# Reciprocity
# Lecture 4, *Psychological GT and Experiments*

Pierpaolo Battigalli
Bocconi University

Summer School Soleto 2025

**Abstract**

Reciprocity theory assumes that people wish to be kind towards those they perceive to be kind, and unkind towards those they perceive to be unkind. Rabin (1993) argues that the *kindness is based on intentions*: the **kindness** of $i$ towards $j$ is measured by the difference between how much $i$ expects to make $j$ earn and an "equitable payoff" of $j$. Hence, *kindness depends on* ($1^{st}$-order) *beliefs*, making this a PGT model. Here I present the *theory of sequential reciprocity* of Dufwenberg & Kirchsteiger (2004) for leader-follower game forms. I also consider a variation, *negative reciprocity theory*, and a 'hold-up' experiment about it (Dufwenberg, Smith & Van Essen 2013). Finally, I hint at a dynamically consistent version of sequential reciprocity theory.

# Introduction

- We studied how:
  - guilt avoidance can make agents keep materially costly promises;
  - frustration and anger can make agents carry out materially costly threats.
- Both effects are also promoted by **reciprocity**, the action tendency of being kind (resp. unkind) towards those whom we perceive as kind (resp. unkind) with us.
- The idea that people wish to be (un)kind towards those they perceive to be (un)kind is age-old. Early academic discussions can be found in anthropology, sociology, social psychology, biology, and economics (see references Akerlof 1982, and in the surveys by B&D 2022 and Sobel 2005).
- Rabin (1993) argues that the *kindness is based on intentions*: the **kindness** of $i$ towards $j$ is measured by the difference between how much $i$ expects to make $j$ earn and an "equitable payoff" of $j$. Hence, *kindness depends on* ($1^{st}$-order) *beliefs.*

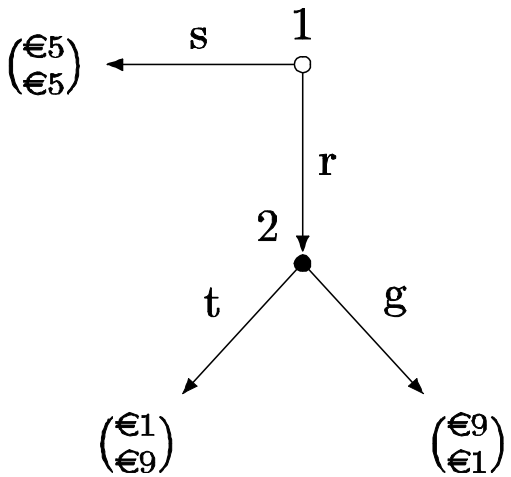
# Modeling kindness in leader-follower (LF) game forms

- In an **LF** game form, first pl. 1 (L) chooses $a_1 \in A_1$, next pl. 2 (F) chooses $a_2 \in A_2(a_1)$. (Let $A_2(a_1) = \{\text{wait}\}$ if $a_1$ is a terminating action.)
- The **kindness** of 1 towards 2 when choosing $a_1$ depends on its intended effects given the $1^{st}$-order belief $\alpha_{12}$:
  - let $A_1^*$ [resp. $A_2^*(a_1)$] denote the set of 1's [resp. 2's] actions that cannot lead to materially Pareto-dominated outcomes [in the examples below, $A_1^* = A_1$, $\forall a_1$, $A_2^*(a_1) = A_2(a_1)$]; recall that $\mathbb{E}_{\alpha_{12}}(\pi_2|a_1) = \sum_{a_2 \in A_2(a_1)} \pi_2(a_1, a_2)\, \alpha_{12}(a_2|a_1)$;
  - then, for any given $\bar{a}_1 \in A_1$, the **leader's kindness** is

$$\kappa_{12}(\bar{a}_1, \alpha_{12}) = \mathbb{E}_{\alpha_{12}}(\pi_2|\bar{a}_1) - \frac{1}{2}\left(\max_{a_1 \in A_1^*} \mathbb{E}_{\alpha_{12}}(\pi_2|a_1) + \min_{a_1 \in A_1^*} \mathbb{E}_{\alpha_{12}}(\pi_2|a_1)\right)$$

    where $\frac{1}{2}(...)$ is the "**equitable payoff**" (see discussion in Dufwenberg & Kirchsteiger 2019).
- **Follower's kindness** of $\bar{a}_2$ given $a_1$ [belief-indep.]: $\kappa_{21}(a_1, \bar{a}_2) = \pi_1(a_1, \bar{a}_2) - \frac{1}{2}\left(\max_{a_2 \in A_2^*(a_1)} \pi_1(a_1, a_2) + \min_{a_2 \in A_2^*(a_1)} \pi_1(a_1, a_2)\right)$

# Kindness in the Dictator mini-Game with Outside Option




- Consider the following Dictator mini-Game with an Outside Option (DmG-OO):
  - To **g**ive (**t**ake) if 1 **r**eached is kind (unkind): $\kappa_{21}(\mathrm{r},\mathrm{g}) = 9 - \frac{1}{2}(1+9) = 4 = -\kappa_{21}(\mathrm{r},\mathrm{t})$.
  - *Is* **r***eaching kind or unkind*? Pl. 1 is **kind** towards pl. 2 when **r**eaching, if he does so with the *intention* of making pl. 2 get, in expectation, more than the "equitable payoff": Let $p = \alpha_{12}(\mathrm{t}|\mathrm{r})$; since $\kappa_{12}(\mathrm{r},p) = 9p + (1-p) - \frac{1}{2}(5 + 9p + (1-p)) = 4p - 2$, **r**eaching is kind (unkind) if $p > \frac{1}{2}$ ($p < \frac{1}{2}$).

# Modeling reciprocity (leader-follower game forms)

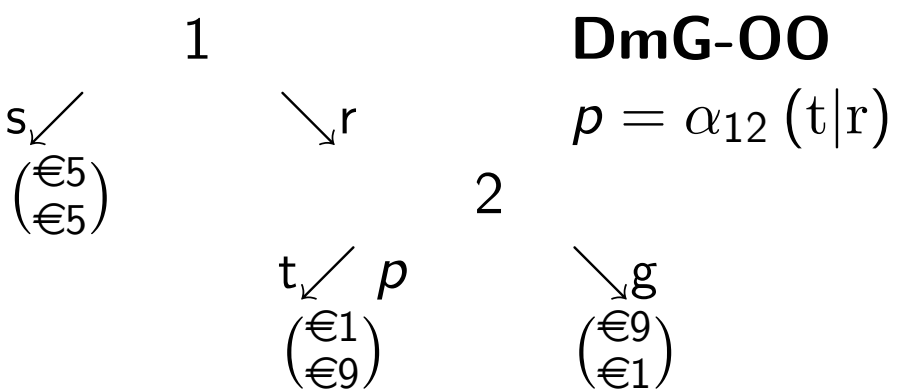
- **Reciprocity** is the *action tendency* of meting (un)kindness with (un)kindness.
- Such action tendency is captured by the following psychological utility functions [recall, only the kindness of pl. 1 (leader) is belief-dependent]:

$$u_i\left(a_1, a_2, \alpha_{12}\right) = \pi_i\left(a_1, a_2\right) + \theta_i \kappa_{12}\left(a_1, \alpha_{12}\right) \kappa_{21}\left(a_1, a_2\right).$$

- The follower must consult his (conditional) $2^{nd}$-order belief $\beta_{21}\left(\cdot | a_1\right)$ to maximize the expected utility of his response to $a_1$:

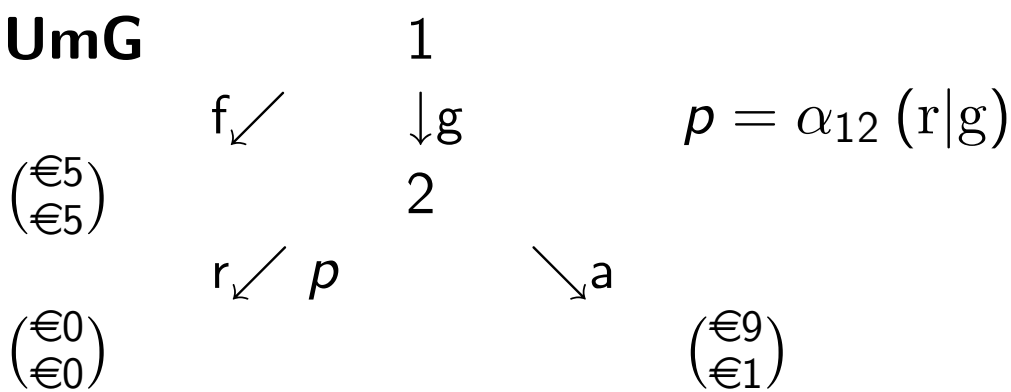
$$\max_{a_2 \in A_2(a_1)} \left[\pi_2\left(a_1, a_2\right) + \theta_2 \mathbb{E}_{\beta_{21}}\left(\kappa_{12}\left(a_1, \alpha_{12}\right) | a_1\right) \kappa_{21}\left(a_1, a_2\right)\right]$$

# Reciprocity in the Dictator mini-Game with Outside Option



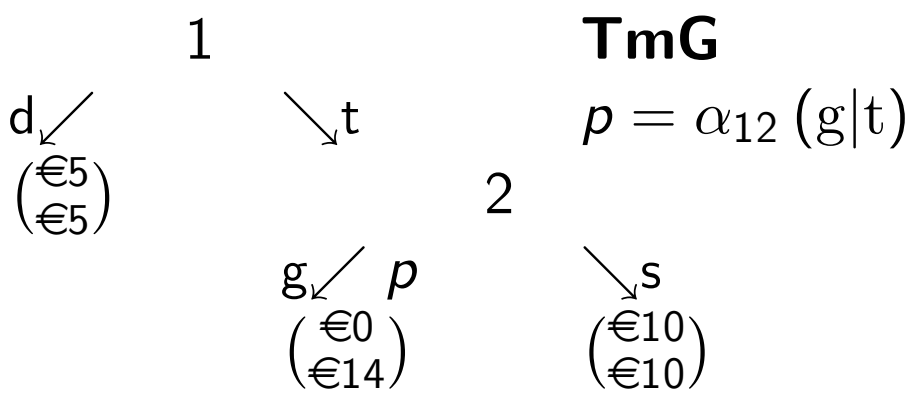

- $\kappa_{21}(\mathrm{r},\mathrm{g}) = 4 = -\kappa_{21}(\mathrm{r},\mathrm{t})$.
- Let $q = \mathbb{E}_{\beta_{21}}(\widetilde{p}|\mathrm{r})$. Since $\kappa_{12}(\mathrm{r},p) = 4p-2$, then $\mathbb{E}_{\beta_{21}}(\kappa_{12}(\mathrm{r},\widetilde{p})|\mathrm{r}) = 4q-2$, and
  - $\bar{u}_{2,\mathrm{r}}(\mathrm{g};\beta_{21}) = 1+\theta_2(4q-2)4$,
  - $\bar{u}_{2,\mathrm{r}}(\mathrm{t};\beta_{21}) = 9+\theta_2(4q-2)(-4)$;
  - $\bar{u}_{2,\mathrm{r}}(\mathrm{g};\beta_{21}) > \bar{u}_{2,\mathrm{r}}(\mathrm{t};\beta_{21})$ IFF $\theta(4q-2) > 1$ ONLY IF $q > \frac{1}{2}$ (for high $\theta_2$, pl. 2 wants to surprise pl. 1, **g**iving if 1 expects him to **t**ake).
  - This implies there is no "pure equilibrium" where pl. 1 correctly anticipates how 2 would respond and 2 understands this.

# Reciprocity vs Anger in the Ultimatum mini-Game




- $\kappa_{12}(\mathrm{g}, p) = (1-p) - \frac{1}{2}[5 + (1-p)] = -2 - \frac{1}{2}p$.
- Let $q = \mathbb{E}_{\beta_{21}}(\widetilde{p}|\mathrm{g})$, then
  - $\bar{u}_{2,\mathrm{g}}(\mathrm{r}, \beta_{21}) = \theta_2\left(-2 - \frac{1}{2}q\right)\left(0 - \frac{9}{2}\right) = \theta_2\left(9 + \frac{9}{4}q\right)$,
  - $\bar{u}_{2,\mathrm{g}}(\mathrm{a}, \beta_{21}) = 1 + \theta_2\left(-2 - \frac{1}{2}q\right)\left(9 - \frac{9}{2}\right) = 1 - \theta_2\left(9 + \frac{9}{4}q\right)$
  - $\bar{u}_{2,\mathrm{g}}(\mathrm{r}, \beta_{21}) > \bar{u}_{2,\mathrm{g}}(\mathrm{a}, \beta_{21})$ IFF $2\theta_2\left(9 + \frac{9}{4}q\right) > 1$ IFF $\theta_2\left(18 + \frac{9}{2}q\right) > 1$ IF $\theta_2 > \frac{1}{18}$.
- If $p = 1$, pl. 1 deems 2 unkind given **g**. For $\theta_1$ large (how large?), 1 makes the **g**reedy offer to harm 2, who reciprocates **r**ejecting even if he expected. This *"miserable equilibrium"* is impossible according to the FA model: 2 does not feel angry if he expected **g**.

# Reciprocity in the Trust mini-Game




- Can reciprocity support cooperative behavior? Yes, because **t**rust is a kind action, independently of $1^{st}$-order belief $p = \alpha_{12}(\mathrm{g}|\mathrm{t})$, and to **s**hare is a kind reply:
  - $\kappa_{12}(\mathrm{t}, p) = 14p + 10(1-p) - \frac{1}{2}[14p + 10(1-p) + 5] = 2p + \frac{5}{2}$;
  - $\kappa_{21}(\mathrm{t}, \mathrm{s}) = 10 - \frac{1}{2}(0 + 10) = 5 = -\kappa_{21}(\mathrm{t}, \mathrm{g})$.
- **Note**: pl. 2 has the *lowest incentive to* **s***hare* if *he believes that* $p = 0$, i.e., that *pl. 1 trusted him to* **s***hare*. Let $q = \mathbb{E}_{\beta_{21}}(\widetilde{p}|\mathrm{t})$,
  - $\bar{u}_{2,\mathrm{t}}(\mathrm{s}, \beta_{21}) = 10 + \theta_2\left(2q + \frac{5}{2}\right)5$,
  - $\bar{u}_{2,\mathrm{t}}(\mathrm{g}, \beta_{21}) = 14 + \theta_2\left(2q + \frac{5}{2}\right)(-5)$;
  - compute the threshold $\hat{\theta}_2$ such that pl. 2 certainly **s**hares if $\theta_2 > \hat{\theta}_2$.

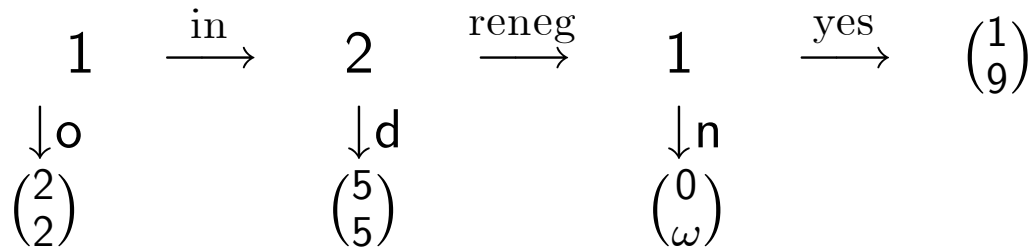
# Negative reciprocity

- Dhaene & Bouckaert (2010) provide experimental evidence on DK's reciprocity theory [Attanasi et al. 2013/2024 study theoretically and experimentally guilt aversion + reciprocity]. Here I consider a variation and experimental evidence about it. According to *negative reciprocity theory*, players meet unkindness with unkindness, but (positive) kindness does not matter. Let $[x]^- = \min\{0, x\}$, then

$$\begin{aligned} u_1(a_1, a_2, \alpha_{12}) &= \pi_1(a_1, a_2) + \theta_1 \kappa_{12}(a_1, \alpha_{12}) [\kappa_{21}(a_1, a_2)]^-, \\ u_2(a_1, a_2, \alpha_{12}) &= \pi_2(a_1, a_2) + \theta_2 [\kappa_{12}(a_1, \alpha_{12})]^- \kappa_{21}(a_1, a_2). \end{aligned}$$

- Dufwenberg, Smith & Van Essen (DS&V-E, 2013) derive interesting predictions about *hold-up problems* by extending negative reciprocity theory to 3-stage game forms where:
  - pl. 1 can **in**vest in a relationship (non-binding contract), or stay **o***ut*,
  - pl. 2 can **d***eliver* (comply with the contract) or *renego*tiate, holding 1 up,
  - pl. 1 can *accept* (**y**es) or *reject* (**n**o).

# Negative reciprocity: Hold-up mini-Game




- $\omega$ is the value for pl. 2 after a rejection and depends on *residual rights of control*:
  - if pl. 1 provided a service, he cannot take it back, $\omega$ is high;
  - if pl. 1 produced a good (having no value for him), he can keep it, $\omega$ low.
- According to the *residual rights of control*, negative reciprocity can make rejection an effective threat (if $\omega < 5$) and promote cooperation $(\mathrm{in}, \mathrm{d})$, or not (if $\omega > 5$).

# Negative reciprocity: experiment of Hold-up mini-Game

- DS&V-E (2013) designed an *experiment* with the following parameterized (treatment-dependent) game form (show-up fee of \$2):

$$\begin{array}{ccccccc} 1 & \xrightarrow{\text{in}} & 2 & \xrightarrow{\text{r}} & 1 & \xrightarrow{\text{yes}} & \binom{1}{9} \\ \downarrow \text{o} & & \downarrow \text{d} & & \downarrow \text{n} & & \\ \binom{2}{2} & & \binom{5}{5} & & \binom{0}{\omega} & & \end{array}$$

(different labels are used in the experiment)

- $\omega =$ \$2+value for pl. 2 after a rejection:
  - if pl. 1 provided a *service*, he cannot take it back, $\omega =$ \$10 (**High Game**)
  - if pl. 1 produced a *good* (of no value for him), he can keep it, $\omega =$ \$2 (=fee, **Low Game**)
- According to the *residual right of control*, negative reciprocity yields rejection ($\omega = 2$) promoting cooperation $(\text{in}, \text{d})$, or not (if $\omega = 10$).

# Hold-up mini-Game: predictions for the High Game

$$\begin{array}{ccccccc} 1 & \xrightarrow{\text{in}} & 2 & \xrightarrow{\text{r}} & 1 & \xrightarrow{\text{yes}} & \binom{1}{9} \\ \downarrow \text{o} & & \downarrow \text{d} & & \downarrow \text{n} & & \\ \binom{2}{2} & & \binom{5}{5} & & \binom{0}{10} & & \omega = 10 \end{array}$$

- **r** is *unkind* after **in**, but rejecting **r** would be a gift of \$1 (=10-9) to pl. 2! Hence, **yes.**
- If pl. 2 anticipates this, he renegotiates with **r** (even if he deems **in** kind, only unkindness is supposed to matter)**.**
- If pl. 1 anticipates this, he stays **o**ut.
- (Same solution as backward induction with CK of utility=money.)
- Thus, we expect *high rates* of **o**ut, **r, yes**.

# Hold-up mini-Game: predictions for the Low Game

$$\begin{array}{ccccccc} 1 & \xrightarrow{\text{in}} & 2 & \xrightarrow{\text{r}} & 1 & \xrightarrow{\text{yes}} & \binom{1}{9} \\ \downarrow \text{o} & & \downarrow \text{d} & & \downarrow \text{n} & & \\ \binom{2}{2} & & \binom{5}{5} & & \binom{0}{2} & & \omega = 2 \end{array}$$

- **r** is *unkind* towards pl. 1 after **in**, rejection hurts pl. 2 a lot $(-7)$ and pl. 1 a little $(-1)$, for *high enough* $\theta_1$, pl. 1's reply is **n**o.
- *If* pl. 2 is afraid of rejection he **d**elivers/complies (**in** is kind if 1 expects compliance, but only unkindness is supposed to matter).
- *If* pl. 1 anticipates this, he goes **in.**
- Thus, we expect *high rates* of **in**, **d**, **n**o (the opposite of the High Game).

# Experimental Design

- Between-subjects experiment (treatments $\omega = 2$ and $\omega = 10$, with 1 ECU=1\$).
- Subjects were randomly assigned to roles 1 and 2 and played 5 times *in the same role* against changing co-players (hoping to induce some convergence to an equilibrium).
- 5 sessions with 6(H)+6(L) subjects randomly assigned to roles (3 changing pairs in each of H and L) playing 5 rounds:
  $5 \times 3 \times 5 = 75$ observed plays (terminal histories).

# Experimental Results (aggr. freq.s in the 75 H/L-plays)

- **High Game**

$$1 \xrightarrow[30/75]{\text{in}} 2 \xrightarrow[27/30]{\text{t}} 1 \xrightarrow[27/27]{\text{yes}} \binom{1}{9}$$

$$\frac{45}{75}\downarrow \text{o} \qquad \frac{3}{30}\downarrow \text{p} \qquad \frac{0}{27}\downarrow \text{n}$$

$$\binom{2}{2} \qquad \binom{5}{5} \qquad \binom{0}{10} \qquad \omega = 10$$

- **Low Game**

$$1 \xrightarrow[57/75]{\text{in}} 2 \xrightarrow[37/57]{\text{t}} 1 \xrightarrow[23/37]{\text{yes}} \binom{1}{9}$$

$$\frac{18}{75}\downarrow \text{o} \qquad \frac{20}{57}\downarrow \text{p} \qquad \frac{14}{37}\downarrow \text{n}$$

$$\binom{2}{2} \qquad \binom{5}{5} \qquad \binom{0}{0} \qquad \omega = 0$$

- The **null hypothesis** of *treatment-independent* behavior (differences due to randomness) is *rejected* (see pp 9-10 in DS&V-E). The *difference* is in the *predicted direction*.

# Reciprocity and dynamic consistency

- According to the general theory of Dufwenberg & Kirchsteiger (DK, 2004), *reciprocity is a reactive action tendency*. This is modeled with players having *different psychological utility functions at different nodes of the game*, which may yield *dynamic inconsistency* of preferences (cf. DK 2004, Battigalli, Corrao & Dufwenberg 2019).
- Such *dynamic inconsistency* may be psychologically plausible, but it *is not a necessary feature of the intuitive notion of reciprocity*.
- I present below a *dynamically consistent model of reciprocity* for general game forms (like DK, I restrict attention for simplicity to game forms with observed actions).

# A dynamically consistent model of reciprocity

- **Kindness of** $i$ (at the beginning of the game): I take as given that the equitable payoff of $j$ from $i$'s perspective is determined by some belief-dependent function $\pi_j^e(\alpha_{i,-i})$ (e.g., as in DK). The kindness of $i$ towards $j$ is

$$\kappa_{ij}(\alpha_i) = \mathbb{E}_{\alpha_i}(\pi_j) - \pi_j^e(\alpha_{i,-i})$$

- **Note:** In LF game forms $u_2(a_1, a_2, \alpha_1) =$

$$\begin{aligned} & \pi_2(a_1, a_2) + \theta_2 \kappa_{12}(a_1, \alpha_{12})(\pi_1(a_1, a_2) - \pi_1^e(a_1)) \\ = \; & \pi_2(a_1, a_2) + \theta_2 \kappa_{12}(a_1, \alpha_{12})\pi_1(a_1, a_2) - \underbrace{\theta_2 \kappa_{12}(a_1, \alpha_{12})\pi_1^e(a_1)}_{\text{independent of } a_2}. \end{aligned}$$

Thus,

$$\begin{aligned} & \arg \max_{a_2 \in A_2(a_1)} \bar{u}_{2,a_1}(a_2, \beta_{21}) \\ = \; & \arg \max_{a_2 \in A_2(a_1)} \pi_2(a_1, a_2) + \theta_2 \mathbb{E}_{\beta_{21}}(\kappa_{12}(a_1, \alpha_{12}) | a_1)\pi_1(a_1, a_2). \end{aligned}$$

# A dynamically consistent model of reciprocity

- Given the previous observation, I propose to model reciprocity concerns with

$$u_i(z, \alpha_{-i}) = \pi_i(z) + \sum_{j \neq i} \theta_{ij} \kappa_{ji}(\alpha_j) \pi_j(z),$$

  a kind of *"state-dependent" utility function*, which yields *dynamically consistent conditional preferences*.

- By standard dynamic programming arguments (*One-Deviation Principle*), $\alpha_{i,i}$ ($i$'s strategy) *maximizes* $\mathbb{E}_{\alpha_{i,i},\beta_{i,-i}}(u_i|h)$ *starting from every* $h \in H$ *IFF* $\alpha_{i,i}$ *is an intrapersonal equilibrium* given $\beta_{i,-i}$, that is, IFF, for every $h \in H$ and $a_i^* \in A_i(h)$,

$$\alpha_{i,i}(a_i^*|h) > 0 \Rightarrow a_i^* \in \arg\max_{a_i \in A_i(h)} \bar{u}_{i,h}(a_i, \alpha_{i,i}, \beta_{i,-i}),$$

  where $\bar{u}_{i,h}(a_i, \alpha_{i,i}, \beta_{i,-i}) = \mathbb{E}_{\alpha_{i,i},\beta_{i,-i}}(u_i|h, a_i)$.

# References


- 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 (2022): "Belief-Dependent Motivations and Psychological Game Theory," *J. Econ. Lit.*, 60, 833-882.
- Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity".*Games Econ. Behav.*, 47, 268-298.
- Dufwenberg, M., A. Smith, and M. Van Essen. (2013): "Hold-up: With a Vengeance". *Econ. Inq.*, 51, 896-908.

# Additional references on reciprocity


- 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.
- Akerlof, G. (1982): "Labour Contracts as a Partial Gift Exchange," *Q. J. Econ.*, 97, 543-69.
- Dhaene, G., and J. Bouckaert (2010): "Sequential Reciprocity in Two-Player, Two-Stage Games: An Experimental Analysis," *Games Econ. Behav.*, 70, 289-303.
- Dufwenberg, M., and G. Kirchsteiger (2019): "Modelling Kindness". *J. Econ. Behav. Organ.*, 167, 228-234.
- Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *Am. Econ. Rev.*, 8, 1281-1302.
- Sobel, J. (2005): "Interdependent Preferences and Reciprocity," *J. Econ. Lit.*, 43, 396-440.