Guilt in Games

P. Battigalli and M. Dufwenberg (AER, 2007)

Presented by Luca Ferocino

March 21st, 2014
“Guilt is a cognitive or an emotional experience that occurs when a person realizes or believes - accurately or not - that he or she has compromised his or her own standards of conduct or has violated a moral standard, and bears significant responsibility for that violation”

What Is the Paper About?

Framework

The paper gives a contribution to the literature of *psychological game theory*, in which players’ utilities are allowed to depend on beliefs.

Questions Addressed:

1. How can guilt be modeled?
2. How are human interaction and economic outcomes influenced?

Aim of the paper

Develop a “portable” theory of guilt aversion and show how to solve for sequential equilibria.
Outline

1 Game Theoretic Preliminaries

2 Two Concepts of Guilt Aversion
   • Simple Guilt
   • Guilt from Blame

3 Equilibrium Analysis

4 Results and Examples
Take 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 behavior, $j$ 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 analyze equilibria when players are motivated, in part, by a desire to avoid guilt.
We consider finite extensive game trees, mathematical structures given by:

$$\Gamma = \langle I, C, m, (\bar{A}_i, A_i(\cdot))_{i \in I} \rangle$$

where:
- $I$ is the set of players
- $T$ the set of nodes with distinguished root $t_0$
- $Z$ is the set of terminal nodes
- $X = T \setminus Z$ is the set of non-terminal histories
- Let $X$ be partitioned into subsets $X_i$ of decision nodes $\forall i \in I$ and a set of chance nodes $X_c$
Information Structure and Material Consequences

Information Structure

- An information set is defined by a “history”, $h$. The information structure of player $i$ is given by the set $H_i \subset 2^T$ (i.e. $H_i$ is a partition of $T$)

- The information set containing node $t$ is denoted $H_i(t)$: set of histories that are not prevented by node $t$

- The information structure $H_i$ satisfies perfect recall

Material Consequences

- Material consequences are determined by a function $m_i : Z \rightarrow \mathbb{R}$, $i \in I$

- Material payoffs are denoted by $m_i = m_i(z)$

- Assume: $m_i(z') \neq m_i(z'') \implies H_i(z') \neq H_i(z'')$: $i$ observes his material payoff
Strategies

Pure Strategies

- A pure strategy \( s_i \) specifies a contingent choice \( \forall h \in H_i \) where \( i \) is active \( (h \subset X_i) \)
- \( S_i \) denotes the set of pure strategies of \( i \)
- \( S = S_c \times \prod_{i \in I} S_i \), \( S_{-i} = S_c \times \prod_{j \neq i} S_j \)
- \( \forall i, \forall h, S_i(h) \) denotes the set of \( i \)'s strategies allowing \( h \).
- A strategy profile \( s \in S \) yields an end node \( z(s) \)

Behavior Strategies

- A behavior strategy for player \( i \) is an array of probability measures \( \sigma_i(\cdot|h), h \in H_i, h \in X_i \)
- \( \sigma_i(a|h) \) is the probability of choosing action \( a \) at \( h \)
- By perfect recall, can compute the conditional prob. \( Pr_{\sigma_i}(s_i|h), h \in H_i \)
Beliefs

- $\forall h \in H_i, \alpha_i(\cdot|h) \in \Delta(S_{-i}(h))$ denotes player $i$’s belief about the strategies of co-players (and chance)

- $\alpha_i = (\alpha_i(\cdot|h))_{h \in H_i}$ is the system of first order beliefs of player $i$

- Player $i$ also holds at each $h \in H_i$ a second order belief $\beta_i(h)$ about the first order belief system $\alpha_j$ of each co-player $j$, a third order belief $\gamma_i(h)$ about second order beliefs, and so on...

Beliefs at different histories are not mutually independent $\implies$ They must satisfy Bayes Rule and common certainty that Bayes Rule holds
The Concept of Disappointment

Given his strategy $s_j$ and initial first-order beliefs $\alpha_j(\cdot|h^0)$, player $j$ forms expectations about his material payoff:

$$E_{s_j,\alpha_j}(m_j|h^0) = \sum_{s_{-j}} \alpha_j(s_{-j}|h^0)m_j(z(s_j, s_{-j}))$$

For any end node $z$ consistent with $s_j$, the measure of how much $j$ is “let down” is given by:

$$D_j(z, s_j, \alpha_j) = \max \{0, E_{s_j,\alpha_j}[m_j|h^0] - m_j(z)\}$$
Simple Guilt

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 $\alpha_j$, then he could derive how much of $D_j(z, s_j, \alpha_j)$ is due to his behavior:

$$G_{ij}(z, s_{-i}, \alpha_j) = D_j(z, s_j, \alpha_j) - \min_{s_i} D_j(z(s_i, s_{-i}), s_j, \alpha_j)$$

Definition

Player $i$ is affected by simple guilt toward player $j$ if he has belief-dependent preferences represented by a utility function of the form:

$$u_i^{SG}(z, s_{-i}, \alpha_{-i}) = m_i(z) - \sum_{j \neq i} \theta_{ij} G_{ij}(z, s_{-i}, \alpha_j)$$

where $s_{-i} \in S_{-i}(z)$, $\theta_{ij} \geq 0$. $\theta_{ij}$ reflects $i$’s guilt sensitivity.

We assume that $i$ maximizes the expected value of $u_i^{SG}$, given his beliefs.
The Concept of Blame

- The concept of “Guilt from Blame” assumes that a player cares about others’ inferences regarding the extent to which he is willing to let them down.

- Given $s_i$ and initial beliefs $\alpha_i(\cdot | h^0)$ and $\beta_i(h^0)$, **ex-ante** $i$ expects to let $j$ down in the amount:

$$G^0_{ij}(s_i, \alpha_i, \beta_i) = E_{s_i, \alpha_i, \beta_i}[G_{ij} | h^0] = \sum_{s_{-i}} \alpha_i(s_{-i} | h^0) G_{ij}(z(s_i, s_{-i}), s_{-i}, \beta^0_{ij}(h^0))$$

where $\beta^0_{ij}(h^0)$ denotes the initial (point) belief of $i$ about the initial belief $\alpha_j(\cdot | h^0)$. 
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, alternatively, how much $j$ “blames” $i$.

**Definition**

Player $i$ is affected by **guilt from blame** if he dislikes being blamed; $i$’s preferences are represented by:

$$u_i^{GB}(z, \alpha_{-i}, \beta_{-i}, \gamma_{-i}) = m_i(z) - \sum_{j \neq i} \theta_{ij} E_{\alpha_j, \beta_j, \gamma_j}[G_{ij}^0|H_j(z)]$$

We assume that $i$ maximizes the **expected value** of $u_i^{GB}$, given his beliefs (up to the 4th order).
Assessments and Consistency

Assessment

An assessment is a profile \((\sigma, \alpha, \beta, \ldots) = (\sigma_i, \alpha_i, \beta_i, \ldots)_{i \in I}\) specifying behavior strategies and first- and higher-order beliefs.

Consistent Assessment

An assessment is consistent if there is a strictly positive sequence \(\sigma^k \rightarrow \sigma\) such that for all \(i \in I, h \in H_i, s_{-i} \in S_{-i}(h),\)

\[
\alpha_i(s_{-i}|h) = \lim_{k \rightarrow \infty} \frac{Pr_{\sigma_c}(s_c) \prod_{j \neq i} Pr_{\sigma_j^k}(s_j)}{\sum_{s'_{-i} \in S_{-i}(h)} Pr_{\sigma_c}(s'_{c}) \prod_{j \neq i} Pr_{\sigma_j^k}(s'_j)}
\]

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

Definition

Fix a profile of utility functions of the form $u_i(z, s_{-i}, \alpha, \beta, \ldots)$. A consistent assessment $(\sigma, \alpha, \beta, \ldots)$ is a sequential equilibrium (SE) if each measure $Pr_{\sigma_i}(\cdot|h)$ assigns positive conditional probability only to conditional expected payoff maximizing strategies:

$$Pr_{\sigma_i}(s_i|h) > 0 \iff s_i \in \arg \max_{s_i' \in S_i(h)} E_{s_i', \alpha_i, \beta_i, \ldots}[u_i|h]$$

for all $i \in I, h \in H_i, s_i \in S_i(h)$

It can be proved that every psychological game with simple guilt, or guilt from blame, has a Sequential Equilibrium
Proposition (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.

Consider a “material payoff game” with utility functions \(u_i \equiv m_i\). In any two-player game form without chance moves (no need simultaneous moves), for every pure strategy, consistent assessment \((s, \alpha, \beta, \ldots)\), every \(i\) and \(s'_i\),

\[
G^0_{ij}(s'_i, \alpha_i, \beta_i) = \max\{0, m_j(z(s)) - m_j(z(s'_i, s_{-i}))\} = E_{\alpha_j, \beta_j, \gamma_j}[G^0_{ij} | H_j(z(s'_i, s_{-i}))]
\]
A Three-Player “Steal” Game

Not Steal

Steal

B

(0, 0, 2) (0, 2, 0) (2, 0, 0) (1, 1, 0)
Suppose A and B are symmetrically affected by guilt towards C:
\[ \theta_{AC} = \theta_{BC} = \theta \]

Suppose now that C correctly believes w.p. 1 that A and B are going to play \((\text{Not Steal}, \text{Not Steal})\)

Then, if \(1 < \theta < 2\), \((\text{Not Steal}, \text{Not Steal})\) is a SE with SG but not with GB

Why? Because C cannot be sure about whom to blame!
GB Equilibria in Three-Player “Steal” Game

- Let $\hat{\alpha}^i = \alpha_C(a_i = \text{steal}|m_C = 0)$ be the ex-post marginal prob. that $i$ deviates, as assessed by $C$.

- By consistency, $C$ thinks that two deviations are infinitely less likely than one, i.e. $(1 - \hat{\alpha}^B - \hat{\alpha}^C) = 0$.

- Therefore, $\hat{\alpha}^B + \hat{\alpha}^C = 1$, and $\hat{\alpha}^i \leq 1/2$ for at least one $i$.

- This player has no incentive to Steal only if

\[ 2 - \theta \times 2\hat{\alpha}^i \leq 0 \implies \theta \geq 1/\hat{\alpha}^i \geq 2 \]
Proposition (2)

In any simultaneous-move game form without chance moves, for any parameter profile \((\theta_{ij})_{i,j \in I, j \neq i}\), 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.

Intuition:

- Fix a simultaneous game form and a SE \((s, \alpha, \beta, \ldots)\) of the material payoff game.
- If \(i\) deviates from \(s_i\), then he weakly decreases his payoff.
- If no deviations occur, no player \(j\) is let down, while by deviating each player \(i\) who deviates increases in expectation the absolute value of each negative component of his psychological payoff function.
Suppose $0 < x < 3$, $0 < y \leq 2$. Then $[(\text{Stop, Grab}), \text{stop}]$ is the only SE of the MP game, yielding $(x, 2)$

This outcome is not supported by any SE of the psychological game with SG for high enough $\theta_{AB}$
Suppose that Ann correctly guesses that Bob expects $2.

At history (Cont., cont.), by choosing Grab, she would let Bob down in the amount: \( \max\{0, [2 - 0]\} = 2 \)

Her guilt will be given by: \( \max\{0, [2 - 0]\} - \max\{0, [2 - 3]\} = 2 \)

Ann would prefer to Share iff \( \theta_{AB} > 3/2 \)

Anticipating this, Bob would continue after Cont. and Ann would deviate to Cont. at \( h^0 \)

Proposition (2) does not extend to sequential game forms for SG
Suppose that Ann correctly guesses that Bob expects $2

\[ [(Stop, Grab), \text{stop}] \text{ is a SE of the game with guilt from blame} \]

Suppose at \( h^0 \) Ann deviates to \( \text{Cont.} \). \( \Rightarrow \) the blame by Bob on Ann

\[ m_B(\text{Stop}) - m_B(\text{Cont.}, \text{stop}) = 2 - y > 0 \]

The result holds independently of what happens afterwards, so Ann

has no incentive to deviate

**Proposition (1) does not extend to sequential game forms**
A Result About Efficiency

Proposition (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 $j \in I$ s.t. $m_j(z) < m_j(z^*)$. Then for sufficiently high guilt sensitivities $(\theta_{ij})_{i,j \in I, j \neq i}$ there is a SE of the game with SG that yields $z^*$ with prob. one.
Conclusions

- The paper provides the tools to analyze guilt and its impact on individual choices from a theoretical point of view.

- It provides two different definitions of guilt:
  - Simple Guilt
  - Guilt from Blame

- It develops a general theory of guilt aversion and shows how to solve for SE in psychological games.

- This framework is potentially useful for analyzing also other phenomena:
  - disappointment
  - regret
  - anger
  - ...
Thank You!

Q & A