Introduction to Psychological Game Theory
Lecture 1, *Psychological GT and Experiments*

Pierpaolo Battigalli
Bocconi University and IGIER

Summer School Soleto 2025

Abstract

Psychological Game Theory (PGT) is a generalization of traditional Game Theory (GT) whereby the utility of outcomes, or—more generally—of whatever actions are taken in the game, may depend on players' endogenous beliefs (i.e., beliefs that depend on the strategic analysis of the game). This generalization allows to incorporate in game-theoretic analysis belief-dependent motivations related, for example, to reciprocity concerns, emotions, and image concerns (Battigalli & Dufwenberg 2009, 2022; Battigalli, Corrao & Dufwenberg 2019).

Introduction

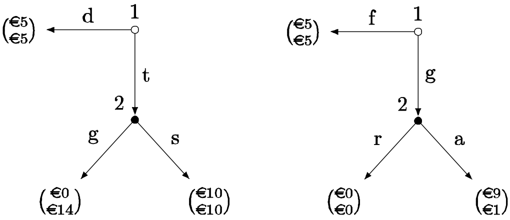
- Credible promises/threats and reliable communication are essential for cooperation.
- According to standard theory, credibility (incentive compatibility) is related to the value of future interactions.
- But often people cooperate, keep their word, and communicate truthfully even when this is not incentivized by future interactions.
- Emotions like guilt, anger, shame and pride can make people act against their selfish material interests in ways that are often (not always) beneficial to achieve cooperation.
- Many emotions are triggered by beliefs, including beliefs about the beliefs of others (higher-order beliefs).
- Emotions affect behavior in two ways:
  - *direct*: induced action tendencies (e.g., frustration-aggression⇒carry out threats);
  - *indirect:* anticipated feelings (valence) modify incentives (e.g., keep costly promises to avoid guilt).

- By letting psychological utility in games depend on endogenous beliefs we can model such phenomena.

$$
\begin{array}{lcl}
\text{Beliefs} & \longrightarrow & \text{Emotions} \\
\uparrow & \searrow & \downarrow \\
\text{Experience} & \longleftarrow & \text{Actions}
\end{array}
$$

- We develop a methodology and illustrate it with some examples/applications.
- We adopt a *subjective* notion of *rationality:* (sequential) best reply to subjective beliefs, with psychological motivations.
- *Caveat:* We do not consider biases, cognitive limitations, and bounded computational abilities, nor do we model how emotions can interfere with cognition.

Stylized dilemmas with implicit threats or promises



- The Ultimatum mini-Game and the Trust mini-Game are very simple game forms representing stylized social dilemmas:
  - **Ultimatum mini-Game (form):** Fear of rejection may make pl. 1 choose the **f**air allocation. Is the (possibly implicit) *threat* of **r**ejection credible? Yes, if pl. 2 *expected* the **f**air allocation and is sufficiently prone to *anger* (Battigalli, Dufwenberg & Smith 2019).
  - **Trust mini-Game (form):** Hope that pl. 2 would **s**hare may make pl. 1 **t**rust. Is the (possibly implicit) *promise* to **s**hare credible? Yes, if pl. 2 *thinks* pl. 1 *expected* him to **s**hare and is *guilt averse*.

Motivations & Examples

The following is *in*consistent with standard social preferences (e.g., inequity or lying aversion), but consistent with the PGT framework and models:

- **Psychology**:
  - desire to live up to others' expectations to avoid guilt feelings (Baumeister *et al.* 1994, Tangney 1995);
  - frustration-aggression hypothesis (Dollard *et al.* 1939, Frijda 1993);
  - moral behavior to avoid the feeling of shame (Tangney 1995).

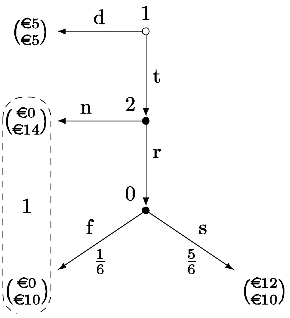
Motivations & Examples (continue)

- **Facts (casual evidence, empirics,** see also survey by Charness & Fehr 2025**):**
  - Non-returning customers give tips.
  - Low offers are often rejected leaving money on the table.
  - Unexpected losses by home football/soccer teams are associated with increased domestic violence (Card & Dahl, 2011) or violent crime (Munyo & Rossi 2013).
- **Facts (experimental):**
  - **Trust mini-Game**: *correlation between sharing and 2^{nd}-order beliefs of sharing*; *treatments effects despite no change in the traditional game form represention*, which neglets information of inactive players (Charness & Dufwenberg 2006, Tadelis 2011, Attanasi *et al.* 2025).
  - **Ultimatum mini-Game**: *Rejections correlate with* (manipulated) *initially expected offers* (Sanfey, 2009; Xiang *et al.*, 2013, with fMRI; Aina *et al.*, 2020).
  - **Lying/truth-telling** is *not* categorical, i.e., *"all or nothing"* (Fischbacher & Föllmi-Heusi 2013), it *depends on the payoffs of receivers* (Gneezy, 2005; Battigalli *et al.*, 2013) *and on exposure to*

Formal setting: one-period, sequential game forms

- **Player set:** $I_0 = I \cup \{0\}$, $i \in I$ are **personal** players, 0 is **chance**.
- **Tree** of **histories:** $\bar{H}$ (*finite,* each prefix of each $h \in \bar{H}$ belongs to $\bar{H}$ as well, including the **empty history** $\varnothing$).
    - Z, set of **terminal** histories/nodes (game over); H, set of **non-terminal** histories/nodes (including root $\varnothing$); $\bar{H} = H \cup Z$; $Z(h) = \{z \in Z : h \prec z\}$, terminal successors of h.
    - $\iota : H \rightrightarrows I_0$ is the **active-players correspondence**; $H_i = \{h : i \in \iota(h)\}$, histories where i is active.
    - $A(h) = \times_{i \in \iota(h)} A_i(h)$ is the set of possible **action profiles** given h.
- **Chance probabilities:** $p_0 = (p_0(\cdot|h))_{h:0 \in \iota(h)}$, with $p_0(\cdot|h) \in \Delta(A_0(h))$.
- **Observed actions:** *active players observe earlier choices.*
- **Terminal information:** $\mathcal{P}_i$ is a *partition* of Z describing what i observes *ex post* ($\mathcal{P}_i(z)$ denotes the cell containing z).
- **Material payoffs:** $\pi_i : Z \to Y_i$ $(i \in I)$, e.g., monetary ($Y_i \subseteq$€$\mathbb{R}$).

Example: TmG with random outcome



- Omitting unncessary parentheses:
  - $H_1 = \{\varnothing\}$, $A_1(\varnothing) = \{\mathrm{d}, \mathrm{t}\}$, $H_2 = \{\mathrm{t}\}$, $A_2(\mathrm{t}) = \{\mathrm{n}, \mathrm{r}\}$, $H_0 = \{(\mathrm{t}, \mathrm{r})\}$, $A_0(\mathrm{t}, \mathrm{r}) = \{\mathrm{f}, \mathrm{s}\}$, $Z = \{\mathrm{d}, (\mathrm{t}, \mathrm{n}), (\mathrm{t}, \mathrm{r}, \mathrm{f}), (\mathrm{t}, \mathrm{r}, \mathrm{s})\}$;
  - $p_0(\mathrm{s}|(\mathrm{t}, \mathrm{r})) = \frac{5}{6}$, $\pi_i(\mathrm{d}) =$€5 $(i = 1, 2)$, $\pi_1(\mathrm{t}, \mathrm{n}) =$ €0 $= \pi_1(\mathrm{t}, \mathrm{r}, \mathrm{f})$, $\pi_1(\mathrm{t}, \mathrm{r}, \mathrm{s}) =$€12, $\pi_2(\mathrm{t}, \mathrm{n}) =$ €14, $\pi_2(\mathrm{t}, \mathrm{r}, \mathrm{f}) =$€10 $= \pi_2(\mathrm{t}, \mathrm{r}, \mathrm{s})$;
  - Players observe their monetary payoffs and have perfect recall: $\mathcal{P}_1 = \{\{\mathrm{d}\}, \{(\mathrm{t}, \mathrm{n}), (\mathrm{t}, \mathrm{r}, \mathrm{f})\}, \{(\mathrm{t}, \mathrm{r}, \mathrm{s})\}\}$, $\mathcal{P}_2$ finest part. of Z.

Formal setting: beliefs & psychological utility

- **Trait-types:** Θ_i, set of types=personal traits of $i \in I$.
- **First-order beliefs:** set Δ_i^1 of belief systems $\alpha_i = (\alpha_i(\cdot|h))_{h \in H \cup \mathcal{P}_i}$ s.t. $\alpha_i(\cdot|h) \in \Delta(\Theta_{-i} \times Z(h))$; given $h \prec h'$ (h prefix of h'), write $\alpha_i(\theta_{-i}, h'|h) = \alpha_i(\{\theta_{-i}\} \times Z(h')|h)$ and $\alpha_i(h'|h) = \alpha_i(\Theta_{-i} \times Z(h')|h)$, with this,
  - *chain rule*: if $(h, a', a'') \in \bar{H}$, $\alpha_i((h, a', a'')|h) = \alpha_i((h, a', a'')|(h, a'))\alpha_i((h, a')|h)$,
  - *self* vs *others indep.*: what i thinks about others' types and simultaneous actions is independent of his action; hence, $\alpha_i(\theta_{-i}, (h, a)|h) = \alpha_{i,i}(a_i|h) \times \alpha_{i,-i}(\theta_{-i}, a_{-i}|h)$.
- **Psy-utility:** $u_i : \Theta_i \times Z \times \Delta^1 \to \mathbb{R}$ with $\Delta^1 = \times_{j \in I}\Delta_j^1$;
  - $u_i(\theta_i, z, \alpha)$, utility of z for type θ_i given $\alpha = (\alpha_j)_{j \in I}$;
  - **note**: *i does not know α_{-i}* (she consults her 2^{nd}-ord. beliefs to decide);
  - **note**: there are *private values* (standard situation in experiments).

Examples: guilt and disappointment

Let $[x]^+ = \max\{x, 0\}$, $\mathbb{E}_{\alpha_i}(\pi_i) = \sum_{z \in Z} \pi_i(z)\, \alpha_i(z|\varnothing)$ (initially expected payoff), $\mathbb{R}_+$=non-negative real n.

- **Guilt aversion**
  - $u_i(\theta_i, z, \alpha) = \pi_i(z) - \sum_{j \neq i} \theta_{ij} \left[\mathbb{E}_{\alpha_j}(\pi_j) - \pi_j(z)\right]^+$, $\theta_i = (\theta_{ij})_{j \neq i} \in \mathbb{R}_+^{I \setminus \{i\}}$,
  - θ_{ij}=how much i dislikes letting j down,
  - u_i does not depend on α_i; hence, *own-plan independence* (plan=$\alpha_{i,i}$).
- **Disappointment aversion**
  - $u_i(\theta_i, z, \alpha) = \pi_i(z) - \theta_i \left[\mathbb{E}_{\alpha_i}(\pi_i) - \pi_i(z)\right]^+$, $\theta_i \in \mathbb{R}_+$;
  - u_i depends on the whole α_i (including $\alpha_{i,i}$); hence *own-plan dependence*.

Examples: image concerns

Fix function $V : Z \to \mathbb{R}$, then $\mathbb{E}_{\alpha_i}(V|h) = \sum_{z \in Z(h)} V(z)\,\alpha_i(z|h)$ denotes the conditional expectation of V given h.

- **Image concerns: good/bad behavior**
  - Z_i^G (resp. Z_i^B), paths where i took **good** (resp. **bad**) **actions**, $\mathbf{I}_i^G : Z \to \{0,1\}$ indicator fun. of Z_i^G ($\mathbf{I}_i^B$ similar),
  - $u_i(\theta_i, z, \alpha) = \pi_i(z) + \sum_{j \neq i} \theta_{ij} \left[\mathbb{E}_{\alpha_j}\left(\mathbf{I}_i^G | \mathcal{P}_j(z)\right) - \mathbb{E}_{\alpha_j}\left(\mathbf{I}_i^B | \mathcal{P}_j(z)\right)\right]$.
  - $\theta_{ij} \geq 0$, how much i cares about the opinion of j.
- **Image concerns: good/bad traits**
  - $\theta_i = \left(\theta_i^{\mathbf{I}}, \theta_i^{\mathbf{R}}\right)$, $\theta_i^{\mathbf{I}} \geq 0$: **intrinsic**-motivation trait, $\theta_i^{\mathbf{R}} = \left(\theta_{ij}^{\mathbf{R}}\right)_{j \neq i} \in \mathbb{R}_+^{I \setminus \{i\}}$: **reputational**-motivation trait,
  - $u_i(\theta_i, z, \alpha_j) = \pi_i(z) + \theta_i^{\mathbf{I}} \left[\mathbf{I}_i^G(z) - \mathbf{I}_i^B(z)\right] + \sum_{j \neq i} \theta_{ij}^{\mathbf{R}} \mathbb{E}_{\alpha_j}\left(\widetilde{\theta}_i^{\mathbf{I}} | \mathcal{P}_j(z)\right)$,
  - where $\widetilde{\theta}_i^{\mathbf{I}}$ denotes a trait of i unknown to (uncertain for) j.
- u_i does not depend on α_i; hence, *own-plan indep.* (plan=$\alpha_{i,i}$).

Best replies, rational planning

- **Second-order beliefs:** Δ_i^2 set of 2^{nd}-order belief systems $\beta_i = (\beta_i(\cdot|h))_{h\in H}$ s.t.
  - $\beta_i(\cdot|h) \in \Delta\left(\Theta_{-i} \times Z(h) \times \Delta^1\right)$, the *chain rule* and *self* vs *others independence* hold;
  - derive 1^{st}-order beliefs α_i by "marginalization".
- **Expected utility of actions:** For $h \in H_i$, $a_i \in A_i(h)$, $\bar{u}_{i,h}(a_i;\beta_i) = \mathbb{E}_{\beta_i}(u_i|h, a_i)$.
- **Local best replies:** $a_i^* \in \arg\max_{a_i\in A_i(h)} \bar{u}_{i,h}(a_i;\beta_i)$.
- **Rational planning:** Given $\alpha_{i,i}$ derived from β_i, for all $h \in H_i$ and $a_i^* \in A_i(h)$, $\alpha_{i,i}(a_i^*|h) > 0 \Rightarrow a_i^* \in \arg\max_{a_i\in A_i(h)} \bar{u}_{i,h}(a_i;\beta_i)$ (intrapersonal equilibrium).

Exercises

Consider the **Trust mini-Game** with **perfect terminal information** ($\mathcal{P}_i(z) = \{z\}$ for every $i \in I$ and $z \in Z$).

- **Exercise:**
  - Let $Z_2^G = \{(\mathrm{t},\mathrm{s})\}$, $Z_2^B = \{(\mathrm{t},\mathrm{g})\}$ (sharing is good, grabbing is bad).
  - Consider **image concerns** of pl. 2 **for traits**, with $\Theta_2 = \left\{0, \bar{\theta}_2^{\mathbf{I}}\right\} \times \left\{0, \bar{\theta}_2^{\mathbf{R}}\right\}$, $\bar{\theta}_2^{\mathbf{I}}, \bar{\theta}_2^{\mathbf{R}} > 0$.
  - $\beta_2(\cdot|\mathrm{t})$ assigns probability $\frac{1}{2}$ to α_1' and α_1'', which are such that $\mathbb{E}_{\alpha_1'}\left(\widetilde{\theta_2^{\mathbf{I}}}|(\mathrm{t},\mathrm{g})\right) = \mathbb{E}_{\alpha_1''}\left(\widetilde{\theta_2^{\mathbf{I}}}|(\mathrm{t},\mathrm{g})\right) = 0$, $\mathbb{E}_{\alpha_1'}\left(\widetilde{\theta_2^{\mathbf{I}}}|(\mathrm{t},\mathrm{s})\right) = \frac{1}{2}\bar{\theta}_2^{\mathbf{I}}$, and $\mathbb{E}_{\alpha_1''}\left(\widetilde{\theta_2^{\mathbf{I}}}|(\mathrm{t},\mathrm{s})\right) = \bar{\theta}_2^{\mathbf{I}}$ [α_1' deems 0 and $\bar{\theta}_2^{\mathbf{I}}$ equally likely given (t,s), α_1'' is certain of $\bar{\theta}_2^{\mathbf{I}}$ given (t,s)].
  - Find values of $\bar{\theta}_2^{\mathbf{I}}$ and $\bar{\theta}_2^{\mathbf{R}}$ such that pl. 2's best reply is to **s**hare, and values of $\bar{\theta}_2^{\mathbf{I}}$ and $\bar{\theta}_2^{\mathbf{R}}$ such that 2's best reply is to **g**rab.

References

- Battigalli, P. (2025): *Mathematical Language and Game Theory.* Typescript, Bocconi University. [Downloadable from webpage]
- Battigalli, P., E. Catonini, and N. De Vito (2025): *Game Theory: Analysis of Strategic Thinking*. Typescript, Bocconi University. [Downloadable from webpage]
- Battigalli, P., R. Corrao, and M. M. Dufwenberg (2019): "Incorporating Belief-Dependent Motivation in Games," *J. Econ. Behav.Organ.*, 167, 185-218.
- Battigalli, P., and M. Dufwenberg (2009): "Dynamic Psychological Games," *J. Econ. Theory*, 144, 1-35.
- Battigalli, P., and M. Dufwenberg (2022): "Belief-Dependent Motivations and Psychological Game Theory," *J. Econ. Lit.*, 60, 833-882.

Additional references

- Aina, C., P. Battigalli, and A. Gamba (2020): "Frustration and Anger in the Ultimatum Game: An Experiment," *Games Econ. Behav.*, 122, 150-167.
- Attanasi G., P. Battigalli, E. Manzoni, and R. Nagel (2025): "Disclosure of Belief-Dependent Preferences in the Trust Game," *Econ. Theory*, 10.1007/s00199-025-01645-5.
- Battigalli P., G. Charness and M. Dufwenberg (2013): "Deception: The Role of Guilt," *J. Econ. Behav. Organ.*, 93, 227-232.
- Battigalli P., M. Dufwenberg and A. Smith (2019): "Frustration, Aggression, and Anger in Leader-Follower Games," *Games Econ. Behav.*, 117, 15-39.
- Baumeister, R., A. Stillwell, and T. Heatherton (1994): "Guilt: An Interpersonal Approach". *Psychol. Bull.*, 115, 243-267.

Card, D. and G. B. Dahl (2011): "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior," *Q. J. Econ.*, 126, 103–143.

Charness G. and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74, 1579-1601.

Dollard, J., L. Doob, N. Miller, O. Mowrer, and R. Sears (1939): *Frustration and Aggression*. Yale University Press, New Haven, CT.

Dufwenberg, M., and M. Dufwenberg Jr., (2018): "Lies in disguise–A theoretical analysis of cheating," *J. Econ. Theory,* 175, 248-264

Fischbacher, U., and F. Follmi-Heusi (2013): "Lies in disguise: An experimental study on cheating," *J. Europ. Econ. Ass.*, 11, 525–547.

Charness, G., and E. Fehr (2025): "Social Preferences: Fundamental Characteristics and Economic Consequences," *J. Econ. Lit.*, 63, 440–514.

Frijda, N. H. (1993): "The Place of Appraisal in Emotion," *Cogn. Emot.*, 7, 357–387.

Gneezy, U. (2005): "Deception: The role of consequences," *Am. Econ. Rev.*, 95, 384–394.

Gneezy, U., A. Kajackaite, and J. Sobel (2018): "Lie Aversion and the Size of a Lie," *Am. Econ. Rev.*, 108, 419–453.

Munyo, I. and M. Rossi (2013): "Frustration, euphoria, and violent crime," *J. Econ. Behav. Organ.*, 89, 136–142.

- Sanfey, A. (2009): "Expectations and Social Decision-Making: Biasing Effects of Prior Knowledge on Ultimatum Responses," *Mind and Society* 8, 93–107.
- Tadelis S. (2011): "The Power of Shame and the Rationality of Trust," typescript, UC Berkeley.
- Xiang, T., T. Lohrenz, and R. Montague (2013): "Computational Substrates of Norms and Their Violations during Social Exchange," *J. Neuroscience*, 33, 1099–1108.