends to accept player 1's offer (i.e., if player 1 plays Y). So player 1 of type w_L compares w_L to G.

Casual empiricism confirms a result akin to propositions 2 and 3. That is, we sometimes do not ask people for favors because we feel that we may be bothering them and therefore they may help us grudgingly out of guilt. So player 2 plays NH with positive probability which is equivalent to playing R with positive probability in propositions 2 and 3.

Suppose player 1 feels guilty if he does not offer to help, although player 2 has made no request. Indeed, playing NH is a signal from player 2 to player 1 that she believes, with a sufficiently high probability, that his social type is w_L. Then knowing that player 2 will reject his offer, player 1 will make an offer precisely for this reason and thereby assuage his guilt. In this case, player 1's offer is akin to a costless action in a cheap-talk game. This also accords with casual empiricism where we sometimes make offers to people who we know will not accept our offer and we, indeed, do not want them to accept our offer. [27]

In what follows and without loss of generality, I stick to the original formulation of the game where player 1 is the first mover.

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

[27] A clear example of this was the April 4, 1991 episode of *Seinfeld* titled "The Apartment". Mrs. Hudwalker, a tenant in one of the apartments where Jerry is also a tenant, dies and Jerry proposes to Elaine to take the newly vacant and very cheap apartment just above his own. Later, he realized that it was a big mistake after talking to George. He now wanted to withdraw the proposal but could not because he will feel guilty. However, someone offers the superintendent $10000 per month for the apartment. Jerry then insincerely encourages Elaine to take the apartment although he knew that Elaine could not afford to pay such a huge monthly rent and would therefore reject his proposal. He is able to assuage or *deconstruct* his guilt by telling himself that Elaine would never know that he (Jerry) did not want her to have the apartment after the original proposal. To the extent that TV shows are reflections of parts of our real lives, this *Seinfeld* episode clearly shows how people make insincere offers to assuage their guilt.

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 $\rho_H \overline{v} - (1- \rho_H)\theta > 0$ 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 α) 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 α) 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."[28]

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 θ, 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. An example

[28] 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).

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 absence of affirmative action.[29]

In a politically-correct equilibrium, w_L types mimick (pool with) w_H types as in propositions 2 and 3. And in a separating equilibrium they deviate from the politically-correct equilibrium if w_L is sufficiently high as in proposition 1 (i.e., $(1-\sigma)w_L > G$).[30] In the latter case, all equilibrium politically-correct language is sincere and in the former case, some politically-correct language is insincere.

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 equilibria 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 player 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 utility from politically correct language *per se*. For example, people derive

[29] The point is not that people do not find such language offensive. There are definitely people who do. My focus is on those who do not find such language offensive, so long as it is sincere. 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) do 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.

[30] Bernheim (1994) finds that people with sufficiently extreme preferences will deviate from social norms.

utility from others restraining their use of racist, anti-semitic, sexist, and homophobic language, even if they know that these people harbor such thoughts. Indeed, Fish (1994) argues that some restriction on free speech is desirable for precisely this reason. However, if people do not value political correctness (i.e., insincerity) *per se*, then it could be welfare reducing as in the present model. [31]

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.

[31] 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 therefore different from the argument in this paper.

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 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 (see Sobel, 1985).

On the preceding point, whether a person accepts a potentially insincere offer depends on the costs of insincerity (i.e., the value of θ). 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 (i.e., $\theta < 0$). 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 expected cost 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.

Propositions 2 and 3 imply that player 2 rejects offers when θ is sufficiently high. A very high θ may be the characteristic of a person with a very high sense of identity or self-image,[32] 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 *pretend* to like her imposes a cost on her similar to the cost stemming from a loss of identity in Akerlof and Kranton (2000).[33] If so, the rejection of player1's offer when θ 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.[34] 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 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.[35] 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

[32] 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).

[33] For other recent economic models of identity, see Darity, Mason, and Stewart (2006) and Benabou and Tirole (2006b).

[34] On a related point, see Case (2003).

[35] 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.

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 the 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 extend an offer if his 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 extend an offer if $(1-\sigma)w_H - \sigma k > -G$. This holds if k is sufficiently small. Indeed, in propositions 1 and 4 there is no rejection of offers in equilibrium, so the cost of rejection will have no effect.

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, θ, 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.

Akerlof, G., and Kranton, R. (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives* 19: 1167-1201.

Andreoni, J. (1990). Impure Altruism and Donations to a Public Good: A Theory of Warm-Glow Giving. *Economic Journal* 100: 464:477.

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.

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).*

Fish, S. (1994). *There is no Such Thing as Free Speech: and it is a Good Thing, Too.* New York: Oxford University Press.

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

Geanakoplos, J., Pearce, D., and Stachetti, 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

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.

Kreps, D., and Wilson, R. (1982). Sequential Equilibria. *Econometrica* 50: 863-894.

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.

Sobel, J. (1985). A Theory of Credibility. *Review of Economic Studies* 52: 557-573.

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.