2008

Intentions, Guilt and Social Interactions

J. Atsu Amegashie

Intentions, Guilt and Social Interactions

J. Atsu Amegashie*

"Sincerity makes the very least person to be of more value than the most talented hypocrite." – Charles Spurgeon[1]

"To give real service you must add something which cannot be bought or measured with money, and that is sincerity and integrity." – Douglas Adams[2]

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. Intention is the basis for the legal distinction between murder and manslaughter and partly explains the attitudes of certain groups towards racial profiling. For murder or manslaughter, 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. For racial profiling, a traveler at an airport or a motorist who is searched by the police may react differently depending on

---

1. Quote available at Brainy Quote, http://www.brainyquote.com/quotes/authors/c/charles_spurgeon.html (last visited Feb. 1, 2008).

2. Quote available at Brainy Quote, http://www.brainyquote.com/quotes/authors/d/douglas_adams.html (last visited Feb. 1, 2008).

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 rather 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 would not have been offered the job if you were not a minority; or someone gives you a present because they felt obligated to do so, not because they really wanted to give you a present. If your boss, supervisor, or professor tells you to feel free to come talk to her anytime you encounter problems in your work, would 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 whom 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 neighbors. 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*.

Faruk Gul and Wolfgang Pesendorfer show that intentions can also be modeled as stemming from interdependent type preferences: preferences over someone's physical or social characteristics (e.g., race, height, gender, personality).[3] In the social interaction in the present paper, the responder has interdependent type preferences because he cares about the proposer's social type (i.e., whether the proposer is prosocial or not). 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 or characteristic.

Psychological game theory pioneered by John Geanakoplos, David Pearce, and Ennio Stacchetti[4] models intentions as beliefs about beliefs, where players have belief dependent preferences. There have been important extensions and analysis of this approach.[5]

This paper has elements of psychological games and interdependent type preferences. Both approaches to modeling intentions in games are complementary. Indeed, John Searle argues, in a very influential philosophical work, that sincerity is linked to a person's state of mind (i.e., his beliefs).[6] Insofar as this paper is concerned with the sincerity of a person's behavior or altruism, psychological game theory, a model with interdependent type preferences, or some combination gives an appropriate analytical framework.

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

---

3. *See* Faruk Gul & Wolfgang Pesendorfer, The Canonical Type Space for Interdependent Preferences, 2 (July 2007) (unpublished manuscript, on file with author), *available at* http://www.princeton.edu/~pesendor/ interdependent.pdf.

4. John Geanakoplos et al., *Psychological Games and Sequential Rationality*, 1 GAMES & ECON. BEHAV. 60, 61 (1989).

5. *See* Matthew Rabin, *Incorporating Fairness into Game Theory and Economics*, 83 AM. ECON. REV. 1281, 1284 (1993); Armin Falk et al., *On the Nature of Fair Behavior*, 41 ECON. INQUIRY 20, 20 (2003); Gary Charness & Martin Dufwenberg, *Promises and Partnerships*, 74 ECONOMETRICA 1579, 1580 (2006); Armin Falk et al., *Testing Theories of Fairness—Intentions Matter* 7 (Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 63, 2000), *available at* http://www.iew.unizh.ch/wp/ iewwp063.pdf.

6. JOHN R. SEARLE, SPEECH ACTS: AN ESSAY INTO THE PHILOSOPHY OF LANGUAGE 54–71 (1969).

beneficial trades (i.e., sincere offers), although she is uncertain about the proposer's social type. Equilibriums with insincere offers are also possible. I also discuss the implications of insincerity aversion for altruism, political correctness, 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. Section 3 presents a simple model of social interaction under incomplete information in a dynamic psychological game and characterizes its equilibriums. 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 Roy Baumeister, Arlene Stillwell, and Todd Hearther 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."[7] 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."[8]

Building on a well-known idea in psychology, Gary Charness and Martin Dufwenberg 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.[9] 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.[10]

In a related contribution, Peter Huang examines how guilt can motivate securities professionals' behavior in their fiduciary relationships with their clients and the legal implications of guilt

---

7. Roy F. Baumeister et al., *Guilt: An Interpersonal Approach*, 115 PSYCHOL. BULL. 243, 243 (1994).

8. *Id.*

9. Charness & Dufwenberg, *supra* note 5, at 1580.

10. *Id.*

for the regulation of securities professionals.[11] In a different but related context, Peter Huang and Ho-Mou Wu examine how remorse can lead to better social order.[12] 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.[13]

A very important difference in the present paper is that I argue 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.[14] 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 if, in certain contexts, she may care about sincerity, while in other contexts, she may not. She may have an *intrinsic* value for sincerity in certain situations, as the quote by Douglas Adams at the beginning of this paper suggests, but may have an *instrumental* value for sincerity in other situations. This kind of cost-benefit calculus by such sincerity pragmatists is alluded to by the Nobel laureate, Albert Camus when he opined: "[H]ow can sincerity be a condition of friendship? A taste for truth at any cost is a passion which spares nothing."[15] In the same vein, John Kang 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.[16] I return to this issue when I discuss political correctness and other applications in section 4.

---

11. *See* Peter Huang, *Trust, Guilt and Securities Regulation*, 151 U. Pa. L. Rev. 1059, 1059 (2003).

12. Peter Huang & Ho-Mou Wu, *More Order Without More Law: A Theory of Social Norms and Organizational Cultures*, 10 J.L. Econ. & Org. 390, 394–97 (1994).

13. *See id.*; Charness & Dufwenberg, *supra* note 5, at 1585.

14. *See* Charness & Dufwenberg, *supra* note 5; Huang, *supra* note 11; Huang & Wu, *supra* note 12.

15. Quote available at Brainy Quote, http://www.brainyquote.com/quotes/authors/a/albert_camus.html (last visited Feb. 1, 2008).

16. *See* John M. Kang, *The Uses of Insincerity: Thomas Hobbes's Theory of Law and Society*, 15 Law & Literature 371 (2003) [hereinafter Kang, *Uses*]; John M. Kang, *The Case for Insincerity*, 29 Stud. L. Pol. & Soc'y 143 (2003) [hereinafter Kang, *Cases*].

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[17] only one player is guilt-averse or only one player's guilt aversion is relevant to 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 the offers 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 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. 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.[18] 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, Claire Hill and Erin O'Hara examine how the law should intervene to either promote more accurate trust levels or to mitigate the costs of

---

17. *See* Charness & Dufwenberg, *supra* note 5; Huang, *supra* note 11; Huang & Wu, *supra* note 12.

18. Interview with Claire A. Hill, Professor of Law, University of Minnesota Law School, in Minneapolis, Minn. (April 6, 2007).

mistaken assessments in contractual and non-contractual relationships.[19]

Ian Ayres and Greg Klass present a lucid and interesting examination of the legal implications of insincere promises and misrepresented intent.[20] A promise is insincere if the promisor never intended to fulfill the promise. According to them, a promisee cares about the sincerity of the promisor because breach-of-contract damages are not fully compensatory. If such damages were fully compensatory, a promisee will not care about the sincerity of the promisor.[21] This is consistent with our earlier point that a person may be insincerity-averse because dealing with an insincere person is costly.[22] However, Claire Hill and Erin O'Hara observe that full compensation for breach-of-contract damages may lead to excessive levels of trust in contracting relationships.[23]

It is conceivable that in formal and financial matters of the kind analyzed by Peter Huang,[24] 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.

---

19. *See* Claire Hill & Erin A. O'Hara, *A Cognitive Theory of Trust* 2, 32 (Minnesota Legal Studies, Research Paper No. 05-51, 2005), *available at* http://ssrn.com/abstract=869423.

20. *See* Ian Ayres & Greg Klass, *Promissory Fraud Without Breach*, 2004 WIS. L. REV. 507, 508. *See generally* IAN AYRES & GREG KLASS, INSINCERE PROMISES: THE LAW OF MISREPRESENTED INTENT (2005).

21. AYRES & KLASS, INSINCERE PROMISES, *supra* note 20, at 61.

22. To be sure, any moral hazardous behavior in a principal-agent relationship could be considered 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 as 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.

23. Hill & O'Hara, *supra* note 19, at 31–32.

24. Huang, *supra* note 11, at 3–7, 33–35.

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, player 1 and player 2.[25] Player 1 has the option of proposing to help player 2 in some activity. Player 1 could be either prosocial, in which case he enjoys helping player 2, or he could be asocial, in which case he grudgingly helps player 2 out of guilt. Player 2 does not know for sure whether player 1 is prosocial or asocial.

Furthermore, player 1 feels guilty if he does not offer to help player 2. Player 1's guilt depends on the extent to which he believes that he has disappointed player 2.[26] The higher player 1 believes is the level of player 2's disappointment, the more guilt player 1 has.

An offer is insincere if player 1 is asocial and it is sincere if player 1 is prosocial. If player 2 believes that player 1 genuinely wants to help her, she gets a positive payoff if she accepts player 1's offer. If she believes that player 1's offer is insincere, she incurs a psychic cost if she accepts player 1's offer. Thus, player 2 will not knowingly accept an insincere offer. However, since she does not know for sure whether player 1 is prosocial or asocial, there is some possibility that she might accept insincere offers or reject sincere ones.

Player 1 has two actions: offer to help or do not offer to help. Player 2 has two actions: accept or reject an offer from player 1. The game is sequential. Player 1 is the first-mover and player 2 is the second-mover.

Player 1 need not show that his offer is out of guilt if he is asocial. 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

---

25. I use male pronouns for player 1 and female pronouns for player 2.

26. Pierpaolo Battigalli & Martin Dufwenberg, *Guilt in Games*, 97 Am. Econ. Rev. 170, 174–75 (2007). The formulation of guilt in this paper follows the analysis of Pierpaolo Battogalli and Martin Dufwenberg.

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.

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 he is asocial, he might offer to help player 2 and if player 2 rejects, 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. This is what Baumeister et al. refer to as the *deconstruction* of guilt. [27]

Of course, if player 2 could 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. [28] However, due to incomplete information, player 2 cannot in general be certain of the insincerity of player 1's offer. Therefore, due to incomplete information, the rejection of an insincere offer does not make player 1 feel guilty.[29]

3.1 EQUILIBRIUM ANALYSIS

I looked for psychological perfect Bayesian equilibriums (PPBE) of this game, defined as a perfect Bayesian equilibrium (PBE) with the additional requirement that players' endogenous first-order and higher-order beliefs are correct in equilibrium.[30]

---

27. *See* Baumeister et al., *supra* note 7, at 259 (noting that by focusing not on the implications of one's actions, one can escape guilt); *Seinfeld: The Apartment, infra* note 37.

28. 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. *See* discussion *infra* Parts 3.1.1, 3.1.2.

29. It may sometimes appear that what we refer to as *guilt* should actually be called *shame*. Accordingly, one may argue that what Pierpaolo Battigalli and Martin Dufwenberg refer to as *guilt from blame* should be called *shame*. *See* Battigalli & Dufwenberg, *supra* note 26, at 172. In order not to get bogged down by semantics, we do not make this distinction.

30. A perfect Bayesian equilibrium is a strategy profile that is sequentially rational given a system of beliefs that is obtained using Bayes rule. Julio Gonzalez-Diaz & Miguel A. Melendez-Jimenez, A Formal Definition of Perfect Bayesian Equilibrium for Extensive Games, 2 (June 7, 2007) (unpublished manuscript, on file with author), *available at* http://www.kellogg.northwestern.edu/faculty/gonzalezdiaz/papers/Perfect_Bayesian.pdf.
Bayes' rule provides a mathematical approach to how one should change his or her existing beliefs in light of new evidence. *In Praise of Bayes*, ECONOMIST, Sept. 30, 2000, at 83.

All proofs can be found in the technical appendix to this paper.[31]

3.1.1 Simple Guilt

Player 2's disappointment is a function of the difference between her expected payoff and her actual payoff. In this *simple guilt* formulation, player 1 feels guilty as a result of the disappointment felt by player 2, even if player 2 does not blame him for his actions.[32]

The following proposition holds in this game:

Proposition 1: If player 1 suffers from simple guilt and player 2 does not expect any insincere offers; player 1 believes that player 2 does not expect any insincere offers; and 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.

We can also state the following proposition:

Proposition 2: If player 1 suffers from simple guilt, then there exists a PPBE in which player 1 always offers to help player 2. Player 2 rejects player 1's offer if her valuation of sincere offers is sufficiently small and accepts player 1's offer if her valuation of sincere offers is sufficiently high.

3.1.2 Guilt from Blame

In addition to *simple guilt*, Pierpaolo Battigalli and Martin Dufwenberg also consider another formulation of guilt, in which a player who has disappointed another player feels guilty depending on the extent to which the affected player blames him for his actions.[33] They refer to this as *guilt from blame*.[34]

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.[35] Unlike *guilt from blame*, player 1 blames

---

31. More elaborate arguments can be found in a version of this paper at http://www.uoguelph.ca/~jamegash/intentions_interaction_psychological.pdf.

32. *See* Battigalli & Dufwenberg, *supra* note 26, at 171 (noting that with simple guilt a player cares about how much he lets another player down).

33. *See* Battigalli & Dufwenberg, *supra* note 26, at 175.

34. *Id.* at 172.

35. *Id.* at 171–72.

himself under *simple guilt* for not offering to help player 2, even if he is asocial. This makes sense because it is not uncommon for people to feel guilty, blame themselves for having certain antisocial preferences, or for being of an antisocial type, even if they do not change their preferences.

It is implicitly assumed that for the same level of disappointment incurred by player 2, player 1's guilt sensitivity 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 different 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. Therefore, *guilt from blame* is only possible when player 1 does not offer to help player 2.

In the case of *guilt from blame*, the proof of the equilibrium of this social interaction is very straightforward. 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 he is asocial, or he would only insincerely offer to help her. Moreover, since player 2 dislikes insincere offers, she does not get disappointed and so does not blame player 1.[36] Under *guilt from blame*, player 1 will not extend insincere offers; hence, all offers are sincere. The proposition below then follows:

Proposition 3: If player 1 suffers from guilt from blame, then there is a unique PPBE in which all offers are sincere and no offers are rejected.

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

---

36. 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.

request by asking player 1 for help or makes no such request. Player 1's response to a request for help is yes or no. Player 1 does not feel guilty if player 2 does not ask for help and he does not offer to help.

Casual empiricism confirms a result akin to proposition 2. 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 sometimes does not ask for help, which is equivalent to the cases in which she rejects player 1's offer in proposition 2.

Suppose player 1 feels guilty if he does not offer to help, although player 2 has made no request. Indeed, not asking for help is a signal from player 2 to player 1 that she believes, with a sufficiently high probability, that he is asocial. 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.[37]

In what follows and without loss of generality, I will continue with 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 a PBE where sincere offers are never rejected.

However, proposition 3, relative to the propositions under *simple guilt*, is even more interesting. Unlike *simple guilt*, proposition 3 shows that under *guilt from blame*, there cannot be

---

37. 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. However, Jerry could not tell Elaine that he did not want her to live in the same building because he would feel guilty. Luckily, Kramer found someone who could offer the superintendent $10,000 per month for the apartment; a sum that Jerry knew Elaine could not afford. Jerry is able to assuage or *deconstruct* his guilt by telling himself that Elaine would never know that he 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 that people do not only extend insincere offers to assuage their guilt, but they also do so hoping that such offers will not be accepted. *Seinfeld: The Apartment* (NBC television broadcast Apr. 4, 1991).

equilibriums 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.

Note that player 1 offers to help player 2 if he believes that player 2 expects an offer and will be sufficiently disappointed otherwise. This accords very well with casual empiricism. The emotional cost (i.e., guilt) of disappointing player 2 *coupled with* player 2's *expectations* could force player 1 to be kind to her, although he 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.

Unlike the equilibriums in propositions 1 and 3, the equilibrium in proposition 2 involves 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 an insincere offer and player 1 believes that player 2 expects an insincere offer, then these beliefs coupled with a high guilt sensitivity 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 with low guilt sensitivity yield an equilibrium with only sincere offers.

However, even if guilt aversion breeds insincerity, is that necessarily a bad thing? Not really. As Gary Charness and Martin Dufwenberg demonstrate, guilt aversion and verbal promises can create commitment power that may foster trust and cooperation.[38] Peter Huang makes a similar point;[39] however, in

---

38. Charness & Dufwenberg, *supra* note 5, at 1594–95.
39. Huang, *supra* note 11, at 1084 (noting that due to loyalty to her client, a broker would have such guilt aversion that the broker would not misbehave).

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 section 2, insincerity-induced disutility is less likely in financial matters of the kind analyzed by Peter Huang.[40]

One can adapt the guilt aversion model to political correctness[41] as follows: When player 1 is prosocial, he gets a benefit from using politically-correct language (e.g., affirmative action is a good policy). When he is asocial, he prefers to use politically incorrect language, which imposes a cost 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 the social norms of using appropriate language or the expectations of one's peers. This is the cost of guilt in the model. Player 2 derives a benefit if player 1's use of politically correct language is sincere, and a cost, 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 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.[42]

From the above propositions, it is easy to obtain an equilibrium with sincere politically correct language, as in proposition 1, and an equilibrium with politically correct language that may be insincere, as in proposition 2.

---

40. *See id.* at 1059–95.

41. Glenn Loury 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." Glen C. Loury, *Self-Censorship in Public Discourse: A Theory of "Political Correctness" and Related Phenomena*, 6 RATIONALITY & SOC'Y 428, 435 (1994).

42. 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.

Political correctness may have the disadvantage that people are more likely to be suspicious of each other's intentions. Hence, a decrease in social interactions akin to the positive probability of rejections in the equilibriums in proposition 2 may occur. Again, an insincere behavior, such as political correctness, need not be a bad thing even if it causes people to be suspicious of others' intentions. One thing missing from the model is that player 2 does not derive any disutility from not receiving an offer (i.e., a disutility from being rejected). If she did, then we could argue that she derives utility from an offer 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 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, Stanley Fish argues that some restriction on free speech is desirable for precisely this reason.[43]

However, if people do not value political correctness (i.e., insincerity) *per se*, then it could be welfare reducing as in the present model. To be sure, there are certain situations in which people prefer insincerity. For example, they may want their peers to not use racial slurs and instead use politically correct language. However, these same people may dislike insincerity in other situations. Player 2 may not want player 1 to offer help if his offer 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, John Kang and Judith Shklar forcefully argue that insincerity is necessary for mutually peaceful co-existence in a democracy.[44]

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

---

43. STANLEY FISH, THERE'S NO SUCH THING AS FREE SPEECH: AND IT'S A GOOD THING, TOO 111 (1994) (noting that it may be politically right to regulate use of certain hateful language).

44. *See* JUDITH N. SHKLAR, ORDINARY VICES 246 (1984); *see also* Kang, *Uses, supra* note 16 at 386; Kang, *Cases, supra* note 16 at 154.

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. 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. However, 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 Ian Ayres and Greg Klass's 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.[45]

The analysis has been based on the assumption that player 1 incurs no cost if his offer is rejected. However, sometimes we do not invite certain people into closer relationships not because we do not like them, but because we are not sure if it is appropriate to offer to help them. By keeping the relationship at the original lower level, we do upset current dynamics. Indeed, a rejection could push the relationship to a much lower level.

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).

To quote Judith Shklar,

> The 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 of social respect. Only some of us think so. But most of us always act as if we

---

45. AYERS, *supra* note 20, at 62.

really did believe it, and that is what counts.[46]

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.

TECHNICAL APPENDIX

Consider two people, 1 and 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. 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.

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 and α is a positive parameter that captures player 1's sensitivity to guilt. 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.

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.

After observing his social type, player 1 has two actions: offer

---

46. SHKLAR, *supra* note 44, at 77.

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.

The players have common priors. All this information is common knowledge. In what follows, I assume that player 2 has intrinsic value for sincerity.

Equilibrium Analysis

I look for psychological perfect Bayesian equilibriums (PPBE) of this game.

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, player 2 computes the posterior probabilities

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

and

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

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

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

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,

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

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.

Simple Guilt

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.[47]

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

---

47. As in Battigalli and Dufwenberg, *supra* note 26, I consider beliefs, at most, of the fourth order.

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 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

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

where $\alpha > 0$ is common knowledge and, as defined before, measures player 1's sensitivity to guilt. The formulation in (6) makes this game a psychological game where player 1 has belief-dependent preferences about player 2's updated belief about his social type.

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.

Since player 1 knows θ and knows the distribution of v, 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 $D_2 = \max[\rho_H \bar{v} - (1 - \rho_H)\theta, 0]$, where

$$\rho_H \bar{v} - (1 - \rho_H)\theta = \int_{\tilde{v}}^{\bar{v}} U_2(\cdot)dF(v).$$

I characterize the equilibriums 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. Player 1 also

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 proposition 1 in the text.[48]

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

Now suppose that $\lambda_1 = \lambda_2 = \lambda^* = 1$. Also, assume that $-(1 - \sigma)w_L > - \alpha(\rho_H\overline{v} - (1- \rho_H)\theta)$, where $\rho_H = p$ and $\rho_H\overline{v} - (1- \rho_H)\theta > 0$. This gives proposition 2 in the text.

Note that we do not have to worry about out-of-equilibrium beliefs in any of the equilibriums above. Suppose that in proposition 2, 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 proposition 1, 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) = $ 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 $U_2(e^*(\lambda_1), \lambda_1) > 0$.[49]

Player 1 of type w_L extends an offer hoping that player 2 will

---

48. In proving this 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 $G = \alpha D_2 = 0$.

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

MINN. J.L. SCI. & TECH.   [Vol. 9:2

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 $(1 - \sigma)w_L \leq G$. 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.