Beyond Preferences: Mental Representations in Games*

Latest Version

Andrea Amelio† Nicola Gennaioli‡ Salvatore Nunnari§

November 13, 2026

Abstract

We offer a theory of context effects in games. A player recognizes what the game is about – morality, self-interest, or joint gain – by recalling a category of similar experiences, which draws her attention to a norm, her own slice, or the pie. In dictator, ultimatum, and trust games, we change context by cuing experiences and by framing the game. As the model implies, measured similarity between treatment contexts and the categories predicts changes in reasoning, beliefs, and transfers. This explains why a market frame lowers transfers in dictator and ultimatum games but raises them in the trust game, where it cues joint gain, and why it reverses the effect of an option to take on dictator giving. Shifts in representation have a sizable explanatory power.

Keywords: Mental Representations, Memory, Behavioral Game Theory, Social Preferences, Social Norms, Framing, Dictator Game, Ultimatum Game, Trust Game.
JEL Codes: C72, C91, D83, D91.

*This study was approved by Bocconi Research Ethics Committee (EA000782) and pre-registered on AsPredicted (https://aspredicted.org/db9fi8.pdf and https://aspredicted.org/p53he8.pdf).

†Bocconi University; andrea.amelio@unibocconi.it.
‡Bocconi University; nicola.gennaioli@unibocconi.it.
§Bocconi University; salvatore.nunnari@unibocconi.it.

1 Introduction

Strategic behavior depends on context and history. Exposure to war fosters ingroup cooperation for years (Bauer et al., 2016). Young men from violent areas read ambiguous encounters as threats; teaching them to reinterpret such situations cuts violence (Heller et al., 2017). Adding a price crowds out morality: fining late daycare pickups makes parents later, paying for blood lowers donations (Gneezy and Rustichini, 2000; Mellström and Johannesson, 2008). Incentives differ across these settings, but this cannot be the whole story: behavior also shifts when the rules of the game are fixed. Ultimatum offers are higher where everyday production relies on cooperation (Henrich et al., 2001) and lower when the game is framed as a buyer-seller exchange (Hoffman et al., 1994). Dictator giving falls when the option to take is added (Bardsley, 2008; List, 2007). Where does this instability come from?

In game theory, players read the rules as written, so instability of the equilibrium set comes from changing tastes. Existing accounts accordingly allow tastes or norms to depend on game frames (Fehr and Schmidt, 1999; Charness and Rabin, 2002; Krupka and Weber, 2013; Ellingsen and Mohlin, 2026). This fits behavior, but does not say why a context activates one taste or norm rather than another, nor when the effect should be stronger.

We offer a cognitive theory of this process. Before strategizing, a player must recognize what the interaction is about. Is dividing a sum like sharing a windfall with a friend, bargaining with a stranger, or investing in a joint project? Following Bordalo et al. (2026), a game's rules and context cue a category of similar situations, which supplies a *mental representation*: a motive for playing, and a belief about recalled counterparts. Given this representation, the player chooses optimally. We model three everyday settings. *Moral* situations focus attention on a sharing norm; *Self-interest* and *Win-Win* focus on different *material* payoffs: the former on one's own slice, the latter on the size of the pie. Context shifts a game's similarity across experiences and thus its representation; Behavior then changes because the player solves a different problem. Stable heterogeneity in behavior also follows: different lived experiences shape representations under the same cue.

In the model, a context resembling a category raises the share of players who reason in that category, moving transfers toward that category's optimum in the game. Two restrictions on context effects follow. Direction: within a game, the sign of the effect is set by the category the context resembles, since optima generally differ across categories; across games, two categories can have similar optima in one game and very different ones in another, predicting game-specific effects of the same context. Strength: within a game, of two contexts the one more similar to a category moves more players to it; across games, the same context resembles some games more than others, so its effect is larger in those games.

We apply this structure to the dictator (DG), ultimatum (UG), and trust (TG) games. A transfer affects one's own slice and the pie differently in each, causing optima to vary: Moral transfers most in the DG, where the pie is fixed, but Win-Win transfers most in the UG and TG, where giving raises expected output; Self-interest transfers least everywhere. This mapping, confirmed in our Control condition, yields directionality: at given beliefs, a context cuing Self-interest lowers transfers in every game, while one cuing Win-Win lowers them relative to Moral in the DG *but* raises them in the UG and TG. It also yields strength: a context more similar to DG than TG should affect the former game more strongly. Beliefs enter the same structure. A context that cues a different counterpart changes the belief; one that leaves her untouched does not. A context raising expected cooperation should then lower UG transfers and raise TG transfers, on top of its effect through categories. Categories, beliefs, and transfers co-move, with directions and strengths set by similarity.

We test the model in a pre-registered Prolific experiment ($N \approx 12{,}750$) that varies context in two treatments. We measure the model's inputs. For similarity, LLMs rate how much the context of a treatment resembles each category description or the game, signing the direction and strength of effects. For representations, we classify subjects' open-ended reasons for their decision as Moral, Self-interest, or Win-Win, and elicit their beliefs about the counterpart. For experiences, DG subjects recall similar situations before choosing.

The first treatment manipulates memory by varying which story participants summarize

before playing. *Bonus* is about sharing a bonus with a teammate; *Aid* is about sharing poverty relief with a needy acquaintance. The game is identical across conditions. LLMs rate Aid as more similar to Self-interest than Bonus, fixing the direction: transfers should fall in every game. They also rate both stories, which divide a fixed sum, as most similar to the DG and least to the TG, which fixes the strength. Categories follow: in the DG and UG, Aid moves 11.5 and 10.4 percent of subjects from Moral to Self-interest; in the TG nothing moves. Beliefs are unchanged: the stories say nothing about the counterpart, so they do not affect recall about her. Following categories and beliefs, transfers fall by 5.6 and 2.4 points of the budget in the DG and UG, and not at all in the TG. A story about financial need lowers giving only in games similar to it, by fostering recall of self-interested situations.

The second treatment changes the description: a *Market* frame recasts the same game, with identical payoffs, as an exchange between a data provider and a broker. As intuition suggests, LLMs rate the frame as reducing similarity to Moral, but also predict that the displaced reasoning goes more to Self-interest in the DG and UG and more to Win-Win in the TG. Categories follow: the Moral share falls by 24 to 43 points in every game, and Win-Win's gain relative to Self-interest is largest in the TG. Beliefs now move, because the frame redescribes the counterpart: the exchange brings to mind people who take profitable deals and honor them, so players expect more acceptance in the UG and more return in the TG. Categories and beliefs move transfers in the predicted opposite directions: they fall by 10.9 points in the DG, where Self-interest pushes toward keeping; fall by 21.9 in the UG, where the rise in Self-interest and in expected acceptance lowers offers in every category; and rise by 14.5 in the TG, where Win-Win reasoning and optimism push toward sending. The frame raises TG surplus by 15% but destroys a third of UG surplus: optimism is excessive in the UG. A market does not make people less prosocial; it changes what they recognize.

Holding the mapping from representations to choices constant, measured shifts in categories and beliefs reproduce the direction and much of the size of treatment effects on average actions for both players, and the shifts in mass at the modal transfers, 50:50 under Moral

and zero under Self-interest.

We conclude with two analyses in the DG. First, we test heterogeneity. Before choosing, subjects describe a situation similar to each of four hypothetical allocations; since all face the same cues, variation in recall reflects differences in databases. Consistent with this, participants who recall Moral situations for more allocations transfer more, and more often justify their chosen transfer on Moral grounds. At the same time, these subjects switch to material categories and cut their transfer more when they later play the Market DG, showing the power of cues. Second, we apply our model to a known framing effect: giving is higher when the dictator starts from \$8/\$4 with an option to take (LT) than from \$12/\$0 (DG) (Krupka and Weber, 2013). Direction and strength again organize the result. In Control, LT is rated more similar to Moral than DG. Accordingly, Moral reasoning and transfers are higher in LT. Under Market, where taking means raising a price, the frame is more similar to LT, so it should displace Moral more strongly there. Accordingly, the gap *reverses*: in LT transfers are now lower. Many see raising the price from \$8 to \$10 as a Win-Win situation.

We offer and test a theory of context in games that builds on the memory and attention framework of Bordalo et al. (2026). It relates to growing work on cognition (e.g., Bordalo, Gennaioli and Shleifer, 2012, 2020; Ba, Bohren and Imas, 2025; Oprea and Woodford, 2026; Graeber, Roth and Zimmermann, 2024).[1] One contribution is to offer a method to restrict the direction and strength of context effects on reasoning and choice across games. These restrictions cannot be explained by a general tendency of a markets to weaken social norms, by ex-post rationalization of a low transfer in material terms, or by experimenter demand: none predicts different signs and strengths across games based on similarity. Predictive power comes from measured similarity, the driver of representations, and from joint restrictions in its link to transfers, beliefs and reasons measured using open-ended text (Haaland et al., 2025). Following cognitive science (Howard and Kahana, 2002; Bhatia, 2017), we measure

[1] In case-based decision theory (Gilboa and Schmeidler, 1995), experiences are used under structural ignorance; in Jehiel (2005), analogies are used under incomplete information. In our model, attention and memory matter even under full information, and similarity varies with salient cues.

similarity with LLM ratings, which track human ones (Marjieh et al., 2024). This is feasible at scale, captures broad patterns rather than idiosyncratic differences across human raters, and is independent of the outcome: the LLM sees only the texts subjects read and the category descriptions, and is never told or asked how people behave.

Our second contribution is to offer a unification for seemingly contradictory findings of heterogeneous and unstable strategic behavior. This includes evidence emerged from a large body of work on social preferences and norms (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016), on intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006), on markets and morality (Hoffman et al., 1994; Liberman, Samuels and Ross, 2004; Dufwenberg, Gächter and Hennig-Schmidt, 2011; Ellingsen et al., 2012; Falk and Szech, 2013; Bartling, Weber and Yao, 2015; Cohn, Fehr and Maréchal, 2014), and others. In the Conclusion we discuss how experiences and changing cues can reconcile different findings. We finally contribute to behavioral game theory (Stahl and Wilson, 1995; Camerer, 2003; Eyster and Rabin, 2005). Closest to us, Bacharach (2006) and Sugden (2003) study team versus individual reasoning in games. We provide a general foundation for modes of reasoning: recognition selects a category, shaping behavior via attention to different payoffs.

Section 2 presents the model, Section 3 describes the experiment and the measurements. Section 4tests our predictions and quantifies the explanatory power of representations. Section 5.1 studies heterogeneity across and within people in the DG and the reversal of the taking effect. Section 6offers concluding thoughts.

2 The Model

Section 2.1 shows how representations shape the transfer in each game, a link we validate in our Control condition. Section 2.2 endogenizes representations, giving our predictions. Proofs are in Appendix A.1.

2.1 How Representations Shape Behavior

The Game. Player 1 has unit income and sends a share $t \in [0,1]$ to Player 2, who observes t and selects $x \in X$. Profile (t,x) sets total income $\Pi(t,x)$, "the pie", and Player 2's slice $\Pi_2(t,x)$, leaving Player 1 with $\Pi_1 = \Pi - \Pi_2$. It also sets the moral cost $\lambda\,(t-t^*)^2/2$ of deviating from norm t^*. Relative to distributional models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) utility depends on attention to these payoffs.

A *representation* is a pair $\mathbf{r} = (\boldsymbol{\alpha}, x^*)$: attention weights $\boldsymbol{\alpha} \equiv (\alpha_\Pi, \alpha_{\Pi_2}, \alpha_{t^*})$ on the unit simplex and a belief x^* about Player 2. Given $\mathbf{r}$, Player 1 maximizes

$$u_1(t;\mathbf{r}) = \alpha_\Pi \cdot \Pi(t,x^*) - \alpha_{\Pi_2} \cdot \Pi_2(t,x^*) - \alpha_{t^*} \cdot \frac{\lambda}{2}\,(t-t^*)^2\,. \tag{1}$$

A player with $\alpha_{t^*} = 0$ and $\alpha_\Pi = \alpha_{\Pi_2}$ cares only about her own slice; one with $\alpha_{t^*} > 0$ feels the norm; one with $\alpha_\Pi > \alpha_{\Pi_2}$ values the pie beyond her share of it. In distributional models these weights are a trait and x^* an equilibrium belief. Here they are produced by recognition: the category Player 1 retrieves supplies $\boldsymbol{\alpha}$ and, with it, a belief x^* from recalled counterparts. Based on recollections, the same person is thus selfish in one context and generous in another, and holds different beliefs about the same counterpart.

Player 2 has nothing to forecast: her representation is a vector of attention weights on the pie, Player 1's slice, and a fair response norm $x(t)$ to transfer t, $\boldsymbol{\alpha_2} \equiv (\alpha_\Pi, \alpha_{\Pi_1}, \alpha_x)$. We focus on Player 1, but also report the results for Player 2 (model predictions for her are in Appendix A.2).

Categories. Player 1 draws $\boldsymbol{\alpha}$ from one of three categories, $\mathrm{C} \in \mathbf{C} = \{M, S, W\}$, that differ in how much they weight the norm against different material payoffs.

- *Moral* (M): the situation is about fairness, as when sharing a windfall with a friend. The norm is attended to as much as money and the pie matters only through one's own slice: $\boldsymbol{\alpha}_M = (1/3, 1/3, 1/3)$.
- *Self-interest* (S): the situation is about selfishness, as when bargaining with a stranger.

Attention to money $m \in (2/3, 1)$ goes to one's own slice: $\boldsymbol{\alpha}_S = (m/2,\, m/2,\, 1-m)$.

- *Win-Win* (W): the situation is about value creation, as when investing in a joint project. The extra attention to money goes to the pie: $\boldsymbol{\alpha}_W = (m - 1/3,\, 1/3,\, 1-m)$.

We assume that the belief x^* does not vary across categories, which isolates the effect of $\boldsymbol{\alpha}$; we then derive what belief changes imply and test them against measured beliefs. Belief x^* can be wrong, and Section 4.2 compares it to Player 2's actual behavior.

Optimal Transfer. Maximizing (1) with the weights of category C, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{C}}$, an interior transfer solves

$$t_{\mathrm{C}} = t^* + \frac{\alpha_{\Pi} \cdot \dfrac{\partial \Pi(t_{\mathrm{C}}, x^*)}{\partial t} - \alpha_{\Pi_2} \cdot \dfrac{\partial \Pi_2(t_{\mathrm{C}}, x^*)}{\partial t}}{\lambda \cdot \alpha_{t^*}}. \tag{2}$$

Player 1 departs from the norm by weighing what an extra unit of t adds to the pie (scaled by α_{Π}) against what it hands to Player 2 (scaled by α_{Π_2}), and departs more the less she attends to the norm (the denominator). Our paradigmatic games, Dictator (DG), Ultimatum (UG), and Trust (TG), differ in these derivatives and in the norm t^*, which we set at equal split.

DG: Player 2 is inactive, the pie is fixed, $\partial\Pi/\partial t = 0$, $\partial\Pi_2/\partial t = 1$, and $t^* = 1/2$.

UG: Player 2 accepts ($x = 1$) or rejects ($x = 0$, both earn zero). With believed acceptance $x^*(t) = a + bt$, the expected pie and slice obey $\partial\Pi/\partial t = b$, $\partial\Pi_2/\partial t = a + 2bt$, and $t^* = 1/2$.

TG: Player 2 receives $(1+r)t$ and returns a share x; the pie is $1 + rt$, so $\partial\Pi/\partial t = r$, Player 2's slice obeys $\partial\Pi_2/\partial t = (1 - x^*)(1 + r)$, and the norm equalizes expected payoffs, $t^*(x^*) = [\max\{1, 1 + (1 - 2x^*)(1 + r)\}]^{-1}$.

We assume $0 < a \leq b$ (in UG $t = 0$ is believed to be accepted with low probability compared to $t = 1$), $(1 + r)x^* < 1$ (in TG Player 1 expects back less than what Player 2 receives), and $r \geq 1$ (in our design $r = 2$). The first two hold in Control: elicited beliefs give $a = 0.13$ and $b = 1.23$, and TG senders expect back 92% of what they send. Inserting game-parameters in (2), we obtain the following.

Proposition 1 (Optimal Transfer by Category)**.** *At interior optima, categories shape the optimal transfer. In* DG*,* $t_M^{\mathrm{DG}} > t_W^{\mathrm{DG}} > t_S^{\mathrm{DG}}$*. For* $\mathrm{G} \in \{\mathrm{UG}, \mathrm{TG}\}$*,* $t_W^{\mathrm{G}} > t_M^{\mathrm{G}} > t_S^{\mathrm{G}}$*.*

Self-interest gives the smallest transfer in all games. Win-Win depends on the game: it transfers less than Moral in DG, where the pie is fixed, but more in UG and TG, where a larger transfer avoids destructive rejection or exploits the multiplier. A joint material gain concern can produce a larger transfer than morality.

Transfers intuitively react to beliefs in all categories: in UG t drops in expected acceptance, in TG it rises with expected return. If beliefs covary with categories in the natural way, $x_S^* \leq x_M^* \leq x_W^*$, the TG ranking is reinforced, while patterns in UG depend on whether the level a or the slope b changes. This is one reason why measuring beliefs is key. Welfare follows the ranking: total income is lowest and inequality highest under Self-interest, and the pie is largest under Win-Win.

Relation to Social Preference Models. Since $\Pi = \Pi_1 + \Pi_2$, the utility function in (1) can be written as

$$u_1(t; \mathbf{r}) = \alpha_\Pi \cdot \Pi_1(t, x^*) + (\alpha_\Pi - \alpha_{\Pi_2}) \cdot \Pi_2(t, x^*) - \alpha_{t^*} \cdot \frac{\lambda}{2} (t - t^*)^2 . \tag{3}$$

Player 1 puts weight α_Π on her own slice, $\alpha_\Pi - \alpha_{\Pi_2}$ on Player 2's slice, and α_{t^*} on the norm; in (1), Player 2's slice enters negatively only because the pie contains it too. Familiar models correspond to particular weights. Without the norm term ($\alpha_{t^*} = 0$), $\alpha_\Pi = \alpha_{\Pi_2}$ is the selfish player, $\alpha_\Pi > \alpha_{\Pi_2}$ gives linear altruism (Andreoni and Miller, 2002) or, equivalently with two players, a concern for efficiency, as in the social-welfare preferences of Charness and Rabin (2002), and $\alpha_\Pi < \alpha_{\Pi_2}$ gives spite, their competitive preferences. With the norm term ($\alpha_{t^*} > 0$) and $\alpha_\Pi = \alpha_{\Pi_2}$, the player is inequity averse. In DG, where $t - t^* = -(\Pi_1 - \Pi_2)/2$, the norm term is a quadratic loss from inequality, so (3) is the motivation function of Bolton and Ockenfels (2000), a smooth and symmetric counterpart of the piecewise-linear loss of Fehr and Schmidt (1999). The norm term is likewise a quadratic loss from inequality, for the

proposed split in UG and for expected payoffs in TG. Read as a loss from departing from what is appropriate, it is the quadratic loss around a fairness ideal of Cappelen et al. (2007) and a single-peaked case of the norm function of Krupka and Weber (2013). The three categories correspond to three configurations of these preference weights. Moral has fixed weights, while a single common parameter, $m \in (2/3, 1)$, determines the Self-interest and Win-Win profiles. Their weights on own slice, Player 2's slice, and the norm are (1/3, 0, 1/3) under Moral, the inequity-averse player; $(m/2,\ 0,\ 1-m)$ under Self-interest, the same player with more weight on her own slice and less on the norm, who tends to the selfish player as $m \to 1$; and $(m-1/3,\ m-2/3,\ 1-m)$ under Win-Win, the only category with a positive weight on Player 2's slice. Reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) has no role for Player 1, who moves first; for Player 2 it enters through the response norm $x(t)$, which can reward or punish the transfer she receives (Appendix A.2). What recognition changes is the source of the weights.

2.2 How Representations Are Determined

Following Bordalo et al. (2026), Player 1 represents the game by matching its description to situations she faced in the past. We set up the model and derive how a recent experience in memory and a change in description shift $\mathbf{r}$.

Description, Database, Similarity. A description is a vector $\mathrm{G}_\delta, \mathrm{G} \in \{\mathrm{DG}, \mathrm{UG}, \mathrm{TG}\}$ that records actions, payoffs, and context (player roles, setting, etc.). Context shapes attention to payoffs via recognition.[2] Player 1's database files each experience in a category $\mathrm{C} \in \mathbf{C}$. An experience vector e records attention $\boldsymbol{\alpha}_e$ to payoffs, and context κ_e (e.g. sharing a windfall with a friend encodes high attention to the norm t^*, the friend's name, location, etc.). It has recency-weighted frequency f_e. Experiences average into their category's prototype $\boldsymbol{\alpha}_\mathrm{C}$, and aggregate into C's frequency weight $f_\mathrm{C} = \sum_{e \in \mathrm{C}} f_e$. Experiences differ across

[2] In Bordalo et al. (2026), description also records payoffs' sensory prominence $\boldsymbol{\alpha}_\delta$. We abstract from it.

people, which is the model's source of heterogeneity, as we later test.

The similarity between the game's attention and description $(\boldsymbol{\alpha}, \mathrm{G}_\delta)$ and a category C is a bounded function equal to:

$$\mathcal{S}\big[(\boldsymbol{\alpha}, \mathrm{G}_\delta), \mathrm{C}\big] = 1 - \big\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\mathrm{C}\big\|^2 - d_\mathrm{C}, \tag{4}$$

where $d_\mathrm{C} = \sum_{e \in \mathrm{C}} (f_e / f_\mathrm{C}) \cdot d(\mathrm{G}_\delta, \kappa_e)$ is the average context distance between the game and the experiences in the category. Similarity falls in the distance between Player 1's attention in the game and the category and in their context distance.

Categorization. Player 1 selects attention $\boldsymbol{\alpha}$ and a category $\mathrm{C} \in \mathbf{C}$ to maximize the weighted similarity between the game and a category, $f_\mathrm{C} \cdot \mathcal{S}[(\boldsymbol{\alpha}, \mathrm{G}_\delta), \mathrm{C}] + \epsilon_\mathrm{C}$ where ϵ_C is an i.i.d. unit scale extreme-value shock. She thus adopts the category's payoff weights $\boldsymbol{\alpha} = \boldsymbol{\alpha}_\mathrm{C}$, and chooses category C in game G with probability:

$$\Pr(\mathrm{C}|\mathrm{G}) = \frac{\exp\big[f_\mathrm{C} \cdot S(\mathrm{G}_\delta, \mathrm{C})\big]}{\sum_{\mathrm{C}'} \exp\big[f_{\mathrm{C}'} \cdot S(\mathrm{G}_\delta, \mathrm{C}')\big]}, \tag{5}$$

where $S(\mathrm{G}_\delta, \mathrm{C}) \equiv 1 - d_\mathrm{C}$ is the average similarity between the game description and the category. A category is ceteris paribus more likely to be used when memory holds more recent experiences of it and the description resembles them.

Proposition 2 (Categorization)**.** *Ceteris paribus, Equation (5) implies that* C *is more likely used in game* G *(and* $\mathrm{C}' \neq \mathrm{C}$ *less) if:*

(i) ***Memory.*** *Player 1 is more exposed to experiences in* C*, where for* $e \in \mathrm{C}$*:*

$$\frac{\partial \Pr(\mathrm{C}|\mathrm{G})}{\partial f_e} = S(\mathrm{G}_\delta, \kappa_e) \cdot \Pr(\mathrm{C}|\mathrm{G}) \left[1 - \Pr(\mathrm{C}|\mathrm{G})\right] > 0. \tag{6}$$

(ii) ***Context.*** *The game's description* G_δ *is more similar to* C*, with:*

$$\frac{\partial \Pr(\mathrm{C}|\mathrm{G})}{\partial S(\mathrm{G}_\delta, \mathrm{C})} = f_\mathrm{C} \cdot \Pr(\mathrm{C}|\mathrm{G}) \left[1 - \Pr(\mathrm{C}|\mathrm{G})\right] > 0. \tag{7}$$

Our first treatment raises exposure to an experience e, as in (i). Its effect depends on

the category to which e is attached. Empirically, we capture it with similarity: e raises the frequency of C by $\partial f_e = S(\kappa_e, C)$, giving the direction of the cue. By (6), the strength of the effect in a game is mediated by the similarity $S(\mathrm{G}_\delta, \kappa_e)$ between e and the game. Recalling e cues the category it resembles, and does so most in the games that resemble e.

Our second treatment changes the description G_δ. Its effect depends on which categories the game becomes more similar to, the term $\partial S(\mathrm{G}_\delta, C)$ in (7), which we again proxy with measured similarity. This gives direction and strength: a new description cues the categories it resembles, most where the change in similarity is largest. The strength is scaled by the category's frequency f_C, which we do not measure directly but proxy in Section 5.1.

Together with Propositions 1 these effects pin down how context shapes t. Changing context τ should move the average transfer in game G as follows:

$$\frac{\partial \bar{t}^{\mathrm{G}}}{\partial \tau} = \sum_{\mathrm{C}' \in \mathbf{C}} \frac{\partial \Pr(\mathrm{C}'|\mathrm{G})}{\partial \tau} \cdot t^{\mathrm{G}}_{\mathrm{C}'}, \tag{8}$$

where $\tau \in \{f_e, S(\mathrm{G}_\delta, \mathrm{C})\}$. Context moves the average transfer in G toward the optimum t^G_C of the categories it resembles and away from that of the categories it does not. Predictions on direction and strength follow. Because t^G_C is game-specific, the same context can lower $\bar{t}^G$ in one game and raise it in another. At constant beliefs, a shift from Moral to Win-Win lowers $\bar{t}$ in DG, where the pie is fixed, but raises it in UG and TG. Because two categories may prescribe similar transfers, as Moral and Win-Win do for a generous UG offer, a shift between them need not move $\bar{t}^G$ at all. Thus, effects are strongest when context reallocates reasoning across categories whose optimal transfers differ most, and in the games to which context is most similar. A context changing the counterpart can then also change transfers through beliefs, which we can sign because we measure beliefs separately. In sum, the model restricts the direction and strength of context effects on categories, beliefs, and transfers across games based on measurable inputs. We now develop our tests.

3 Experimental Design and Measurement

Section 3.1 describes the games and our two treatments, Section 3.2 how we measure representations, similarity, and experiences. Section 3.3 derives our predictions.

3.1 Games and Treatments

Pairs of subjects on Prolific are randomly matched to play DG, UG, or TG (Forsythe et al., 1994; Güth, Schmittberger and Schwarze, 1982; Berg, Dickhaut and McCabe, 1995); the outcome is implemented for one pair in ten. In a game, actions and payoffs are identical across four conditions. Only memory or description varies.

Control. The description is abstract. In DG, Player 1 is "provisionally allocated" \$12 and Player 2 \$0, and Player 1 may transfer any amount. UG adds that Player 2 accepts or rejects, with both earning \$0 after rejection. In TG, Player 1 receives \$6 and decides how much to send; the amount is tripled ($r = 2$) and Player 2 decides how much to send back.

Changing Memory: Aid versus Bonus. Subjects read and summarize a short story, then play the game in its abstract description. A protagonist decides whether to keep an endowment or share it with someone who has a claim on it. In *Bonus*, a team leader considers sharing a performance bonus with a teammate who "made a significant contribution to the team's success". In *Aid*, a man "going through a rough financial patch" receives government assistance and considers sharing it with a struggling neighbor whose application was refused. Both stories end on "What should he do?" without mentioning the chosen action or the game. We designed Aid, with its financial need and distant counterpart to favor material interest over Moral compared to Bonus (whose close teammate sustains Moral). How do stories however relate to Self-interest and Win-Win, and how to different games? This is where measured similarity and the model help.

Changing Description: Market versus Control Player 1 is a *data provider* who "produces survey data through online questionnaires like the one you are in", Player 2 a *broker* who

resells the data. In DG and UG a computer sets a starting price of \$12, which the provider may lower by any amount; the broker sells the data to a research team for a fixed \$12 and pays the provider the price she set, so lowering the price by t is equal to transferring t to Player 2. In DG player 2 is forced to buy at the set price. In UG the broker accepts or rejects the price. In TG the provider has six units of data worth \$1 each to the research team; she decides how many to sell directly at \$1 and how many to pass to the broker, whose buyer pays \$3 per unit, and the broker then decides how much revenue to transfer back. We designed Market to cue material interest and anonymous traders, reducing the incidence of Moral. But again, the model and measured similarity are needed to sign direction and strength on other categories and across games.

The two comparisons are Aid versus Bonus (Control has no reading task) and Market versus Control. The experiment was pre-registered on AsPredicted (#261370 for the DGs, #286187 for the UGs and TGs), with 400 participants per between-subject dictator cell, 600 per within-subject dictator design, and 600 per role in each UG and TG cell. Participants failing an attention check or failing comprehension questions twice are screened out. The Player 1 sample has 1,604 DG, 1,603 LT, 2,400 UG, and 2,400 TG observations; the Player 2 sample has 2,402 UG and 2,346 TG.[3]

3.2 Measuring Representations, Similarity, and Experiences

Representations. Subjects write open-ended text on the reason for their chosen t (Haaland et al., 2025). Following the pre-registration, an LLM classifies each answer as Moral, Self-interest, Win-Win, or unclassified, the last accounting for 1–3% of responses in Control (human validation on 3,990 responses gave 76% agreement for both Players, with Cohen's κ at 0.64, see Appendix B.1). The model receives a description of each category (Appendix B.1): the logic that defines it, the language in which people express it, and examples. A Moral reason is about what is right and treats the other as someone with a claim: "I

[3]TG receivers are recruited only for senders who send a positive amount. Between- and within-subject DG samples are analyzed separately.

wanted to be fair", "I didn't want to be greedy". A Self-interest reason is about what the participant gets: "I don't know these other people...I only know I need 12$", "They might keep everything, so I sent little". A Win-Win reason is about making the pie larger: "It's a good investment since the amount increases", "We increase profits if I send more".[4] Categories are defined by reasons; they mention no amounts and are identical across DG, UG, and TG. Thus, two categories may give similar actions in a game and the same category can give different actions in different games, as in Proposition 1. We also classify reasons by *social proximity*: whether the counterpart is construed as close (peer, teammate, friend) or distant (Appendix B.1). Social proximity increases giving (Hoffman, McCabe and Smith, 1996; Bohnet and Frey, 1999; Charness and Gneezy, 2008; Small and Loewenstein, 2003); here we examine how context affects its presence in stated reasons.

In UG and TG, subjects also report the belief component of $\mathbf{r}$: the probability that their offer is accepted or the amount they expect back. To compare beliefs across subjects who choose different t, we also elicit a *hypothetical belief* at a common action equal to one third of the endowment; Player 2 participants report their hypothetical response to the same action.[5]

Similarity. Categories and stories are not tied to games, but their similarity to games is a key input of the model, which we measure. For Market versus Control, we measure the similarity $S(\mathrm{G}_\delta, \mathrm{C})$ between a description $\delta \in \{\text{Market,Control}\}$ and each category C. This allows us to compute a proxy for direction and strength based on Equation (7), $\partial \Pr(\mathrm{C}|\mathrm{G}) = \partial_\delta S(\mathrm{G}_\delta, \mathrm{C}) \cdot f_C \cdot \Pr(\mathrm{C}|\mathrm{G})[1 - \Pr(\mathrm{C}|\mathrm{G})]$. For Aid versus Bonus, recalling e increases the frequency of C based on their resemblance $\partial f_e = S(\mathrm{C}, \kappa_e)$. We can then proxy for direction and strength using Equation (6), $\partial \Pr(\mathrm{C}|\mathrm{G}) = S(\mathrm{C}, \kappa_e) \cdot S(\mathrm{G}_\delta, \kappa_e) \Pr(\mathrm{C}|\mathrm{G})[1 - \Pr(\mathrm{C}|\mathrm{G})]$. The latter also needs the similarity of e to the game $S(\mathrm{G}_\delta, \kappa_e)$: the story is more powerful

[4]The prompt tells logics apart when words coincide: "both benefit" is Moral when it means a fair outcome and Win-Win when it means more for both. Each response receives the category of its main reason.

[5]Both elicitations are unincentivized. Given the number of questions, including open-ended responses, we use a simple elicitation to limit survey length and respondent burden. Danz, Vesterlund and Wilson (2022) document behavioral challenges with theoretically incentive-compatible belief elicitation, showing that the binarized scoring rule can generate systematic reporting distortions.

in games that are more similar to it.

To compute similarities, we give GPT-6 the text of a game's instructions G_δ and ask it to divide 1,000 similarity points among the categories; averaging 500 runs gives $S(G_\delta, C)$ for each game and condition. We analogously measure the similarities $S(\kappa_e, C)$ and $S(G_{\text{Control}}, \kappa_e)$ for each story $e \in \{\text{Aid, Bonus}\}$ whose text we give to the LLM. LLMs read what participants read and the descriptions the classifier uses. They see no behavior, so similarity restricts predictions before the game is played.

Experiences. In DG we measure experiences: before choosing, a subset of dictators describes a real or imagined situation that resembles each of four hypothetical allocations, $t \in \{0, 1/3, 1/2, 2/3\}$. We classify the descriptions into our three categories. All subjects in a condition face the same cues, so differences in what they recall measure differences in their databases f, before the choice is made.[6] We do not elicit recall in UG and TG because a hypothetical outcome there would specify Player 2's response, affecting measured beliefs.

3.3 From Model to Predictions

Categories and Transfers. Figure 1 reports Player 1's representations and t in Control. The games are recognized differently (left panel). In DG and UG, Moral dominates (68.4% and 76.6% of classified responses), Self-interest takes most of the rest, and Win-Win is negligible (0.8% and 3.3%). In TG, Moral falls to 43.8% and Win-Win rises to 37.0%. The high share of people focusing on moral concerns is in line with work emphasizing the importance of fairness, but there is interesting instability across games: in TG even more subjects focus on joint returns. Our model explains why: the multiplier makes surplus prominent, so subjects attend to it.

Categories are associated with transfers as in Proposition 1. People focused on their own

[6]Another advantage is that participants describe the hypothetical situations before choosing, so these descriptions cannot rationalize their choice ex post. As we show below, they reproduce the relevant patterns in stated reasons, providing a pre-choice check on our main measure of reasoning.

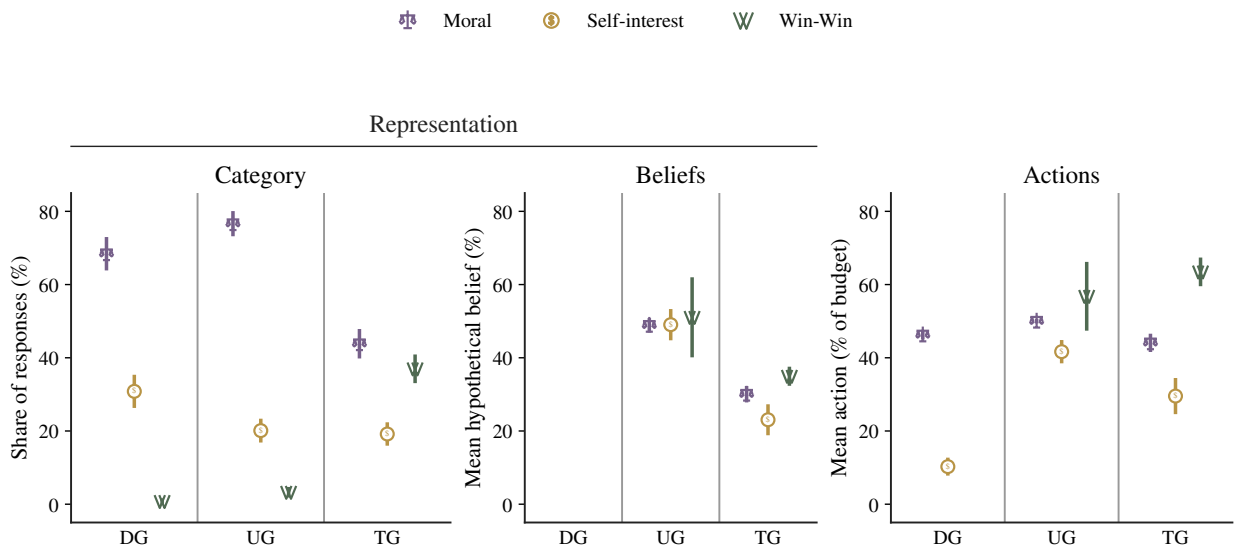


Figure 1: Player 1 representations and t in Control. Left: share of classified responses by category and game. Middle: believed acceptance of a one-third offer (UG) and of a one-third return after a one-third send (TG), by category. Right: mean t by category, as a share of the budget transferred (DG), offered (UG), or sent (TG). Symbols are estimates, vertical lines 95% confidence intervals. Win-Win has three observations in DG and is omitted from the right panel there.

payoff (Self-interest) transfer least in every game (right panel): 10.3% of the budget against 46.2% for people focused on fairness (Moral) in DG, 41.7% against 50.0% in UG, and 29.5% against 44.1% in TG. The Moral–Self-interest gap shrinks from 36 points in DG to 8 in UG, where rejection tames selfish proposers. The largest transfers in UG and TG come from people focused on joint gain (Win-Win): 56.8% in UG ($N = 19$) and 63.5% in TG. Focus on fairness is thus generally associated with higher t, but where the pie must be preserved or enlarged, a material motive such as joint gain is an even stronger engine.

The middle panel shows beliefs. In UG, they do not vary much: all categories expect a one-third offer acceptance rate of roughly 49.0%; Win-win subjects are slightly more optimistic but insignificantly so (they only account for 3% of UG subjects). In TG, beliefs instead vary: Self-interest expect returns 23.1%, people focused on fairness (Moral) expect 30.0%, those focused on joint returns (Win-Win) are most optimistic at 35.0%. Subjects project their own "cooperation" onto the counterpart, reinforcing $t_{\mathrm{C}}^{\mathrm{G}}$.

We conclude with Player 2. Her categories attach attention to material payoffs and to a

norm $x(t)$. Because the latter can reward or punish the received t, for Player 2 we separate a Moral Good from a Moral Bad category. Moral Good is the modal one in UG and TG (56% and 55%), Win-Win accounts for about 25%, and Moral Bad and Self-interest are less common. Categories separate behavior. In UG, almost all receivers with Moral Good, Self-interest, or Win-Win accept, against 24% of those with Moral Bad, which prevails after $t = 0$. In TG, Moral Good and Win-Win return 38% and 32%, against 10% for Self-interest and 4% for Moral Bad (Appendix C.1).

Similarity in Our Treatments. The second model input is similarity, which signs each treatment's effect on categories. We set the baseline incidence $\Pr(\mathrm{C}|\mathrm{G})$ at the Control share. For the memory treatment, we plug into (6) our proxies for the two similarities it calls for, $\partial f_e = S(\kappa_e, \mathrm{C})$ and $S(\mathrm{G}_{\mathrm{Control}}, \kappa_e)$, for each story $e \in \{\mathrm{Aid}, \mathrm{Bonus}\}$, and take the difference across stories; this gives the predicted shift in each category in each game, $\partial_{\text{Aid-Bonus}} \Pr(\mathrm{C}|\mathrm{G})$. For the description treatment, we plug into (7) the measured change in $S(\mathrm{G}_\delta, \mathrm{C})$ from Control to Market, with f_{C} set to one for every category, which gives $\partial_{\text{Market-Control}} \Pr(\mathrm{C}|\mathrm{G})$. (The calculations are in Appendix Table 5).

These predictions should be read as directions and relative strengths, not as point forecasts, for three reasons: i) we do not observe f_{C}, ii) similarity ratings are more reliable in their ordinal than in their cardinal content, and iii) Equations (6) and (7) impose a specific functional form. These limitations bear mainly on the size of a predicted shift, much less on its sign or ranking across categories and games (which depend largely on how similarities are ordered, and these orderings are what Section 4 tests).[7]

Figure 2 has a simple reading. A story about financial need (Panel A) should cue situations of need, in which one looks after one's own payoff; Accordingly, it is predicted to reduce Moral by 2.7 p.p. in DG, 1.8 in UG, and 1.5 in TG, mostly in favor of Self-interest. Crucially, being about dividing a fixed sum, the story resembles mostly DG most and least TG, where

[7] An exercise with concrete vignettes in place of category descriptions reproduces the key orderings: Self-interest gains under Aid, less in TG, and Win-Win gains under Market in TG (Appendix B.3.1).

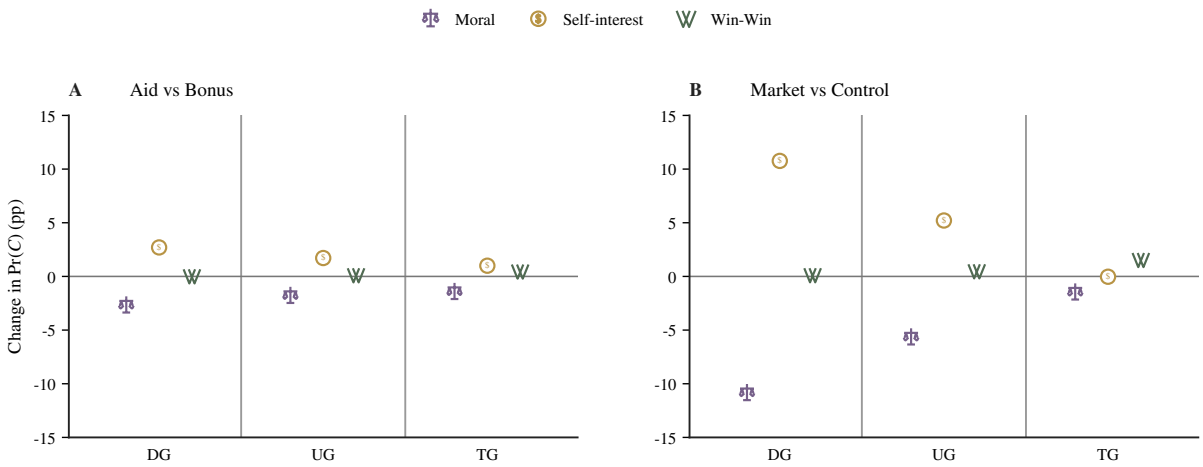


Figure 2: Predicted category shifts from Control shares and measured similarities (pp). A: Aid minus Bonus; B: Market minus Control. Symbols denote Moral, Self-interest, and Win-Win. Appendix Table 5 reports 95% bootstrap intervals.

a growing pie stands out. A market frame (Panel B) should bring to mind material concerns; Accordingly, Moral is predicted to drop by 10.9 and 5.7 p.p. in DG and UG, mostly in favor of Self-interest. Crucially, a transaction can be a haggle or a joint venture, and in TG, where sending multiplies the pie, the ratings say that the drop in Moral is absorbed by Win-Win rather than Self-interest. Finally, the model predicts that the Market frame should affect categories more than stories: a change in the game description is of course most similar to the game itself. By Proposition 1, these shifts in reasoning sign the direction of shifts in t.

- Aid compared to Bonus should:

  P1) reduce Moral and increase Self-interest, most in DG and least in TG;

  P2) reduce the average transfer $\bar{t}^{\mathrm{G}}$, most in DG and least in TG.

- Market compared to Control should:

  P3) reduce Moral in all games, increasing Self-interest in G=DG,UG, Win-Win in TG;

  P4) reduce $\bar{t}^{\mathrm{G}}$ in DG and UG and increase it in TG.

There are three predictions on strength. Stories should be strongest in DG and weakest in TG (P1 and P2). Stories should reduce transfers mostly in DG because selfish senders in UG are optimally more generous: $t_M^{\mathrm{DG}} - t_S^{\mathrm{DG}} > t_M^{\mathrm{UG}} - t_S^{\mathrm{UG}}$. Third, the market treatment should be stronger than stories.

4 Testing the Model

Sections 4.1 and 4.2 test our predictions, Section 4.3 evaluates the quantitative power of representations.

4.1 Changing Memory: Aid vs. Bonus

Figure 3 shows the pdf of transfers. In DG, and less clearly in UG, Aid moves mass from the equal split toward zero and small transfers; in TG, there is no sizable effect. We now test the model using measured representations and average transfers.

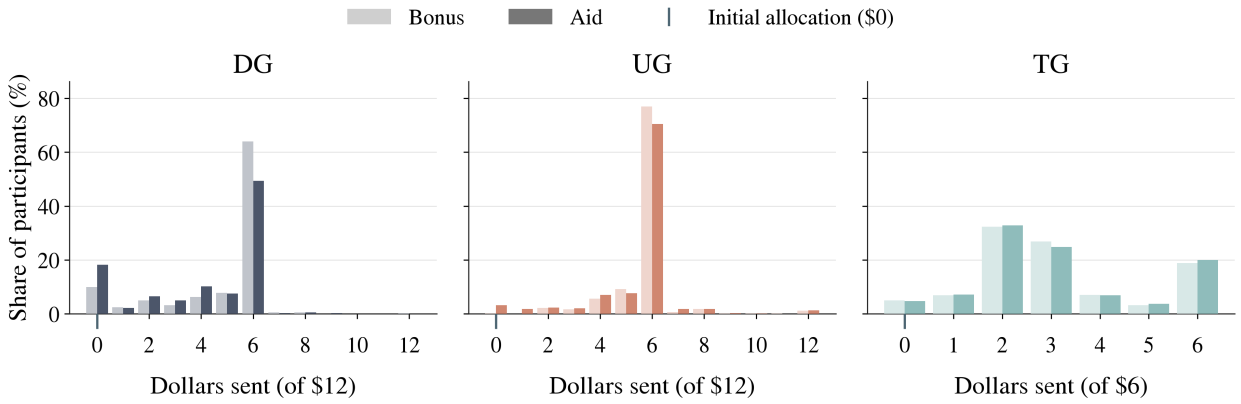


Figure 3: Distributions of Player 1's t, Bonus versus Aid. Bars show shares of participants at each dollar amount (light: Bonus; dark: Aid). Kolmogorov–Smirnov tests reject equality in DG ($D = 0.15$, $p < 0.001$) but not in UG ($D = 0.07$, $p = 0.12$) or TG ($D = 0.02$, $p = 1.00$).

Category shares confirm P1. In the left panel of Figure 4, Aid raises Self-interest by 11.5 and 10.4 p.p. in DG and UG, while Moral falls by 11.7 and 9.7. No category moves in TG, where the story's fit is weakest. Measured social proximity moves the same way: Aid lowers

the share of subjects who describe a close counterpart by 18 points in DG, 12 in UG, 0.4 in TG (Appendix C.2). This is memory at work. A story of financial need makes situations of need top of mind; when the game that follows resembles them, dividing a fixed sum, they are recalled, and the player sees her own payoff where she would have seen a fairness problem. What the game is about has not changed; what it brings to mind has.

Transfers confirm P2. In the right panel, as Self-interest replaces Moral, $\bar{t}$ falls by 5.6 points of the budget in DG and 2.4 in UG (both $p < 0.001$), more in DG, where $t_M^{\mathrm{G}} - t_S^{\mathrm{G}}$ is widest (this is comparable to the 6.7 DG–LT gap of Section 5.2). There is no effect in TG, where categories did not move. Focusing on one's own payoff causes lower t, as Proposition 1 says, so transfers fall where the financial need is retrieved. In TG it is not, so neither reasoning nor giving moves. Direction and strength are disciplined by similarity.

Beliefs about Player 2 do not move (middle panel): the stories say nothing about how she will behave, and her description is unchanged. Stories work through attention. Player 2 is trickier to analyze because she reacts to Player 1's chosen t. Consider, however, hypothetical behavior. Relative to Bonus, Aid raises acceptance of a one-third offer in UG, while the amount sent back in TG is unaffected. Representations move in tandem: in UG Aid reduces Moral Good and raises Moral Bad and Self-interest. These movements are consistent with the model's predictions and measured similarity for Player 2 (see Appendix C.1 and Table 11).

Consistent with P1 and P2, a story making financial need salient moves reasons and behavior toward the category it cues and in the games it resembles. Experimenter demand cannot explain this: both stories describe a moral dilemma and behavior moves asymmetrically across games. A changing game-specific social norm (Krupka and Weber, 2013) cannot either, since the game is fixed.

4.2 Changing Description: Market vs Control

Figure 5 shows the pdf of transfers. Market's effects are large: it nearly empties the equal split in DG and UG, moving mass to zero and small transfers; it moves mass toward the full

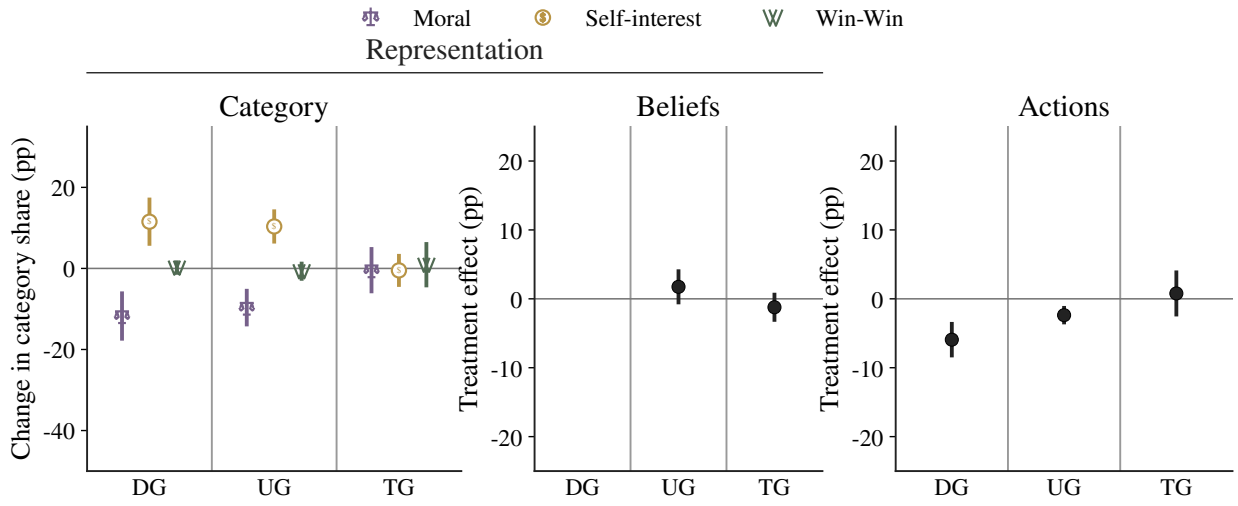


Figure 4: Aid versus Bonus treatment effects for Player 1. Left: effects on category shares. Middle: effects on hypothetical beliefs at the one-third reference action. Right: effects on t in percentage points of the budget. Dots are point estimates, vertical lines 95% confidence intervals. Classified responses; with all responses, estimates change by less than half a point.

send in TG. We now test our predictions on representations and average transfers.

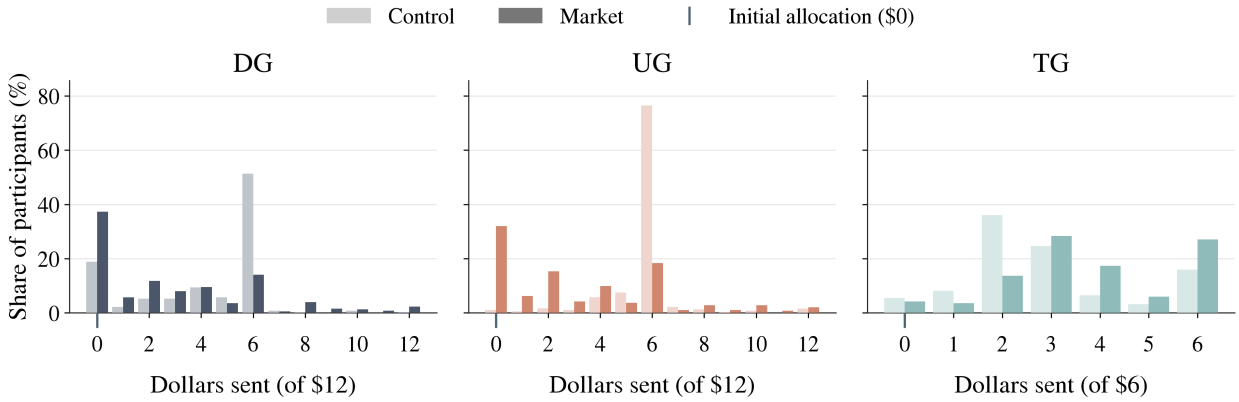


Figure 5: Distributions of Player 1's t, Control versus Market. Bars show shares of participants at each dollar amount (light: Control; dark: Market). Kolmogorov–Smirnov tests reject equality ($D = 0.32$, 0.58, 0.28 in DG, UG, TG; all $p < 10^{-17}$).

Consistent with P3, the left panel of Figure 6 shows that Moral falls in every game, by 27.4, 42.7, and 23.8 p.p. in DG, UG, and TG, and that Win-Win captures the bulk of the adjustment in TG, rising by 18.3 points against 5.5 for Self-interest. Unlike P3, Win-Win also rises more than Self-interest in DG and UG (16.9 and 25.9 against 10.5 and 16.8), which

may reflect unobserved category frequencies, since we set $f_C = 1$.[8] The ranking across games is nonetheless the one similarity predicts. Seeing a market prompts attention to material payoffs, but what the market is about differs by game: in DG it is the division of a fixed price, and Self-interest gains most relative to its Control share; in TG it is a venture whose proceeds depend on what one commits, and joint profit attracts most of the displaced attention.

Market also redescribes Player 2, and so moves the belief component of the representation. Proposers in UG are 16.9 p.p. more confident that a one-third offer is accepted, and senders in TG expect 10.5 p.p. more of the tripled amount back, with beliefs rising within every category. As recall implies, a market changes whom Player 1 thinks she is dealing with: a party who accepts any deal with a positive surplus and delivers on a joint project.

The right panel shows the consequences for transfers, in opposite directions across games as P4 requires. In DG, where the pie is fixed, the focus on one's own share lowers $\bar{t}^{DG}$ by 10.9 points. In UG, optimism about receivers lowers offers in every category, and $\bar{t}^{UG}$ falls by 21.9 points. In TG, the focus on joint profit and the optimism about returns both push toward sending, and $\bar{t}^{TG}$ *rises* by 14.5 points (all $p < 0.001$). One cue, read through the categories each game resembles and the counterpart it brings to mind, produces three different effects, each signed by similarity in advance.

Markets are blamed for reducing prosocial behavior (Falk and Szech, 2013; Bartling, Weber and Yao, 2015; Gneezy and Rustichini, 2000; Hoffman et al., 1994). But no account in which markets simply weaken social preferences or fairness predicts opposite signs in DG and UG versus TG. Measured social proximity makes the point: Market lowers the share of subjects who describe a close counterpart by 13 to 18 p.p. in all three games, which should lower generosity everywhere (Hoffman, McCabe and Smith, 1996), yet t rises in TG. Trust rises through mutual benefit, not kindness (Isoni and Sugden, 2019).

[8] Forcing unclassified text into the categories changes results little (Appendix B.1). The pattern also appears in pre-choice descriptions of hypothetical allocations in DG: Market reduces Moral and social proximity and raises Win-Win (Appendix Figure 11 and Tables 4 and 15).

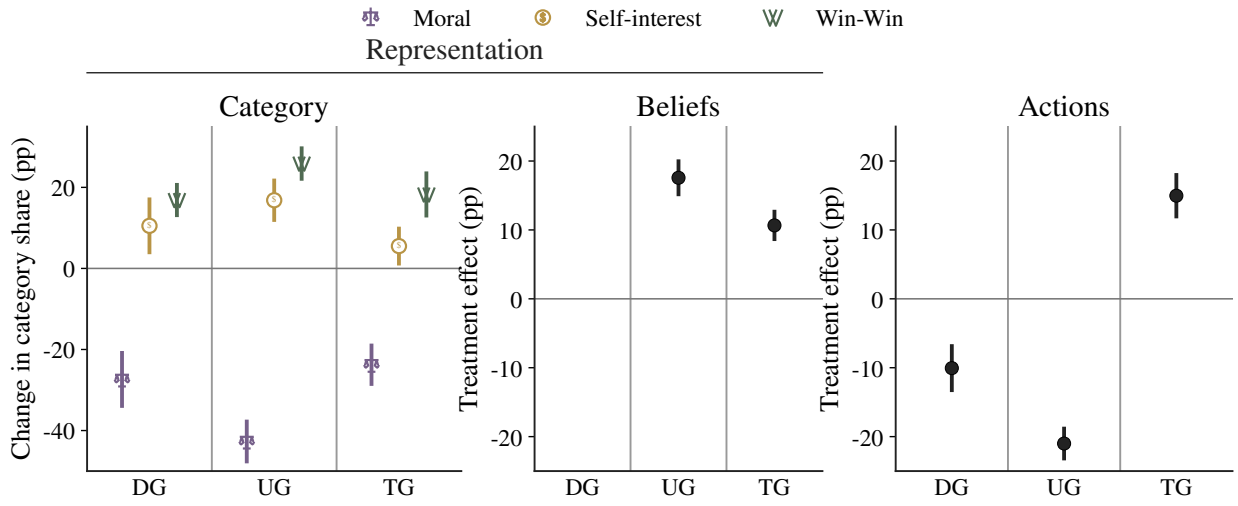


Figure 6: Market versus Control treatment effects for Player 1. Left: effects on category shares. Middle: effects on hypothetical beliefs at the one-third reference action. Right: effects on t in percentage points of the budget. Dots are point estimates, vertical lines 95% confidence intervals. Classified responses; with all responses, estimates change by less than one point.

Are Beliefs Rational? Game theory assumes rational expectations. In our model beliefs are recalled, so they can be wrong, and the errors shape allocations. Player 1's beliefs move in the right direction: receivers do become more "cooperative" under Market, their hypothetical acceptance and share returned rising by 10.5 and 10.7 points (Appendix C.1).[9] But Player 1 overreacts. At the chosen transfer, UG proposers underestimate acceptance by 17.6 points in Control and overestimate it by 9.1 under Market, a 26.6-point swing to optimism; TG senders are only slightly optimistic, by 2.6 points, and the change from Control is negligible (Figure 7, Panel A).[10] Market weakens moral reasoning, but it backfires when optimistic proposers make low offers that receivers find insulting.

Hoffman et al. (1994) showed that framing UG as a transaction between a seller and a buyer lowers offers, but not why; in DG, with 24 pairs per cell, they found no significant effect of the same frame, where we find one with about 400 participants per cell. Our theory

[9]Receivers' representations also move as the model predicts. In UG, Market lowers Moral Good by 39 p.p. and raises Self-interest by 29 and Moral Bad by 10; in TG, it lowers Moral Good by 12 and raises Win-Win by 13, consistent with similarity (Appendix Table 11).

[10]At the one-third action the direction is the same: UG pessimism falls from 13.8 to 7.4 points, and TG errors are +2.5 and +2.3 (Appendix Figure 14). Aid versus Bonus leaves errors unchanged in both games.

and tests explain these effects: one cue, a market frame, shapes recognition and so produces different representations and behavior across games. It reduces Moral but it also raises senders' focus on joint profit and their expected cooperation, with opposite effects across games. Grosskopf (2003); Fischbacher, Fong and Fehr (2009) show that competition among UG respondents lowers offers and raises acceptance, and attribute this to receivers' reduced power to destroy output. We obtain the same effect with a single receiver, by making a market salient, and through senders' excess optimism. These belief effects matter because they shape welfare, to which we turn next.

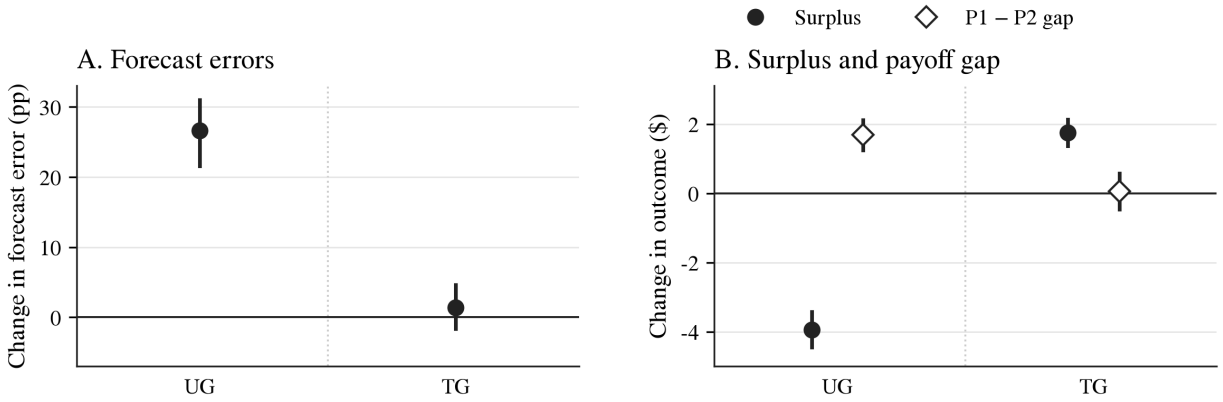


Figure 7: Market minus Control effects on forecast errors, social surplus, and inequality. Panel A: changes in first movers' forecast errors, their belief minus the mean realized second-mover response at the chosen action (acceptance in UG; returned share in TG); positive values indicate a shift toward optimism. Panel B: changes in realized surplus per pair (circles) and in the signed Player 1 minus Player 2 earnings gap (diamonds). Vertical lines are 95% confidence intervals, from a participant bootstrap in Panel A and Welch intervals in Panel B.

Surplus and Inequality. Panel B shows the allocative consequences. In UG, overoptimistic proposers meet receivers less accepting than they expect: acceptance falls from 94.7% to 61.8%, realized surplus per pair from \$11.36 to \$7.42, and the earnings gap between players rises from \$0.08 to \$1.76 ($p < 0.001$). In TG, more sending raises surplus from \$11.81 to \$13.55, by 15%, with no change in the signed gap (\$0.24 versus \$0.29).[11] Market

[11]The absolute TG gap rises from \$2.27 to \$3.19 ($p < 0.001$). Aid versus Bonus shows the same signs, attenuated: UG acceptance falls from 94.7% to 91.2% and surplus from \$11.36 to \$10.94 ($p = 0.018$); TG surplus is unchanged ($p = 0.75$).

creates value where joint gain pays and destroys it where the imagined counterpart fails to materialize. Preference- or norm-based accounts cannot deliver this: a shared norm is correctly anticipated, so it changes behavior without systematic errors. Representations change motives and beliefs together, and the belief errors are where welfare is lost.

4.3 How Much Do Representations Explain?

We assess the ability of shifts in representations alone to explain choice. We thus allow treatments to change only the share of people using Moral, Self-interest, and Win-Win and beliefs of Player 1. The mapping of category shares and beliefs to actions is held constant. We first study average behavior in Equation 8, we then consider specific modes.

Average Behavior. For each game and comparison (Market-Control or Aid-Bonus), we pool the two conditions and estimate $a_i^{\mathrm{G}} = a_0 + \boldsymbol{\beta}' \mathbf{r}_i^{\mathrm{G}} + e_i^{\mathrm{G}}$, where a_i^{G} is the transfer of subject i in game G or the acceptance/return of receiver i in the UG or TG. $\mathbf{r}_i^{\mathrm{G}}$ is the player's representation: category indicators and, for Player 1, hypothetical beliefs. For Player 2 we must hold the response function $x_{\mathrm{C}}^{\mathrm{G}}(t)$ constant. To do so, we include her hypothetical response and the fitted Player 1 action faced.[12] To predict treatment effects, we average each regressor in the treatment and baseline samples and apply the estimated coefficients to both sets of averages. The difference between the fitted mean choices is

$$\overline{\widehat{a}}_{\text{Treatment}} - \overline{\widehat{a}}_{\text{Baseline}} = \widehat{\boldsymbol{\beta}}' \left(\overline{\mathbf{r}}_{\text{Treatment}} - \overline{\mathbf{r}}_{\text{Baseline}} \right),$$

where $\overline{\mathbf{r}}_\tau$ is the average representation (the averages of indicators are category shares) and $\overline{\widehat{a}}_\tau$ is the corresponding fitted mean choice. This extends equation (8) to include beliefs and,

[12] We estimate Player 1's action from her category and hypothetical belief using the two conditions pooled. For each observed action level, we average the fitted values among Player 1s who chose that action; every Player 2 facing it receives the same value. This mapping is common to both conditions. Using instead the actual action faced or indicators for each action level preserves the signs of all four Player 2 predictions. For Market in UG, the predicted fall in acceptance is 28.0 or 32.4 points, respectively, compared with 26.4 in the benchmark and 32.8 observed; the other three predictions change by less than 0.6 points. Appendix D.1 reports the benchmark coefficients; the pre-registrations specified representation–behavior regressions and a decomposition of mean effects, which we extend here (Appendix D.2).

for Player 2, the hypothetical response and action faced.

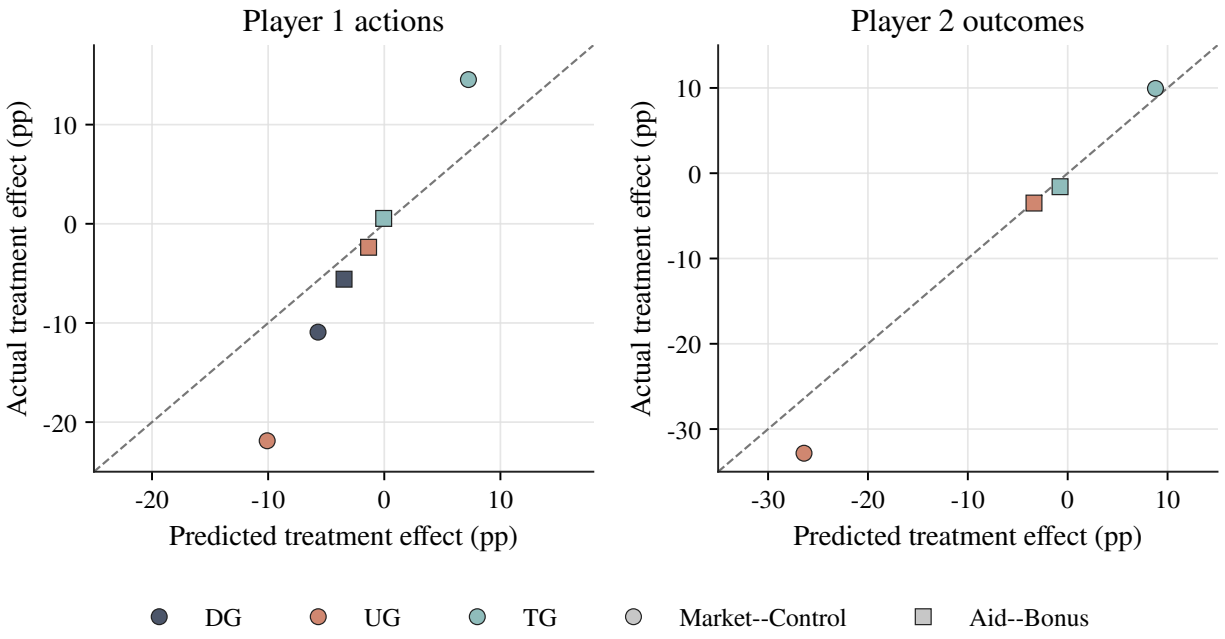


Figure 8: Predicted versus actual treatment effects on actions. Left: Player 1's share sent in DG, UG, and TG. Right: Player 2's acceptance in UG and share returned in TG. Each point is one game-by-comparison effect in percentage points; circles are Market minus Control, squares Aid minus Bonus, colors identify games. The dashed line marks exact prediction.

Figure 8 compares predicted vs. observed effects. The predicted direction is right, including Market's opposite effects on UG and TG. The effects of stories in TG are within two points of zero. A regression of actual on predicted effects gives $R^2 = 0.94$ across the ten, and 0.994 and 0.999 within Player 1 and Player 2.[13] Receiver predictions are close in size: Market lowers UG acceptance by 32.8 points against 26.4 predicted and raises the TG return by 9.9 against 8.8. For Player 1 the predicted Market effects are -5.7, -10.1, and $+7.2$ points in DG, UG, and TG against observed effects of -10.9, -21.9, and $+14.5$. Changes in measured representations alone reproduce the sign and relative size of context effects, and about half of their magnitude.[14] The remainder is behavior changing within categories, a margin the

[13]Estimating the choice functions on all four conditions rather than separately on the two conditions in each comparison preserves the predicted direction of all eight significant effects and gives $R^2 = 0.92$ across the ten (Figure 15 in the Appendix).

[14]Folding unclassified reasons into C leaves results nearly unaffected (-5.7, -9.3, $+7.3$). The pre-registered Oaxaca–Blinder decomposition, which lets choices within representation cells differ across conditions, at-

model, in which attention is fully specified by the category, does not have.

Salient Modes. Does reallocation across ways of seeing a game also reproduce the pdf of t, not only its mean? We group t into seven bins, the salient shares $t \in \{0, 1/3, 1/2, 1\}$ and the intervals between them. We estimate a linear probability model for each bin on the pooled conditions with the same regressors as in the previous test. We then predict the change in each bin's probability, $\widehat{\Delta f}(t)^{\mathrm{G}} = \widehat{\boldsymbol{\beta}}_t'(\bar{\mathbf{r}}_{\text{Treatment}} - \bar{\mathbf{r}}_{\text{Baseline}})$, where t denotes the bin and $f(t)^{\mathrm{G}}$ its density in game G.

Figure 9 compares observed and predicted changes in the bins. Market affects different modes in different games: in DG and UG it removes mass from $t = 1/2$ and adds it at $t = 0$ and $t < 1/3$; in TG it moves mass from shares at $t \leq 1/3$ to $t \geq 1/2$. The prediction gets the direction of each shift right, e.g. an 8-point rise at $t = 1$ in TG against 11 observed, and obtains (but understates) the loss of half-budget transfers in DG and UG. Across seven bins, representations capture 45%, 51%, and 34% of Market's distributional shift in DG, UG, and TG, and 51% and 56% of the Aid shift in DG and UG, measured as the reduction in total-variation error relative to predicting no shift. Categories are associated with salient modes of behavior (fairness to 50:50, selfishness to 0, Win-Win in TG to 1) and changes in representations allow us to track the changing prevalence of these modes.

A game is not what its rules say but what it brings to mind. We offer a theory of this process in which recognition is shaped by experiences and similarity to cues. We obtain two broad results. The first is a theory of context effects and a method for predicting their direction and strength. The second is a mechanism for history (experiences) and context (cues) to shape strategic behavior in surprising ways: a story that puts financial need top of mind makes people more selfish in contexts that resemble the story, a market makes them more selfish where the pie is fixed but more materially cooperative where it can grow. As we argue in the conclusion, this mechanism can unify a range of apparently contradictory

tributes 39%, 19%, and 47% of the Market effect in DG, UG, and TG, and 65% and 55% of DG and UG Aid effects, to changes in cell shares (Appendix D).

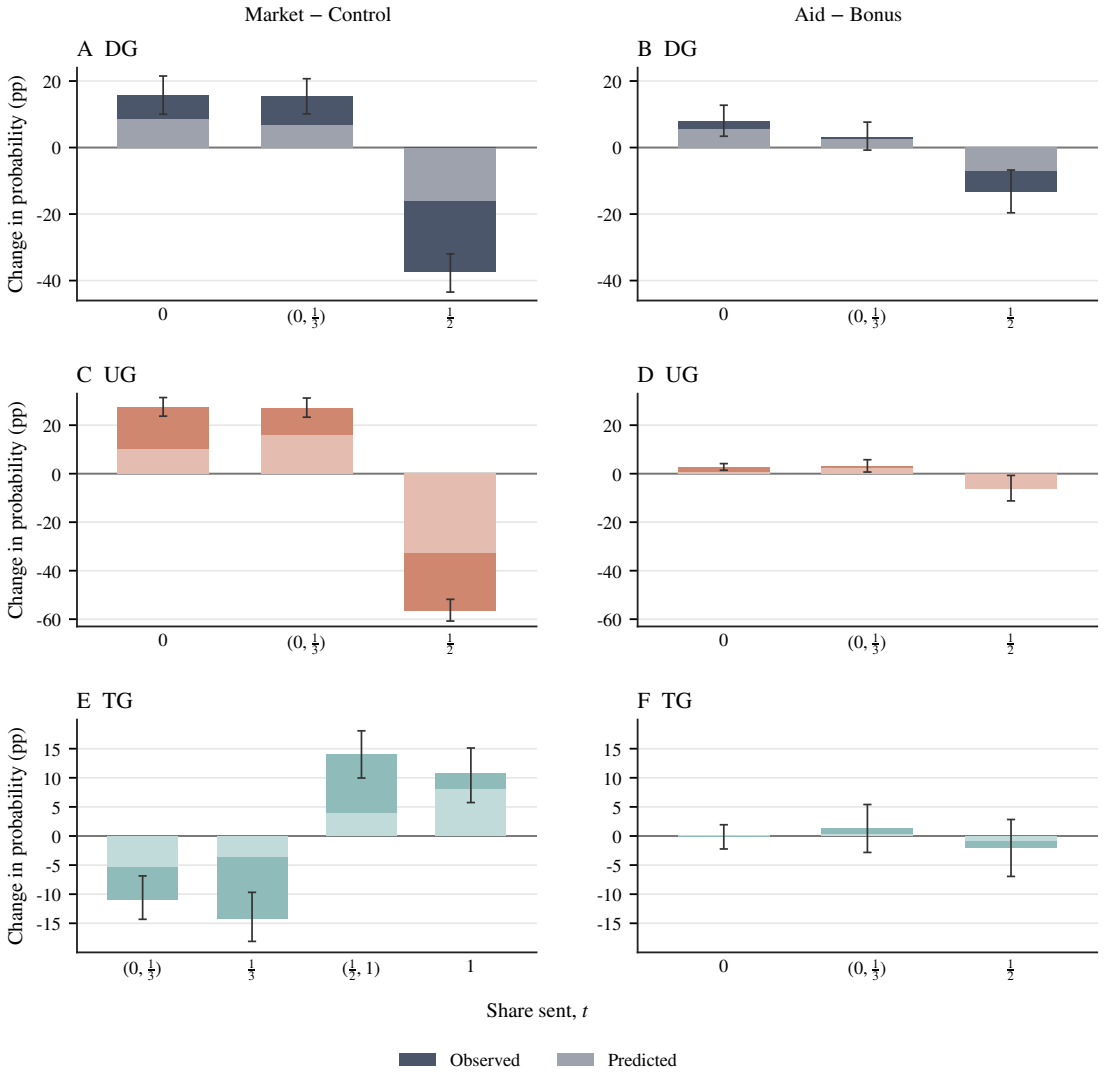


Figure 9: Changes in the distribution of Player 1's share sent. Dark bars show observed changes in bin probabilities, in percentage points; lighter bars overlaid on them show the predictions from representations. Vertical lines are 95% participant-bootstrap confidence intervals for the observed changes (1,000 draws). Columns compare Market with Control and Aid with Bonus; rows show DG, UG, and TG, with the same vertical scale within each game. The Market column displays the principal movements; the Aid–Bonus column displays $t = 0$, $(0, 1/3)$, and $t = 1/2$ in every game. In TG, $t = 1/2$ means half the initial budget. Predictions use category indicators and, in UG and TG, hypothetical beliefs.

patterns of strategic interaction in the field, our original motivation.

5 Heterogeneity and Instability in the DG

We conclude with two exercises in the DG. The first studies the database as a source of strategic heterogeneity across people. The second applies our approach to a well-known framing effect in dictator taking, showing how representations shed light on this puzzle.

5.1 Heterogeneity Across and Within People

Using the recall task in DG we study: i) heterogeneity from experiences under Control but also ii) market-induced instability. We begin with 800 participants whose first DG choice was in Control and then follow the 300 who later played the same DG under Market.

Heterogeneity in Control. Before their first choice, all 800 participants describe situations brought to mind by four hypothetical allocations (Section 3.2).[15] Our control sample exhibits significant variation in recall: 18.8% describe at most one allocation as a Moral situation, 27.5% describe two, and 53.8% describe three or four. This is related to representations and choice: among subjects describing at most one allocation as Moral, 60.0% of classified post-choice reasons are Moral, against 74.0% among those who describe three or four; the same groups send 31.5% and 38.5% of the budget.

To study this issue holding t constant, we use all Control descriptions to estimate:

$$\log \frac{\Pr(\mathrm{C}_{it} = M)}{\Pr(\mathrm{C}_{it} \in \{S, W\})} = \mu + \gamma_t + f_{i\mathrm{M}}, \tag{9}$$

where γ_t captures the hypothetical allocation and $f_{i\mathrm{M}}$ the participant's propensity to retrieve Moral rather than Self-interest or Win-Win. The index is available for 755 participants

[15] The sample comprises 200 between-subject DG Control participants, 300 who played Control first in the DG Control–Market study, and 300 who played DG first in the DG–LT study. The recall task precedes the choice for 200 of the 403 DG and 200 of the 400 LT dictators in Control and for all 400 dictators in each Market cell. Those who complete it transfer and reason like those who do not (DG: 34.4% versus 35.6% of the budget, $p = 0.57$; 67.0% versus 69.8% Moral, $p = 0.55$).

with at least one classified description.[16]

Recall links to representations and choice (see Appendix Table 3). A one-standard-deviation increase in $f_{i\mathrm{M}}$ predicts a 3.4 p.p. higher probability of Moral reason ($p = 0.047$) and a 1.4 p.p. higher share sent ($p = 0.052$). In the model this variation reflects the database: people more familiar with moral reasoning are more likely to employ it in the game, focusing on fairness norms and making larger transfers.

Instability under Market. Does this baseline heterogeneity interact with cues? If so, how? To answer this question we focus on the 300 subjects who played Control first and then played Market. In Market, Moral recall falls from 2.32 to 0.93 descriptions out of four. Moral reasons fall from 70.6% to 53.9%, and the share sent from 37.0% to 26.5%. Critically, the change reaches subjects with strongly Moral recall in Control. Among the 157 who describe three or four allocations as Moral in Control, the average Moral count falls from 3.22 to 1.09 under Market; 43% recall no Moral situation at all.

We next study heterogeneity in within-subject Market effects. We relate each participant's change from Control to Market to her Control index $f_{i\mathrm{M}}$. In the low, middle, and high index groups, the share sent falls by 4.5, 13.2, and 12.2 percentage points, respectively.[17] A regression of the within-person change on $f_{i\mathrm{M}}$ shows that a one-standard-deviation increase in the index predicts a 5.1-point larger fall in the share sent ($p = 0.026$, $N = 278$). The Control index also has little relation to Moral reasoning under Market (+0.8 points, $p = 0.813$). The model explains this pattern through cued recall: a context emphasizing material payoffs is especially effective at moving people who would otherwise bring fairness to mind. Their moral recall in Control does not identify a fixed and unchanging moral type. These participants often recall Moral situations and act generously, but can abruptly change their reasoning and choice when the same game cues a different representation.

[16]The participant effects are penalized, with the penalty chosen by cross-validation. Unclassified descriptions are omitted from the logit.

[17]Many participants have identical fitted index values, so exact terciles would split ties. We instead group participants below, at, and above the modal index, yielding groups of 88, 90, and 100.

5.2 A Known Framing Effect and Its Reversal

We close with two payoff-equivalent dictator games. In DG, our Control condition, Player 1 is given \$12 and may give to Player 2. In LT, Player 1 is given \$8, Player 2 is given \$4, but Player 1 may give or take. The set of feasible allocations is identical, so a stable utility predicts the same behavior in the two games. It is however well known that in LT dictators are reluctant to take (List, 2007), which in our comparison should imply $\bar{t}^{\mathrm{LT}} > \bar{t}^{\mathrm{DG}}$, consistent with Krupka and Weber (2013). This is true in our experiment: LT participants leave 6.7 p.p. more of the budget to Player 2 than DG participants.

Can representations be at work? Intuitively, giving Player 1 the option to take may bring to mind the "unfairness" of taking, increasing her moral concern compared to DG. We apply the similarity-based method of Figure 2 to the DG–LT comparison (Appendix Table 6). In Control, similarity predicts a 1.2 p.p. higher Moral share in LT than DG. The observed Moral share is 6.6 p.p. higher (Figure 10, left panel, Control). The positive Moral gap coincides with the positive transfer gap shown in the right panel.[18] The mental association between a subtle manipulation of the choice set and lived experiences offers a theory for changing social norms in Krupka and Weber (2013).

Being produced by a mental association, this force can be unstable: the possibility of cutting into Player 2's initial entitlement is legitimate in some situations. We thus compare LT and DG again under Market. In DG, the initial price is \$12 and the provider can lower it; in LT it is \$8 and she can raise or lower it. It seems intuitive that many view raising a price to exploit market power as legitimate (and convenient). But what does similarity predict? Strikingly, it predicts a reversal: LT should have fewer Moral reasons (1.7 p.p.) than DG. Measured representations move in that direction: LT has 6.2 p.p. fewer Moral reasons (Figure 10, left panel, Market). The prediction assigns the relative gain to Self-interest, whereas the observed gain is Win-Win (Appendix Table 6); both, however, imply

[18]The DG pre-registration also proposed a finer classification of pre-choice recollections which included "Theft". In Control, allocations in which Player 1 keeps all \$12 prompt recall of "Theft" for 43.5% of LT participants (taking Player 2's \$4), against 5.5% of DG participants (giving nothing).

the same behavior: a lower Moral share and transfer in Market LT than in Market DG.

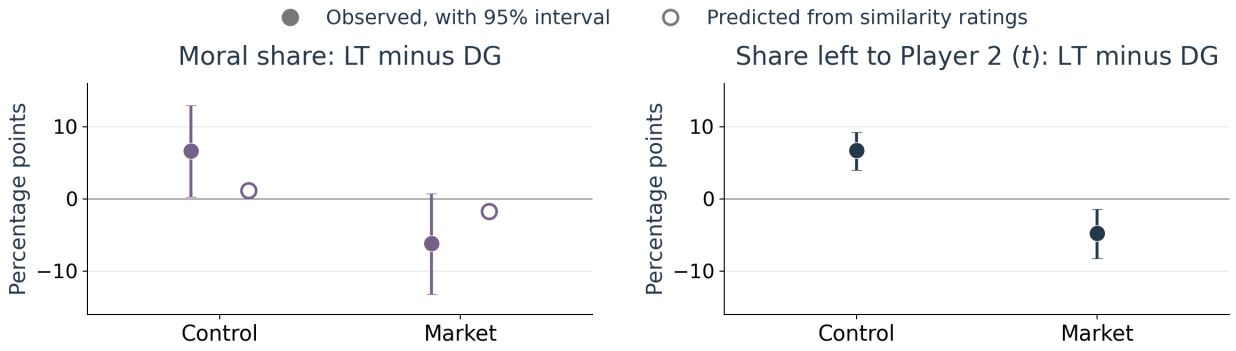


Figure 10: Representations and the DG–LT reversal. Left: LT minus DG in the Moral share, observed (filled) and predicted from similarity ratings (open). Right: observed LT minus DG in the share of the budget left to Player 2. Each panel shows Control and Market; values are percentage points and error bars are 95% intervals. In Market, a lower share left to Player 2 means a higher final price. Classified sample: 399 DG and 389 LT in Control, 329 and 347 in Market.

Consistent with the reversal in Moral reasoning, Market also reverses the transfer gap. In LT, 73% of providers raise the \$8 price, with 25% choosing \$10; in DG, 67% discount from \$12, with 25% charging \$6 or less. LT's mean final price is higher (\$9.57 versus \$9.00, $p = 0.006$), and Player 2 obtains 4.8 points less of the budget than in DG.[19] Raising a price from \$8 to \$10 still allows for a Win-Win situation in which both parties profit, while the initial price of \$12 in DG invites a discount. Representations explain why an option to take prompts morality and hence generosity in one setting, but the reverse in another.

6 Conclusions

In game theory players read the rules as written, so their tastes determine the equilibrium set. We argue instead that, before strategizing, a player recognizes the interaction by recalling similar experiences. These may come from an entirely different domain, but fall into broad categories that recur in social life, and in our subjects' own words: sharing a windfall with

[19]The comparisons in this section use participants with classified reasons. The reversal also holds when unclassified respondents are included: LT leaves Player 2 6.4 p.p. more than DG in Control, but 4.3 p.p. less under Market ($p = 0.006$ for the Market LT–DG gap).

a friend, where the point is fairness; bargaining with a stranger, where the point is one's own payoff; investing with a partner, where the point is the size of the pie. The recalled category shapes which game the player plays and what she believes about the counterpart. What follows is a cognitive theory of context effects with measurable inputs: the category a context evokes (similarity), the memory database (recollections before choosing), and the game played (reasoning about choice). Because each category entails distinct actions, context cues have predictable effects on the shares of reasoning types and on behavior. We conclude by discussing the implications of our approach for the unification of stable and unstable social behavior, for policy, and game theory.

Unifying social preferences and instability. Two bodies of work on social behavior sit uneasily together. One shows instability: Hoffman et al. (1994) found that ultimatum offers fall under a buyer-seller frame, and concluded that fairness is a belief prompted by the setting rather than a taste. The other shows stability: social preferences estimated from choices sort people into a few types that are stable over time and predict behavior out of sample (Fehr, Epper and Senn, 2023), and the field now treats social preferences as preferences like any other (Fehr and Charness, 2025). We offer a way to reconcile these findings. Because most interactions are with family, friends, and colleagues, cooperative experiences weigh heavily at recall even in anonymous games, creating genuinely social behavior. Likewise, different experiences can cause different people to retrieve different categories from the same cue, and behave differently. Herz and Taubinsky (2018) provide experimental evidence that transaction histories shape fairness judgments: buyers exposed to higher prices subsequently regard higher prices as fairer. This may explain why ultimatum offers differ across Jerusalem, Ljubljana, Pittsburgh, and Tokyo (Roth et al., 1991), where community projects are funded collectively a public goods game is recognized as one (Ensminger, 2004); where everyday production relies on cooperation, offers are higher (Henrich et al., 2001); Or why where exchange with strangers is routine, fairness toward strangers rises (Henrich et al., 2010), as the old claim that commerce teaches strangers to see each other as partners would have it

(Hirschman, 1982; Jha, 2013).

Instability, then, does not mean that social behavior is ephemeral, it just reflects another cognitive factor: changing cues. A context that changes the similarity of a game to the categories in memory changes which one is retrieved, at constant rules and constant database. This is what the buyer-seller frame does, and our Market treatment shows that it does not simply weaken social preferences, as the received view holds (Falk and Szech, 2013; Bowles and Polanía-Reyes, 2012): it displaces Moral everywhere, but where the interaction can create value it recruits Win-Win, the intention for mutual benefit of Sugden (2018) that Isoni and Sugden (2019) invoke to explain trust without kindness, and cooperation rises among strangers. The converse holds too: a Moral motive can divert attention from joint returns, as when fairness forbids market-clearing prices after a shortage (Kahneman, Knetsch and Thaler, 1986) or when a transaction is repugnant because a Moral situation is recognized where a Win-Win one is on offer (Roth, 2007).

This helps organize the mixed effects of market framing across settings. Ellingsen et al. (2012) find lower cooperation in a prisoner's dilemma called the Stock Market Game than in the same game called the Community Game, whereas Dreber et al. (2013) find no effect of social framing on dictator giving.[20] Direction depends on the category that replaces Moral and on the game. Rivalry cues Self-interest, which implies the lowest transfer in each game. Mutually beneficial exchange cues Win-Win, which lowers transfers relative to Moral where the pie is fixed and raises them where giving creates surplus.

Strength depends on how much the frame changes similarity. A label may change these associations less than a redescription of roles and actions, potentially explaining why labels leave dictator giving unchanged (Dreber et al., 2013) while our frame does not. A frame that

[20]Other examples of mixed framing effects include lower cooperation under a price-competition description (Engel and Rand, 2014), lower ultimatum offers but unchanged dictator giving under buyer-seller framing (Hoffman et al., 1994), and unchanged one-shot public good contributions under Wall Street and Community labels (Bernold et al., 2023). The Stock Market label also has no effect in sequential or unilateral versions of the prisoner's dilemma (Ellingsen et al., 2012). Related evidence on market interaction finds greater willingness to accept money for the death of a mouse (Falk and Szech, 2013), but no erosion of moral behavior once repetition is held constant (Bartling, Fehr and Özdemir, 2023).

redescribes the counterpart can also move beliefs. This channel affects choices only where the counterpart acts, so a label can change play in a simultaneous dilemma without changing dictator giving. Our model disciplines the direction and strength of these effects, and our experiment tests these restrictions across games.

Outside the laboratory, a cue that raises similarity to a frequent, socially proximate interaction, such as a donation request made in person rather than by mail, increases giving (DellaVigna, List and Malmendier, 2012), while a fine turns a favor into a purchase and reduces it (Gneezy and Rustichini, 2000). Experiences and cues also interact, because a cue retrieves what the database holds. Calling a public goods game a "community" project raised contributions and beliefs among students in St. Gallen, where the word names a lived institution, and lowered them in Bonn, where it carries negative connotations, while the two pools did not differ under a neutral label (Dufwenberg, Gächter and Hennig-Schmidt, 2011). In our terms, a stable type is a database and an unstable behavior is a cue meeting it; the inconsistency of social attitudes across settings noted by Levitt and List (2007) is the second seen without the first.[21]

Policy. If motives and beliefs follow recognition, the cues that institutions carry are instruments. Our Market frame raises surplus in the TG, where it cues joint gain, and destroys about a third of it in the UG, where proposers meet receivers less accepting than they expect. Some familiar successes and failures read the same way. Open-source software and online encyclopedias are produced by strangers who see a joint venture, while paying for blood reduced donations (Mellström and Johannesson, 2008). Many people see trade as zero-sum, a view that travels in families and predicts support for protection (Chinoy et al., 2026). Costly strikes are more likely when each side frames the dispute as self-serving fairness rather than as a deal to be made (Babcock, Wang and Loewenstein, 1996; Babcock and Loewenstein, 1997). The model suggests two remedies, one for each input. Change the

[21] Accounts that put frames in beliefs (Ellingsen et al., 2012; Dreber et al., 2013) or let people infer duties by analogy (Ellingsen and Mohlin, 2026) fit this logic. Retrieval supplies belief and motive together, is predicted by similarity to context and personal experience, and is measured by the reasons people give, not only by the allocations they choose.

description so that it resembles the category wanted: letting blood donors give the payment to charity kept the act a gift and restored supply; describing a trade agreement through the jobs it creates on both sides works the same way. Or change the database: lived experience of fair or mutually beneficial situations, as when rich students acquire poor classmates (Rao, 2019), fosters generosity later. Measured similarity, reasoning, and recall can help design settings that cue the category wanted, and say in advance which category a setting will cue.

Game theory. Recognition may explain several puzzles. The winner's curse may persist because bidders do not recognize adverse selection: they see a contest to win or a value to guess. In the finitely repeated prisoner's dilemma, cooperation starts high, collapses after a defection, and rarely returns (Embrey, Fréchette and Yuksel, 2018); reputation can rationalize the early cooperation, but the abruptness of the collapse suggests a change of representation, from partner to stranger, that a defection cues and later cooperation does not undo. Equilibrium also requires mutually consistent representations, which recognition does not guarantee: our proposers in the Market UG expected a cooperating counterpart who was not there. Consistency requires a shared cue that retrieves a shared category, which is how we read the norms of Kahneman, Knetsch and Thaler (1986), regular patterns of behavior that everyone expects everyone to follow, often arbitrary in content. A focal point is a cue that retrieves a norm (Schelling, 1960). Both can be modeled and measured with our tools, and both capture what a game is to the person playing it.

References

Andreoni, James, and John Miller. 2002. "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." *Econometrica*, 70(2): 737–753.

Babcock, Linda, and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives*, 11(1): 109–126.

Babcock, Linda, Xianghong Wang, and George Loewenstein. 1996. "Choosing the Wrong Pond: Social Comparisons in Negotiations That Reflect a Self-Serving Bias." *Quarterly Journal of Economics*, 111(1): 1–19.

Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory.* Princeton University Press. Edited by Natalie Gold and Robert Sugden.

Ba, Cuimin, J. Aislinn Bohren, and Alex Imas. 2025. "The Role of Representational and Computational Complexity in Belief Formation." *AEA Papers and Proceedings*, 115: 631–636.

Bardsley, Nicholas. 2008. "Dictator Game Giving: Altruism or Artefact?" *Experimental Economics*, 11(2): 122–133.

Bartling, Björn, Ernst Fehr, and Yagiz Özdemir. 2023. "Does Market Interaction Erode Moral Values?" *Review of Economics and Statistics*, 105(1): 226–235.

Bartling, Björn, Roberto A. Weber, and Lan Yao. 2015. "Do Markets Erode Social Responsibility?" *Quarterly Journal of Economics*, 130(1): 219–266.

Bauer, Michal, Christopher Blattman, Julie Chytilová, Joseph Henrich, Edward Miguel, and Tamar Mitts. 2016. "Can War Foster Cooperation?" *Journal of Economic Perspectives*, 30(3): 249–274.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, 10(1): 122–142.

Bernold, Elizabeth, Elisabeth Gsottbauer, Kurt A. Ackermann, and Ryan O. Murphy. 2023. "Accounting for Preferences and Beliefs in Social Framing Effects." *Frontiers in Behavioral Economics*, 2: 1147492.

Bhatia, Sudeep. 2017. "Associative Judgment and Vector Space Semantics." *Psychological Review*, 124(1): 1–20.

Bohnet, Iris, and Bruno S. Frey. 1999. "Social Distance and Other-Regarding Behavior in Dictator Games: Comment." *American Economic Review*, 89(1): 335–339.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166–193.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice Under Risk." *Quarterly Journal of Economics*, 127(3): 1243–1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *Quarterly Journal of Economics*, 135(3): 1399–1442.

Bordalo, Pedro, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer. 2026. "A Cognitive Theory of Reasoning and Choice." *Quarterly Journal of Economics*, 141(3): 1921–1963.

Bowles, Samuel, and Sandra Polanía-Reyes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature*, 50(2): 368–425.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton, NJ:Princeton University Press.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review*, 97(3): 818–827.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117(3): 817–869.

Charness, Gary, and Uri Gneezy. 2008. "What's in a Name? Anonymity and Social Distance in Dictator and Ultimatum Games." *Journal of Economic Behavior and Organization*, 68(1): 29–35.

Chinoy, Sahil, Nathan Nunn, Sandra Sequeira, and Stefanie Stantcheva. 2026. "Zero-Sum Thinking and the Roots of US Political Differences." *American Economic Review*, 116(3): 1052–1096.

Cohn, Alain, Ernst Fehr, and Michel André Maréchal. 2014. "Business Culture and Dishonesty in the Banking Industry." *Nature*, 516(7529): 86–89.

Danz, David, Lise Vesterlund, and Alistair J. Wilson. 2022. "Belief Elicitation and Behavioral Incentive Compatibility." *American Economic Review*, 112(9): 2851–2883.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, 127(1): 1–56.

Dreber, Anna, Tore Ellingsen, Magnus Johannesson, and David G. Rand. 2013. "Do People Care about Social Context? Framing Effects in Dictator Games." *Experimental Economics*, 16(3): 349–371.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2): 268–298.

Dufwenberg, Martin, Simon Gächter, and Heike Hennig-Schmidt. 2011. "The Framing of Games and the Psychology of Play." *Games and Economic Behavior*, 73(2): 459–478.

Ellingsen, Tore, and Erik Mohlin. 2026. "A Model of Social Duties." *Journal of Political Economy*, 134(5): 1307–1346.

Ellingsen, Tore, Magnus Johannesson, Johanna Mollerstrom, and Sara Munkhammar. 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior*, 76(1): 117–130.

Embrey, Matthew, Guillaume R. Fréchette, and Sevgi Yuksel. 2018. "Cooperation in the Finitely Repeated Prisoner's Dilemma." *Quarterly Journal of Economics*, 133(1): 509–551.

Engel, Christoph, and David G. Rand. 2014. "What Does "Clean" Really Mean? The Implicit Framing of Decontextualized Experiments." *Economics Letters*, 122(3): 386–389.

Ensminger, Jean. 2004. "Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* , ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr and Herbert Gintis, 356–381. Oxford:Oxford University Press.

Eyster, Erik, and Matthew Rabin. 2005. "Cursed Equilibrium." *Econometrica*, 73(5): 1623–1672.

Falk, Armin, and Nora Szech. 2013. "Morals and Markets." *Science*, 340(6133): 707–711.

Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2): 293–315.

Fehr, Ernst, and Gary Charness. 2025. "Social Preferences: Fundamental Characteristics and Economic Consequences." *Journal of Economic Literature*, 63(2): 440–514.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, Thomas Epper, and Julien Senn. 2023. "The Fundamental Properties, Stability and Predictive Power of Distributional Preferences." Working Paper No. 439, Department of Economics, University of Zurich.

Fischbacher, Urs, Christina M. Fong, and Ernst Fehr. 2009. "Fairness, Errors and the Power of Competition." *Journal of Economic Behavior & Organization*, 72(1): 527–545.

Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton. 1994. "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior*, 6(3): 347–369.

Gilboa, Itzhak, and David Schmeidler. 1995. "Case-Based Decision Theory." *Quarterly Journal of Economics*, 110(3): 605–639.

Gneezy, Uri, and Aldo Rustichini. 2000. "A Fine Is a Price." *Journal of Legal Studies*, 29(1): 1–17.

Graeber, Thomas, Christopher Roth, and Florian Zimmermann. 2024. "Stories, Statistics, and Memory." *Quarterly Journal of Economics*, 139(4): 2181–2225.

Grosskopf, Brit. 2003. "Reinforcement and Directional Learning in the Ultimatum Game with Responder Competition." *Experimental Economics*, 6(2): 141–158.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization*, 3(4): 367–388.

Haaland, Ingar K., Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2025. "Understanding Economic Behavior Using Open-Ended Survey Data." *Journal of Economic Literature*, 63(4): 1244–1280.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2017. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *Quarterly Journal of Economics*, 132(1): 1–54.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer,

and John Ziker. 2010. "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science*, 327(5972): 1480–1484.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies." *American Economic Review*, 91(2): 73–78.

Herz, Holger, and Dmitry Taubinsky. 2018. "What Makes a Price Fair? An Experimental Study of Transaction Experience and Endogenous Fairness Views." *Journal of the European Economic Association*, 16(2): 316–352.

Hirschman, Albert O. 1982. "Rival Interpretations of Market Society: Civilizing, Destructive, or Feeble?" *Journal of Economic Literature*, 20(4): 1463–1484.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review*, 86(3): 653–660.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 7(3): 346–380.

Howard, Marc W., and Michael J. Kahana. 2002. "When Does Semantic Similarity Help Episodic Retrieval?" *Journal of Memory and Language*, 46(1): 85–98.

Isoni, Andrea, and Robert Sugden. 2019. "Reciprocity and the Paradox of Trust in Psychological Game Theory." *Journal of Economic Behavior & Organization*, 167: 219–227.

Jehiel, Philippe. 2005. "Analogy-Based Expectation Equilibrium." *Journal of Economic Theory*, 123(2): 81–104.

Jha, Saumitra. 2013. "Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia." *American Political Science Review*, 107(4): 806–832.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market." *American Economic Review*, 76(4): 728–741.

Kimbrough, Erik O., and Alexander Vostroknutov. 2016. "Norms Make Preferences Social." *Journal of the European Economic Association*, 14(3): 608–638.

Krupka, Erin L., and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3): 495–524.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21(2): 153–174.

Liberman, Varda, Steven M. Samuels, and Lee Ross. 2004. "The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." *Personality and Social Psychology Bulletin*, 30(9): 1175–1185.

List, John A. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy*, 115(3): 482–493.

Marjieh, Raja, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L. Griffiths. 2024. "Large Language Models Predict Human Sensory Judgments across Six Modalities." *Scientific Reports*, 14(1): 21445.

Mellström, Carl, and Magnus Johannesson. 2008. "Crowding Out in Blood Donation: Was Titmuss Right?" *Journal of the European Economic Association*, 6(4): 845–863.

Oprea, Ryan, and Michael Woodford. 2026. "Behavioral Signatures of Cognitive Noise." NBER Working Paper 35714.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5): 1281–1302.

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review*, 109(3): 774–809.

Roth, Alvin E. 2007. "Repugnance as a Constraint on Markets." *Journal of Economic Perspectives*, 21(3): 37–58.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, 81(5): 1068–1095.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* Cambridge, MA:Harvard University Press.

Small, Deborah A., and George Loewenstein. 2003. "Helping a Victim or Helping the Victim: Altruism and Identifiability." *Journal of Risk and Uncertainty*, 26(1): 5–16.

Stahl, Dale O., and Paul W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior*, 10(1): 218–254.

Sugden, Robert. 2003. "The Logic of Team Reasoning." *Philosophical Explorations*, 6(3): 165–181.

Sugden, Robert. 2018. *The Community of Advantage: A Behavioural Economist's Defence of the Market.* Oxford:Oxford University Press.

A Model Details

A.1 Proofs of Propositions

Throughout, attention weights are strictly positive, in the ultimatum game $0 < a \leq b$, and in the trust game $(1+r)x^* < 1$ and $r \geq 1$. In every game the objective in (1) is strictly concave in t: $\partial^2 u_1/\partial t^2$ equals $-\alpha_{t^*}\lambda$ in the dictator and trust games, where Π and Π_2 are linear in t, and $-2b\,\alpha_{\Pi_2} - \alpha_{t^*}\lambda$ in the ultimatum game, where $\Pi_2 = (a+bt)\,t$. The maximizer on $[0,1]$ is therefore unique and equals the unconstrained solution censored to $[0,1]$; comparative statics are stated at interior optima, where the censoring does not bind.

Proof of Proposition 1. Write $(\alpha_\Pi, \alpha_{\Pi_2}, \alpha_{t^*})$ for the weights of a category. The first-order condition of (1) is $\alpha_\Pi\,\partial\Pi/\partial t - \alpha_{\Pi_2}\,\partial\Pi_2/\partial t - \alpha_{t^*}\lambda\,(t - t^*) = 0$, which is (2). In each game we solve it, insert the weights $\boldsymbol{\alpha}_M = (1/3, 1/3, 1/3)$, $\boldsymbol{\alpha}_S = (m/2, m/2, 1-m)$, and $\boldsymbol{\alpha}_W = (m-1/3, 1/3, 1-m)$, and compare the three transfers. Every comparison turns on $3m-2 > 0$.

Dictator game. With $\partial\Pi/\partial t = 0$ and $\partial\Pi_2/\partial t = 1$, (2) gives $t_{\mathrm{C}} = \frac{1}{2} - \alpha_{\Pi_2}/(\lambda\alpha_{t^*})$, that is,

$$t_M = \frac{1}{2} - \frac{1}{\lambda}, \qquad t_W = \frac{1}{2} - \frac{1}{3(1-m)\lambda}, \qquad t_S = \frac{1}{2} - \frac{m}{2(1-m)\lambda}.$$

Since $1 < 1/[3(1-m)] < m/[2(1-m)]$ for $m \in (2/3, 1)$, $t_M > t_W > t_S$. The transfer is interior for $\alpha_{\Pi_2}/\alpha_{t^*} < \lambda/2$ and zero otherwise.

Ultimatum game. With $\partial\Pi/\partial t = b$ and $\partial\Pi_2/\partial t = a + 2bt$, (2) is linear in t_{C} and solves to

$$t_{\mathrm{C}} = \frac{\lambda\alpha_{t^*}/2 + b\,\alpha_\Pi - a\,\alpha_{\Pi_2}}{\lambda\alpha_{t^*} + 2b\,\alpha_{\Pi_2}}, \tag{10}$$

that is,

$$t_M = \frac{\lambda/2 + b - a}{\lambda + 2b}, \qquad t_W = \frac{3(1-m)\lambda/2 + (3m-1)b - a}{3(1-m)\lambda + 2b}, \qquad t_S = \frac{(1-m)\lambda + m(b-a)}{2(1-m)\lambda + 2mb}.$$

Cross-multiplying, $(t_W - t_M)\,[3(1-m)\lambda + 2b]\,(\lambda + 2b) = (3m-2)\,[\lambda(b-a) + 2b^2]$, which is positive because $a \leq b$, and $(t_M - t_S)\,(\lambda + 2b)\,[2(1-m)\lambda + 2mb] = (3m-2)\,a\lambda$, which is positive because $a > 0$. Hence $t_W > t_M > t_S$.

Trust game. With $\partial\Pi/\partial t = r$ and $\partial\Pi_2/\partial t = (1-x^*)(1+r)$, (2) gives

$$t_{\mathrm{C}}(x^*) = t^*(x^*) + \frac{r\,\alpha_{\Pi} - (1-x^*)(1+r)\,\alpha_{\Pi_2}}{\lambda\alpha_{t^*}}, \tag{11}$$

so that, at the common belief x^*, $t_W - t_M = (3m-2)\,[r-1+(1+r)x^*]/[3(1-m)\lambda]$ and $t_M - t_S = (3m-2)\,[1-(1+r)x^*]/[2(1-m)\lambda]$. The first difference is positive if and only if $x^* > (1-r)/(1+r)$, which holds for every $x^* \geq 0$ because $r \geq 1$; the second is positive because $(1+r)x^* < 1$. □

Beliefs and transfers. The paragraph that follows Proposition 1 states how transfers react to beliefs. In the ultimatum game, differentiating (10) gives

$$\frac{\partial t_{\mathrm{C}}}{\partial a} = -\frac{\alpha_{\Pi_2}}{\lambda\alpha_{t^*} + 2b\,\alpha_{\Pi_2}} < 0, \qquad \frac{\partial t_{\mathrm{C}}}{\partial b} = \frac{\lambda\alpha_{t^*}(\alpha_{\Pi} - \alpha_{\Pi_2}) + 2a\,\alpha_{\Pi_2}^2}{(\lambda\alpha_{t^*} + 2b\,\alpha_{\Pi_2})^2} > 0, \tag{12}$$

the second because $\alpha_{\Pi} \geq \alpha_{\Pi_2}$ in every category: the offer falls in the believed acceptance level a and rises in the believed slope b. In the trust game, the norm $t^*(x) = \big[\max\{1,\, 1 + (1-2x)(1+r)\}\big]^{-1}$ is continuous and non-decreasing in x, with $\mathrm{d}t^*/\mathrm{d}x = 2(1+r)\big[1+(1-2x)(1+r)\big]^{-2} > 0$ for $x < \frac{1}{2}$ and $\mathrm{d}t^*/\mathrm{d}x = 0$ for $x > \frac{1}{2}$, so by (11)

$$\frac{\partial t_{\mathrm{C}}}{\partial x^*} = \frac{\mathrm{d}t^*}{\mathrm{d}x} + \frac{(1+r)\,\alpha_{\Pi_2}}{\lambda\alpha_{t^*}} > 0, \tag{13}$$

and the send rises in the believed return. Finally, write x_{C} for a belief that differs across categories. Since each t_{C} is increasing in its belief, $x_W \geq x_M$ gives $t_W(x_W) \geq t_W(x_M) > t_M(x_M)$, and $x_S \leq x_M$ with $(1+r)x_S < 1$ gives $t_S(x_S) < t_M(x_S) \leq t_M(x_M)$: the trust-game ranking of Proposition 1 is preserved.

Proof of Proposition 2. For a given category C, the maximand $f_{\mathrm{C}}\big[1 - \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathrm{C}}\|^2 - d_{\mathrm{C}}\big] + \epsilon_{\mathrm{C}}$ is maximized over $\boldsymbol{\alpha}$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{C}}$, with value $V_{\mathrm{C}} + \epsilon_{\mathrm{C}}$, where $V_{\mathrm{C}} \equiv f_{\mathrm{C}}(1 - d_{\mathrm{C}})$. Player 1 picks the category with the largest $V_{\mathrm{C}} + \epsilon_{\mathrm{C}}$, and i.i.d. type-I extreme-value shocks give the logit probabilities in (5). Writing $\Pr(\mathrm{C}) = e^{V_{\mathrm{C}}}/\sum_{\mathrm{C}'} e^{V_{\mathrm{C}'}}$, $\partial\Pr(\mathrm{C})/\partial V_{\mathrm{C}} = \Pr(\mathrm{C})\big(1 - \Pr(\mathrm{C})\big) > 0$ and $\partial\Pr(\mathrm{C}')/\partial V_{\mathrm{C}} = -\Pr(\mathrm{C})\Pr(\mathrm{C}') < 0$ for $\mathrm{C}' \neq \mathrm{C}$. For item (i), $\partial V_{\mathrm{C}}/\partial f_{\mathrm{C}} = 1 - d_{\mathrm{C}} > 0$ because $d_{\mathrm{C}} < 1$; adding an experience of weight w to C raises f_{C} by w and leaves every $V_{\mathrm{C}'}$

unchanged, so the log odds $\ln[\Pr(\mathrm{C})/\Pr(\mathrm{C}')] = V_{\mathrm{C}} - V_{\mathrm{C}'}$ rise by $w(1-d_{\mathrm{C}})$. For item (ii), $\partial V_{\mathrm{C}}/\partial(-d_{\mathrm{C}}) = f_{\mathrm{C}} > 0$, so the derivatives of the probabilities with respect to $-d_{\mathrm{C}}$ are those with respect to V_{C} multiplied by f_{C}. □

Welfare and Inequality

Let $\Pi^{\mathrm{G}}(\mathrm{C})$ denote expected total income in game G when Player 1 uses category C, evaluated at her belief: $\Pi^{\mathrm{DG}} = 1$, $\Pi^{\mathrm{UG}}(\mathrm{C}) = a + b\,t_{\mathrm{C}}$, and $\Pi^{\mathrm{TG}}(\mathrm{C}) = 1 + r\,t_{\mathrm{C}}$. Let $D^{\mathrm{G}}(\mathrm{C}) \equiv \Pi_1 - \Pi_2$ denote the expected payoff gap, in the ultimatum game conditional on acceptance and in the trust game at an actual returned share x_a.

Proposition 3 (Welfare and Inequality)**.** *At interior optima:*

(i) ***Total income.*** *In* DG *it equals one under every category. In* UG *and* TG *it increases in t_{C}, so by Proposition 1 it is lowest under Self-interest and highest under Win-Win.*

(ii) ***Inequality.*** *In* DG *and* UG*, $D^{\mathrm{G}}(\mathrm{C}) = 1 - 2t_{\mathrm{C}}$: the gap is largest under Self-interest, and in* UG *smallest under Win-Win. In* TG*, $D^{\mathrm{TG}}(\mathrm{C}) = 1 - t_{\mathrm{C}}\big[1 + (1+r)(1-2x_a)\big]$ decreases in the send if and only if $x_a < (2+r)/[2(1+r)]$, and the same ranking then holds.*

(iii) ***Beliefs.*** *In* UG*, a higher believed acceptance level a raises total income at the sender's belief but lowers her offer, so for any actual acceptance schedule increasing in t realized total income falls with optimism. In* TG*, a higher believed return raises the send and total income.*

Proof of Proposition 3. (i) $\Pi^{\mathrm{DG}} = 1$ does not depend on C; $\partial\Pi^{\mathrm{UG}}/\partial t_{\mathrm{C}} = b > 0$ and $\partial\Pi^{\mathrm{TG}}/\partial t_{\mathrm{C}} = r > 0$, and Proposition 1 orders t_{C}. (ii) In the dictator game, and in the ultimatum game conditional on acceptance, $D = (1-t_{\mathrm{C}}) - t_{\mathrm{C}}$, decreasing in t_{C}. In the trust game $D^{\mathrm{TG}} = (1-t_{\mathrm{C}}) + x_a(1+r)t_{\mathrm{C}} - (1-x_a)(1+r)t_{\mathrm{C}} = 1 - t_{\mathrm{C}}\big[1+(1+r)(1-2x_a)\big]$, affine in

t_C with slope $-[1+(1+r)(1-2x_a)]$, which is negative if and only if $x_a < (2+r)/[2(1+r)]$. (iii) By (12), $\partial\Pi^{UG}/\partial a = 1 + b\,\partial t_C/\partial a = (\lambda\alpha_{t^*} + b\,\alpha_{\Pi_2})/(\lambda\alpha_{t^*} + 2b\,\alpha_{\Pi_2}) > 0$, while for an actual schedule $\bar{x}(t)$ with $\bar{x}' > 0$, $\partial\bar{x}(t_C)/\partial a = \bar{x}'(t_C)\,\partial t_C/\partial a < 0$. In the trust game $\partial\Pi^{TG}/\partial x^* = r\,\partial t_C/\partial x^* > 0$ by (13). □

A.2 Player 2's Behavior and Player 1's Beliefs

The main text treats Player 2 only through Player 1's belief about her: an acceptance schedule $x^*(t) = a + bt$ in the ultimatum game and a returned share x^* in the trust game. This appendix models Player 2 herself, as the mirror image of Player 1. Facing a transfer t, she chooses a response x to maximize a utility of the form of Equation (1) with the roles reversed,

$$u_2(x;t) = \alpha_\Pi \cdot \Pi(t,x) - \alpha_{\Pi_1} \cdot \Pi_1(t,x) - \alpha_x \cdot \frac{\lambda}{2} \times \begin{cases} \left(t - \frac{1}{2}\right)^2 x & \text{in UG,} \\ \left(x - x^e(t)\right)^2 & \text{in TG,} \end{cases} \tag{14}$$

where $\boldsymbol{\alpha}_2 = (\alpha_\Pi, \alpha_{\Pi_1}, \alpha_x)$ are her attention weights on the pie, on Player 1's slice, and on the norm (in this appendix, α_Π is Player 2's weight on the pie). The norm is Player 1's, equal payoffs, with the same λ. In UG, $x = 1$ is accepting and $x = 0$ rejecting: rejecting equalizes payoffs at zero and carries no loss, while accepting carries Player 1's own norm term, evaluated at the offer that Player 2 lets stand. In TG, x is the share of the output $(1+r)t$ that she returns, and $x^e(t)$ is the share that equalizes payoffs, derived below. Her categories are also Player 1's, with Player 1's slice in place of Player 2's and the same m: Moral, with weights (1/3, 1/3, 1/3); Self-interest, with $(m/2,\ m/2,\ 1-m)$; and Win-Win, with $(m - 1/3,\ 1/3,\ 1-m)$. We derive Player 2's response in each game and rank it across categories, then let her categorize the game from a description that includes t. Finally, we relate Player 1's belief to this behavior: the belief is the response of the Player 2s she imagines, whose weights $\hat{\boldsymbol{\alpha}}_2 = (\hat{\alpha}_\Pi, \hat{\alpha}_{\Pi_1}, \hat{\alpha}_x)$ come from Player 1's own memory and need not equal $\boldsymbol{\alpha}_2$.

Player 2's Response. In UG, accepting an offer t gives Player 2 a pie of one and leaves Player 1 a slice $1-t$, while rejecting gives zero to both. By (14), she accepts if and only if

$$\alpha_\Pi - \alpha_{\Pi_1}(1-t) \;\geq\; \alpha_x \frac{\lambda}{2}\Big(t-\frac{1}{2}\Big)^2. \tag{15}$$

Every category has $\alpha_\Pi \geq \alpha_{\Pi_1}$. For offers up to the equal split, the left side rises in t and the right side falls, and the condition holds at $t=\frac{1}{2}$, so Player 2 accepts if and only if the offer is at least her *minimum acceptable offer*,

$$\underline{t} = \max\left\{0,\; \frac{1}{2} + \frac{\alpha_{\Pi_1}}{\lambda\alpha_x} - \sqrt{\Big(\frac{\alpha_{\Pi_1}}{\lambda\alpha_x}\Big)^2 + \frac{2\alpha_\Pi - \alpha_{\Pi_1}}{\lambda\alpha_x}}\right\} < \frac{1}{2}, \tag{16}$$

which is zero if and only if $\alpha_\Pi - \alpha_{\Pi_1} \geq \lambda\alpha_x/8$.[22] The minimum acceptable offer rises in the norm weight α_x and falls in the pie weight α_Π.

In TG, Player 2 holds the output $(1+r)t$ and is a dictator over it: the pie $1+rt$ is fixed whatever she returns, so α_Π is idle and she trades her own slice against the norm. Since she moves last, equal payoffs prescribe the return that equalizes realized payoffs, the share

$$x^e(t) = \frac{2+r}{2(1+r)} - \frac{1}{2(1+r)\,t}$$

of the output. This is the inverse of Player 1's equal-payoff send, $t^*(x^e(t)) = t$. It rises in the send, from zero at $t = 1/(2+r)$ to one half at $t=1$; below $t=1/(2+r)$ no return equalizes payoffs and the norm asks for none. Maximizing (14), where $\Pi_1 = 1-t+x(1+r)t$, gives

$$x(t) = \max\Big\{0,\; x^e(t) - \frac{\alpha_{\Pi_1}}{\lambda\alpha_x}\,(1+r)\,t\Big\}. \tag{17}$$

We call α_{Π_1}/α_x Player 2's *selfishness ratio*, the counterpart of Player 1's $\alpha_{\Pi_2}/\alpha_{t^*}$ in (2). When it is low, the returned share is close to $x^e(t)$ and rises in the send, at every t if $\alpha_{\Pi_1}/(\lambda\alpha_x) < 1/[2(1+r)^2]$; as it grows, the returned share falls to zero at every send.

Proposition 4 (Optimal Response by Category)**.** *Categories shape Player 2's response. In* UG*, the minimum acceptable offers satisfy* $\underline{t}_M > \underline{t}_S > \underline{t}_W \geq 0$*. In* TG*, at every send* t*, the*

[22] The loss is symmetric around the equal split, so offers above one half carry it too. Player 2 accepts all of them if $\lambda\alpha_x \leq 8\alpha_\Pi$; otherwise she rejects offers close to one, which leave Player 1 too little.

returned shares satisfy $x_M(t) \geq x_W(t) \geq x_S(t)$, and each inequality is strict when its larger side is positive.

Proof of Proposition 4. Ultimatum game. Dividing (15) by α_x, Player 2 accepts if and only if $G_{\mathrm{C}}(t) \geq \frac{\lambda}{2}\left(t-\frac{1}{2}\right)^2$, where $G_{\mathrm{C}}(t) \equiv [\alpha_\Pi - \alpha_{\Pi_1}(1-t)]/\alpha_x$ is her material gain from accepting per unit of norm weight under category C: $G_M(t) = t$, $G_S(t) = mt/[2(1-m)]$, and $G_W(t) = (m - 2/3 + t/3)/(1-m)$. For $m \in (2/3, 1)$ and $t \in (0, 1]$,

$$G_S(t) - G_M(t) = \frac{(3m-2)\,t}{2(1-m)} > 0, \qquad G_W(t) - G_S(t) = \frac{(3m-2)(2-t)}{6(1-m)} > 0.$$

Since $G_M(0) = G_S(0) = 0 < \lambda/8$, Moral and Self-interest reject a zero offer, so $\underline{t}_M > 0$ and $\underline{t}_S > 0$. At $t = \underline{t}_M$ the condition holds with equality under Moral and strictly under Self-interest; both sides are continuous in t, so Self-interest also accepts offers just below $\underline{t}_M$ and $\underline{t}_S < \underline{t}_M$. The same argument at $t = \underline{t}_S$ gives $\underline{t}_W < \underline{t}_S$. *Trust game.* By (17), the returned share is the larger of zero and an expression that falls in the selfishness ratio α_{Π_1}/α_x. This ratio is 1 under Moral, $1/[3(1-m)]$ under Win-Win, and $m/[2(1-m)]$ under Self-interest, and $1 < 1/[3(1-m)] < m/[2(1-m)]$ for $m \in (2/3, 1)$, as in the proof of Proposition 1. Hence $x_M(t) \geq x_W(t) \geq x_S(t)$, strictly when the larger side is positive. □

Intuitively, in UG, Win-Win accepts the most offers because rejecting destroys the pie she attends to, and Moral the fewest because she attends most to the norm that an unequal offer violates. In TG Player 2 is a dictator over the output, and her ranking is that of Player 1 in DG (Proposition 1).

Relation to Social Preference Models. As for Player 1, the utility function of our Player 2 has an analogue in standard social preference models, and what recognition changes is the source of the weights. In Fehr and Schmidt (1999), the responder accepts a share $t < \frac{1}{2}$ if and only if $t - \alpha(1-2s) \geq 0$, that is, if it is at least $\alpha/(1+2\alpha)$, where α is her aversion to disadvantageous inequality. In Bolton and Ockenfels (2000), she accepts if and only if the offer is at least a threshold at which she is indifferent between it and equal payoffs of zero.

In both, the proposer faces a distribution of thresholds, which is also how an acceptance probability arises here. With $\alpha_\Pi = \alpha_{\Pi_1}$, condition (15) is the responder's rule under the motivation function of Bolton and Ockenfels (2000), with α_{Π_1} and $\lambda\alpha_x$ as their weights on own payoff and on equality. In the trust game Player 2 is a dictator over the output, and her return in (17) is the equalizing return less a term in her selfishness ratio, as the dictator's share under the motivation function of Bolton and Ockenfels (2000) is one half plus the ratio of the two weights.

An Alternative Norm. A norm of returning half of a target amount X gives the opposite slope in the trust game. Returning y, Player 2 then maximizes $\alpha_\Pi(1+rt) - \alpha_{\Pi_1}(1-t+y) - \alpha_x\frac{\lambda}{2}\big(y/X - \frac{1}{2}\big)^2$, whose solution is $y/X = \frac{1}{2} - \alpha_{\Pi_1}X/(\lambda\alpha_x)$, censored at zero. With the amount sent as target, $X = t$, the returned share of the output is $x(t) = \big[\frac{1}{2} - \alpha_{\Pi_1}t/(\lambda\alpha_x)\big]/(1+r)$; with the output as target, $X = (1+r)t$, it is $x(t) = \frac{1}{2} - \alpha_{\Pi_1}(1+r)t/(\lambda\alpha_x)$. Under both, the returned share is affine and *decreasing* in the send.[23] The two norms are therefore separable in the data. Returned shares rise with the send among Moral Good and Win-Win receivers (Appendix C.1), which the equal-payoffs norm predicts and the alternative norm cannot produce

Player 2's Categorization. As in Section 2.2, Player 2 recognizes the game by categorizing its description. Her description differs from Player 1's in one way: it contains Player 1's action. She faces (G_δ, t), a game in which she has just been offered t or entrusted $(1+r)t$. The probability that she recognizes the situation as an instance of category C_2 is $q_{\mathrm{C}_2}(t)$, a logit as in (5) with Player 1's action as a contextual cue. The category supplies her weights $\boldsymbol{\alpha}_2$, and her response is that of (16) and (17) at those weights. The classification of Section 3 records four categories, because a Player 2 who attends to the norm reaches one of two verdicts, depending on the action she faces. When equal payoffs call for rejecting the offer

[23] The total-wealth target, $X = 1 + rt$, is ill behaved: at small sends the norm asks for more than the output itself, forcing full return below a threshold and a kinked schedule. The amount-sent and output targets are the tractable specifications.

or for returning nothing, the reason she gives condemns Player 1, and it is recorded as *Moral Bad*; when they call for accepting or for an equalizing return, the reason she gives is fairness, recorded as *Moral Good*. Self-interest and Win-Win reasons are recorded as for Player 1.

Player 2's average response is the mixture

$$\bar{x}(t) = \sum_{\mathrm{C}_2} q_{\mathrm{C}_2}(t)\, x_{\mathrm{C}_2}(t), \tag{18}$$

where $x_{\mathrm{C}_2}(t)$ is the average response of the Player 2s in category C_2: in the ultimatum game, the share whose minimum acceptable offer is at most t; in the trust game, the returned share. Player 1's action works through two channels,

$$\bar{x}'(t) \;=\; \underbrace{\sum_{\mathrm{C}_2} q_{\mathrm{C}_2}(t)\, x'_{\mathrm{C}_2}(t)}_{\text{margin within categories}} \;+\; \underbrace{\sum_{\mathrm{C}_2} q'_{\mathrm{C}_2}(t)\, x_{\mathrm{C}_2}(t)}_{\text{composition across categories}} \quad .$$

The second channel has two sources. The action is a cue, so it can change the category Player 2 retrieves. And, when Moral Good and Moral Bad are counted as separate categories, as the classification does, it includes the change of the norm's verdict: a low offer is condemned and a fair one approved, so the share of Moral Bad falls with t and that of Moral Good rises. The composition term is then positive, and the aggregate schedule is steeper than the within-category margins alone; it can be steeper than *every* category schedule. Two consequences follow. First, the mixture is measurable: Player 2's reasons should covary with the action faced, condemnation concentrated at low offers and cooperation at generous sends, which her stated reasons check directly (Appendix C.1). Second, treatments reach Player 2 through the same two levers as Player 1, memory and description, but nothing forces the two sides to move equally, which is the asymmetry between the treatment effects on Player 1 and on Player 2 documented in Section 4.2 and Appendix C.1.

Player 1's Beliefs and Forecast Errors. Player 1's belief is the response of the Player 2s she imagines. In the ultimatum game, the believed acceptance $x^*(t)$ is the share of them whose minimum acceptable offer, given by (16) at the weights $\hat{\boldsymbol{\alpha}}_2$ that Player 1 attributes to

them, is at most t. The linear schedule of the main text, $x^*(t) = a + bt$ wherever it is below one, is the case in which a share a of them accepts any offer and the minimum acceptable offers of the others are uniformly spread, with density b. The believed schedule is therefore a representation of Player 2's motives. Imagining Player 2s who attend more to the pie, or more to their own slice relative to the norm, lowers their minimum acceptable offers and raises believed acceptance at every offer; imagining Player 2s who attend more to the norm does the opposite.

In the trust game, the belief of the main text is a constant share x^*, and no weights make the response in (17) a constant positive share: where it is positive, its slope in the send is $1/[2(1+r)t^2] - \alpha_{\Pi_1}(1+r)/(\lambda\alpha_x)$, which vanishes at one send at most. A constant share is what the alternative norm delivers as the selfishness ratio goes to zero. We therefore keep x^* as a primitive in the trust game. The data support the constant: the two beliefs elicited from each sender imply a believed slope with a median of zero in every category, while the shares that Moral Good and Win-Win receivers actually return rise with the send (Appendix C.1). Player 1 is thus mistaken about the shape of Player 2's response, not only about her weights. The constant belief also matters for Player 1's own norm: since $t^*(x^e(t)) = t$, a Player 1 who expected an equalizing return would find every send above $1/(2+r)$ consistent with her norm, which would no longer pin down the transfer.

Forecast errors follow. Player 1's belief is retrieved from Player 1's memory given her description of the game; Player 2's behavior is retrieved from Player 2's memory given a description that includes the action. Nothing aligns the two retrievals, so systematic, context-dependent forecast errors are a structural feature of the model, and a treatment that moves the two descriptions differently moves *mean* forecast errors, the treatment-level result of Section 4.2. Learning under a stable context is then convergence of the Player 2 that Player 1 imagines toward the Player 2 she faces.

B Measurement and Validation

B.1 Classification Rules, Examples, and Validation

Representation categories. The classifier assigns each stated reason to one mutually exclusive category using the reason explicitly given, rather than inferring a motive from the chosen amount. For Player 1, Moral covers fairness, equality, generosity, and doing what is right; Win-Win covers joint gains, productive cooperation, and returns from the interaction; Self-interest covers protecting or maximizing one's own payoff, including mistrust or fear of receiving nothing. A vague answer or a restatement of the choice without a reason is No clear justification. When several motives appear, the classifier follows the reason emphasized by the respondent. Its tie-breakers distinguish a fair "both benefit" outcome from a larger joint surplus, and a risky investment with joint upside from self-protection. For Player 2, the moral category is divided into Moral Good (fairness or positive reciprocity) and Moral Bad (unfairness or punishment); the other categories follow the same distinction between joint gains, own payoff, and unclear reasoning. The complete Player 1 and Player 2 prompts, including examples and tie-breakers, are available online.

Social-proximity coding. A separate prompt codes the counterpart evoked by a response from 0 to 4: no recipient mentioned; an abstract stranger; an anonymous peer such as "we both"; a teammate or coworker; and a friend or family member. Mentions of a shared task take precedence over generic peer language, and an explicit personal bond takes precedence over both. The analysis groups levels 2–4 as high proximity and levels 0–1 as low proximity. The full prompt is available online.

Following the pre-registered procedure, the classification of the reasons question—produced by GPT-4.1 from a prompt containing the written classification rules—was validated against human coders. For each of the three classification batches—Player 1 in the story conditions, Player 1 in Market and Control, and Player 2 in all four conditions—two disjoint valida-

tion subsamples were drawn, stratified on the LLM-assigned category, and coded by hand by two research assistants, each responsible for one subsample throughout and blind to the condition and to the game, using the same classification rules given to the LLM; in total, human coders classified 3,990 responses, and no response was coded in more than one subsample. Table 1 reports agreement rates and Cohen's κ by batch, subsample, and condition. Agreement is high for Player 1's four-category scheme (three substantive categories plus the residual; 79–80% pooled across subsamples, with κ between 0.63 and 0.69) and lower for Player 2's finer five-category scheme (71%, $\kappa = 0.57$), where disagreements concentrate at the *Moral Good*/*Win-Win* boundary. The LLM labels in the validation subsamples coincide with the final classification used in the analysis files.

Forced classification of residual reasons. For the Market–Control check in Section 4.2, 49 of 2,003 Control answers and 260 of 2,000 Market answers by Player 1 have no clear justification under the original classification. To fold these 309 answers into the three substantive categories, we first retain 95 reference answers for which an earlier forced LLM label agrees with at least one of two research assistants' forced labels. We give these examples and the category definitions to Claude Opus 4.8, which classifies the remaining 214 answers without a residual option. Folding all 309 assignments into the sample changes no Market–Control category effect's sign across DG, LT, UG, and TG; the largest absolute change is 4.97 percentage points, for Self-interest in DG.

B.2 Pre-Choice Descriptions of Hypothetical Allocations

We classify the pre-choice descriptions with Claude Opus 4.8, using the same Moral, Self-interest, and Win-Win distinctions as for stated reasons but asking about the situation recalled rather than the reason for a choice. The same classification is used for the between- and within-subject descriptions. Section 5.1 uses all four descriptions from each participant's first DG choice in Control: 200 between-subject participants, 300 who played Control first

Table 1: **Validation of the LLM Classification Against Human Coders**

	Subsample 1			Subsample 2		
	N	Agreement	κ	N	Agreement	κ
Player 1, story conditions (Aid/Bonus)						
Bonus	300	0.84	0.67	300	0.82	0.62
Aid	300	0.81	0.65	300	0.76	0.56
All conditions	600	0.82	0.66	600	0.79	0.59
Player 1, Market/Control						
Control	300	0.91	0.83	300	0.76	0.58
Market	300	0.79	0.71	298	0.70	0.59
All conditions	600	0.85	0.78	598	0.73	0.61
Player 2, all conditions						
Control	200	0.70	0.54	200	0.73	0.55
Market	199	0.66	0.56	200	0.71	0.61
Bonus	197	0.68	0.48	199	0.67	0.42
Aid	197	0.72	0.58	200	0.81	0.69
All conditions	793	0.69	0.55	799	0.73	0.59
Overall	$N = 3,990$, agreement $= 0.76$, $\kappa = 0.64$					

Notes: Human coders classified free-text answers to the reasons question using the same classification rules given to the LLM (GPT-4.1). Each batch comprises two disjoint validation subsamples drawn stratified on the LLM-assigned category (for Player 1, 100 responses per game–condition cell across DG, UG, and TG; for Player 2, roughly 100 per game–condition cell across UG and TG in all four conditions); each subsample was coded by one of two research assistants, blind to the condition and to the game. Agreement is the share of responses assigned to the same category; κ is Cohen's kappa over the full category set (four categories for Player 1, five for Player 2), including the residual "No clear justification".

in the DG Control–Market study, and 300 who played DG first in the DG–LT study. The within-person follow-up uses the 300 Control-first participants' later Market choices and descriptions. Of the 800 Control participants, 755 have at least one substantively classified description and enter the index analysis.

A regularized binary logit estimates a participant-specific Moral-versus-Non-Moral index $f_{i\mathrm{M}}$, net of common differences across the hypothetical allocations. Non-Moral combines Self-interest and Win-Win and is the reference category; descriptions without a substantive category are excluded. Only participant effects are penalized, and the penalty is chosen by four-fold cross-validation. The index is estimated once from first-choice Control descriptions

Table 2: **Representative Open-Ended Answers, by Category (Control Condition)**

Category	Game	Response
Panel A: Player 1		
Moral	DG	"I am not a greedy person. it just seems fair to split this unexpected money evenly"
Moral	UG	"Fair is fair and I could only hope they did the same, type of thing."
Moral	TG	"Splitting it equally between us so we both get the same seemed the fairest way, and also even if they decide not to send anything back I still get something"
Self-interest	DG	"I don't know these other people. I don't know their story, I don't know who they are, all I know is that I need that $12 so I'm going to choose myself"
Self-interest	UG	"It's strategic...I'm trying to keep a bit more for myself, many responders will still accept this because some money is better than no money..."
Self-interest	TG	"It felt like a safe amount as if they sent back $0 I'd still be able to keep some money."
Win-Win	UG	"Actually i love team work. I love it when two or more people work together with honesty and transparency to achieve a common goal which benefits both party."
Win-Win	TG	"Because I hope for the best in people and that they will want to share the bonus. (It is in all our best interests.)"
Panel B: Social proximity of the evoked counterpart		
High proximity	UG	"It's split evenly and is most fair. I don't get more and they don't get more. We are equals."
Low proximity	TG	"If player 2 gives back it will be tripled and I will end up with more."
Panel C: Player 2		
Moral Good	UG	"Because both amounts were the same and I believe that the sum is a fair proposition!"
Moral Bad	TG	"They still earned more money than me so I would not send them back anything"
Self-interest	UG	"It makes sense to accept the offer since while not completely fair, it wasn't that unfair as well. Not accepting means that I get nothing, so the offer I got was much better."
Win-Win	TG	"Player 1 has $1 left from the original $6 so I sent $7 back so we both have $8 total"

Notes: Answers to the open-ended reasons question, control condition, reproduced verbatim (including typos), drawn at random within category-by-game cells among responses of typical length.

and standardized using its variation across those participants.

Table 3 reports the relation of this Control index to choices and reasons. The pooled Control regressions include indicators for the three studies. The within-person regressions use the same index to predict each of the 300 Control-first participants' Control and Market outcomes, and their transfer change. The action sample contains 755 pooled Control participants and 278 in the follow-up; the corresponding Moral-reason samples omit unclassified post-choice answers. Standard errors are heteroskedasticity-robust and treat the estimated index as fixed.

Table 3: Pre-choice Moral recall and subsequent behavior in DG

Outcome	Slope	s.e.	p	N
Panel A: First-choice Control, three samples pooled				
Share sent in Control	+1.4	0.7	0.052	755
Moral reason in Control	+3.4	1.7	0.047	736
Panel B: Control first, Market second				
Share sent in Control	+2.7	1.2	0.026	278
Share sent in Market	-2.4	1.7	0.163	278
Market − Control share sent	-5.1	2.3	0.026	278
Moral reason in Market	+0.8	3.3	0.813	229

Notes: Slopes are percentage points per standard deviation of the Moral-versus-Non-Moral Control retrieval index estimated on 800 first-choice DG Control participants. Forty-five participants with no substantively classified hypothetical description have no index. Panel A includes study indicators. Panel B uses the 300 Control-first participants in the DG Control–Market study who have a subsequent Market choice. Regressions use HC3 standard errors and treat the fitted index as fixed. Reasons without a substantive classification are omitted from reason regressions.

Treatment effects across the four descriptions. Because the task precedes the choice, the four descriptions of each participant can be compared across conditions without selecting on the chosen transfer. We estimate linear probability models of each category indicator on the between-subject DG and LT sessions, with hypothetical-allocation fixed effects and standard errors clustered by participant (Figure 11). In DG, Market reduces Moral descriptions by 43.7 percentage points, increases Self-interest by 6.4, and increases Win-Win by 37.2 among substantively classified descriptions (N = 1,548 from 512 participants); high social

proximity (levels 2–4) falls by 21.0 points across all 2,400 descriptions (578 participants). In LT, Market reduces Moral by 42.6 points and raises Win-Win by 42.2, while Self-interest barely changes (+0.5 points, $p = 0.835$; $N = 1{,}538$ from 529 participants); high social proximity falls by 17.2 points. In both games the frame moves the descriptions as it moves the stated reasons of Section 4.2: away from Moral and from close counterparts, toward Win-Win. The residual category (No clear justification) also rises under Market, from 15.1% to 45.7% of all descriptions in DG and from 21.8% to 43.0% in LT, so the three category contrasts refer to responses conditional on substantive classification, not to unconditional category shares.

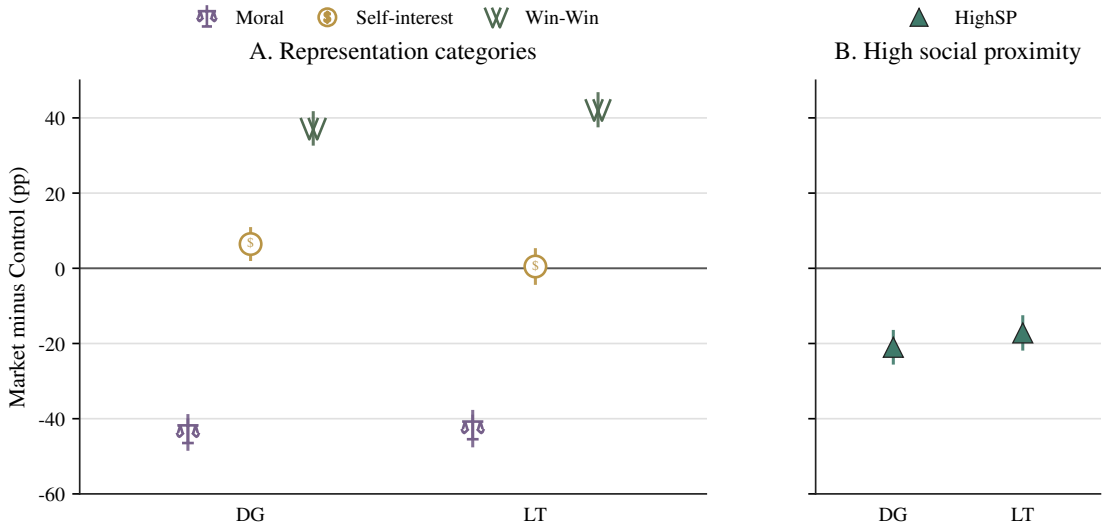


Figure 11: **Pre-choice hypothetical-allocation representations, Market versus Control, DG and LT.** Panel A reports changes in Moral, Self-interest, and Win-Win classification rates among substantively classified descriptions; Panel B reports changes in high social proximity (levels 2–4) using all descriptions. Estimates come from linear probability models with hypothetical-allocation fixed effects; intervals are 95% confidence intervals based on participant-clustered standard errors. Panel A: 1,548 descriptions from 512 participants in DG and 1,538 from 529 in LT. Panel B: 2,400 descriptions from 578 participants in DG and 2,400 from 581 in LT.

Description nearest the chosen allocation. As a complementary check, Table 4 retains each participant's pre-choice description at the hypothetical allocation closest to her eventual choice. Among substantively classified descriptions, Market reduces Moral and raises Win-Win in both DG and LT, as in the analysis of all four descriptions. Table 15 reports social

proximity for the same selected descriptions.

Table 4: Pre-choice representation categories by game and frame: memory attached to the hypothetical allocation closest to the participant's eventual choice

	Control		Market	
	DG	LT	DG	LT
Moral	75.6	77.8	37.1	30.6
Self-interest	22.2	16.8	24.9	23.6
Win-Win	2.2	5.4	38.1	45.8
Classified N	180	167	197	216
Unclassified (% of all)	10.0	16.5	50.7	46.0
All observations	200	200	400	400

Notes: Entries in the first three rows are percentages among substantively classified memories and sum to 100 within columns, matching the main-text convention for the reasons measure. The hypothetical-allocation descriptions were elicited before the choice. For each participant, the retained memory is the one attached to the hypothetical allocation closest to the eventual choice. In Control, the joint M/S/W distribution does not differ significantly between DG and LT ($p = 0.158$). Relative to Control, Market changes the joint distribution in both DG and LT ($p < 0.001$ in each game).

B.3 The Context–Category Similarity Module

Similarity inputs and predicted shifts. Figure 2 combines Control category shares with three 500-agent GPT-6 Luna exercises on story–category, game–category, and absolute story–game similarity. Table 5 gives the predicted shifts and 95% bootstrap intervals.

Table 5: **Predicted Category Shifts from Control**

Game	Moral	Self-interest	Win-Win
A: Aid–Bonus			
DG	-2.72	+2.70	+0.01
	[-2.96, -2.46]	[2.45, 2.94]	[0.01, 0.02]
UG	-1.82	+1.71	+0.11
	[-1.99, -1.66]	[1.56, 1.86]	[0.10, 0.13]
TG	-1.46	+1.00	+0.46
	[-1.60, -1.32]	[0.91, 1.09]	[0.38, 0.54]
B: Market–Control			
DG	-10.88	+10.76	+0.12
	[-11.61, -10.12]	[10.01, 11.48]	[0.10, 0.14]
UG	-5.69	+5.21	+0.48
	[-6.22, -5.16]	[4.72, 5.69]	[0.39, 0.58]
TG	-1.51	-0.03	+1.54
	[-2.21, -0.78]	[-0.31, 0.27]	[0.67, 2.36]

Notes: Entries are Aid minus Bonus (Panel A) or Market minus Control (Panel B) predicted changes in category shares, in percentage points. Predictions apply the first-order effects in Equations (6) and (7) at each game's observed Control shares, including cross-category effects. Panel A adds $S(\kappa_e, C)S(\mathrm{G}_{\mathrm{Control}}, \kappa_e)$ for each story e; Panel B changes $S(\mathrm{G}_\delta, C)$ with $f_C = 1$ for every category. Similarity scores are divided by 1,000. Brackets give 95% intervals from 5,000 draws resampling agents independently within each of three 500-agent GPT-6 Luna exercises, holding Control shares fixed. Source: the `eq4_similarity_probabilities` exercise in the replication package.

We apply the same game–category rating exercise to the Control and Market LT instructions. Table 6 compares LT with DG in each frame, reporting both similarity ratings and their implied category-share changes. LT's predicted Moral advantage in Control becomes a disadvantage under Market; the relative similarity gain goes mainly to Self-interest.

Table 6: **LT–DG Differences in Similarity and Predicted Category Shares**

Category	Control: LT–DG	Market: LT–DG	Market minus Control
Similarity ratings (pp)			
Moral	+6.6	-2.3	-8.8
	[+4.8, +8.4]	[-4.2, -0.3]	[-11.8, -5.9]
Self-interest	+1.3	+10.7	+9.4
	[-0.2, +2.8]	[+8.7, +12.7]	[+6.7, +12.1]
Win-Win	-7.9	-8.5	-0.6
	[-9.1, -6.6]	[-10.1, -6.9]	[-2.4, +1.2]
Predicted category shares (pp)			
Moral	+1.19	-1.75	-2.94
	[+0.53, +1.85]	[-2.52, -0.99]	[-4.06, -1.80]
Self-interest	-1.09	+3.60	+4.69
	[-1.75, -0.45]	[+2.82, +4.37]	[+3.55, +5.80]
Win-Win	-0.10	-1.85	-1.75
	[-0.11, -0.08]	[-2.20, -1.51]	[-2.10, -1.42]

Notes: Similarity entries are percentage-point differences in the mean allocation of 1,000 similarity points across Moral, Self-interest, and Win-Win. Predicted category changes apply the first-order probability calculation of Figure 2 to the LT–DG similarity differences, using observed DG category shares within each frame as baselines and setting $f_C = 1$. The final column subtracts the Control LT–DG gap from the Market LT–DG gap. Each target contains the game instructions and Player 1 decision screen. The DG and LT exercises use the same prompt and category descriptions, with separate groups of 500 GPT-6 Luna agents; each agent rates both Control and Market within its exercise. Brackets give 95% intervals from 5,000 bootstrap draws, resampling agents separately by exercise and preserving their paired frame ratings; baseline category shares are held fixed.

B.3.1 Similarity to Concrete Vignettes

As a check using concrete situations, we rate context texts against 18 vignettes crossing DG, UG, and TG game families with Moral, Self-interest, and Win-Win reasoning and personal or anonymous settings. The targets are the standalone Aid and Bonus stories and the Control and Market instructions with the actual Player 1 decision screen for each game. Each of 500 independently tasked GPT-6 Luna agents allocates 1,000 points across the six same-game vignettes for each game target and across all 18 vignettes for each story. We sum points by the vignettes' preassigned category. The rating prompt and vignettes, including their categories, are available online.

Table 7 shows that Aid receives 15.3 percentage points more Self-interest similarity and

17.3 points less Win-Win similarity than Bonus. Market reduces Moral similarity by 7.4 points in DG and 8.0 points in UG, while raising Self-interest in DG and Win-Win in UG. In TG, Market shifts weight from Self-interest (−4.8 points) toward Win-Win (+4.3 points), with a small Moral contrast. These directions broadly match the abstract-category exercise.

Table 7: **Robustness: Similarity to Concrete Vignettes**

Contrast	Moral	Self-interest	Win-Win
Aid–Bonus	+1.9	+15.3	-17.3
	[0.7, 3.2]	[13.9, 16.8]	[-18.7, -15.9]
DG: Market–Control	-7.4	+5.5	+1.9
	[-8.8, -6.0]	[3.7, 7.3]	[0.4, 3.3]
UG: Market–Control	-8.0	+1.2	+6.8
	[-9.5, -6.6]	[-0.4, 2.8]	[5.5, 8.2]
TG: Market–Control	+0.4	-4.8	+4.3
	[-0.7, 1.5]	[-6.3, -3.2]	[3.0, 5.7]

Notes: Mean within-rater treatment contrasts in percentage points of the 1,000-point allocation, with 95% Student-t intervals across 500 separately tasked GPT-6 Luna runs in brackets. Each target is rated against concrete vignettes, and scores are aggregated into Moral, Self-interest, and Win-Win based on the vignettes' preassigned categories. Stories are rated against all 18 vignettes; each Control/Market game target is rated against six vignettes from that same game. Intervals describe variability across model runs, not sampling error for participants.

Direct story–game similarity. In the third rating exercise for Equation (5), 500 GPT-6 Luna agents independently rate each story against each Control game description and decision screen on a 0–1,000 scale. Table 8 shows a stronger fit to DG and UG than to TG for both stories, supporting the contextual-distance comparison in Section 3.3. The rating prompt is available online.

Table 8: **Direct Similarity of Stories to Control Game Descriptions**

Story	DG	UG	TG
Aid	635.8	502.0	390.5
Bonus	666.5	541.7	422.3

Notes: Mean absolute similarity points on a 0–1,000 scale. Each story–game pair is rated independently by 500 GPT-6 Luna agents. Game targets are the Control instructions and Player 1 decision screen. These are the story–game ratings used in the equation (5) similarity calculation.

C Additional Experimental Evidence

C.1 Additional Analyses on Receivers

This appendix confronts the model of Player 2 in Appendix A.2 with the data. In Control, we report how receivers' reasons vary with the action they face, how the share they return varies with this action, and how their response is ranked across categories (Proposition 4). We then turn to the treatments: how Market and Aid change receivers' reasons and responses, and how well similarity ratings predict the changes in reasons.

Receiver Representations and the Action Faced. Equations (15) and (18) predict that receivers' reasons covary with the action faced. In the control ultimatum game they do: *Moral Bad* is the modal category among the 6 receivers who faced a zero offer (4 of them) and virtually disappears at offers just below or at the equal split (2% of 482), where *Moral Good* reaches 62%; linear-probability slopes of the category indicators on the offer are +0.86 (Moral Good) and −0.79 (Moral Bad), both $p < 0.001$. Of the control-sample aggregate acceptance slope of 0.87, the within-category margin contributes 0.35 and the composition term 0.52: six tenths of the response to the offer comes with a change in the reasons receivers give, from Moral Bad at low offers to Moral Good at fair ones. In the model these are the two verdicts of one norm: equal payoffs condemn a low offer and approve a fair one (Appendix A.2). In the trust game composition accounts for 14% of the aggregate return slope, so returns respond to the send mostly within categories. The verdict changes there too, where equal payoffs stop asking for a return: *Moral Bad* is 31.5% of classified reasons at sends up to one quarter of the endowment, 8.8% between one quarter and five twelfths, and under 1% above.

Return Schedules: Equal Payoffs versus the Alternative Norm. Table 9 reports the within-category schedules (Panel A, acceptance; Panel B, returns). Returned shares *rise* with the send among *Moral Good* receivers (slope +0.28, s.e. 0.03, $N = 303$) and *Win-Win*

receivers (+0.28, s.e. 0.05, $N = 150$), which the alternative norm of Appendix A.2, whose schedules decline with the send, cannot produce, while *Self-interest* receivers sit near the selfish benchmark (slope -0.05). The equal-payoffs return $x^e(t)$ of Appendix A.2 fits both low-selfishness categories best from the reference send upward; below $t = 1/(2 + r)$, where equal payoffs ask for no return, receivers still return a positive share. Senders' believed schedules show the opposite pattern: the two elicited belief points per sender imply believed slopes with a median of exactly 0.00 in every category, matching the constant-share specification of the main text. Senders thus attribute to the receiver a flat schedule while facing receivers who equalize, which locates the trust-game forecast-error wedge of Section 4.2 in the receiver that senders imagine.

Table 9: **Second-Mover Schedules by Category (Control Sample)**

Category	Intercept	(SE)	Slope in t	(SE)	N
Panel A: UG acceptance, $p_c(t) = a_c + b_c\, t$					
Moral good	0.755	(0.117)	0.455	(0.223)	332
Moral bad	-0.064	(0.091)	1.106	(0.405)	29
Self-interest	0.898	(0.096)	0.143	(0.190)	81
Mutual Benefit / Cooperation	0.959	(0.042)	0.068	(0.072)	150
Panel B: TG returned share, $s_c(t)$					
Moral good	0.227	(0.022)	0.282	(0.033)	303
Moral bad	-0.062	(0.030)	0.452	(0.168)	42
Self-interest	0.122	(0.052)	-0.049	(0.072)	58
Mutual Benefit / Cooperation	0.185	(0.032)	0.278	(0.053)	150

Notes: Control sample. OLS of the second mover's realized response (UG: acceptance of the offer faced; TG: share of the tripled amount returned) on the first mover's action t (share of the budget), by stated category. Robust (HC1) standard errors. The offers and sends faced are endogenous to the matched first mover; hypothetical responses at the one-third reference action give the treatment-clean counterpart in Figure 13.

Rankings across Categories. Proposition 4 ranks Player 2's response across categories at a given action. The fitted schedules of Table 9 are ordered as it predicts: acceptance of a zero offer is 0.96 for Win-Win, 0.90 for Self-interest, and 0.76 for Moral Good, and the returned share is highest for Moral Good, then Win-Win, then Self-interest at every

send. Those schedules use the actions receivers faced, which differ across categories. The hypothetical response to the one-third action, which every receiver reports, compares categories at a common action (Table 10); here Moral pools Moral Good and Moral Bad. In the Control trust game, the share returned is 31.4% under Moral, 26.3% under Win-Win, and 12.2% under Self-interest: the predicted ranking, with each pairwise difference statistically significant at the 1% level. In the Control ultimatum game, a one-third offer is accepted by 57.1% of Moral receivers, 65.3% of Win-Win receivers, and 84.0% of Self-interest receivers. Moral receivers accept least, as predicted: 26.9 points less often than Self-interest receivers ($p < 0.001$) and 8.3 points less often than Win-Win receivers, a difference in the predicted direction that is not statistically significant ($p = 0.079$). Against the proposition, in which Win-Win accepts most, Win-Win receivers accept 18.6 points less often than Self-interest receivers ($p = 0.001$). This may reflect the boundary between Win-Win and Moral Good reasons, where the classification of Player 2 is least reliable (Appendix B.1). Under Market, Self-interest receivers again accept more often than Moral receivers and return less than the other two categories; the acceptance rates of Win-Win and Self-interest receivers, and the returns of Moral and Win-Win receivers, are within sampling error of each other.

Treatment Shifts and Receiver Responses. In the model, the action faced is part of the receiver's description of the game (equation (18)). The Control results above show why this matters: low UG offers more often elicit Moral Bad reasons and are more likely to be rejected. Under Market, acceptance of the same one-third offer rises by 10.5 percentage points (Figure 13), yet realized acceptance falls by 33 points because proposers make much lower offers. Market also shifts receiver reasons from Moral Good toward Self-interest in UG and toward Win-Win in TG (Figure 12). Thus the frame and the action it induces both shape Player 2's representation and response.

The two shifts in reasons differ in what they explain (Table 10). In UG, the shift toward Self-interest, the category that accepts a one-third offer most often, accounts for 5.6 of the

Table 10: **Player 2's Hypothetical Response at the One-Third Action, by Category**

	Control			Market			Market − Control	
Category	Mean	(SE)	*N*	Mean	(SE)	*N*	Diff.	(SE)
Panel A: UG, acceptance of a one-third offer (%)								
Moral	57.1	(2.6)	361	66.0	(3.4)	191	+8.9	(4.3)
Moral Good	58.7	(2.7)	332	63.5	(4.9)	96	+4.8	(5.6)
Moral Bad	37.9	(9.2)	29	68.4	(4.8)	95	+30.5	(10.3)
Win-Win	65.3	(3.9)	150	75.2	(3.7)	137	+9.8	(5.4)
Self-interest	84.0	(4.1)	81	78.8	(2.6)	241	−5.1	(4.9)
All receivers	63.0	(2.0)	600	73.5	(1.8)	600	+10.5	(2.7)
Panel B: TG, share returned after a one-third send (%)								
Moral	31.4	(0.9)	345	40.5	(0.9)	254	+9.1	(1.3)
Moral Good	33.4	(1.0)	303	41.4	(0.8)	238	+8.0	(1.3)
Moral Bad	17.1	(2.1)	42	26.6	(4.9)	16	+9.5	(5.3)
Win-Win	26.3	(1.2)	150	41.9	(1.1)	223	+15.7	(1.6)
Self-interest	12.2	(2.1)	58	24.2	(2.9)	73	+12.0	(3.6)
All receivers	28.2	(0.7)	573	39.0	(0.7)	580	+10.7	(1.0)

Notes: Every second mover reports a hypothetical response to a one-third offer (UG) or send (TG); second movers who faced that action answered for it only, and that answer is used (Appendix D.1). Categories are those of the reason given for the realized response. Moral pools Moral Good and Moral Bad; All receivers includes responses with no clear justification. The last two columns report the Market minus Control difference in means and its unequal-variance standard error.

10.5 points by which its acceptance rises; the other 4.9 points arise within categories. In TG, the shift toward Win-Win cannot account for the 10.7-point rise in the share returned after a one-third send, because Win-Win receivers return less than Moral receivers in Control: the change in category shares contributes −0.6 points and the change within categories 11.3, with the largest rise among Win-Win receivers, from 26.3% to 41.9%. The model, in which the category fixes the response, does not have this margin.

Receiver Similarity and Category Shifts. We extend the similarity exercise to Player 2 using 500 blinded GPT-6 Luna agents. Each agent rates eight decision situations: UG and TG receivers under Control, Market, Bonus, and Aid. A situation combines the receiver instructions and decision screen with a first-mover action drawn from that condition's empirical distribution; Aid and Bonus also include their respective stories. Agents allocate

1,000 similarity points across Moral Good, Moral Bad, Self-interest, and Win-Win, the four labels of the classification, so the two verdicts of Moral are rated separately. We use differences in average normalized similarity scores to calculate first-order category-probability shifts at the observed Control shares, setting the unobserved category-frequency weights to one (Appendix Table 11).

Table 11: **Player 2: Observed and Similarity-Implied Category Shifts**

		Market–Control		Aid–Bonus	
Game	Category	Actual	Similarity	Actual	Similarity
UG	Moral Good	−39.21	−10.03	−5.45	−2.20
	Moral Bad	+11.80	+0.97	+3.59	+0.39
	Self-interest	+28.67	+2.40	+6.01	+1.17
	Win-Win	−1.26	+6.66	−4.16	+0.64
TG	Moral Good	−11.52	−7.72	−9.30	−1.74
	Moral Bad	−4.69	−0.25	−0.68	+0.21
	Self-interest	+2.78	+0.24	+0.72	+0.79
	Win-Win	+13.42	+7.73	+9.26	+0.74

Notes: Entries are treated-minus-baseline changes in category probabilities (percentage points). Actual columns use shares among substantively classified Player 2 reasons. Similarity columns use the first-order approximation $100p_C(\Delta S_C - \sum_K p_K \Delta S_K)$, where p_C is the observed Control share and ΔS_C is the treated-minus-baseline difference in mean similarity ratings divided by 1,000. Category-frequency weights are set to $f_C = 1$. Similarity ratings average 500 GPT-6 Luna agents per situation, with first-mover actions sampled separately from each treatment's observed distribution. These predictions are directional, not quantitatively calibrated.

The exercise predicts the direction of 13 of 16 observed category shifts. In Market, it correctly predicts the decline in Moral Good in both games, the gains in Moral Bad and Self-interest in UG, and especially the Win-Win gain in TG. In Aid, it predicts the shift away from Moral Good toward Moral Bad and Self-interest in UG, as well as the Moral Good decline in TG. The magnitudes are not calibrated: notably, Market's predicted Win-Win gain in UG is not observed, while Aid's large Win-Win increase in TG is substantially understated.

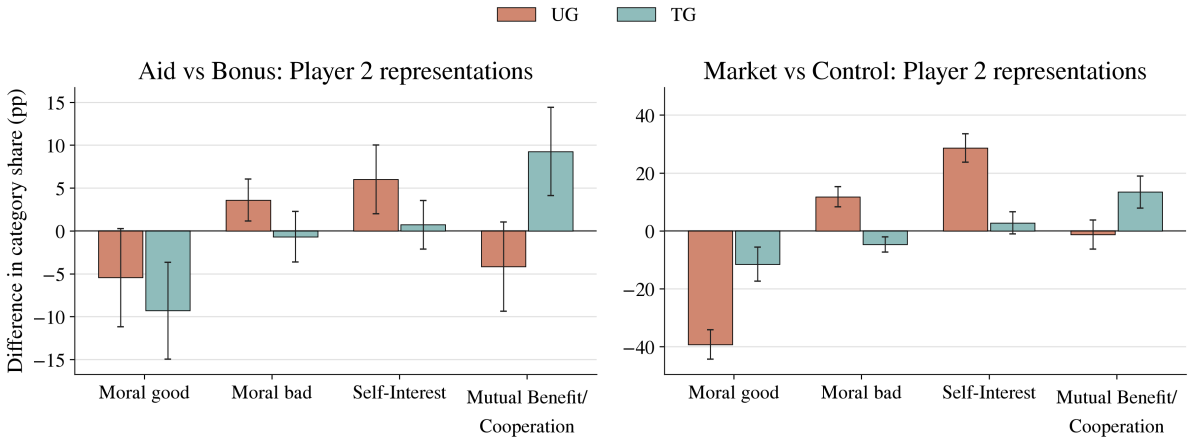


Figure 12: Treatment-induced shifts in Player 2 category shares. Panels report Aid vs Bonus and Market vs Control. Error bars show 95% confidence intervals.

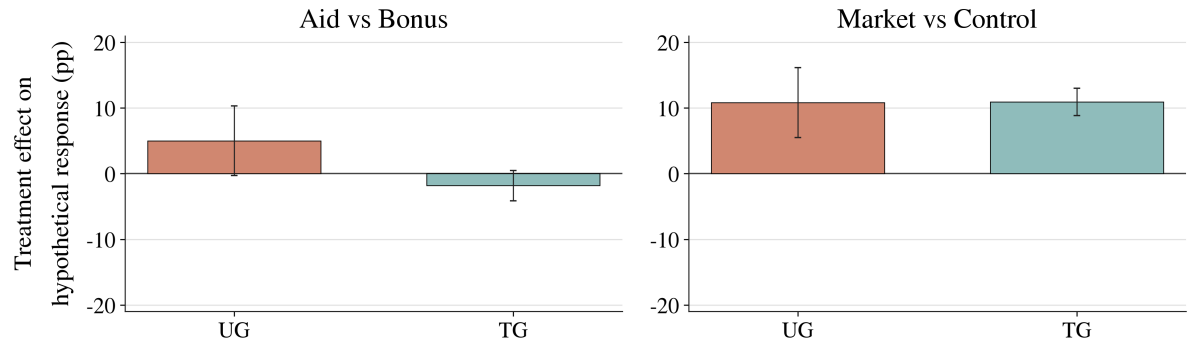


Figure 13: Treatment effects on second movers' hypothetical responses at the common one-third reference action. Outcomes are the acceptance rate of a one-third offer in the ultimatum game and the share of the multiplied amount returned after a one-third send in the trust game. Panels report Aid minus Bonus and Market minus Control. Error bars show 95% confidence intervals.

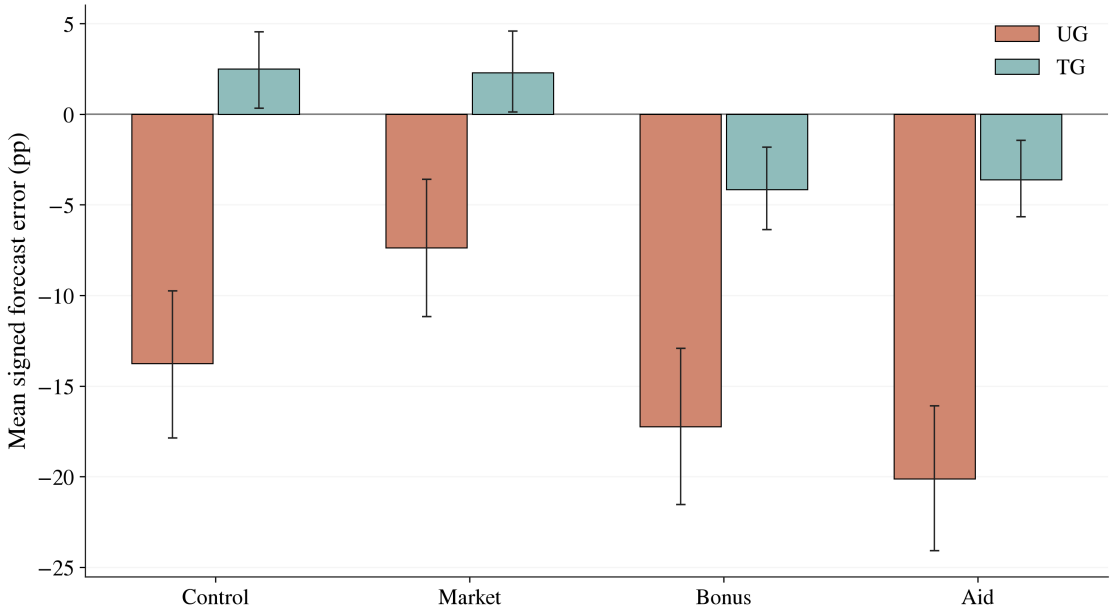


Figure 14: Forecast errors by treatment, hypothetical beliefs at the common one-third reference action, using the corresponding mean hypothetical Player 2 response as the benchmark. Positive values indicate optimism and negative values pessimism. Error bars show bootstrap 95% confidence intervals. Figure 7 reports treatment contrasts at chosen actions.

C.2 Social Proximity Evidence

Alongside the representation category, the classification of stated reasons records the social proximity of the counterpart evoked by each response: no mention of the recipient, abstract stranger, anonymous peer, teammate/coworker, or friend. We group anonymous peer, teammate/coworker, and friend as high social proximity, and no mention and abstract stranger as low social proximity.

Social proximity, representation categories, and behavior. Table 12 reports the pooled relationship between social proximity and Player 1 representation categories. High social proximity appears in 50.6% of Moral reasons and 40.6% of Win-Win reasons, but only 8.1% of Self-interest reasons. Within each representation category, high-proximity reasons are also associated with larger actions. In the Control condition, the difference in mean actions between high- and low-proximity reasons is 17.5 percentage points in DG, 2.9 points in UG, and 10.1 points in TG. These are descriptive associations because social proximity is measured from participants' reasons rather than experimentally assigned.

Table 12: **Player 1: Counterpart Proximity within Stated Reason Category**

	Moral	Self-interest	Mutual Benefit/Cooperation
Panel A: Distribution of counterpart proximity (column shares)			
Low social proximity	49.4%	91.9%	59.4%
High social proximity	50.6%	8.1%	40.6%
Panel B: Mean action, share of the budget			
Low social proximity	44.3%	23.6%	57.0%
High social proximity	47.0%	34.6%	63.2%

Notes: Pooled across all games and conditions for Player 1, classified sample. Panel A reports the distribution of counterpart proximity within each stated reason category (columns sum to 100%). Panel B reports the mean action (share of the budget transferred, offered, or sent) in each proximity-by-category cell. High social proximity: the stated reason refers to the counterpart as at least an anonymous peer (anonymous peer, teammate or coworker, friend); low: no mention of the counterpart or an abstract stranger.

Treatment effects on social proximity. Because condition assignment is randomized, we can also ask whether the contextual manipulations change the proximity of the counterpart evoked in participants' reasons. Table 13 reports the share of Player 1 reasons classified as high social proximity and the corresponding treatment contrasts. Relative to Bonus, Aid reduces high social proximity by 18.2 percentage points in DG and 12.2 points in UG, while producing essentially no change in TG (−0.4 points). This is the proximity pattern referenced in Section 4.1. The Market frame also lowers high social proximity relative to Control in all three games, by 13–18 percentage points.

Table 13: **Treatment Effects on High Social Proximity in Player 1 Reasons**

Game	Baseline	Treatment	Difference	95% CI
Panel A: Aid versus Bonus				
DG	48.7%	30.5%	−18.2	[−24.9, −11.5]
UG	58.3%	46.1%	−12.2	[−17.8, −6.5]
TG	42.6%	42.2%	−0.4	[−6.1, 5.2]
Pooled	50.0%	40.7%	−9.3	[−12.8, −5.9]
Panel B: Market versus Control				
DG	31.6%	14.0%	−17.6	[−23.5, −11.7]
UG	42.1%	28.9%	−13.2	[−18.8, −7.5]
TG	38.0%	19.7%	−18.3	[−23.4, −13.2]
Pooled	37.9%	21.7%	−16.2	[−19.4, −13.0]

Notes: The first two columns report the share of classified Player 1 reasons with high social proximity. In Panel A, Baseline is Bonus and Treatment is Aid; in Panel B, Baseline is Control and Treatment is Market. Difference is Treatment minus Baseline, in percentage points. High social proximity means that the stated reason refers to the counterpart as an anonymous peer, teammate/coworker, or friend; low social proximity means no mention of the recipient or an abstract stranger. The sample is restricted to substantive-category responses with nonmissing social-proximity coding. Sample sizes (Baseline/Treatment) are 390/400 in DG, 587/588 in UG, and 589/588 in TG for Panel A; and 399/329, 582/501, and 584/563 for Panel B. Confidence intervals use the normal approximation for differences in proportions.

Pre-choice evidence and second movers. The dictator-game hypothetical-allocation task provides a pre-choice measure of the social situations participants associate with different allocations. In the Control condition, high social proximity in these descriptions is associated with keeping $1.70 less on average (Table 14), and the Market frame sharply shifts the distribution toward no mention of the recipient and away from socially closer counterparts (Table 15).

Table 14: Allocation kept by the dictator and allocation-specific outcomes, by social proximity (hypothetical-allocation measure): OLS estimates, control condition. High SP indicates a socially close counterpart. Standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$.

	Mean Allocation	=12	>8	=8	=6	=4
High SP	-1.703^{***}	-0.228^{***}	-0.299^{***}	-0.028	0.357^{***}	0.006
	(0.220)	(0.034)	(0.039)	(0.034)	(0.047)	(0.019)
Constant	8.101^{***}	0.246^{***}	0.351^{***}	0.145^{***}	0.364^{***}	0.035^{***}
	(0.144)	(0.022)	(0.026)	(0.022)	(0.031)	(0.013)
Observations	400	400	400	400	400	400
R^2	0.131	0.101	0.126	0.002	0.125	0.000

Table 15: Social proximity of the pre-choice description nearest the eventual choice, DG and LT pooled. Column shares by condition sum to 100%. p-values come from two-sided tests of equal proportions with continuity correction.

Social proximity	Market	Control	p-value
No mention of recipient	75.88%	28.75%	< 0.001
Abstract stranger	10.75%	28.25%	< 0.001
Anonymous peer	5.00%	18.25%	< 0.001
Teammate / coworker	2.13%	6.50%	< 0.001
Friend	6.25%	18.25%	< 0.001

C.3 Belief Sensitivity of Player 1 Actions

Table 16 reports the category-specific slopes relating Player 1's chosen share to her hypothetical belief at the one-third reference action. Control-only regressions are shown alongside estimates pooling the four conditions with condition fixed effects. The reported slope comparisons are descriptive associations; categories and beliefs are not randomly assigned.

Table 16: **Belief Sensitivity of Actions, by Representation Category**

		Control		All conditions (condition FE)	
Game	Category	Slope	(SE)	Slope	(SE)
UG	Moral	-0.005	(0.021)	-0.018	(0.016)
	Self-interest	-0.109	(0.089)	-0.087	(0.045)
	Mutual Benefit/Coop.	-0.191	(0.308)	-0.141	(0.081)
TG	Moral	0.299	(0.084)	0.315	(0.038)
	Self-interest	0.390	(0.129)	0.424	(0.076)
	Mutual Benefit/Coop.	0.285	(0.096)	0.178	(0.048)

Notes: OLS slopes of the Player 1 action (share) on the hypothetical belief at the one-third reference action, estimated separately by representation category; robust (HC1) standard errors. "All conditions" pools the four conditions with category-specific condition fixed effects. Two-sided Moral–Self-interest slope comparisons in Control (HC1 Wald): UG $p = 0.25$, TG $p = 0.55$. Slope-equality F-tests across categories: UG control $p = 0.432$, UG pooled $p = 0.132$, TG control $p = 0.790$, TG pooled $p = 0.011$. No pairwise slope difference is significant at the 5% level in Control or pooled UG. In pooled TG, Win-Win differs from Moral ($p = 0.024$) and Self-interest ($p = 0.006$), while Moral and Self-interest do not differ ($p = 0.20$). Model predictions: negative slopes in the UG, positive in the TG for every category (the equal-payoff norm anchor moves with the believed return); magnitude largest for Self-interest, whose material channel adds to the common anchor channel. Surplus attention does not enter the sensitivity, so Mutual Benefit/Cooperation should differ from Moral in level rather than slope. The UG Mutual Benefit/Cooperation cell is very small and should be read accordingly.

C.4 Within-Subject Evidence from the Dictator Game

The dictator-game pre-registration (AsPredicted #261370) included three within-subject designs of 600 participants each (599 complete pairs in the DG market design), all without stories: (i) game fixed at DG, frame varying between abstract and market; (ii) game fixed at LT, frame varying between abstract and market; (iii) frame fixed at abstract, game varying between DG and LT. Order was randomized; order effects are small—within condition, mean shares sent differ by at most 2.5 percentage points between first and second decisions—and order enters all specifications below.

Table 17 summarizes switching behavior. When the same participant plays DG and LT under the abstract frame, 46.2% change their choice (share sent) and 22.7% change their representation category; when the frame switches between abstract and market with the game fixed, 73.1% (DG) and 82.3% (LT) change their choice and over half change their category. In all three designs, choice switching is strongly predicted by category switching: participants whose representation category changes are 14–44 percentage points more likely to change their choice than participants whose category is stable (linear-probability estimates: 0.44, robust SE 0.04, in the DG/LT design; 0.19, SE 0.04, and 0.14, SE 0.03, in the two market designs; all $p < 0.001$). Choice switching counts any change in the share sent, so it includes small adjustments; the conditional contrast between category switchers and stayers is the informative statistic.

Tables 18 and 19 report within-subject regressions with participant fixed effects. Holding the participant fixed, moving from LT to DG lowers the share sent by 3.3–3.6 percentage points, and moving from the abstract to the market frame lowers it substantially; representation categories predict behavior within participant, with self-interested representations associated with significantly lower transfers.

Table 17: **Within-Subject Switching in the Dictator Game**

		Share switching			P(choice switch | category)	
Comparison	N pairs	Choice	Category	Social proximity	Switched	Stayed
DG vs LT (abstract frame)	600	46.2%	22.7%	40.5%	80.1%	36.2%
Market vs Control (DG)	599	73.1%	51.6%	66.4%	82.2%	63.4%
Market vs Control (LT)	600	82.3%	55.0%	62.0%	88.5%	74.8%

Notes: Each participant makes two dictator-game decisions. In the first row, both decisions use the abstract (control) frame and the game varies (KW vs LT); in the second and third rows, the game is fixed and the frame varies (abstract vs market). "Choice" switching is any change in the share sent; "Category" switching is a change in the reason-based representation category (Moral, Self-interest, Mutual Benefit/Cooperation, No clear justification); "Social proximity" switching is a change in the social-proximity class of the representation. The last two columns report the probability of a choice switch conditional on the representation category switching or staying fixed.

Table 18: **Within-Subject Regressions**: Control sample, LT vs. KW

	(1)	(2)	(3)	(4)
	Share Sent P1			
KW	-0.036***	-0.035***	-0.033***	-0.033***
	(0.011)	(0.011)	(0.010)	(0.010)
KW x LTFirst	-0.026	-0.026	-0.033**	-0.033**
	(0.016)	(0.016)	(0.015)	(0.015)
HighSP		0.011		-0.004
		(0.015)		(0.015)
Mutual Benefit / Cooperation			-0.070	-0.070
			(0.071)	(0.072)
Self-interest			-0.091***	-0.093***
			(0.026)	(0.026)
No clear justification			-0.053	-0.054
			(0.039)	(0.040)
Participant fixed effects	Yes	Yes	Yes	Yes
Observations	1200	1200	1200	1200
Participants	600	600	600	600
R^2	0.867	0.868	0.877	0.877

Outcome is DG share sent by Player 1, defined as $1 - \texttt{allocation}/12$. Models use the within-control LT/KW sample only. All specifications include participant fixed effects and cluster standard errors by participant. Baseline text category is Moral. HighSP equals 1 if `sp_num`¿1. LTFirst equals 1 when LT was shown before KW; its main effect is collinear with the participant fixed effects and therefore not identified. Estimated without fixed effects (standard errors clustered by participant), the LTFirst level effect is +0.019 (SE 0.015). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 19: **Within-Subject Regressions**: Market vs. Control within subject

	(1)	(2)	(3)	(4)
	Share Sent P1			
Market	-0.160***	-0.139***	-0.135***	-0.126***
	(0.017)	(0.018)	(0.018)	(0.019)
Market x ControlFirst	0.016	0.017	0.017	0.017
	(0.025)	(0.025)	(0.024)	(0.024)
HighSP		0.113***		0.062***
		(0.021)		(0.021)
Mutual Benefit / Cooperation			-0.076**	-0.073**
			(0.033)	(0.033)
Self-interest			-0.197***	-0.177***
			(0.022)	(0.023)
No clear justification			-0.084**	-0.068**
			(0.034)	(0.034)
Participant fixed effects	Yes	Yes	Yes	Yes
Observations	2398	2398	2398	2398
Participants	1199	1199	1199	1199
R^2	0.597	0.614	0.642	0.647

Outcome is DG share sent by Player 1, defined as $1 - \texttt{allocation}/12$. Models use the within-subject Market vs. Control sample pooled across the fixed-game designs. All specifications include participant fixed effects and cluster standard errors by participant. Baseline text category is Moral. HighSP equals 1 if sp_num¿1. ControlFirst equals 1 when Control was shown before Market; its main effect is collinear with the participant fixed effects and therefore not identified. Estimated without fixed effects (standard errors clustered by participant), the ControlFirst level effect is +0.009 (SE 0.013). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D Treatment Effects and Choice Distributions

D.1 Benchmark Regressions

This appendix reports the full benchmark regressions summarized in Section 4.3. The regressions are estimated separately by game and pre-registered treatment comparison. Player 1 specifications include category indicators and, where applicable, hypothetical beliefs. Player 2 specifications additionally account for the first-mover action faced.

Table 20: **Benchmark Regressions: Player 1, Aid vs Bonus, Categories + Beliefs HP**

	(1) KW action	(2) UG action	(3) TG action	(4) UG beliefs	(5) TG beliefs
Self-interest	-0.337***	-0.121***	-0.119***	-0.158***	-0.015
	(0.010)	(0.009)	(0.021)	(0.015)	(0.012)
Mutual Benefit / Cooperation	0.011	0.060***	0.276***	0.046*	0.093***
	(0.033)	(0.016)	(0.015)	(0.027)	(0.009)
No clear justification	-0.245***	-0.042*	-0.060	-0.222***	0.020
	(0.036)	(0.023)	(0.051)	(0.039)	(0.029)
Hypothetical beliefs		-0.052***	0.270***	0.354***	0.645***
		(0.015)	(0.038)	(0.025)	(0.022)
Constant	0.457***	0.514***	0.348***	0.618***	0.090***
	(0.005)	(0.008)	(0.015)	(0.014)	(0.009)
Observations	801	1200	1200	1200	1200
R^2	0.599	0.170	0.333	0.203	0.489

Standard errors in parentheses. Baseline category is Moral. Sample restricted to Bonus and Aid. UG and TG equations include hypothetical beliefs in addition to category dummies; KW action uses category dummies only. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Reference beliefs of one-third choosers. First movers who chose exactly one third of the budget answered only the question about their actual action, and the analysis uses that answer as their belief at the one-third reference; second movers who faced a one-third action are treated in the same way. Treating these reference beliefs as missing instead would drop 5.0% of UG and 18.4% of TG first movers, and unevenly across conditions, since one third is the modal Control action in TG (22.8% of Control senders against 8.7% of Market senders). Table 24 reports the belief-dependent Player 1 results of the main text with and without

Table 21: **Benchmark Regressions: Player 2, Aid vs Bonus, Categories + Beliefs HP**

	(1) UG outcome	(2) TG outcome	(3) UG outcome + fitted share sent P1	(4) TG outcome + fitted share sent P1
Moral bad	-0.839***	-0.244***	-0.717***	-0.179***
	(0.025)	(0.016)	(0.030)	(0.015)
Mutual Benefit / Cooperation	0.015	-0.000	0.024**	0.002
	(0.012)	(0.009)	(0.012)	(0.008)
No clear justification	-0.000	-0.064**	0.016	-0.038
	(0.036)	(0.027)	(0.036)	(0.025)
Self-interest	-0.043***	-0.152***	0.000	-0.137***
	(0.016)	(0.017)	(0.017)	(0.015)
Hypothetical acceptance	0.049***		0.055***	
	(0.011)		(0.011)	
Hypothetical share sent P2		0.607***		0.608***
		(0.020)		(0.018)
Predicted share sent P1			1.631***	0.672***
			(0.222)	(0.041)
Constant	0.937***	0.193***	0.151	-0.165***
	(0.010)	(0.009)	(0.107)	(0.023)
Observations	1200	1191	1200	1191
R^2	0.504	0.582	0.526	0.659

Standard errors in parentheses. Baseline category is Moral good. Sample restricted to Bonus and Aid. Columns (3) and (4) add fitted share sent P1, computed from the comparison-specific Player 1 action regression in the same specification by averaging fitted values within each observed Player 1 donation level. UG equations include hypothetical acceptance and TG equations include hypothetical share sent P2; columns (3) and (4) additionally include fitted share sent P1. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 22: **Benchmark Regressions: Player 1, Market vs Control, Categories + Beliefs HP**

	(1) KW action	(2) UG action	(3) TG action	(4) UG beliefs	(5) TG beliefs
Self-interest	-0.299***	-0.156***	-0.061***	-0.093***	0.004
	(0.016)	(0.015)	(0.020)	(0.014)	(0.011)
Mutual Benefit / Cooperation	-0.082***	-0.154***	0.227***	-0.069***	0.071***
	(0.027)	(0.019)	(0.017)	(0.018)	(0.009)
No clear justification	-0.211***	-0.189***	0.045	-0.150***	0.060***
	(0.025)	(0.021)	(0.036)	(0.020)	(0.019)
Hypothetical beliefs		-0.147***	0.358***	0.297***	0.720***
		(0.026)	(0.036)	(0.025)	(0.019)
Constant	0.418***	0.539***	0.339***	0.616***	0.083***
	(0.010)	(0.016)	(0.017)	(0.015)	(0.009)
Observations	803	1200	1200	1200	1200
R^2	0.324	0.174	0.289	0.141	0.591

Standard errors in parentheses. Baseline category is Moral. Sample restricted to Control and Market. UG and TG equations include hypothetical beliefs in addition to category dummies; KW action uses category dummies only. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 23: **Benchmark Regressions: Player 2, Market vs Control, Categories + Beliefs HP**

	(1) UG outcome	(2) TG outcome	(3) UG outcome + fitted share sent P1	(4) TG outcome + fitted share sent P1
Moral bad	-0.896***	-0.242***	-0.713***	-0.185***
	(0.032)	(0.019)	(0.035)	(0.019)
Mutual Benefit / Cooperation	-0.051**	-0.011	-0.014	-0.011
	(0.024)	(0.009)	(0.023)	(0.009)
No clear justification	-0.232***	-0.014	-0.140***	-0.003
	(0.052)	(0.020)	(0.051)	(0.019)
Self-interest	-0.292***	-0.133***	-0.150***	-0.134***
	(0.023)	(0.014)	(0.026)	(0.013)
Hypothetical acceptance	0.070***		0.104***	
	(0.020)		(0.019)	
Hypothetical share sent P2		0.588***		0.574***
		(0.023)		(0.022)
Predicted share sent P1			1.847***	0.547***
			(0.165)	(0.053)
Constant	0.926***	0.195***	0.140*	-0.113***
	(0.019)	(0.010)	(0.073)	(0.032)
Observations	1200	1153	1200	1153
R^2	0.425	0.521	0.480	0.561

Standard errors in parentheses. Baseline category is Moral good. Sample restricted to Control and Market. Columns (3) and (4) add fitted share sent P1, computed from the comparison-specific Player 1 action regression in the same specification by averaging fitted values within each observed Player 1 donation level. UG equations include hypothetical acceptance and TG equations include hypothetical share sent P2; columns (3) and (4) additionally include fitted share sent P1. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

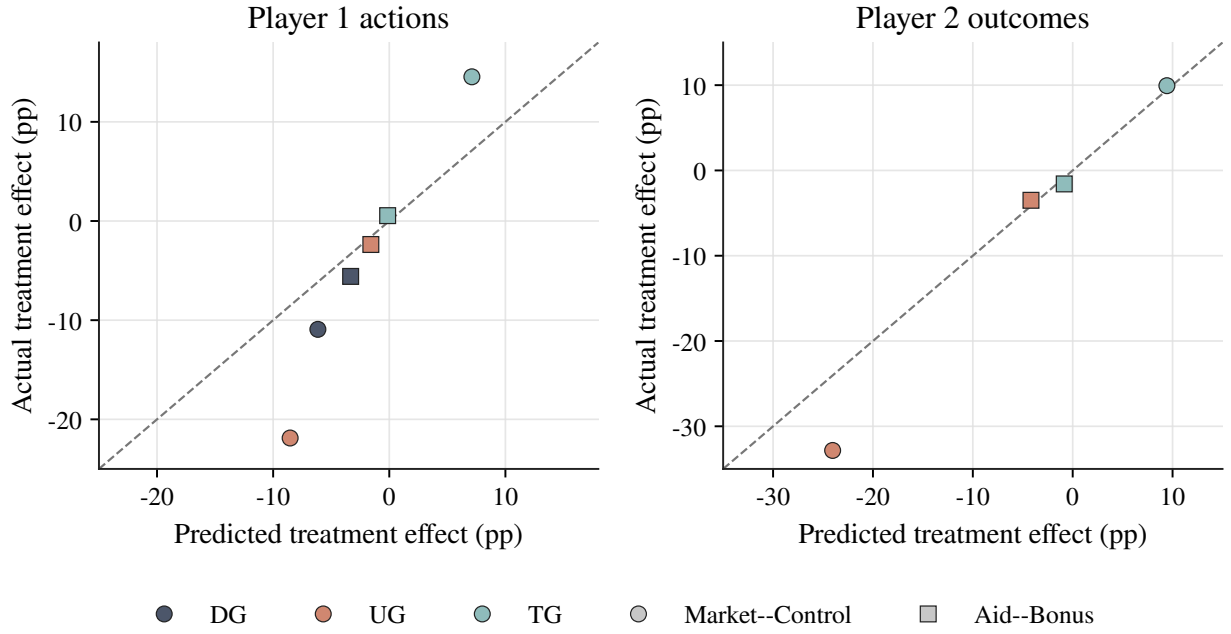


Figure 15: Predicted versus actual treatment effects on actions when the choice functions are estimated on all four conditions pooled within each game and player. Outcomes, regressors, and treatment contrasts are as in Figure 8; the fitted Player 1 action used for Player 2 is also estimated on all four conditions. The dashed line marks exact prediction.

these participants. Dropping them moves the benchmark predictions by at most 0.8 points and the effects of Market on reference beliefs by at most 0.3 points, and changes the share of the Market effect attributed to cell composition in the decomposition from 19% to 16% in UG and from 47% to 59% in TG.

Table 24: Belief-dependent results with and without the one-third choosers

		Full sample		Excluding one-third choosers	
Statistic	Game	Market − Control	Aid − Bonus	Market − Control	Aid − Bo
Observed effect on share sent (pp)	UG	-21.9	-2.4	-21.9	-
Observed effect on share sent (pp)	TG	+14.5	+0.5	+14.5	+
Benchmark prediction (pp)	UG	-10.1	-1.4	-10.7	-
Benchmark prediction (pp)	TG	+7.2	-0.1	+8.0	+
Effect on the reference belief (pp)	UG	+16.9	+1.6	+17.2	+
Effect on the reference belief (pp)	TG	+10.5	-1.2	+10.3	-
Oaxaca share of cell composition	UG	19%	55%	16%	5
Oaxaca share of cell composition	TG	47%	—	59%	
Benchmark observations	UG	1,200	1,200	1,137	1,
Benchmark observations	TG	1,200	1,200	1,011	9

Notes: Player 1. The full sample uses the chosen-action belief of first movers who chose exactly one third as their reference belief (the paper's specification); the other columns drop them, as an earlier version of the analysis did. Observed effects use all responses. Benchmark predictions apply the coefficients of a regression of the share sent on category indicators and the reference belief, estimated on the two compared conditions pooled, to the change in regressor means (the Player 1 benchmark of Figure 8). Belief effects are mean differences across all responses. Oaxaca shares are the representation component of the symmetrized decomposition with category × belief-tercile cells (Table 25); a dash marks a TG story effect too small for an informative share.

D.2 Symmetrized Decompositions of Mean Treatment Effects

The pre-registered analysis specified a complementary, regression-free decomposition of mean treatment effects into changes in the distribution of representations and changes in behavior conditional on representation. We use a symmetrized Shapley/Oaxaca–Blinder decomposition, which averages the two possible orders of changing composition and conditional behavior.

We implement the decomposition twice in UG and TG. Representation cells combine the stated category with one of three quantile groups of either the hypothetical belief at the common one-third reference action or the chosen-action belief. DG has no belief elicitation, so both specifications use categories alone. Both specifications use the same classified sample, so the two columns differ only in the belief used to define cells.

The pre-registrations specify representation categories alone for UG and TG and social-proximity levels for the dictator game. The belief groups extend those specifications to include the representation's second measured component. Applying the specifications literally on the same samples assigns 12% of the Market effect and 51% of the Aid–Bonus effect in UG to category composition. Dictator-game cells defined by social proximity assign 36% of the Market effect and 50% of the story effect to composition. In the TG Market comparison, categories alone account for 25% of the effect, compared with 47% when hypothetical beliefs are included and 57% when chosen-action beliefs are included.

Table 25 reports both category-and-belief decompositions. With hypothetical beliefs, representation composition accounts for 19% of the Market effect in UG, 47% in TG, and 55% of the story effect in UG. The corresponding shares with chosen-action beliefs are 12%, 57%, and 66%. The dictator-game shares are 39% for Market and 65% for Aid versus Bonus under both definitions.

Chosen-action beliefs are evaluated at an endogenous action, so their decomposition is a descriptive robustness exercise rather than a clean split of an upstream representation channel. Under either definition, Market primarily changes UG behavior within representation

cells, while the belief component contributes substantially more to the TG Market effect. The TG story effect is only +0.8 percentage points in this sample, making the corresponding percentage decompositions uninformative.

Table 25: **Symmetrized Decomposition of Treatment Effects: Hypothetical and Actual Beliefs**

			Hypothetical beliefs			Actual beliefs			
Comparison	Game	Δ mean	Repr.	Behav.	Repr. %	Repr.	Behav.	Repr. %	N
Market vs Control	DG	-0.101	-0.039	-0.061	39%	-0.039	-0.061	39%	728
		(0.018)	(0.016)	(0.020)		(0.016)	(0.020)		
Market vs Control	UG	-0.210	-0.040	-0.170	19%	-0.025	-0.185	12%	1083
		(0.012)	(0.012)	(0.017)		(0.011)	(0.017)		
Market vs Control	TG	+0.150	+0.070	+0.080	47%	+0.085	+0.065	57%	1147
		(0.016)	(0.011)	(0.016)		(0.011)	(0.015)		
Aid vs Bonus	DG	-0.059	-0.039	-0.021	65%	-0.039	-0.021	65%	790
		(0.013)	(0.010)	(0.008)		(0.010)	(0.008)		
Aid vs Bonus	UG	-0.024	-0.013	-0.011	55%	-0.016	-0.008	66%	1175
		(0.007)	(0.003)	(0.006)		(0.004)	(0.006)		
Aid vs Bonus	TG	+0.008	+0.006	+0.002	—	-0.002	+0.010	—	1177
		(0.017)	(0.011)	(0.014)		(0.012)	(0.012)		

Notes: Symmetrized (Shapley/Oaxaca–Blinder) decomposition of the difference in mean Player 1 actions between conditions into a component due to the distribution of representation cells and a component due to behavior conditional on the cell. In the ultimatum and trust games, cells interact the stated category with a quantile group of the belief; cutoffs are computed within game on the two conditions compared, with tied values kept together. The first specification uses hypothetical beliefs at the common one-third action. The second uses actual (chosen-action) beliefs; for trust-game senders who chose zero, this belief refers to the hypothetical case of sending one sixth. Both specifications use the same classified sample, so the mean difference and N are common across columns. Participants who chose exactly one third were not asked the hypothetical question separately; their chosen-action belief serves as their belief at the one-third action. The dictator game has no belief elicitation, so both specifications use categories alone and are identical. Actual beliefs condition on an endogenous chosen action; their decomposition is therefore descriptive and does not isolate an upstream representation channel. Cells empty in one condition contribute their observed mean to the representation component. Bootstrap standard errors in parentheses ($B = 1{,}000$ participant-level draws within game × condition strata; belief groups and both decompositions recomputed in each draw). Representation shares are point estimates; they are not reported for the trust-game story comparison, where the total effect is +0.008 (bootstrap SE 0.017).

D.3 Distributional Accounting and the DG–LT Comparison

Table 26 reports predicted and observed changes in the mean and dispersion of Player 1 choices across all eight game-by-comparison cells. The Market comparison includes both dictator games, complementing the DG–LT reversal in Section 5.2: reweighting category-conditional behavior by the observed category shares predicts a Market effect on the mean transfer of −9.9 percentage points in LT against −4.5 in DG, compared with observed effects of −21.5 and −10.1. Figure 9 in the main text displays the principal bin-level shifts for DG, UG, and TG.

Table 26: **Predicted versus Actual Treatment Effects on Means and Dispersion**

		Δ Mean (pp)		Δ SD	
Comparison	Game	Predicted	Actual	Predicted	Actual
Aid vs Bonus	DG	-3.9	-5.9	+0.022	+0.026
	LT	-4.0	-5.0	+0.023	+0.028
	UG	-1.3	-2.4	+0.017	+0.048
	TG	+0.3	+0.8	+0.001	+0.005
Market vs Control	DG	-4.5	-10.1	+0.019	+0.059
	LT	-9.9	-21.5	+0.010	+0.003
	UG	-7.3	-21.0	+0.043	+0.149
	TG	+4.2	+15.0	+0.019	+0.003

Notes: Player 1, classified responses only. Predicted values reweight the category-conditional action distributions (means and variances pooled across the two conditions of each comparison) by each condition's category shares; the predicted effect is the mixture implication of the representation shift alone. Actual values are the raw between-condition differences.

E Experimental Instructions

E.1 Dictator Game, Abstract Frame (KW, with LT Variants in Brackets)

The LT version differs only where indicated: the bracketed text marked LT replaces the amount or the paragraph that precedes it.

> In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. You and your match are provisionally allocated a certain amount of dollars, adding to a total of **$12**. You will have the chance to decide **whether and how to change this original allocation**.
>
> The allocation chosen will be implemented for **1 participant out of 10**. **This means that your decisions affect a bonus payment you and the other participant may receive. It is thus in your best interest to choose the allocations that you truly prefer**.
>
> [new page]
>
> You were provisionally allocated **$12** [LT: **$8**]. The other participant was provisionally allocated **$0** [LT: **$4**].
>
> You have the option to make a transfer to the other participant. You can transfer any amount between $0 and $12. [LT: You can transfer any amount between −$4 and $8. If you choose a negative dollar transfer to the other participant, this means that you take that amount *from the other participant*.]
>
> We now present four hypothetical final allocations and we ask you to describe a real-world situation that you find similar to each. The situation can involve two people or groups. It can be an experience you or somebody you know lived, or it could be fictional.
>
> Then, you may choose the amount of your transfer.

E.2 Dictator Game, Market Frame (KW, with LT Variants in Brackets)

The LT version differs only where indicated: the bracketed text marked LT replaces the amount or the paragraph that precedes it.

> In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. You take the role of **data provider (producer)**, and the other participant takes the role of **broker (intermediary)**. As a data provider, you produce survey data through online questionnaires like the one you are in. A computer will randomly determine an initial price for your data. You will then have the opportunity to **adjust this price** before it is sent to the broker.
>
> The broker will sell your data to the research team for a fixed price of \$12. You will receive the final price you set. Your decision affects a bonus payment. Specifically, 1 in 10 participant pairs will be randomly selected, and for those selected:
>
> - The **data provider (you)** receives a bonus payment equal to the final price you set.
> - The **broker** receives \$12 from the research team for the sold data and pays the provider the set price, keeping the difference.
>
> [new page]
>
> The computer selected a **starting price of \$12** [LT: **\$8**] for the data.
>
> You may change the price by any amount between −\$12 and \$0. A negative change means that you decrease the price. [LT: You may change the price by any amount between −\$8 and +\$4. A positive change means that you increase the price, while a negative change means that you decrease the price.]

E.3 Ultimatum Game, Abstract Frame

In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. One of you will be assigned to be Player 1, and the other will be assigned to be Player 2.

Player 1 is provisionally allocated $12. **Player 1** will have the chance to decide **whether and how to change this original allocation**.

After seeing the proposed allocation, **Player 2** will decide **whether to accept or reject it**. If Player 2 **accepts**, the proposed **allocation will be implemented**. If Player 2 **rejects**, **both players will receive $0**.

The final allocation will be implemented for 1 participant out of 10. This means that your decision affects a bonus payment you and the other participant may receive.

[new page]

Please choose how much of the $12 you would like to transfer to Player 2.

If the transfer is accepted, it will be implemented.

E.4 Ultimatum Game, Market Frame

In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. One of you will be assigned to be **data provider (producer)**, and the other will be assigned to be **broker (intermediary)**.

The data provider produces survey data through online questionnaires like the one you are in. A computer determines an **initial price** for your data. If you are assigned the role of **data provider**, you will have the opportunity to **adjust this initial price** before it is sent to the broker.

The broker can sell the data to the research team for a fixed price of **$12**.

After seeing the proposed price set by the data provider, the **broker will decide whether to accept or reject it**.

- If the broker **accepts**, the data will be sold for $12.
- The **data provider** receives a bonus equal to the final price set.
- The **broker** receives $12 from the research team, pays the provider the set price, and keeps the difference.
- If the broker **rejects**, the sale does not take place and **both participants receive $0**.

The final outcome will be implemented for 1 participant pair out of 10. This means that your decision affects a bonus payment you and the other participant may receive.

[new page]

The computer selected a **starting price of $12.**

You may change the price by any amount between −$12 and $0. A negative change means that you decrease the price.

If the price is accepted by the broker, the sale will be implemented.

E.5 Trust Game, Abstract Frame

In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. One of you will be assigned to be **Player 1**, and the other will be assigned to be **Player 2**.

Player 1 will receive $6, and Player 2 will receive $0. **Player 1** will decide **how much of the $6 to send** to Player 2. **Any amount sent will be multiplied by 3** before being received by Player 2.

After receiving this amount, **Player 2** will decide **how much of it to send back to Player 1**. The amount sent back will not be changed.

The final allocation will be implemented for 1 participant out of 10. This means that your decision affects a bonus payment you and the other participant may receive.

[new page]

Please choose how much of the $6 you would like to send to Player 2. The remaining amount will be yours.

Any amount you send will be multiplied by 3 before being received by Player 2.

Player 2 will then decide how much of the amount received to send back to you.

E.6 Trust Game, Market Frame

In this task, you are matched with another Prolific participant. You will never know the identity of the person with whom you are paired, and this person will never know your identity. One of you will be assigned to be the data provider (producer), and the other will be assigned to be the broker (intermediary).

The data provider produces survey data through online questionnaires like the one you are in. The data provider has produced 6 units of data, each valued at $1 by the research team, for a total of $6.

If you are assigned the role of data provider, you will decide how many units to sell to the research team for $1 per unit, and how many units to transfer to the broker.

The broker has access to a specialized buyer who is willing to pay $3 per unit, so the broker's revenue is 3 times the number of units you send.

If you are assigned the role of broker, you sell the units received to the high-value buyer, and your revenue is three times the units of data sent to you. After the sale, as a **broker you will decide whether and how much of the revenue to transfer back to the data provider.**

The final outcome will be implemented for 1 participant pair out of 10. This means that your decision affects a bonus payment you and the other participant may receive.

[new page]

Please choose how much of the $6 worth of data you would like to send to the broker. The remaining amount will be sold to the research team and you will cash in its value.

Any amount you send will be multiplied by 3 before being received by the broker and sold to the research team.

The broker will then decide how much of the revenue received to send back to you.

E.7 Stories

Bonus.

After weeks of intense preparation, Steve was thrilled to learn that his team was awarded a performance bonus for exceeding their sales target. The recognition felt satisfying—a tangible reward for the long hours he'd poured into work and the many nights he'd stayed late fine-tuning his pitch. The director called Steve to discuss the bonus and said: "You guys did great. Since you are the team leader, the full bonus is allocated to you. However, you may choose to give a portion to your teammate Max if you wish. The final decision is yours."

Now Steve faced a choice. He could keep the full bonus, as initially allocated. Or he could give some to Max, knowing that Max made a significant contribution to the team's success. What should he do?

Aid.

James was going through a rough financial patch. He struggled to make ends meet and had no savings left. One morning, he received an unexpected email from a government welfare agency he had applied to months ago. "Congratulations! You have been selected for financial assistance. You will be receiving the full aid amount to help with your expenses."

As James went to the agency to collect, he found Mark, a neighbor. He learned that Mark also applied but wasn't selected. Only James received the funds.

Now James faced a choice. Should he keep all the money for himself? James was tempted because he really needed the money, and Mark was just an acquaintance. Or should he give some of it to Mark, given that he was struggling? What should he do?