Cardiff Economics Working Papers



Working Paper No. E2025/10

Unrewarded Cooperation

Arkady Konovalov, Daniil Luzyanin, and Sergey V. Popov

March 2025

ISSN 1749-6010

Cardiff Business School
Cardiff University
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
business.cardiff.ac.uk

This working paper is produced for discussion purpose. These working papers are expected to be published in due course, in revised form.
Cardiff Economics Working Papers are available online from:
http://econpapers.repec.org/paper/cdfwpaper/ and
https://carbsecon.com/wp/index_abstracts.html
Enquiries: EconWP@cardiff.ac.uk

Unrewarded Cooperation*

Arkady Konovalov Daniil Luzyanin Sergey V. Popov

March 27, 2025

Abstract

Experiment participants in a social dilemma game choose cooperation over defection, even though neither is more beneficial. High levels of cooperation cannot be explained by favorable labels for actions, collusion, k-level reasoning, quantal response behavior, or misplaced optimism about others' actions, but can be rationalized by the Charness and Rabin (2002) preference model. However, cooperation rates fall with changes in payoffs, which cannot be explained by the standard formulation; to account for these results, we introduce a generalization of the model.

Keywords: cooperation; coordination; social preferences.

JEL: C7; C9.

*Konovalov: University of Birmingham, a.konovalov@bham.ac.uk. Luzyanin: University of Birmingham, dan.luzyanin@gmail.com. Popov: Cardiff University, PopovS@cardiff.ac.uk, corresponding author. Popov thanks Cardiff Business School's Seedcorn grant scheme for funding and ethics committee approval. Audiences at Cardiff, Birmingham, Monash, Sydney, Indiana, Illinois, and ECARES gave useful feedback; individual discussions with Chris Colvin, Aleksei Chernulich, Jason Shachat, Ted Turocy, Stefan Penczynski, Arnold Polanski, Michele Garagnani, Birendra Rai, Mallory Avery, Murali Agastya, Jason Tawaya, Andreas Ortmann, Konstantinos Ioannidis, Daniela Puzzello, Ala Avoyan, Caitlyn Trevor, and many many others are greatly appreciated. The usual disclaimer applies.

Cooperation generally refers to the choice to act in a way that benefits the collective, even if it means forgoing potential individual gains. The Prisoner's Dilemma game, where defection is the dominant strategy while mutual cooperation is the Pareto-optimal outcome, is a common example of this trade-off. Incentive structures can influence whether players are more likely to cooperate or defect. Intuitively, increasing the mutual cooperation payoff tends to increase cooperation rates, while a higher "temptation" payoff (reward for one-sided defection) decreases cooperation (Charness et al., 2016, Gächter et al., 2020). If the game is repeated, beliefs about future cooperation can influence choices (Aoyagi et al., 2024, Fréchette and Yuksel, 2017). Significant literature in economics and psychology has shown that intrinsic individual factors, such as social preferences (including altruism), a desire for fairness, or a sense of reciprocity, can also lead people to cooperate (Blanco et al., 2011, Fehr and Charness, forthcoming).

In 2023, a viral Twitter poll offered a simple game: "Choose between a blue pill or red pill. If $> 50\%$ of people choose blue pill, everyone lives. If not, red pills live and blue pills die. Which do you choose?".[1] The results were counter-intuitive to many: 65% of more than 68'000 respondents chose the blue pill. The poll was successfully replicated many times with very similar results.[2] This strategy profile is a Nash equilibrium, as any outcome where the majority almost surely chooses the blue pill. However, the "safe" and seemingly intuitive equilibrium is for every individual to pick the red pill as it guarantees the better outcome independent of everyone else's choices.

There are many potential explanations for choosing to cooperate when there is no benefit in doing so. These include misplaced optimism, risk attitudes, stochastic choice (quantal response), and inequality aversion. To explore potential causes, we use an online laboratory game that mimics the original poll using incentivized choices and small monetary stakes.

		Other players	
		$> 50\% - 1$ cooperate	Otherwise
Player i	Cooperate	R	C
	Defect	R	R

Table 1: The structure of the game. Payoffs for each player with reward R and cooperation failure payoff C.

[1] https://x.com/lisatomic5/status/1690904441967575040

[2] https://www.reddit.com/r/polls/comments/16xpm2m/which_button_will_you_press/,https://www.reddit.com/r/WouldYouRather/comments/12xua1r/presented_with_two_buttons_red_and_blue_if_more/, https://www.reddit.com/r/polls/comments/12t619s/in_front_of_you_appears_red_and_blue_button_if/

We formalize the decision problem in the following way. Consider a game with $N > 2$ players (let us assume N is odd), where each player can either support an action that supports the social welfare (Cooperate) or choose a safe option that only guarantees their own benefit (Defect) (Table 1). Defectors always get a payoff of £R. Cooperators get £R only if more than half of the players cooperate. If less than half of the players cooperate, cooperators receive £C.

This game is close in structure to common social dilemmas such as the Prisoner's Dilemma (PD), public goods game, and Stag Hunt (SH) game, which illustrate various cooperation and coordination problems. An uprising against a totalitarian government can only be successful if enough people participate; a chess championship will be enjoyable if enough people partake; a seminar would be useful if enough people attend. However, in the games traditionally studied in the literature, individual payoffs for cooperation and defection typically differ for different participants and/or outcomes; this discrepancy emphasizes the trade-off between the individual and common goals. In the game we study, the payoff structure offers little space for such trade-offs. First, defection is not additionally rewarded compared to cooperation; in the standard PD game, the temptation payoff for one-sided defection is higher than the reward for mutual cooperation. Second, collective cooperation is not more beneficial than collective defection; in the standard PD game, mutual cooperation is rewarded.

Multiple Nash equilibria arise in this game, including (a) an equilibrium where everyone cooperates; (b) an equilibrium where everyone defects; (c) an equilibrium where one player defects and the rest cooperate; (d) an equilibrium where one player defects with probability p and the rest cooperate, et cetera. There can be other equilibria depending on the number of participants. Any equilibria with a positive probability of cooperation need certainty about the cooperation's success; otherwise, defection as a strategy dominates as cooperation can yield at most as much as defection.

Since before Selten (1975), multiple Nash equilibria in games lead to arguments about why some equilibria are more likely to emerge in the laboratory than others. In this case, equilibrium perfection criteria, such as Trembling Hand Perfection, support the idea that universal defection should be the most intuitive outcome. In all other equilibria, even when strategies are perturbed even slightly, the probability of cooperation will go below certainty. For each player, cooperation would then yield less than defecting.

One potential rationalization of this behavior that is not purely game-theoretic is social preferences. If the individual's utility function takes others' outcomes into account,

cooperation can emerge under certain combinations of beliefs and other-regarding preferences. Many utility functions have been suggested over the years to model such preferences, including the Fehr-Schmidt utility function (Fehr and Schmidt, 1999), Charness-Rabin preferences (Charness and Rabin, 2002), and Bolton-Ockenfels preferences (Bolton and Ockenfels, 2000). All of these share a common approach where others' outcomes linearly enter the individual's utility function. If individuals exhibit such preferences (to a certain degree, as defined by the utility function's parameters), and their beliefs about others being cooperative are sufficiently high, cooperation can emerge in the game where cooperation is unrewarded.

Our contribution to the literature is as follows. First, the results of our incentivized experiment replicate the online polls. We find that 62% of our participants and 90% of groups cooperate. We repeat the game ten times within the same group of participants, and cooperation sustains. We find that Charness-Rabin other-regarding preferences can explain the average levels of cooperation.

Second, we manipulate the rewards for cooperation and defection. To identify whether cooperation can be driven by inequality aversion, we introduce two alternative treatments: High Reward, where the reward is doubled (£2 instead of £1), thus increasing inequality in the outcomes; and Low Cost, where non-cooperation leads to an outcome of £0.5 instead of £0 (thus decreasing inequality). Surprisingly, we find that both treatments significantly decrease cooperation rates (to 55-58% at the group level) which is not consistent with the standard multiplayer Charness-Rabin model.

Finally, we generalize the Charness-Rabin model to rationalize these results by introducing (a) restrictions on the individual utility function and (b) aggregation of the others' payoffs. We find that under specific restrictions, the decrease in cooperation rates in both Low Cost and High Reward treatments can be rationalized.

We follow with a literature review, theoretical foundations, experimental design and the outcomes, then rationalize our findings by extending the Charness and Rabin (2002) model, and conclude with a summary and discussion of implications.

Literature

One similar game is Stag Hunt, where participants allocate limited resources between a public good activity and a private good activity. The key trade-off in these games, in the words of Silva (2024), is the "tension between a payoff superior option (stag) and a less

risky but payoff inferior alternative (hare)". In our game, which can be viewed as a multiplayer version of the SH game, the stag is not more valuable than the hare, so there must be other reasons why experiment participants would choose stag. Dal Bó and Fréchette (2018) argue that stag is a more likely outcome if the stag action is risk dominant, but this is not the case in our environment.

Some experiments use a richer decision space for individual effort choice (Van Huyck et al., 1990), while our action space is binary. The common finding is that participants fail to cooperate efficiently (see the overview in Devetag and Ortmann, 2007), even though cooperation is more attractive than defection. The closest paper to ours is Shurchkov (2013), which considers bank runs. The public good of resisting bank-running, if enough people participate in its production, yields more than participation costs. The paper investigates how participants respond to signals about the difficulty of cooperation and find that they cooperate more than the theory predicts. In our case, cooperation cannot possibly yield more monetary gain than defection, so no amount of misplaced optimism can prevent the defection outcome. Therefore, failure to form expectations correctly is not the driving force behind our excessive participation result.

Gueye et al. (2020) study a public good provision game where one of the player types earns the best possible payoff safely by defecting, and participation in a public good game can yield a worse payoff if other types do not participate. The paper finds that such type contributes in 26% of experiments, especially if payoffs are very unequal. Gueye et al. (2020) argue that what drives this type is maximizing the total payoff instead of individual payoff for other-regarding-preferences reasons. In our setting, however, the total payoff is certain and is already maximized if all players are defecting, and therefore total payoff maximization is unlikely to be the driver behind our excess cooperation.

Others, including Rankin et al. (2000), find excessive cooperation but explain it with the establishment of institutions or non-trivial experimental design features (cf Van Huyck et al., 2018). Cabrales et al. (2007) show that participants do not repeatedly eliminate dominated strategies and do not rely on the optimality of other participants' behavior. Our game does not require high reasoning power from the participants, and insufficient sophistication would only lead to less cooperation, not more.

To explain our results, we turn to theories of social preferences, which introduced other-regarding preferences that suggest why participants might choose actions that are not individually beneficial due to concern for others' wellbeing (Bolton and Ockenfels, 2000, Charness and Rabin, 2002, Fehr and Schmidt, 1999). We find that inequality aversion

cannot be the primary driver of the observed effects, and we build upon the multiplayer version of the Charness-Rabin preferences as it offers the most flexible framework for our case.

Theoretical Framework

We use the multiplayer version of the other-regarding preference model developed by Charness and Rabin (2002). There are N players in the game. Let π_i denote player i's monetary payoff. The player i's utility is

$$V_i(\pi_1, \pi_2, .., \pi_N) = (1-\lambda)\pi_i + \lambda \left(\delta \times \min_{j \in \{1..N\}} (\pi_j) + (1-\delta) \times \sum_{j \in \{1..N\}} \pi_j \right). \quad (1)$$

The individual-level parameter λ reflects the importance of others' outcomes for the player ($\lambda = 0$ reflects strictly selfish behavior, and $\lambda = 1$ is pure altruism). The individual parameter δ weighs the lowest payoff against the total payoff.

We illustrate the rationalization using the case $N = 7$ as implemented in our experiment, which is described in detail in the next section. If four or more players cooperate, everyone receives a payoff of R. If three or fewer players cooperate, they receive C. Defectors always receive a payoff of R. The minimal and total payoff depend on the total number of cooperators (Table 2).

If players knew the choices of the others in advance, their choices would be straightforward. As the game is simultaneous, we define the players' *a priori* beliefs. Let players believe that other players are playing the same mixed strategy independently. Let p denote player i's belief that any other player will cooperate. The binomial distribution formula provides the probability of the cooperative outcome:

$$P_Q = P\left[Q \text{ Cooperators among others} | p\right] = C_6^Q p^Q (1-p)^{6-Q}. \quad (2)$$

The benefit of defection is a higher individual payoff when there are few other cooperators. The benefit of cooperation is a higher payoff when there are exactly three other cooperators. In the mixed strategy equilibrium, cooperation and defection yield the same expected payoff. Combine (1) and (2):

$$(1-\lambda)(R-C)(P_0+P_1+P_2) - \lambda\delta\left(P_3 - P_0\right)(R-C) = \lambda(1-\delta)\left(P_0 + P_1 + P_2 - 3P_3\right)(R-C). \quad (3)$$

# of other cooperators	π_i	$\min_j \pi_j$	$\sum_j \pi_j$
Cooperate			
0	C	C	$6R + C$
1	C	C	$5R + 2C$
2	C	C	$4R + 3C$
3	R	R	$7R$
4	R	R	$7R$
5	R	R	$7R$
6	R	R	$7R$
Defect			
0	R	R	$7R$
1	R	C	$6R + C$
2	R	C	$5R + 2C$
3	R	C	$4R + 3C$
4	R	R	$7R$
5	R	R	$7R$
6	R	R	$7R$

Table 2: Monetary payoffs for player i for both cooperation and defection.

Cancel $(R - C)$ to obtain

$$(1 - \lambda\delta)(P_0 + P_1 + P_2) = \lambda\delta\,(P_3 - P_0) + \lambda(1 - \delta)3P_3. \tag{4}$$

If $p \approx 0.7$ (as observed in the survey data), an indifferent agent would have $\lambda = \frac{2610}{20580 - 11137\delta}$ (see Figure 1 for representation of threshold values of λ and δ for various levels of p). Higher λ than the threshold value would lead to preferring cooperation over defection. For any level of p and δ there is a positive λ that rationalizes cooperation, with the threshold value of λ decreasing as p increases. Importantly, the values of R and C do not affect the threshold position.

Experimental Design

In each round of the experiment, all participants in a group of seven (we used an odd number to avoid ties and the highest number supported by the online platform Gorilla) chose between two abstract shapes, a circle and a triangle, presented on their computer screens.

The "safe" option yielded a guaranteed reward £R. The "cooperative" option yielded

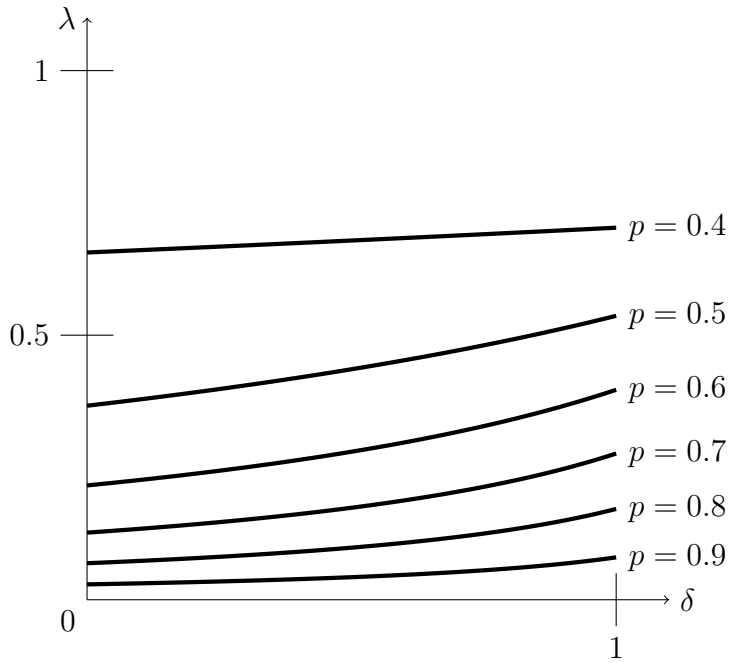


Figure 1: Combinations of λ and δ that rationalize mixed strategies consistent with different levels of p.

the reward only if four or more participants in the group chose that option. If fewer than four participants chose the cooperative option, all of them received a smaller payment £C <£R, and participants who defected received £R. We counterbalanced the action labels (circle and triangle) for the underlying actions (cooperate and defect) across groups. There were no significant differences between the two labels, so all results report pooled outcomes across the two shapes.

Once the participant made their choice, they observed how many participants in the group chose each option, their reward for the round, and the sum of their own earnings across all completed rounds (Figure 2). The participant could not track individual choices and rewards across rounds and only observed the group outcomes. After feedback, the game was repeated.

Each group completed 10 rounds of the game using the same group of participants. At the end of the experiment, the participants were paid the sum of their earnings in all rounds and completed a short demographic questionnaire.

We used Gorilla (`https://gorilla.sc/`) to program the web-based interface. Running the study online rather than in person minimized the possibility of potential previous interactions and ensured complete anonymity. We recruited the participants via Prolific preregistration in groups of seven.

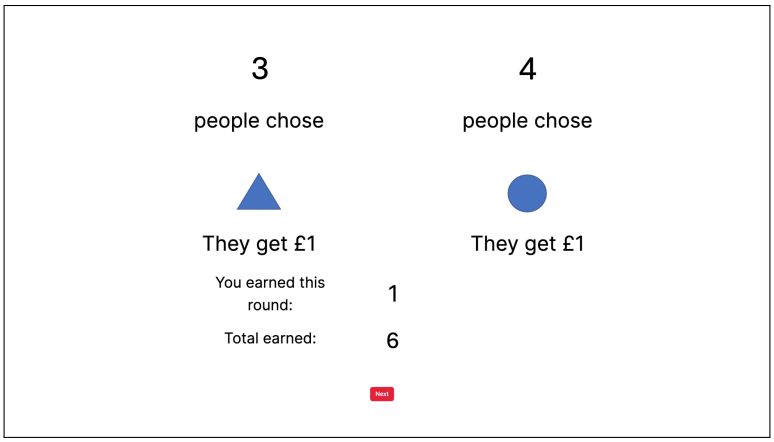


Figure 2: Experimental interface: feedback screen.

Participants

We conducted the experiment in June and July 2024. We recruited 98 participants (age range 19-43, mean age 28.3, 53 identified as women, 41 as men, 3 as non-binary) using Prolific.

All participants provided written informed consent, and the study was approved by the Cardiff Business School Research Ethics Committee. Each session lasted about 10 minutes, and participants earned £12.44 on average (roughly 8 times higher than the average hourly Prolific rate at the time of the study, and approximately 2.5 times the price of the Happy Meal in a UK McDonald's), including a £2 fixed participation fee. The experimental instructions are included in Appendix A.

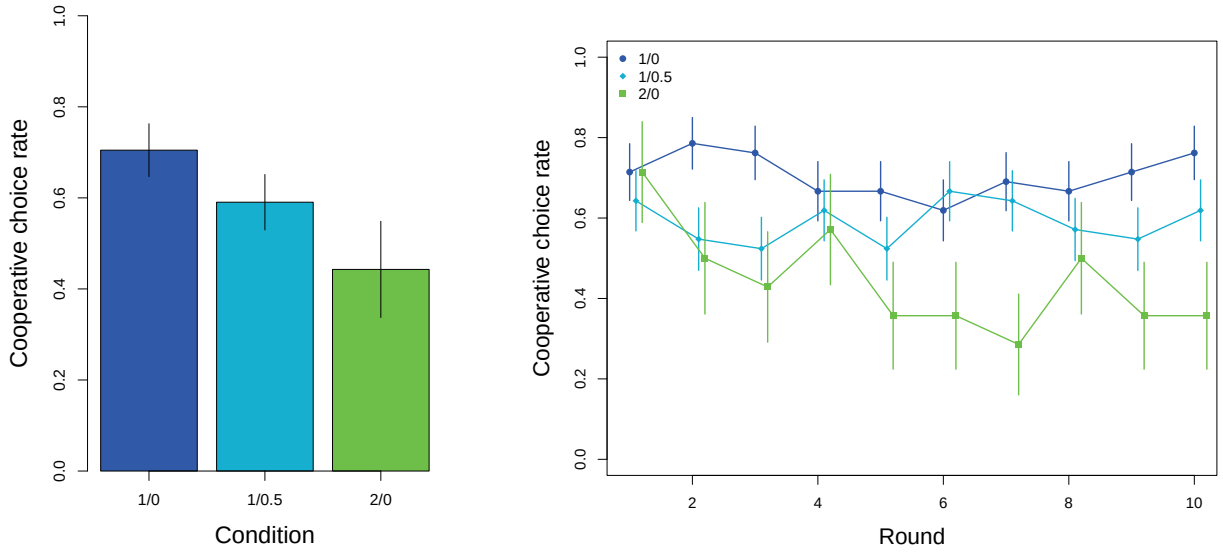
Treatments

To test whether cooperative behavior is driven by inequality aversion, across groups, we manipulated the reward R and losing payment C using the following treatments:
Baseline ($\{R, C\} = \{1, 0\}$, 42 participants): defection yielded £1, successful cooperation yielded £1, failed cooperation yielded £0.
Low Cost ($\{R, C\} = \{1, 0.5\}$, 42 participants): defection and successful cooperation yielded £1, failed cooperation yielded £0.5.
High Reward ($\{R, C\} = \{2, 0\}$, 14 participants): defection and successful cooperation yielded £2, failed cooperation yielded £0.

We avoided negative payments to avoid loss aversion effect (e.g., Rydval and Ortmann (2005) document that losses motivate participants to cooperate more).

Results

Participants cooperate



(a) Overall cooperation split by treatment

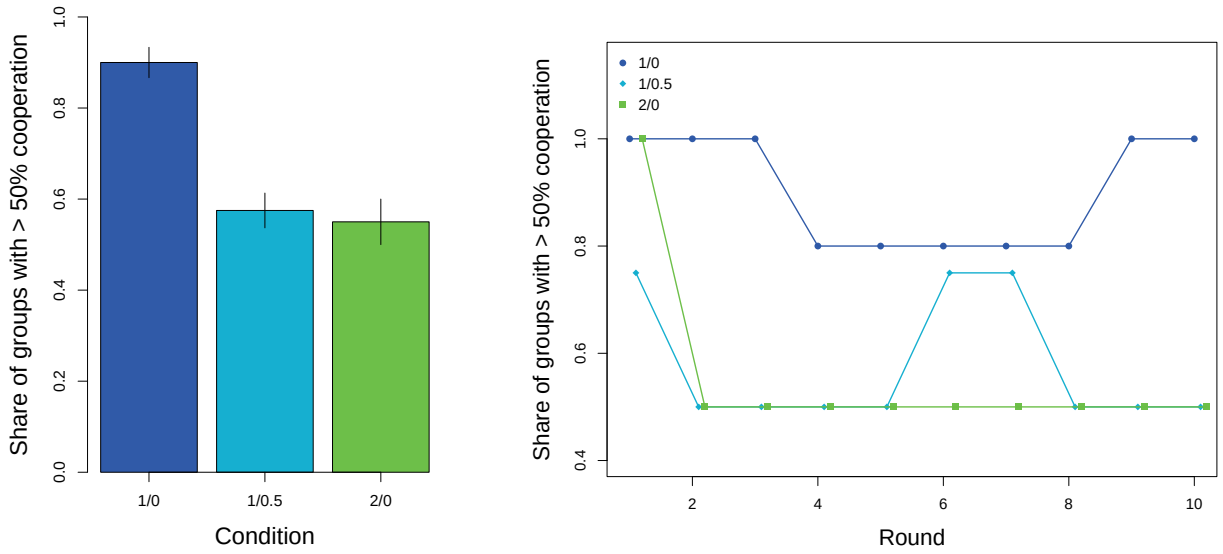
(b) Cooperation rate dynamics across the 10 rounds of the experiment

Figure 3: Individual cooperation rates by treatment (1/0 = Baseline, 2/0 = High Reward, 1/0.5 = Low Cost) and round (1-10). Error bars denote standard errors at the participant level.

First, we review the general cooperation rates. In all sessions and all rounds, participants cooperated in 62% of rounds. Only 18% of participants never cooperated, 36% always cooperated, and the rest changed their behavior across rounds.

Both alternative treatments decreased cooperation rates (Figure 3a). In the Baseline treatment, participants cooperated in 70% of rounds. Increasing the outcome for the defection option (Low Cost, 1/0.5) decreased cooperation rates to 59%, and increasing the cooperation reward (High Reward, 2/0) decreased cooperation rates to 44%. Only the difference between 1/0 and 2/0 treatments was statistically significant (see Table 3).

We then analyzed the number of groups that achieved the cooperative outcome (rounds where more than 50% of participants chose to cooperate; see Figure 4). In the Baseline treatment, 90% of groups cooperated in all 10 rounds of the game, and the rates were 58% and 55% in the Low Cost and High Reward treatments, respectively (Figure 4a). The differences between the Baseline and alternative treatments were statistically significant (using group clustering, $p < 0.001$).



(a) Share of cooperative groups split by treatment

(b) The dynamics of group cooperation across the 10 rounds of the experiment

Figure 4: Share of groups where more than 50% of participants cooperated, split by treatment (1/0 = Baseline, 2/0 = High Reward, 1/0.5 = Low Cost) and round (1-10). Error bars denote standard errors at the group-round level.

Cooperation persists

In each session, the experiment lasted 10 rounds. Although most groups cooperated in round 1, some participants then experimented with their choices in the subsequent rounds. Between the first and second halves of the experiment, 63% of participants maintained a stable cooperation rate, 18% increased their cooperation, and 18% decreased their cooperation rates.

In round 1, 68% of participants cooperated across all experimental treatments. In the Baseline treatment (1/0), 71% of participants cooperated; in the Low Cost treatment (1/0.5), 64% of participants; in the High Reward treatment (2/0), 71% of participants. The differences between conditions were not statistically significant (Figure 3b).

In later rounds, individual cooperation rates changed across conditions (Figure 3b). Overall, cooperation rates remained stable in the Baseline and Low cost treatments and decreased over time in the High Reward treatment (Figure 3b). In the High Reward (2/0) treatment, the cooperation rates decreased over time (Figure 3b), but this decrease was not statistically significant (see Table 3, column 3).

Analyzing at the group level, cooperation rates were high in the first round (in the Baseline and High Reward treatment all groups reached cooperation), but then the share

	(1)	(2)	(3)
(Intercept)	0.70^{***}	0.73^{***}	0.72^{***}
	(0.06)	(0.06)	(0.06)
1/0.5 (Low Cost)	-0.11	-0.11	-0.14
	(0.08)	(0.08)	(0.09)
2/0 (High Reward)	-0.26^{*}	-0.26^{*}	-0.12
	(0.12)	(0.12)	(0.13)
Round		-0.00	-0.00
		(0.00)	(0.01)
1/0.5 x Round			0.00
			(0.01)
2/0 x Round			-0.03
			(0.02)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: OLS regressions of cooperation choice on treatment dummy variables and round number.

of cooperative groups dropped in the alternative treatments and remained stable in all three conditions (Figure 4b).

At the group level, about 10% of groups did not reach the cooperative outcome (>50% choosing to cooperate) in round 1. This proportion increased to 28% on average in subsequent rounds and remained stable (Figure 4b). As outlined before, while 80-100% of groups kept cooperating in the Baseline treatment, only around 50% of groups maintained cooperation in the alternative treatments.

Surprisingly, even when the cooperation level in some groups fell below 50%, some participants kept cooperating in the subsequent rounds, going against the implied dynamics in trembling-hand perfection. To check whether previous reward affected the participants' subsequent choices, we regressed their choices on the previous round's reward and their choice in that previous round (clustering standard errors at the participant level). While the coefficient for the previous choice was significant ($p < 0.001$), indicating a high degree of choice perseverance, the past reward did not have a significant impact on the subsequent choice ($p = 0.26$).

Individual demographics do not predict cooperation

We collected basic demographic data on our participants, including age, gender, language, social environment where they grew up (village, small town, city) and self-perceived fam-

ily wealth. We found that none of the demographic variables predicted the individual cooperation rates, both in terms of statistical or economic effects (Table 4, column 1). We found that age (positive) predicted higher overall reward ($p < 0.05$), while growing up in a rural area negatively predicted the overall reward ($p < 0.05$) (Table 4, column 2).

	Cooperation	Earnings
(Intercept)	0.27	0.49
	(0.27)	(0.30)
Age	0.01	0.02*
	(0.01)	(0.01)
Male	0.03	−0.07
	(0.08)	(0.09)
Non-binary	−0.24	0.13
	(0.24)	(0.27)
Bilingual	0.03	−0.14
	(0.12)	(0.13)
Small town	−0.01	−0.29*
	(0.10)	(0.11)
Village	−0.07	−0.23
	(0.15)	(0.17)
Medium Wealth	0.01	−0.13
	(0.11)	(0.12)
High Wealth	−0.18	−0.06
	(0.18)	(0.20)
Very High Wealth	0.19	−0.77
	(0.43)	(0.48)
R^2	0.05	0.17
Adj. R^2	−0.05	0.08
Num. obs.	97	97

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4: OLS regressions of individual cooperation rates and total earnings on demographic variables.

Rationalization

Given that the standard Charness and Rabin's model predicts the same outcomes for our $\{R, C\} = \{2, 0\}$ (High Reward) and $\{R, C\} = \{1, 0.5\}$ (Low Cost) treatments (neither R nor C are a part of Equation (4)), we now introduce a generalization of the model that can explain the decrease in cooperation rate in both alternative treatments.

Let us introduce a modified Charness and Rabin (2002) utility function:

$$V_i(\pi_1, \pi_2, .., \pi_N) = (1-\lambda)\pi_i + \lambda \cdot g\left(\sum_{j \neq i} h(\pi_j)\right). \tag{5}$$

There are two differences introduced here compared to (1). First, π_i does not enter the second term. Now, $\lambda = 1$ denotes self-sacrifice rather than caring about others as much as one would do for themselves as in the original formulation. This change allows us to simplify further derivations. Second, we replace δ with a combination of an aggregator function $g(\cdot)$ and an individual payoff normalization function $h(\cdot)$.[3] For $h(x) = g^{-1}(x) = x^\rho$ we have a standard CES (constant elasticity of substitution) function; for $\rho = 1$ the function is similar to $\delta = 0$ in the original Charness-Rabin formulation, and with $\rho \to -\infty$ this is similar to $\delta = 1$.

Claim 1. *For $h(x) = g^{-1}(x) = x^\rho$, for every $\rho \neq 0$, a change from $R, C = A, B$ to $R, C = qA, qB$ with $A > B > 0$ and $q > 0$ does not change the threshold λ and therefore should not lead to a change in cooperation reported in Figures 3 and 4.*

# of other Cooperators	Cooperate		Defect	
	π_i	$\sum_{j \neq i} h(\pi_j)$	π_i	$\sum_{j \neq i} h(\pi_j)$
0	C	$6h(R)$	R	$6h(R)$
1	C	$5h(R) + h(C)$	R	$5h(R) + h(C)$
2	C	$4h(R) + 2h(C)$	R	$4h(R) + 2h(C)$
3	R	$6h(R)$	R	$3h(R) + 3h(C)$
4	R	$6h(R)$	R	$6h(R)$
5	R	$6h(R)$	R	$6h(R)$
6	R	$6h(R)$	R	$6h(R)$

Table 5: Modified monetary payoffs for player i

Proof. Table 5 shows the monetary payoffs in the modified payoff game. λ that makes participant i indifferent is governed by

$$(1-\lambda)\overbrace{(R-C)(P_0+P_1+P_2)}^{\text{Benefits of Defection}} = \lambda\overbrace{P_3\left(g(6h(R)) - g(3h(R)+3h(C))\right)}^{\text{Benefits of Cooperation}}. \tag{6}$$

Multiply both R and C by q to obtain

[3] This is similar to Andreoni and Miller (2002), where they consider the elasticity of substitution between own and another player's payoff. Our utility function aggregates payoffs of different other players into a monetary equivalent of own payoff.

$$(1-\lambda)q(R-C)(P_0+P_1+P_2) = \lambda P_3\left(g(6h(qR)) - g(3h(qR)+3h(qC))\right). \tag{7}$$

When $g^{-1}(x) = h(x) = x^\rho$, $g(Qh(qR)) = g(q^\rho Qh(R)) = qg(Qh(R))$ for every $Q > 0$, and therefore if λ solves the equation above for some p, it will solve the equation above no matter how R/C scales. □

This shows that the aggregator function $g(y)$ and normalization function $h(x)$ must be more sophisticated to support changes in cooperation rates with changes in payoffs.

Claim 2. *If $g(y) = y^{1/\kappa}$ and $h(x) = x^\rho$:*

- *reconciling the change in payoffs from Baseline to High Reward with a decrease of cooperation requires $\kappa > \rho$;*
- *reconciling the change in payoffs from Baseline to Low Cost with a decrease of cooperation requires $\kappa > 0$.*

Caveat: due to $C = 0$ in the Baseline treatment, constrain $\rho > 0$.

Proof. Using (6) and (7), consider player i whose λ is such that they are indifferent between cooperation and defection in the Baseline scenario:

$$(1-\lambda)(P_0+P_1+P_2) = \lambda P_3\left(g(6h(1)) - g(3h(1)\right). \tag{8}$$

In the High Reward scenario, they must have

$$(1-\lambda)2(P_0+P_1+P_2) \gtrless \lambda P_3\left(g(6h(2)) - g(3h(2)\right) = 2^{\rho/\kappa}\lambda P_3\left(g(6h(1)) - g(3h(1)\right). \tag{9}$$

We observe less cooperation in the High Reward scenario; this means that the left-hand side, representing individual payoff from defection, must become larger than the right-hand side, representing group payoff from cooperation, which implies $2^{\rho/\kappa} < 2$ or $\kappa > \rho$.

In the Low Cost scenario, the same player i who is indifferent between cooperation and defection in the Baseline scenario should have:

$$(1-\lambda)(P_0+P_1+P_2)) \gtrless \lambda P_3\left(g(6h(1)) - g(3h(1)+3h(0.5)\right). \tag{10}$$

The left-hand side is the same as in the Baseline scenario; let us replace it with the right-hand side of the Baseline scenario and cancel out P_3:

$$g(6h(1)) - g(3h(1)) \gtrless g(6h(1)) - g(3h(1) + 3h(0.5)) \Rightarrow g(3h(1) + 3h(0.5)) \gtrless g(3h(1)). \quad (11)$$

Since $h(\cdot) > 0$ for positive arguments, $g(\cdot)$ needs to be increasing for the left hand side to be larger than right hand side; this implies $\kappa > 0$. □

Claim 2 can be generalized: for instance, the argument supports a large class of beliefs, not just based on a polynomial model with respect to p. Quasi-concavity of $g(\sum h(\cdot))$ represents preferences for equality among payoffs of others, in the spirit of but not mathematically equivalent to Fehr and Schmidt (1999); quasi-convexity will lead to a preference for the opposite. More empirical experiments can help design a better rationalization device.

Conclusion

We considered a simple threshold social dilemma game where defection is riskless, and cooperation is not more profitable than defection. We documented a significant deviation from the most plausible outcome of universal defection. Individuals and groups can maintain cooperation even without explicit reward and in the presence of risk. Our design, which included payoff manipulations, allowed us to rule out some common explanations for these results.

One potential trivial explanation for excessive cooperation is noise in participants' decisions; for instance, online participants can be inattentive and always click on the same response or stimulus (e.g., the triangle). To mitigate this possibility, we counterbalanced the stimulus identity across sessions and the positions of circle and triangle across rounds. We observed no preference for any specific stimulus or side of the screen across participants.

Analogously, quantal response equilibrium (McKelvey and Palfrey, 1995) is unlikely to drive the results as well. The certainty equivalent payoff from cooperation is at most as much as for defecting, making the probability of choosing cooperation 50% at most for any quantal response function linking payoffs and probability choices, yet we observe cooperation rates above 50%.

Another similar possibility is bounded rationality, e.g, cognitive sophistication differ-

ences in participants as modeled by level-k reasoning (Gill and Prowse, 2016) or cognitive hierarchy (Camerer et al., 2004). In these frameworks, k=0 decision makers would randomize between defection and cooperation, forcing $k = 1$ decision makers to defect, making it the only possible choice for $k = 2$ and higher. This would imply that the initial cooperation rate should not exceed 50% if the share of level-0 participants is below 100%.

Another potential explanation is misplaced optimism (Shurchkov, 2013). This could imply that participants would update their beliefs after the first round of the game and switch to defection if they observe enough overall defection. However, even with multiple repetitions, our participants sustain cooperation. Additionally, the results are unlikely to be driven by risk attitudes, as defection is first-order stochastically dominant.

Using payoff manipulation, we investigated the possibility that cooperation is driven by inequality aversion; Ramalingam and Stoddard (2024) and Gueye et al. (2020) find that inequality unequivocally hurts cooperation. We introduced two treatments that either decrease (Low Cost) or increase (High Reward) inequality. We found that both actually lead to a decrease in cooperation. This observation suggests that inequality aversion is not the only driver of cooperation in this game.

The difference between the Baseline and Low Cost treatments is thus consistent with previous findings, while the comparison between the Baseline and High Reward is not. Note that both of the alternative treatments weakly increase the payoffs, increasing the kernel of the payoff matrix (see Kendall, 2022). The income effect thus might dominate the inequality aversion effect—which means that the inequality aversion, if present, does not drive the cooperation overall.

To account for these discrepancies, we introduce a generalization of the Charness-Rabin preference function (Charness and Rabin, 2002) that could accommodate these results. We find that under certain restrictions on the individual utility functions and aggregation of the social payoff, it is possible to observe a decrease in cooperation in both conditions. The restriction suggested by the results of the Low Cost treatment requires the aggregate of other agents' payoffs to be an increasing function. This monotonicity is sufficient to explain a decrease in cooperation when C is increasing, since it decreases the losses of other cooperators when the pivotal agent decides to defect instead of cooperating.

The High Reward motivated restriction is less trivial, but still intuitive. The increase in the reward improves the attractiveness of both defection (as the individual monetary payment improves) and cooperation (as the potential bonus collected by other cooperators increases when the deciding agent is pivotal). For cooperation to decrease, the former

effect must dominate the latter. With power functions, since the individual payoff is linear, the others' aggregate utility needs to increase slower than linearly. Exploring the potential limits of these restrictions requires further empirical work.

References

James Andreoni and John Miller. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753, 2002.

Masaki Aoyagi, Guillaume R. Fréchette, and Sevgi Yuksel. Beliefs in repeated games: An experiment. *American Economic Review*, 114(12):3944–3975, 2024.

Mariana Blanco, Dirk Engelmann, and Hans Theo Normann. A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338, June 2011. ISSN 08998256. doi: 10.1016/j.geb.2010.09.008. URL `https://linkinghub.elsevier.com/retrieve/pii/S089982561000148X`.

Gary E Bolton and Axel Ockenfels. Erc: A theory of equity, reciprocity, and competition. *American economic review*, 91(1):166–193, 2000.

Antonio Cabrales, Rosemarie Nagel, and Roc Armenter. Equilibrium selection through incomplete information in coordination games: an experimental study. *Experimental Economics*, 10(3):221–234, September 2007. ISSN 1386-4157, 1573-6938. doi: 10.1007/s10683-007-9183-z. URL `http://link.springer.com/10.1007/s10683-007-9183-z`.

Colin F Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, 119(3):861–898, 2004.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869, 2002.

Gary Charness, Luca Rigotti, and Aldo Rustichini. Social surplus determines cooperation rates in the one-shot Prisoner's Dilemma. *Games and Economic Behavior*, 100:113–124, November 2016. ISSN 08998256. doi: 10.1016/j.geb.2016.08.010.

Pedro Dal Bó and Guillaume R. Fréchette. On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114, 2018. ISSN 0022-0515.

Giovanna Devetag and Andreas Ortmann. When and why? A critical survey on coordination failure in the laboratory. *Experimental Economics*, 10:331–344, 2007.

Ernst Fehr and Gary Charness. Social Preferences: Fundamental Characteristics and Economic Consequences. *Journal of Economic Literature*, forthcoming. ISSN 0022-0515.

doi: 10.1257/jel.20241391. URL https://www.aeaweb.org/articles?id=10.1257/jel.20241391&from=f.

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

Guillaume R. Fréchette and Sevgi Yuksel. Infinitely repeated games in the laboratory: four perspectives on discounting and random termination. *Experimental Economics*, 20 (2):279–308, June 2017. ISSN 1386-4157, 1573-6938. doi: 10.1007/s10683-016-9494-z.

David Gill and Victoria Prowse. Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6):1619–1676, 2016.

Mamadou Gueye, Nicolas Quérou, and Raphael Soubeyran. Social preferences and coordination: An experiment. *Journal of Economic Behavior & Organization*, 173:26–54, May 2020. ISSN 01672681. doi: 10.1016/j.jebo.2020.02.017.

Simon Gächter, Kyeongtae Lee, and Martin Sefton. Risk, temptation, and efficiency in prisoner's dilemmas. Technical report, CeDEx Discussion Paper Series, 2020.

Ryan Kendall. Decomposing coordination failure in stag hunt games. *Experimental Economics*, 25(4):1109–1145, 2022.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Abhijit Ramalingam and Brock V. Stoddard. Does reducing inequality increase cooperation? *Journal of Economic Behavior & Organization*, 217:170–183, January 2024. ISSN 01672681. doi: 10.1016/j.jebo.2023.10.029.

Frederick W Rankin, John B Van Huyck, and Raymond C Battalio. Strategic similarity and emergent conventions: Evidence from similar stag hunt games. *Games and economic behavior*, 32(2):315–337, 2000.

Ondrej Rydval and Andreas Ortmann. Loss avoidance as selection principle: evidence from simple stag-hunt games. *Economics Letters*, 88(1):101–107, 2005.

Reinhart Selten. A re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4:25–55, 1975.

Olga Shurchkov. Coordination and learning in dynamic global games: experimental evidence. *Experimental Economics*, 16(3):313–334, September 2013. ISSN 1386-4157, 1573-6938. doi: 10.1007/s10683-012-9339-3.

Rui Silva. Coordination in stag hunt games. *Journal of Behavioral and Experimental Economics*, 113:102290, 2024.

John Van Huyck, Ajalavat Viriyavipart, and Alexander L Brown. When less information

is good enough: experiments with global stag hunt games. *Experimental Economics*, 21: 527–548, 2018.

John B. Van Huyck, Raymond C. Battalio, and Richard O. Beil. Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review*, 80(1): 234–248, 1990.

Appendix A: Instructions

Instructions

Thank you for participating! The study today is simple.

You will make decisions along with a group of other real online participants.

All participants will make decisions at the same time, and no one will know what any other person chose before the outcome is revealed.

Your decisions and the decisions of other participants will determine everyone's bonus payoff at the end of the experiment.

Please carefully consider each decision as it will impact your final payment.
If you leave at any time before the end of the experiment, you will not receive your payment.

Next page

Instructions

You will play a game for several rounds.

In each round of the game, each participant will make a simple decision: choose a triangle or a circle



Next page

Instructions

The payment rules for the round are based on everyone's choices.

If more than half of the participants choose the circle, **all** participants will receive £1 for this round.

If less than half of the participants choose the circle:

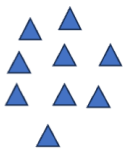
- everyone who chose the triangle receives £1.
- everyone who chose the circle receives nothing for this round.

Next page

Instructions

Let us show several examples. Suppose there are 9 participants online.

Suppose everyone chose a triangle.

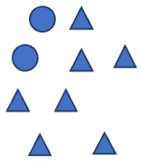


It means less than half of people chose a circle.
Everyone who chose a triangle gets £1.
So every participant gets £1.

Next page

Instructions

Now suppose 2 people chose a circle and 7 people chose a triangle.

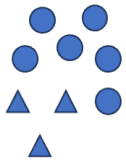


It means less than half of people chose a circle.
Everyone who chose a triangle gets £1.
Everyone who chose a circle gets nothing.

Next page

Instructions

Now suppose 6 people chose a circle and 3 people chose a triangle.

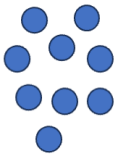


It means more than half of people chose a circle.
Everyone gets £1.
So every participant gets £1.

Next page

Instructions

Now suppose everyone chose a circle.



It means more than half of people chose a circle.
Everyone gets £1.
So every participant gets £1.

Next page

Instructions

In each round of the game, you simply need to click on one of the shapes to make your choice.



Next page

Instructions

You then will see the outcome:

7 people chose	**3 people chose**
	
They get £1.	They get £0.

Next page

Instructions

You will play the game for 10 rounds.

In each round, you will be able to get either £1 or £0.

At the end of the game, you will receive the sum of all of payments from each round.

Good luck and press the button below when you are ready:

I am ready!