

Trust and Guilt Aversion

Lecture 2, *Psychological GT and Experiments*

Pierpaolo Battigalli
Bocconi University and IGER

Summer School Soleto 2025

Abstract

Psychologists argue that “the prototypical cause of *guilt* would be the infliction of harm, loss, or distress” and that 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.” In PGT, we model such guilt avoidance following Battigalli & Dufwenberg (2007): people feel guilty for making others get less than they expected. Since guilt has “negative valence”, people are willing to trade off some personal material gains to decrease the probability with which and the extent to which they let others down. This has important economic implications. I focus on trust and deception.

Introduction: emotions

- For a long time, neither psychologists nor economists paid much attention to emotions and how they shape behavior, although founding figures in psychology like C. Darwin and W. James did (Keltner & Lerner's 2010 handbook chapter).
- Economist J. Elster (1996, 1998) argues that economists have neglected to study the emotions, although “all human satisfaction comes in the form of emotional experiences” (1996, p. 1368). Not recognizing this, economists may fail to get a correct grip on how decisions are formed. [Also JEL surveys of B&D and Charness & Fehr.]
- That view is corroborated by more recent developments in psychology. Keltner & Lerner (2010) argue that “a robust science of emotion ... emerged” (p. 317), indicating that a variety of emotions impacts well-being and behavior.
- The appraisal-tendency approach is often stressed (Lerner & Keltner 2000, 2001): **appraisal tendencies** are goal-directed processes through which emotions exert effects on judgments and decisions.

Introduction: guilt

- Psychologists Baumeister *et al.* (1994) argue that “the prototypical cause of *guilt* would be the infliction of harm, loss, or distress” and that 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” (see p. 245; cf. Tangney 1995).
- Economists assume that people feel guilty for making others get less than they expected. Since guilt has “negative valence”, people are willing to trade off some personal material gains to decrease the probability with which and the extent to which they let others down (Battigalli & Dufwenberg 2007).
- Although some prominent psychologists stress guilt avoidance (see above), psychology mostly focuses on the *action tendency* following the experience of guilt, that is, “repair behavior.” (See, e.g., Silfver 2007 for a discussion, and Attanasi *et al.* 2025 for a model of—*inter alia*—such tendency.)

Modeling guilt avoidance

- To model guilt avoidance in 2-person situations, we can posit the following psychological utility function (the meaning of math. symbols was explained in the introductory lecture, in particular, $[x]^+ = \max\{0, x\}$): let $v_i(\pi_i)$ be the utility of money, with $v'_i > 0$, $v''_i \leq 0$, then

- $u_i(z, \alpha_{-i}) = v_i(\pi_i(z)) - G_i \left([\mathbb{E}_{\alpha_{-i}}(\pi_{-i}) - \pi_{-i}(z)]^+ \right)$, $G'_i \geq 0$,
- e.g., $u_i(z, \alpha_{-i}) = \pi_i(z) - \theta_i [\mathbb{E}_{\alpha_{-i}}(\pi_{-i}) - \pi_{-i}(z)]^+$.

- In n -person situations:

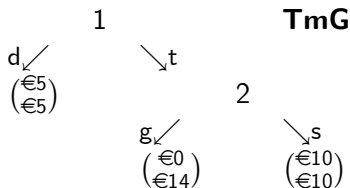
- $u_i(z, \alpha_{-i}) = v_i(\pi_i(z)) - G_i \left(\left([\mathbb{E}_{\alpha_j}(\pi_j) - \pi_j(z)]^+ \right)_{j \neq i} \right)$, $G'_{i,j} \geq 0$,
- e.g., $u_i(z, \alpha_{-i}) = \pi_i(z) - \sum_{j \neq i} \theta_{ij} [\mathbb{E}_{\alpha_j}(\pi_j) - \pi_j(z)]^+$.

- It may be argued that “excessive expectations” (e.g., above equal sharing of max surplus) do not matter (cf. Balafoutas *et al.* 2017):

- let $\bar{\pi}_j$ be the “legitimate limit” to $\mathbb{E}_{\alpha_j}(\pi_j)$, then
- $u_i(z, \alpha_{-i}) = \pi_i(z) - \sum_{j \neq i} \theta_{ij} [\min\{\mathbb{E}_{\alpha_j}(\pi_j), \bar{\pi}_j\} - \pi_j(z)]^+$.

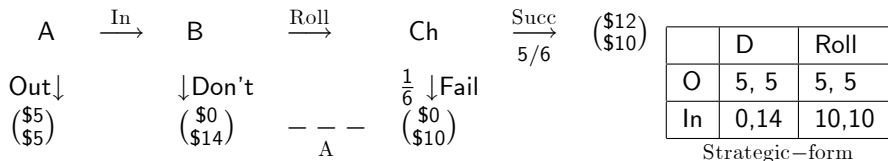
Guilt and Trust

- The Trust mini-Game is a very simple game form representing a stylized social dilemma:



- Would 2 share? Should 1 trust 2? Assume *common knowledge* that $\theta_1 = 0$, but $\theta_2 \geq 0$ is unknown to 1 (cf. Attanasi *et al.* 2016).
 $\mathbb{E}_{\beta_2}(\mathbb{E}_{\alpha_1}(\pi_1) | t)$ = 2's expectation of $\mathbb{E}_{\alpha_1}(\pi_1)$ cond. on t .
 - Pl. 2 prefers to share given trust if $10 > 14 - \theta_2 \mathbb{E}_{\beta_2}(\mathbb{E}_{\alpha_1}(\pi_1) | t)$.
 - Let pl. 1 trust only if $\mathbb{E}_{\alpha_1}(\pi_1) \geq 5$ and let pl. 2 be certain of this (also *after* he observes t), then $\mathbb{E}_{\beta_2}(\mathbb{E}_{\alpha_1}(\pi_1) | t) \geq 5$, and the sharing condition is $10 > 14 - 5\theta_2$, i.e., $\theta_2 > 0.8$.
 - Assuming pl. 1 is certain of this, she trusts if $\alpha_1(\tilde{\theta}_2 > 0.8) > 0.5$.

Trust mini-Game with imperfect monitoring



- Charness & Dufwenberg (CD, 2006) analyze *experimentally* a variation of the TmG where the \$0 payoff of the first mover (pl. A) may be due to bad luck.
 - A-subjects observe *ex post* only their own payoff. This may matter for, e.g., image concerns. CD posit that *it frames the game as a principal-agent problem* with moral hazard.
 - The game is played with the **strategy method**: B-subjects are asked to (covertly) commit in advance to the choice to be made in case A goes In. The resulting strategic form (with average payoffs given In-Roll) is the same as for the TmG shown above.

Second-order beliefs

- In most PGT model, 2nd-order beliefs $\beta_{i,-i}$ are key to understand incentives: best replies depend on $\beta_{i,-i}$ and personal traits θ_i .
Some experiments about guilt try to elicit (=measure) θ_i (e.g., Bellemare *et al.* 2011, Attanasi *et al.* 2025, Cartwright 2019),
most experiments
 - try to elicit key features of $\beta_{i,-i}$ —such as $\mathbb{E}_{\beta_{B,A}} (\mathbb{E}_{\alpha_A} (\pi_A) | \text{In})$ in the TmG above—and analyze correlation with choices.
 - try to manipulate $\beta_{i,-i}$ by changes of the game form that are supposed to move $\beta_{i,-i}$ unambiguously in one direction, e.g., treatments that should move upward the empirical distribution, or average, of $\mathbb{E}_{\beta_{B,A}} (\mathbb{E}_{\alpha_A} (\pi_A) | \text{In})$.
- CD (2006) had many subjects in a room randomly assigned to roles A and B, and anonymously matched to play (the strategic form of) the TmG. After choices were made, they asked
 - A-subjects to guess the proportion of B's who chose Roll;
 - B-subjects to guess the average guess *of the A's who chose In*;
 - “almost correct” answers were rewarded with \$5.

Guilt, trust, and communication: design

- To test GA, CD considered a control version of the TmG and several alternative treatments. The main treatment manipulations in the design were:
 - **Most important:** each B-subject was given the opportunity to send a free-form message to the paired A-subject [e.g., one message was: *"If you choose In, I will choose to roll. This way we both have an opportunity to make more than 5\$!:)"*].
 - *Expected effect:* messages were expected to move $\hat{\pi}_{B,A}(\beta_{B,A}) = \mathbb{E}_{\beta_{B,A}}(\mathbb{E}_{\alpha_A}(\pi_A) | \text{In})$ upward by a self-fulfilling expectation of trust: my message is likely to make you trust me, which increases my $\hat{\pi}_{B,A}(\beta_{B,A})$, which makes you (who understand this) likely to trust me ...
 - To check the **robustness** of the effects of communication, the **Outside-option** payoffs were changed from (\$5,\$5) to (\$7,\$7).
 - **Also, A-subjects** could send **send** a message.
- CD found a significant (i) positive correlation of Roll with (elicited) $\hat{\pi}_{B,A}(\beta_{B,A})$, and (ii) expected effect of communication.

- **Beliefs & Behavior:** CD observe a strong correlation in the expected direction between beliefs and behavior for both A's and B's in all treatments: *B's who chose Roll made significantly higher guesses about A's guesses compared to those who chose Don't* (but this may also follow from false consensus, see Cartwright 2019):
 - (5,5) No Messages, average 2nd-order guess of B's (cond. on In): 54% (given Roll) vs 40% (given Don't);
 - (5,5) Messages, average 2nd-order guess of B's (cond. on In): 73% (given Roll) vs 45% (given Don't).
- **Communication:** More strikingly, C&D observe a *strong effect of communication*: approximately (see Fig 3, p 1587 of CD),
 - (5,5) No Messages: 55% of A's go In, 45% of B's Roll, 22% of pairs choose In-Roll (no correlation);
 - (5,5) Messages: 75% of A's go In, 70% of B's Roll, 50% of pairs choose In-Roll (correlation of choices mediated by messages).

Guilt and deception: game forms, question

- **Cheap-Talk Sender-Receiver (CTSR) game form:** Pl. 1 sends a message $m \in \{m^A, m^B\}$, pl. 2 takes an action $a \in \{A, B\}$, payoffs depend only on a , but *only pl. 1 knows how*. Message m^A (m^B) says “action A (B) gives you more money”. Consider the following 3 cases, where only *pl. 1 knows the payoffs*, *pl. 2 knows nothing*:

$\Delta_i = \pi_i(B) - \pi_i(A)$	action	π_1	π_2
low stakes: $\Delta_1 = 1 = -\Delta_2$	A	\$5	\$6
	B	$\$(5 + 1)$	$\$(6 - 1)$
asymmetric stakes: $10\Delta_1 = 10 = -\Delta_2$	A	\$5	\$15
	B	$\$(5 + 1)$	$\$(15 - 10)$
high stakes: $\Delta_1 = 10 = -\Delta_2$	A	\$5	\$15
	B	$\$(5 + 10)$	$\$(15 - 10)$

- *How do the three cases of CTRS compare in terms of propensity to deceive (lie)?*

Guilt & deception: hypotheses

- CK that $\theta_2 = 0$, but $\theta_1 \geq 0$ is unknown. Let: $Y = [\text{do } A \text{ iff } m^A]$ (**trusting strategy**), $N = [\text{do } A \text{ iff } m^B]$ (**contrarian strategy**), $\Pi_2^X = \mathbb{E}_{\alpha_{2,1}, X}(\pi_2) = 2$'s expected payoff from strat. X given $\alpha_{2,1}$. Assume the following about $\alpha_{2,1}$ and $\beta_{1,2}$ for each case (Why does it make sense? Because pl. 2 cannot distinguish!):

- ① **(H.1. α -Symmetry)** $\alpha_{2,1}$ is s.t.
 $\mathbb{E}_{\alpha_{2,1}, X}(\pi_2 | m^A) = \Pi_2^X = \mathbb{E}_{\alpha_{2,1}, X}(\pi_2 | m^B)$ for each $X \in \{Y, N\}$ (m^A and m^B are perceived as equally truthful/deceiving). Thus Y (resp. N) is the unique BR iff $\Pi_2^Y > \Pi_2^N$ (resp. $\Pi_2^Y < \Pi_2^N$).
- ② **(H.2: Belief in rationality, H.1, and trust)** 1's belief $\beta_{1,2}$ is s.t.
 - ① H.1 certainly holds;
 - ② pl. 2 (receiver) is certainly rational;
 - ③ pl. 2 (receiver) is likely to trust 1: $\mathbb{P}_{\beta_{1,2}}(\Pi_2^Y > \Pi_2^N) > 0.5$.
- ③ **(H.3: β -Symmetry)** Exp. disapp. depends only on realized payoff:
$$\forall x, \mathbb{E}_{\beta_{1,2}} \left([\Pi_2^Y - x]^+ \mid \Pi_2^Y > \Pi_2^N \right) = \mathbb{E}_{\beta_{1,2}} \left([\Pi_2^N - x]^+ \mid \Pi_2^Y < \Pi_2^N \right)$$

Guilt & deception: comparative predictions

Let $D(x)$ denote the expected disappointment of pl. 2, according to 1's belief $\beta_{1,2}$, if 2 gets x . H.2-3 (+technical assumption) imply:

Lemma

Function D is strictly decreasing and convex ($D' < 0$, $D'' > 0$ where differentiable).

Corollary

The ratio $(D(x) - D(x + h)) / h$ is strictly decreasing in $h > 0$.

Given $\beta_{1,2}$, let $\hat{\theta}_1^t = (\pi_1^t(B) - \pi_1^t(A)) / (D(\pi_2^t(B)) - D(\pi_2^t(A)))$, with $t \in \{\text{ls}, \text{as}, \text{hs}\}$: *player 1 (sender) lies in t iff $\theta_1 < \hat{\theta}_1^t$.*

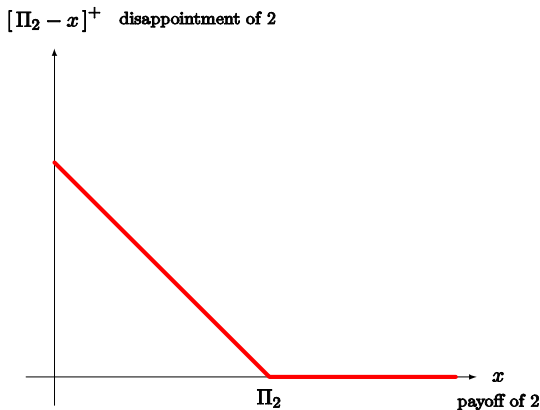
Proposition

Under H.2-3, the thresholds satisfy $0 < \hat{\theta}^{\text{as}} < \hat{\theta}^{\text{ls}} < \hat{\theta}^{\text{hs}}$.

Guilt & deception: intuition for predictions

(1) Disappointment is convex

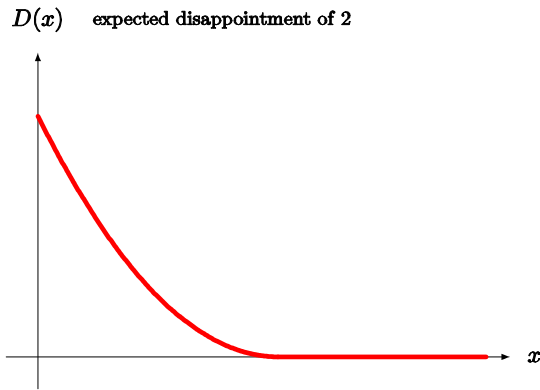
The higher the payoff of 2 the less he is disappointed. Furthermore, disappointment $[\Pi_2^x - x]^+$ is a convex function of realized payoff x .



Guilt & deception: intuition for predictions

(2) Expected disappointment is convex

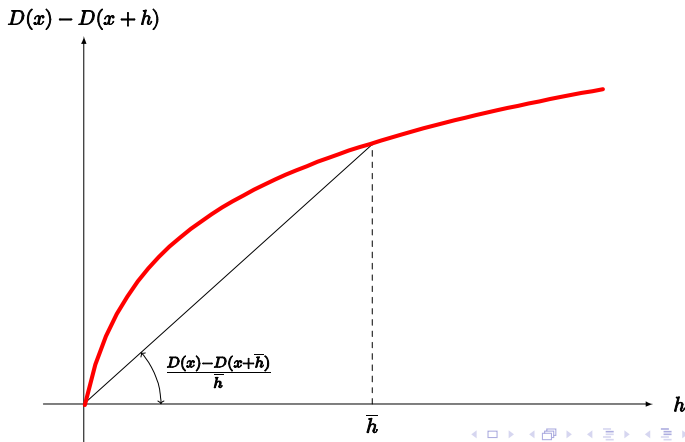
Therefore, $D(x)$, the expected disappointment of 2 according to 1's (2nd-ord.) belief $\beta_{1,2}$ is decreasing and convex in the payoff x of 2.



Guilt & deception: intuition for predictions

(3) Expected disappointment, incremental ratio

Thus, $(D(x) - D(x + h)) / h$ is strictly decreasing in $h > 0$, because $f(h) = D(x) - D(x + h)$ is strictly increasing and concave, with $f(0) = 0$.



Guilt & deception: intuition for predictions

With this, **Proposition 1** follows from the payoff differences in the 3 cases:

- $\hat{\theta}^{\text{as}} < \hat{\theta}^{\text{ls}}$ because

$$\frac{1}{D(5) - D(5 + 10)} < \frac{1}{D(5) - D(5 + 1)}$$

(by (2), D is decreasing);

- $\hat{\theta}^{\text{ls}} < \hat{\theta}^{\text{hs}}$ because

$$\frac{1}{D(5) - D(5 + 1)} < \frac{10}{D(5) - D(5 + 10)}$$

(by (3), $h / (D(x) - D(x + h))$ increasing in h).







Guilt & deception: the role of payoff consequences

- Gneezy (2005) designed a clever *experiment* with 3 (main) treatments:







$\Delta_i = \pi_i(B) - \pi_1(A)$	action	π_1	π_2
1: low stakes: $\Delta_1 = 1 = -\Delta_2$	A	\$5	\$6
	B	$\$(5 + 1)$	$\$(6 - 1)$
2: asymm. stakes: $10\Delta_1 = 10 = -\Delta_2$	A	\$5	\$15
	B	$\$(5 + 1)$	$\$(15 - 10)$
3: high stakes: $\Delta_1 = 10 = -\Delta_2$	A	\$5	\$15
	B	$\$(5 + 10)$	$\$(15 - 10)$

- According to **Proposition 1**, under the stated assumptions about guilt aversion and 2nd-ord. beliefs, *senders tend to lie the least in treatment 2 (as) and the most in treatment 3 (hs)*. The frequencies of lies across treatments are *consistent with this prediction*.





References

-  BATTIGALLI P., G. CHARNESS, & M. DUFWENBERG (2013): "Deception: The Role of Guilt," *J. Econ. Behav. Org.*, 93, 227-232.
-  BATTIGALLI, P., R. CORRAO, & M. DUFWENBERG (2019): "Incorporating Belief-Dependent Motivation in Games," *J. Econ. Behav. Organ.*, 167, 185-218.
-  BATTIGALLI P., & M. DUFWENBERG (2007): "Guilt in Games," *Am. Econ. Rev. (P&P)*, 97, 170-176.
-  — (2022): "Belief-Dependent Motivations and Psychological Game Theory," *J. Econ. Lit.*, 60, 833-882.
-  CARTWRIGHT, E. (2019): "A Survey of Belief-Based Guilt Aversion in Trust and Dictator Games," *J. Econ. Behav. Organ.*, 167, 430-444.
-  CHARNESS, G., & M. DUFWENBERG (2006): "Promises & Partnership," *Econometrica*, 74, 1579-1601.







Additional references: economics

-  ATTANASI, G., P. BATTIGALLI, & E. MANZONI (2016): "Incomplete Information Models of Guilt Aversion in the Trust Game," *Manag. Sci.*, 62, 648-667.
-  ATTANASI, G., P. BATTIGALLI, E. MANZONI, AND R. NAGEL (2019): "Belief-Dependent Preferences and Reputation: Experimental Analysis of a Repeated Trust Game," *J. Econ. Behav. Organ.*, 167, 341-360.
-  ATTANASI, G., P. BATTIGALLI, E. MANZONI, & R. NAGEL (2025): "Disclosure of Belief-Dependent Preferences in a Trust Game," *Econ. Theory*, 10.1007/s00199-025-01645-5.
-  BALAFOUTAS, L., & H. FORNWAGNER (2017): "The Limits of Guilt". *J. Econ. Science Ass.*, 3, 137-148.
-  BELLEMARE, C., A. SEBALD, & M. STROBEL (2011): "Measuring Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models". *J. Applied Econ.* 26: 

Additional references: economics

-  CHARNESSE, G., AND E. FEHR (2025): "Social Preferences: Fundamental Characteristics and Economic Consequences," *J. Econ. Lit.*, 63, 440–514.
-  ELSTER, J. (1998): "Emotions and Economic Theory," *J. Econ. Lit.*, 36, 4774.
-  — (1996): "Rationality and the Emotions". *Econ. J.*, 106: 1386-1397.
-  GNEEZY, U. (2005): "Deception: the Role of Consequences," *Am. Econ. Rev.*, 95, 384-394.

Additional references: psychology

-  BAUMEISTER, R., A. STILLWELL, & T. HEATHERTON (1994): "Guilt: An Interpersonal Approach," *Psy. Bull.*, 115, 243-267.
-  LERNER, J., & D. KELTNER (2000): "Beyond Valence: Toward a Model of Emotion-Specific Influences on Judgement and Choice". *Cogn. Emot.*, 14: 473-93.
-  — (2001): "Fear, Anger, and Risk". *J. Pers. Soc. Psychol.*, 81: 146-159.
-  — (2010): "Emotion," in S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 317–352). John Wiley & Sons, Inc..
-  SILFVER, M. (2007): "Coping with Guilt and Shame: A Narrative Approach," *J. Moral Educ.*, 36, 169-183.
-  TANGNEY, J.P. (1995): "Recent Advances in the Empirical Study of Shame and Guilt," *Am. Behav. Sci.*, 38, 1132-1145.