Guilt, Shame, and Games with Belief-Dependent Preferences

Pierpaolo **Battigalli** (Bocconi University)

May, 8 2008

Abstract

Presented at the workshop *"Understanding Moral Emotions, Perspectives From Cognitive Sciences and Economics".*

I apply the theoretical framework *dynamic psychological games* to model belief-dependent preferences related to guilt and shame. With *simple guilt* preferences, an agent dislikes letting other agents down. With *guilt from blame* preferences an agent dislikes that other agents think that she intended to let them down. With *shame* preferences, an agnet dislikes that others think that she is "bad", or that she chose a "bad action". With the latter two types of preferences the terminal information of agents (what they learn at the end of the interaction) is crucial. Changing this information affects behavior even though agents cannot act on it, because they learn it *ex post*. Such effects cannot be explained by standard game theory, but they occur in the lab.

Beliefs in the utility function help explain "non-standard" behavior, including behavior affected by emotions, in

(1) decision problems: e.g. avoidance of anxiety (Caplin & Leahy, QJE01) or of disappointment → dynamic consistency may be an issue

(2) *interactive decision problems* with other-regarding preferences: e.g. reciprocity (Rabin AER-93, Dufwenberg & Kirchsteiger GEB-04, Falk & Fischbacher GEB-06), conformity/social respect (Bernheim JPE-94, Dufwenberg & Lundholm EJ-01, Tadelis), concern of experts for the emotions of others (Caplin&Leahy EJ-04) → endogenous higher-order beliefs are crucial (Geanakoplos *et al.* GEB-89).

My paper "Dynamic Psychological Games" (DPG, with M. Dufwenberg, forthcoming in JET) provides a theoretical framework covering (1) and (2), but the main focus is (2).

DPG argues that conditional higher-order beliefs, beliefs of others and plans of action should be arguments of the utility function (on top of material consequences). I apply DPG to analyze guilt (see Battigalli & Dufwenberg AER-07) and shame (cf Tadelis-07) in the context of games.

Guilt

Psychologists: "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" (Baumeister *et al*, *Psychological Bulletin*, 1994).

A quite substantial body of experimental evidence on trust games supports the view that players dislike failing to live up to the expectations of others (Dufwenberg & Gneezy, GEB-00, Bacharach *et al.* mimeo-02, Guerra & Zizzo JEBO-04, Charness & Dufwenberg, Econometrica-06, Attanasi & Nagel-07).

Building on previous work on trust games (Dufwenberg JEBO-02), we model guilt in two ways:

Say that agent i *lets* agent j *down* (disappoints j) if as a result of i's choice of strategy j gets a lower material payoff than j initially expected.

Simple guilt
Agent i's guilt may depend on how much i believes he lets j down

Guilt from blame
Agent i's guilt may also depend on how much i believes j believes i intended to let j down.

Shame

Psychologists: "shame is usually dependent on public exposure of one's frailty or failings" (Gehn & Scherer 88).

More generally, shame seems to be related to a concern for the (ex post) opinions of others about us.

In situations where agent i may hurt (or omit to help) agent j, it seems that what matters is whether i's action is revealed to j and hence affects what j thinks of i (Tadelis 07).

Both guilt and shame can be modeled as belief-dependent other-regarding preferences, whereby agents care about the beliefs of others.

Trust game

I consider guilt and shame in the context of a simple interactive situation: Ann has safe option (*Out*), but she can doubles the total *expected* material payoff choosing risky option *In*; Bob can *Grab* all payoff for himself, if he does not Ann can be hurt by chance.

$$\begin{tikzcd}
\text{Ann} \arrow[r, "In"] \arrow[d, "Out"'] & \text{Bob} \arrow[r, "Share"] \arrow[d, "Grab"'] & \text{chance} \arrow[r, "p"] \arrow[d, "1-p"'] & \begin{pmatrix} 2/p \\ 2 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 0 \\ 4 \end{pmatrix} & \begin{pmatrix} 0 \\ 2 \end{pmatrix} &
\end{tikzcd}$$

Trust Game with monetary random payoffs

Strategies of Ann: I, O.

Strategies of Bob: G (*Grab* if *In*), S (*Share* if *In*)

Assume Ann just maximizes her expected material payoff ($u_{Ann} = m_A$). Bob may be affected by guilt of shame.

Relevant beliefs of Ann:
$\alpha = \mathsf{Pr}_{Ann}[\mathsf{S}]$ (initial 1st order belief), $\hat{\alpha} = \mathsf{Pr}[\alpha|\text{observed result}]$ (ex post 1st order belief), $\alpha < \frac{1}{2} \Rightarrow Out$, $\alpha > \frac{1}{2} \Rightarrow In$

Relevant beliefs of Bob:
$\beta = \mathsf{Pr}_{Bob}[\alpha|In]$ (interim 2nd order belief)
$\hat{\beta}_G = \mathsf{Pr}_{Bob}[\hat{\alpha}|In, Grab]$, $\hat{\beta}_S = \mathsf{Pr}_{Bob}[\hat{\alpha}|In, Share]$ (interim 2nd ord.)

Simple guilt aversion: $u_{Bob} = m_{Bob} - \theta^g \,\mathsf{max}\,(0, E_{Ann}[\widetilde{m}_{Ann}] - m_{Ann})$

$$\begin{tikzcd}
\text{Ann} \arrow[r, "In"] \arrow[d, "Out"'] & \text{Bob} \arrow[r, "S"] \arrow[d, "G"'] & \begin{pmatrix} 2 \\ 2 \end{pmatrix} \\
\begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 0 \\ 4 - \theta^g \times 2\alpha \end{pmatrix} &
\end{tikzcd}$$

Trust Game with simple guilt aversion: chance replaced by exp. values

Ex post info. does not matter. Bob's strategy depends on $\beta = \mathsf{Pr}_{Bob}[\alpha|In]$:
$\beta > \frac{1}{\theta^g}$ $(2 > 4 - 2\theta^g\beta) \Rightarrow$S, $\beta < \frac{1}{\theta^g} \Rightarrow$G

Experimental evidence supports this result:

S positively correlated to β, letting Ann send free-form messages to Bob increases α, β and frequency of I, S (Charness & Dufwenberg 06).

B subjects seem to exhibit guilt preferences: reveal relatively high θ^g by answering hypothetical questions; making revealed θ^g common knowledge between Ann and Bob increases α, β and frequency of I, S (Attanasi & Nagel 07).

This experimental work keeps *ex post information structure fixed*.

Simple Shame: $u_{Bob} = m_{Bob} - \theta^s(1 - \hat{\alpha})$

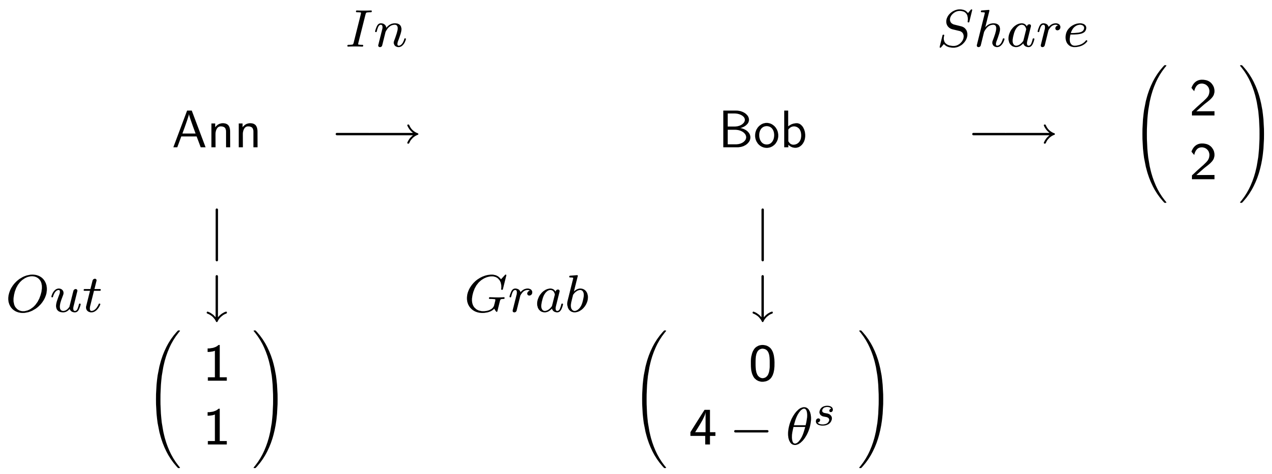
Bob's decision depends on $\hat{\beta}_G$ and $\hat{\beta}_S$, the game looks like this

$$\begin{tikzcd}
\text{Ann} \arrow[r, "In"] \arrow[d, "Out"'] & \text{Bob} \arrow[r, "S"] \arrow[d, "G"'] & \begin{pmatrix} 2 \\ 2 - \theta^s(1 - \hat{\beta}_S) \end{pmatrix} \\
\begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 0 \\ 4 - \theta^s(1 - \hat{\beta}_G) \end{pmatrix} &
\end{tikzcd}$$

Trust Game with simple shame aversion: chance replaced by exp. values

Ex post information structure matters:

Suppose (Bob knows that) Ann has perfect ex post info, then $\hat{\beta}_G = 0$, $\hat{\beta}_S = 1$. The game looks like this:


In
$Share$
Ann
Bob
$\begin{pmatrix} 2 \\ 2 \end{pmatrix}$
Out
$Grab$
$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$
$\begin{pmatrix} 0 \\ 4 - \theta^s \end{pmatrix}$

Now suppose (Bob knows that) Ann only observes m_A:

$$\begin{aligned}
\Pr_{Ann}[S|m_A &= 0] = \frac{\alpha - \alpha p}{1 - \alpha p} \\
\Pr_{Ann}[S|m_A &= 2] = 1 \\
\hat{\beta}_G &= E_{Bob}\left[\frac{\alpha - \alpha p}{1 - \alpha p}|In\right] \\
\hat{\beta}_S &= p + (1-p)\hat{\beta}_G
\end{aligned}$$

Under imperfect ex post information the expected psychological utility of Grab is higher and the expected psychological utility of Share is lower. *Bob is more likely to Grab under imperfect information*, anticipating this *Ann trusts Bob less* (more likely to play *Out*).

Experimental evidence supports this result on the impact of the ex post information structure (Tadelis 2007).

According to standard game theory preferences depend only on actions and random events, and this implies that only the information the players have when they are active may be relevant. Thus, contrary to experimental evidence, standard game theory rules out the impact of ex post information. This shows that observed phenomena explained with belief-dependent preferences cannot be observed with standard game theory (unlike what some "orthodox" theorists claim).

Guilt from blame: Bob dislikes being "blamed" by Ann for the *intention* of causing an unexpectedly low m_A ($m_A < E_{Ann}[\widetilde{m}_A]$). I avoid the formalism here (see Battigalli & Dufwenberg, 2007). Enough to note that Bob "feels bad" if *ex post* Ann believes that he intended to disappoint her.

Bob's intention to disappoint depends on his 2nd order belief β, thus Bob's psychological utility depends on Ann *ex post* (terminal) *third order beliefs*.

Bob's decision depends on his beliefs about the ex post 3rd order beliefs of Ann.

Again, the ex post information structure matters: in the trust game, under imperfect ex post information Bob is more likely to *Grab*, than under perfect ex post information, because a low m_A may be the result of bad luck rather than an intention to disappoint. This is anticipated (lower α).

Shame and ex post beliefs about personality/type

Suppose Bob can be more or less altruistic, e.g.

$$u_{Bob} = m_{Bob} + \theta^a m_{Ann} + *$$

If there is no other term ($* = 0$), standard game theory works. But suppose Bob dislikes to be thought by Ann as a greedy guy, e.g., Bob is "greedy" if $\theta^a < \overline{\theta}$ and

$$u_{Bob} = m_{Bob} + \theta^a m_{Ann} - \theta^s \Pr_{Ann}[\widetilde{\theta}^a < \overline{\theta} | \text{observed result}]$$

Again, we have to use the theory of games with belief-dependent preferences, and the ex post information structure has an impact. In the trust game, under imperfect ex post information Bob is more likely to *Grab*, because a low m_A may be the result of bad luck rather than greed. This is anticipated (lower α).

Conclusions

The impact of emotions in interactive situations can be analyzed with formal models, applying a generalization of standard game theory that allows for belief-dependent preferences. This includes guilt and shame.

More theoretical work to be done:

- identify differences in predictions that help discriminate experimentally between models that seem to have similar qualitative implications (differenr forms of shame, shame and guilt from blame)

- better models of guilt

- model of shame preferences that can be applied to general game forms (interactive situations)

The latter may be harder: is shame more situation specific than shame?

Modeling shame and (some types of) guilt points to the relavance of the ex post information structure, and more generally the information of inactive players, which is excluded by standard game theory.

Experimental evidence tends to support results based on game models with belief-dependent preferences and the relevance of information of inactive players.

References

[1] G. Attanasi and R. Nagel, A survey of psychological games: theoretical findings and experimental evidence, in: A. Innocenti and P. Sbriglia (Eds.), *Games, Rationality and Behaviour. Essays on Behavioural Game Theory and Experiments*, Palgrave McMillan, Houndmills, 2007, pp 204-232.

[2] M. Bacharach, G. Guerra and D.J. Zizzo, The self-fulfilling property of trust: an experimental study, *Theory Dec*. 63 (2007), 349–388.

[3] P. Battigalli and M. Dufwenberg, Guilt in Games, *Amer. Econ. Rev., Papers & Proceedings*, 97 (2007), 170-176.

[4] P. Battigalli and M. Dufwenberg, Dynamic Psychological Games, *J. Econ. Theory*, forthcoming.

[5] D. Bernheim, A Theory of Conformity, *J. Polit. Economy*, 102 (1994), 841-877.

[6] A. Caplin, Fear as a policy instrument: economic and psychological perspectives on intertemporal choice, in: G. Loewenstein, D. Read and R.F. Baumeister (Eds.), *Time and Decision*, Russell Sage Foundation, New York NY, 2003.

[7] A. Caplin and K. Eliaz, AIDS policy and psychology: a mechanism design approach, *RAND J. Econ*. 34 (2003), 631-646.

[8] A. Caplin and J. Leahy, Psychological expected utility theory and anticipatory feelings, *Quart. J. Econ*. 116 (2001), 55-79.

[9] A. Caplin and J. Leahy, The supply of information by a concerned expert, Econ. J. 114 (2004), 487-505.

[10] G. Charness and M. Dufwenberg, Promises and Partnership, *Econometrica* 74 (2006), 1579-1601.

[11] M. Dufwenberg, Marital investment, time consistency and emotions, *J. Econ. Behavior and Organ.* 48 (2002), 57-69.

[12] M. Dufwenberg and U. Gneezy, Measuring beliefs in an experimental lost wallet game, Games Econ. Behav. 30 (2000), 163-182.

[13] M. Dufwenberg and G. Kirchsteiger, A theory of sequential reciprocity, *Games Econ. Behav.* 47 (2004), 268-298.

[14] Dufwenberg, M. and M. Lundholm, Social Norms and Moral Hazard, *Econ. J.* 111 (2001), 506-525.

[15] A. Falk and U. Fischbacher, A theory of reciprocity, *Games Econ. Behav.* 54 (2006), 293-315.

[16] J. Geanakoplos, D. Pearce and E. Stacchetti, Psychological games and sequential rationality, *Games Econ. Behav.* 1 (1989), 60-79.

[17] T.L. Gehm and K.R Scherer, Relating situation evaluation to emotion differentiation: nonmetric analysis of cross-cultural questionnaire data. In: K.R. Scherer (Ed.), *Facets of emotion: recent research*. Lawrence Earlbaum, Hillsdale NJ, 2007, pp 61-77.

[18] G. Guerra and D.J. Zizzo, Trust responsiveness and beliefs, *J. Econ. Behav. Organ.* 55 (2004), 25-30.

[19] M. Rabin, Incorporating fairness into game theory and economics, *Amer. Econ. Rev.* 83 (1993), 1281-1302.

[20] M. Rabin, Psychology and economics, *J. Econ. Lit.* 83 (1998), 11-46.

[21] S. Tadelis, The power of shame and the rationality of trust, mimeo, UC Berkeley, 2007.

[22] J.P. Tangney, Recent advances in the empirical study of shame and guilt, *American Behavioral Scientist* 38 (1995), 1132-1145.