

Guilt in Games

By PIERPAOLO BATTIGALLI AND MARTIN DUFWENBERG*

“A clear conscience is a good pillow.” Why does this old proverb contain an insight? The emotion of *guilt* holds a key. Psychologists report that “the prototypical cause of guilt would be the infliction of harm, loss, or distress on a relationship partner” (Roy Baumeister, Arlene M. Stillwell, and Todd F. Heatherton 1994, 245; June Price Tangney 1995). Moreover, guilt is unpleasant and may affect behavior to render the associated pangs counterfactual. Baumeister, Stillwell, and Heatherton state, “If people feel guilt for hurting their partners ... and for failing to live up to their expectations, they will alter their behavior (to avoid guilt) in ways that seem likely to maintain and strengthen the relationship.” Avoided guilt is the down of the sound sleeper’s bolster.

How can guilt be modeled? How are human interaction and economic outcomes influenced? We offer a formal approach for providing answers. Start with an extensive game form which associates a monetary outcome with each end node. Say that player i lets player j down if as a result of i ’s choice of strategy, j gets a lower monetary payoff than j expected to get before play started. Player i ’s guilt may depend on how much he lets j down. Player i ’s guilt may also depend on how much j believes i believes he lets j down. We develop techniques to analyze equilibria when players are motivated, in part, by a desire to avoid guilt.

The intellectual home for our exercise is what has been called *psychological game theory*. This framework—originally developed by John Geanakoplos, David Pearce, and Ennio Stacchetti (1989) and recently extended by Battigalli and Dufwenberg (2005) (henceforth B&D)—allows players’ utilities to depend on beliefs (about choices, states of nature,

or others’ beliefs) as is typical of many emotions.¹ Our approach formalizes Baumeister, Stillwell, and Heatherton’s (1994) remark that guilt depends on a failure “to live up to [others’] expectations,” and embraces some previous related theoretical and experimental results on “trust games.”² We refer to, e.g., Gary Charness and Dufwenberg (2006) for elucidation on the role of guilt in that specific context, which space constraints prevent us from repeating here as we develop a theory for general games.

I. Game-Theoretic Preliminaries

We consider finite extensive game forms specifying monetary payoffs for each player at each end node. These payoffs describe the material consequences of players’ actions, not their preferences. The players’ utilities will be introduced in Section II.

Let N be the player set, T the set of nodes in the game tree with distinguished root t^0 , and Z the set of end (or terminal) nodes. The set $X = T \setminus Z$ is partitioned into subsets X_i of decision nodes for each $i \in N$ and the set of chance nodes X_c . We let $\sigma_c(\cdot|x)$ denote the strictly positive chance probabilities of the immediate followers of node $x \in X_c$. In our theory, it is important to represent players’ information also at nodes where they are *not* active. Thus, we let the information structure of i be a partition H_i of the whole set T that contains, as a subcollection, the standard information partition of X_i . A typical information set is denoted h . The information set containing node t is denoted $H_i(t)$. The (extended) information structure H_i satisfies *perfect recall*. We also assume that H_i is a refinement of $\{\{t^0\}, X \setminus \{t^0\}, Z\}$. Players know

* Battigalli: Bocconi University, Via Sarfatti 25, Milano 20136, and IEP and IGIER, (e-mail: pierpaolo.battigalli@unibocconi.it); Dufwenberg: Department of Economics, University of Arizona, Tucson, AZ 85721-0108, and Institute for Behavioral Economics (e-mail: martind@eller.arizona.edu). We thank Geir Asheim, Gary Charness, Edoardo Grillo, and Marciano Siniscalchi for comments, and MIUR (Battigalli) and NSF (Dufwenberg) for financial support.

¹ Jon Elster (1998, 49) argues that emotions are characteristically “triggered by beliefs.”

² See Peter H. Huang and Ho-Mou Wu (1994), Dufwenberg (1995, 2002), Dufwenberg and Uri Gneezy (2000), Michael Bacharach, Gerardo A. Guerra, and Daniel J. Zizzo (forthcoming), Guerra and Zizzo (2004), Charness and Dufwenberg (2006), and Giuseppe Attanasio and Rosemarie Nagel (2006).

when they are at the root of the game tree, and they know when the game is over. The material consequences of players' actions are determined by functions $\mathbf{m}_i : Z \rightarrow \mathbb{R}, i \in N$. A typical material payoff is denoted by m_i , as in $m_i = \mathbf{m}_i(z)$. We assume that $\mathbf{m}_i(z') \neq \mathbf{m}_i(z'')$ implies $H_i(z') \neq H_i(z'') : i$ observes his material payoff. Whenever we do not explicitly specify players' terminal information, *the default assumption is that they have the coarsest terminal information consistent with perfect recall and with their material payoff.*

A pure strategy s_i specifies a contingent choice for each $h \in H_i$, where i is active ($h \subset X_i$). We also find it convenient to refer to "pure strategies" of chance, i.e., functions $s_c : X_c \rightarrow T$, which select an immediate successor of each chance node (such strategies are chosen at random according to the mixed representation of $\sigma_c = [\sigma_c(\cdot|x)]_{x \in X_c}$). The set of pure strategies of i is S_i , and we let $S = S_c \times \prod_{i \in N} S_i, S_{-i} = S_c \times \prod_{j \neq i} S_j$. For any h and $i, S_i(h)$ denotes the set of i 's strategies allowing h , and $S_{-i}(h) \subset S_{-i}$ denotes the set of profiles s_{-i} allowing h . A strategy profile $s \in S$ (including s_c) yields an end node denoted $\mathbf{z}(s)$.

We assume players do not actually randomize, but randomized choices—in the form of behavior strategies—enter the analysis as an expression of players' beliefs. A behavior strategy for i is an array σ_i of probability measures $\sigma_i(\cdot|h), h \in H_i, h \subset X_i$, where $\sigma_i(a|h)$ is the probability of choosing action a at h . Given σ_i , we can compute the probability of each pure strategy s_i , denoted $\text{Pr}_{\sigma_i}(s_i)$. By perfect recall, one can compute conditional probabilities $\text{Pr}_{\sigma_i}(s_i|h), h \in H_i$, even if $\text{Pr}_{\sigma_i}(S_i(h)) = 0$.

Conditional on each $h \in H_i$, player i holds an updated, or revised, belief $\alpha_i(\cdot|h) \in \Delta(S_{-i}(h))$ about the strategies of the co-players and chance; $\alpha_i = (\alpha_i(\cdot|h))_{h \in H_i}$ is the system of *first-order* beliefs of i . (Note that we include in α_i also i 's beliefs about chance moves. Later on we impose that these are determined by the objective probabilities σ_c .) Player i also holds, at each $h \in H_i$, a *second-order* belief $\beta_i(h)$ about the first-order belief system α_j of each co-player j , a third-order belief $\gamma_i(h)$ about the second-order beliefs, and so on. For the purposes of this paper, we may assume that higher-order beliefs are degenerate point beliefs. Thus, with a slight abuse of notation we identify $\beta_i(h)$ with a particular

array of conditional first-order beliefs $\alpha_{-i} = [\alpha_j(\cdot|h')]_{j \neq i, h' \in H_j}$. A similar notational convention applies to other higher-order beliefs. Clearly, the beliefs i would hold at different information sets are not mutually independent. They must satisfy Bayes's rule and common certainty that Bayes's rule holds (cf. B&D). In our analysis we consider beliefs at most of the fourth order. Players *initial* beliefs are those held at the information set $h^0 = \{t^0\}$.

II. Two Concepts of Guilt Aversion

Given his strategy s_j and initial first-order beliefs $\alpha_j(\cdot|h^0)$, player j forms an expectation about his material payoff: $E_{s_j, \alpha_j}[m_j|h^0] = \sum_{s_{-j}} \alpha_j(s_{-j}|h^0) \mathbf{m}_j(\mathbf{z}(s_j, s_{-j}))$. For any end node z consistent with s_j , the expression $D_j(z, s_j, \alpha_j) = \max\{0, E_{s_j, \alpha_j}[m_j|h^0] - \mathbf{m}_j(z)\}$ measures how much j is "let down." If at the end of the game i knew the terminal node z , the strategy profile $s_{-i} \in S_{-i}(z)$, and j 's initial beliefs α_j , then he could derive how much of $D_j(z, s_j, \alpha_j)$ is due to his behavior: $G_{ij}(z, s_{-i}, \alpha_j) = D_j(z, s_j, \alpha_j) - \min_{s_i} D_j(\mathbf{z}(s_i, s_{-i}), s_j, \alpha_j)$.

Our first concept draws directly on $G_{ij}(z, s_{-i}, \alpha_j)$. We say i is affected by *simple guilt* toward j if he has belief-dependent preferences represented by a utility function of the form

$$(1) \quad u_i^{SG}(z, s_{-i}, \alpha_j) = \mathbf{m}_i(z) - \sum_{j \neq i} \theta_{ij} G_{ij}(z, s_{-i}, \alpha_j), \\ s_{-i} \in S_{-i}(z), \quad \theta_{ij} \geq 0.$$

The exogenously given parameters θ_{ij} reflect i 's guilt sensitivity. Since i does not know s_{-i} or α_{-i} , and may not even observe z , u_i^{SG} does not represent a utility "experienced" by i . What we assume is that, given his first- and second-order beliefs, i tries to make the expected value of u_i^{SG} as large as possible.³

Whereas with simple guilt a player cares about the extent to which he lets another player down, our second formulation assumes that a player cares about others' inferences regarding

³ Equation (1) yields the same sequential best response correspondence as the slightly simpler function $v_i(z, s_{-i}, \alpha_{-i}) = \mathbf{m}_i(z) - \sum_{j \neq i} \theta_{ij} D_j(z, s_j, \alpha_j)$. We use (1) for two reasons: it is conceptually more appropriate (i cannot be "guilty" for others' behavior), and expression $G_{ij}(z, s_{-i}, \alpha_{-i})$ is needed below to define our second guilt concept.

the extent to which he is willing to let them down. We model this as follows. Given s_i and initial beliefs $\alpha_i(\cdot|h^0)$ and $\beta_i(h^0)$, we first compute how much i expects to let j down:

$$\begin{aligned}
 (2) \quad & G_{ij}^0(s_i, \alpha_i, \beta_i) \\
 &= E_{s_i, \alpha_i, \beta_i}[G_{ij}|h^0] \\
 &= \sum_{s_{-i}} \alpha_i(s_{-i}|h^0) G_{ij}(z(s_i, s_{-i}), s_{-i}, \beta_{ij}^0(h^0)),
 \end{aligned}$$

where $\beta_{ij}^0(h^0)$ denotes the initial (point) belief of i about the initial belief $\alpha_j(\cdot|h^0)$. Now, suppose $z \in Z$ is reached. The conditional expectation $E_{\alpha_j, \beta_j, \gamma_j}[G_{ij}^0|H_j(z)]$ measures j 's inference regarding how much i intended to let j down, or how much j "blames" i . We say i is affected by *guilt from blame* if he dislikes being blamed; i 's preferences are represented by

$$\begin{aligned}
 (3) \quad & u_i^{GB}(z, \alpha_{-i}, \beta_{-i}, \gamma_{-i}) = \mathbf{m}_i(z) \\
 & - \sum_{j \neq i} \theta_{ij} E_{\alpha_j, \beta_j, \gamma_j}[G_{ij}^0|H_j(z)], \quad \theta_{ij} \geq 0.
 \end{aligned}$$

Player i maximizes the expectation of u_i^{GB} , given his beliefs (up to the fourth order).

When we append the functions $(u_i^{SG})_{i \in N}$ (respectively, $(u_i^{GB})_{i \in N}$) to the given extensive game form, we obtain a *psychological game with simple guilt* (respectively *with guilt from blame*).⁴ We assume that the psychological game has complete information. In particular there is common knowledge of the psychological payoff functions (this is clearly farfetched, but incomplete information could be captured by making chance choose the parameters θ_{ij}).

III. Equilibrium Analysis

We adapt to the present framework the sequential equilibrium concept of David M. Kreps and

Robert Wilson (1982). An *assessment* is a profile $(\sigma, \alpha, \beta, \dots) = (\sigma_i, \alpha_i, \beta_i, \dots)_{i \in N}$ specifying behavior strategies and first- and higher-order beliefs. Assessment $(\sigma, \alpha, \beta, \dots)$ is *consistent* if there is a strictly positive sequence $\sigma^k \rightarrow \sigma$ such that for all $i \in N, h \in H_i, s_{-i} \in S_{-i}(h)$,

$$\begin{aligned}
 (4) \quad & \alpha_i(s_{-i}|h) = \\
 & \lim_{k \rightarrow \infty} \frac{\Pr_{\sigma^k}(s_i) \prod_{j \neq i} \Pr_{\sigma_j^k}(s_j)}{\sum_{s'_{-i} \in S_{-i}(h)} \Pr_{\sigma^k}(s'_i) \prod_{j \neq i} \Pr_{\sigma_j^k}(s'_j)},
 \end{aligned}$$

and higher-order beliefs at each information set are correct for all $i \in N, h \in H_i, \beta_i(h) = \alpha_{-i}, \gamma_i(h) = \beta_{-i}, \delta_i(h) = \gamma_{-i}$, and so on.

Fix a profile of utility functions of the form $u_i(z, s_{-i}, \alpha, \beta, \dots)$ (this covers u_i^{SG} and u_i^{GB} as special cases). A consistent assessment $(\sigma, \alpha, \beta, \dots)$ is a *sequential equilibrium (SE)* if each measure $\Pr_{\sigma_i}(\cdot|h)$ assigns positive conditional probability only to conditional expected payoff maximizing strategies: for all $i \in N, h \in H_i, s_i \in S_i(h), \Pr_{\sigma_i}(s_i|h) > 0 \Rightarrow s_i \in \arg \max_{s_i \in S_i(h)} E_{s'_i, \alpha_i, \beta_i, \dots}[u_i|h]$ (this sequential rationality condition is redundant, but well posed, at information sets where i is not active). If the payoff functions depend only on the end node, our definition of SE is equivalent to that of Kreps and Wilson. Adapting an existence proof from B&D, one can show that every psychological game with simple guilt, or guilt from blame, has an SE.⁵

We now list some results and examples about the relationships between SE with simple guilt and guilt from blame, as well as with SE and efficient outcomes of the "material-payoff game" with utility functions $u_i \equiv m_i$. First note that in any two-player game form without chance moves, for every pure-strategy, consistent assessment $(s, \alpha, \beta, \dots)$, every i and s'_i ,

$$\begin{aligned}
 (5) \quad & G_{ij}^0(s'_i, \alpha_i, \beta_i) \\
 &= \max\{0, \mathbf{m}_j(z(s)) - \mathbf{m}_j(z(s'_i, s_{-i}))\} \\
 &= E_{\alpha_j, \beta_j, \gamma_j}[G_{ij}^0|H_j(z(s'_i, s_{-i}))].
 \end{aligned}$$

⁴We build on B&D's framework, not that of Geanakoplos, Pearce, and Stacchetti (2004), which would not allow i 's utility to depend on other players' beliefs, in contrast to (1) and (3), or on updated beliefs, in contrast to (3).

⁵B&D argue that other solution concepts and forward induction reasoning should be explored. For space reasons, we do not pursue this here.

The first equality is an immediate consequence of consistency. The second follows from consistency, perfect recall, and observation of own material payoff. This implies:

OBSERVATION 1: *In any two-player, simultaneous-move game form without chance moves, for any given parameter profile $(\theta_{ij})_{i,j \in N, j \neq i}$, the pure strategy SE assessments of the psychological games with simple guilt and guilt from blame coincide.*

In other games, an SE with simple guilt need not be an SE with guilt from blame, and vice versa. To see this, consider first the following three-player simultaneous-move game form.

Example 1: Cleo (a dummy player) has \$2. Ann and Bob simultaneously decide whether to *steal* from Cleo or to *abstain*. If at least one of them *steals*, Cleo is left with \$0. If only one player *steals*, that player gets \$2. If two players *steal*, they get \$1 each. Ann and Bob are symmetrically affected by guilt toward Cleo: $\theta_{AC} = \theta_{BC} = \theta > 0$. If $1 < \theta < 2$, then the strategy profile *(abstain, abstain)* is an SE with simple guilt but not with guilt from blame. Note the intuition, if Ann or Bob deviates from *(abstain, abstain)* and *steals*, then since Cleo observes only her material payoff of \$0, she cannot be sure whom to blame. With guilt from blame, this shelters the deviator from some pangs under which a player affected by simple guilt must suffer. More formally, let $\hat{\alpha}^i = \alpha_C(a_i = \text{steal} | m_C = 0)$ be the ex post marginal probability that *i* deviated, as assessed by Cleo. By consistency, Cleo thinks two deviations are infinitely less likely than one, hence $\hat{\alpha}^A + \hat{\alpha}^B = 1$ and $\hat{\alpha}^i \leq 1/2$ for at least one *i*. This player has no incentive to *steal*, only if $2 - \theta \times 2\hat{\alpha}^i \leq 0$, that is, $\theta \geq 1/\hat{\alpha}^i \geq 2$. (Note how, with guilt from blame, off-equilibrium-path updated beliefs matter even in simultaneous-move game forms.)

Next, consider a two-player-plus-chance game form with asymmetric information.

Example 2: Ann first observes a chance move with equally likely outcomes *b* or *g*, and then chooses *in* or *out*. If she chooses *out*, Bob (a dummy player) gets \$2. If she chooses *in*, Bob's material payoff depends on chance: \$0 if *b*, \$8

if *g*. Ann always gets \$0 but is affected by guilt toward Bob. Look at the strategy profile (= strategy of Ann's) *(in, in)* (meaning *in* if *b*, *in* if *g*). Clearly this is not an SE with simple guilt. Bob initially expects to get $1/2 \times 0 + 1/2 \times 8 = 4$ —he is thus let down in the (expected) amount $1/2 \times 4 + 1/2 \times 0 = 2$. By deviating to *(out, in)*, Ann can change this to $1/2 \times 2 + 1/2 \times 0 = 1$. This is the unavoidable expected extent to which Bob will be let down. Thus, the expected guilt associated with *(in, in)* is $2 - 1 = 1$, as compared to $1 - 1 = 0$ for strategy *(out, in)*. Since material payoff is not an issue for Ann, she wants to deviate to *(out, in)*. Yet *(in, in)* is an SE with guilt from blame. It is supported by Bob's out-of-path beliefs such that if he got a material payoff of \$2 then he would think it is because Ann plays strategy *(in, out)*.⁶ The expected associated guilt is $[(1/2 \times 4 + 1/2 \times 2) - 1] = 2$, and this is how much Bob blames Ann if he observes a payoff of \$2. If Ann does not deviate, Bob gets a payoff of \$0 or \$8, infers that Ann is playing *(in, in)*, and therefore his blame on Ann is 1, the expected guilt associated with *(in, in)*. Therefore any deviation from *(in, in)* increases Bob's blame in expectation.

OBSERVATION 2: *In any simultaneous-move game form without chance moves, for any parameter profile $(\theta_{ij})_{i,j \in N, j \neq i}$, all the pure strategy SE assessments of the material payoff game are also SE of the psychological games with simple guilt and guilt from blame.*

PROOF:

Fix a simultaneous game form and an SE $(s, \alpha, \beta, \dots)$ of the material-payoff game. Then, if *i* deviates from s_i , he (weakly) decreases his material payoff. Given α , each player *j* expects to get exactly $m_j(s)$. Hence, if no deviations occur, no player *j* is let down. By consistency, this implies that, given (α, β, \dots) , each player *i* (weakly) increases in expectation the absolute value of each negative component of his psychological payoff function if he deviates. Therefore, a deviation by any player (weakly) decreases his total payoff.

⁶ Such a belief is consistent: consider the sequence $\sigma_A^k(in|g) = 1 - k^{-1}$, $\sigma_A^k(in|b) = 1 - k^{-2}$ for $k = 1, 2, \dots$

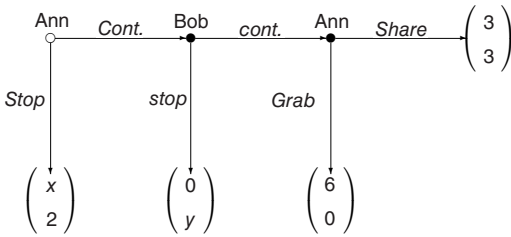


FIGURE 1. A PERFECT INFORMATION GAME FORM

The following parametrized example (Figure 1) shows that Observation 2 does not extend to sequential game forms. The example relates also to Observation 1.

Example 3: Suppose that $0 < x < 3$, $0 < y \leq 2$. Then [(*Stop*, *Grab*), *stop*] is the only SE of the material payoff game depicted in Figure 1 and it yields outcome $(x, 2)$. This outcome is not supportable by any SE of the psychological game with simple guilt if θ_{AB} is high enough. The reason is that if Ann correctly guesses that Bob initially expects \$2, then at history/node (*Cont.*, *cont.*) she would be sure to let Bob down in the amount of 2 by choosing to *Grab*. If $\theta_{AB} > \frac{3}{2}$, Ann would prefer to *Share*. Anticipating this, Bob would continue after *Cont.*, and Ann would deviate to *Cont.* at the beginning of the game. Thus, Observation 2 does not extend to sequential game forms for simple guilt, even if we look only at equilibrium outcomes.

On the other hand, for the same parameter values [(*Stop*, *Grab*), *stop*] is an SE of the game with guilt from blame. The reason is that if Ann does not *Stop*, the blame by Bob on Ann is $\mathbf{m}_2(\text{Stop}) - \mathbf{m}_2(\text{Cont.}, \text{stop.}) = 2 - y$, independently of what happens afterward, because this is how much Ann intended to let Bob down. Therefore, Ann would have no incentive to *Share* if given the opportunity. This shows that Observation 1 does not extend to sequential game forms, even if we look only at equilibrium outcomes.

Now, suppose that $x > 6$ and $y < 0$. The only SE of the material payoff game is [(*Stop*, *Grab*), *cont.*]. If $\theta_{AB} > \frac{3}{2}$, this is not a SE of the game with guilt from blame. Since the equilibrium strategy of Bob is to choose *cont.*, in this case, Ann’s action in the subgame affects the

guilt blamed by Bob on Ann, who would rather *Share*.

We close this section with a result that relates the “materially efficient” outcomes of the game form with the SE of the game with simple guilt.⁷

OBSERVATION 3: Fix a game form without chance moves and let z^* be a terminal node s.t. for all $z \in Z \setminus \{z^*\}$ there is some $j \in N$ s.t. $\mathbf{m}_j(z) < \mathbf{m}_j(z^*)$. Then for sufficiently high guilt sensitivities $(\theta_{ij})_{i,j \in N, j \neq i}$, there is an SE of the game with simple guilt that yields z^* with probability one.

PROOF:

For each node $z \neq z^*$, fix a player $j(z)$ such that $\mathbf{m}_{j(z)}(z) < \mathbf{m}_{j(z)}(z^*)$. Let $\theta^* = \max \{[\mathbf{m}_i(z) - \mathbf{m}_i(z^*)] / [\mathbf{m}_{j(z)}(z^*) - \mathbf{m}_{j(z)}(z)] : i, z \text{ s.t. } \mathbf{m}_i(z) > \mathbf{m}_i(z^*)\}$ (let the maximum over an empty set be 0 by convention). Fix a game with simple guilt such that $\theta_{ij} > \theta^*$ for all $i, j \neq i$. Now consider a modified version of this game: for every player i , information set $h \subset X_i$ of i and action $a \in A(h)$ (where $A(h)$ is the set of feasible actions at h) define a minimal probability $\varepsilon(a|h) \in (0, 1)$ so that $\sum_{a \in A(h)} \varepsilon(a|h) \leq 1$. For any strictly positive behavior strategy profile σ , let $(\alpha^\sigma, \beta^\sigma, \dots)$ denote the unique system of conditional beliefs consistent with σ . An assessment $(\sigma, \alpha^\sigma, \beta^\sigma, \dots)$ is an ε -equilibrium if $\sigma_i(a|h) \geq \varepsilon(a|h)$ for all $i, h \in H_i$ with $h \subset X_i$, $a \in A(h)$, and $\sigma_i(a|h) = \varepsilon(a|h)$ whenever a is not a best response at h against $(\sigma, \alpha^\sigma, \beta^\sigma, \dots)$. It can be shown by standard compactness-continuity arguments that for every vector ε there is at least one ε -equilibrium (see the equilibrium existence proof in B&D for details). Let ε^k be a sequence of minimal probability vectors such that $\varepsilon^k(a|h) \rightarrow 0$ if h is off the z^* -path, and $\varepsilon^k(a|h) \rightarrow 1$ if (h, a) is on the z^* -path. By compactness, there is a sequence of ε^k -equilibria converging to some consistent assessment $(\sigma, \alpha, \beta, \dots)$. Clearly $\text{Pr}_\sigma(z^*) = 1$. By continuity of expected utility in beliefs, $\sigma_i(a|h) > 0$ implies that a is a best reply at h against $(\sigma, \alpha, \beta, \dots)$ if h is off the z^* -path. The choice of $(\theta_{ij})_{i \in N, j \neq i}$ implies that there are no

⁷ We conjecture that the result extends to games with guilt from blame, but we can prove it only under the somewhat restrictive assumption that at the end of the game every player can identify who deviated from any given path.

incentives to deviate from the z^* -path. By the one-shot-deviation principle (which applies to the psychological games considered here; again cf. B&D), sequential rationality is satisfied. Therefore $(\sigma, \alpha, \beta, \dots)$ is an SE.

The next example shows that Observations 2 and 3 do not extend to games with chance.

Example 4: Consider a two-player-plus-chance game form where chance chooses between b and g , with equal probabilities, and Ann simultaneously chooses between actions r and s . Bob is a dummy player. Ann's material payoff is constant, say \$0. Bob's material payoff is \$5 under the "safe" action s , whereas the "risky" action r yields either \$0 if $s_c = b$ or \$12 if $s_c = g$. r is a pure equilibrium of the material-payoff game. It is also "materially efficient" in the sense that deviating to s decreases Bob's expected material payoff. However, r cannot be an equilibrium if $\theta_{AB} > 0$; if there is common belief that r is played, then r yields expected utility $0 - \theta_{AB}[\frac{1}{2} \times 0 + \frac{1}{2} \times (\frac{1}{2} \times 12 + \frac{1}{2} \times 0 - 0)] = -3\theta_{AB}$ to Ann, whereas s yields $0 - \theta_{AB}(\frac{1}{2} \times 12 + \frac{1}{2} \times 0 - 5) = -\theta_{AB} > -3\theta_{AB}$.

IV. Concluding Remarks

We develop a general theory of guilt aversion and show how to solve for sequential equilibria. We hope the approach will prove useful for a variety of applications concerning economic situations where it seems plausible that decision makers are affected by guilt. Contributions to public goods, contractual relationships, and work in teams are natural candidates.

To end on a more general note, psychological game theory provides the intellectual home for our approach. Few previous applications of that framework exist. The most prominent examples concern kindness-based reciprocity (e.g., Matthew Rabin 1993; Dufwenberg and Georg Kirchsteiger 2004), anxiety (Andrew Caplin and John Leahy 2004; Caplin and Leahy 2001), and social respect (B. Douglas Bernheim 1994; Dufwenberg and Michael Lundholm 2001; these authors do not explicitly refer to psychological games, but their work fits the framework of B&D). The usefulness of psychological game theory for studying these diverse kinds of motivation augurs well for the framework's

potential for analyzing other phenomena including disappointment, regret, anger, surprise, shame, and joy.

REFERENCES

- Attanasi, Giuseppe, and Rosemarie Nagel.** 2006. "Actions, Beliefs and Feelings: An Experimental Study on Dynamic Psychological Games." Unpublished.
- Bacharach, Michael, Gerardo A. Guerra, and Daniel J. Zizzo.** Forthcoming. "The Self-Fulfilling Property of Trust? An Experimental Study." *Theory and Decision*.
- Battigalli, Pierpaolo, and Martin Dufwenberg.** 2005. "Dynamic Psychological Games." Bocconi University, Innocenzo Gasparini Institute for Economic Research Working Paper 287.
- Baumeister, Roy F., Arlene M. Stillwell, and Todd F. Heatherton.** 1994. "Guilt: An Interpersonal Approach." *Psychological Bulletin*, 115(2): 243–67.
- Bernheim, B. Douglas.** 1994. "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841–77.
- Caplin, Andrew, and John Leahy.** 2001. "Psychological Expected Utility Theory and Anticipatory Feelings." *Quarterly Journal of Economics*, 116(1): 55–79.
- Caplin, Andrew, and John Leahy.** 2004. "The Supply of Information by a Concerned Expert." *Economic Journal*, 114(497): 487–505.
- Charness, Gary, and Martin Dufwenberg.** 2006. "Promises and Partnership." *Econometrica*, 74(6): 1579–1601.
- Dufwenberg, Martin.** 1995. "Time-Consistent Wedlock with Endogenous Trust." PhD diss. Uppsala University.
- Dufwenberg, Martin.** 2002. "Marital Investments, Time Consistency and Emotions." *Journal of Economic Behavior and Organization*, 48(1): 57–69.
- Dufwenberg, Martin, and Uri Gneezy.** 2000. "Measuring Beliefs in an Experimental Lost Wallet Game." *Games and Economic Behavior*, 30(2): 163–82.
- Dufwenberg, Martin, and Georg Kirchsteiger.** 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2): 268–98.

- Dufwenberg, Martin, and Michael Lundholm.** 2001. "Social Norms and Moral Hazard." *Economic Journal*, 111(473): 506–25.
- Elster, Jon.** 1998. "Emotions and Economic Theory." *Journal of Economic Literature*, 36(1): 47–74.
- Geanakoplos, John, David Pearce, and Ennio Stacchetti.** 1989. "Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1(1): 60–79.
- Guerra, Gerardo, and Daniel John Zizzo.** 2004. "Trust Responsiveness and Beliefs." *Journal of Economic Behavior and Organization*, 55(1): 25–30.
- Huang, Peter H., and Ho-Mou Wu.** 1994. "More Order without More Law: A Theory of Social Norms and Organizational Cultures." *Journal of Law, Economics, and Organization*, 10(2): 390–406.
- Kreps, David M., and Robert Wilson.** 1982. "Sequential Equilibria." *Econometrica*, 50(4): 863–94.
- Rabin, Matthew.** 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5): 1281–1302.
- Tangney, June Price.** 1995. "Recent Advances in the Empirical Study of Shame and Guilt." *American Behavioral Scientist*, 38(8): 1132–45.