Acting *as if*: Self-interested players act as if others will mirror their moves

Matthew Cashman[1], Drazen Prelec[1]*
[1]Massachusetts Institute of Technology

Word count: XXXX excluding abstract, methods, references, and figure legends

Authors Note:

Matthew Cashman, Sloan School of Management, MIT; Drazen Prelec, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences, MIT

Author Contributions:

Matthew Cashman contributed to the design and implementation of experiments, to the analysis plan, and to drafting the manuscript. Drazen Prelec contributed to the design of the experiments and to drafting the manuscript.

**Please address correspondence to Matthew Cashman; matt@cashman.science*

Abstract

Theoretical accounts of cooperation include pro-social motivation, norms and reputation, and cognitive heuristics like team thinking. We provide experimental evidence for a different psychological mechanism, one that, notably, explains cooperation even among the self-interested and does so without external monitoring: quasi-magical thinking. In one-shot Public Goods Games where players move sequentially but do not observe others' moves, we find that contributions to the public good are highest at the beginning and decline as order increases. We interpret this as reflecting differences in players' sense of impact on the collective outcome: Subjective impact is maximal when other players have not yet moved. Three results provide further support for this interpretation: (1) The order effect is generated by players who are acting in their own interests, (2) instructing players to maximize their own payoff increases the order effect, and (3) the order effect is eliminated if the moves of future players, but not of past players, are determined randomly.

Table of Contents

Table of Figures

1 Introduction

Social cooperation without external monitoring is widely regarded as fundamental to human culture, sustaining teamwork, mass political participation, and personal sacrifice for family, tribe or nation. People often face opportunities to incur an individual cost in exchange for a collective benefit, and there is a rich literature exploring the whys and wherefores (Henrich & Muthukrishna, 2021; Rand & Nowak, 2013). For example, a pedestrian can choose to throw litter into the gutter, or he can wait until he comes across a trash bin. A CEO might choose to move assets overseas in order to avoid taxes, or she might choose to avoid chicanery, keep assets domestically, and pay more in taxes—in the end, contributing to the public weal. Each choice involves a tradeoff between what is good for the agent and what is good for the group. This tradeoff is widely studied using Public Goods Games (PGGs, Zelmer, 2003 for a meta-analysis). The PGG is used as a model of human cooperation because of the tradeoff between the benefits accruing to the group via cooperation and the benefits accruing to the individual via defection captures the essence of cooperation problems humans solve on a daily basis. In standard PGGs, it is always better for an individual player to defect no matter what decisions others make, but it is always better for the group if everyone cooperates.

There may, however, be circumstances in which even self-interested players—players for whom there is no tradeoff, players who are just trying to maximize their own payouts—end up cooperating. We investigate this possibility with a subtle variation on the classic one-shot PGG, changing it so that players within a single round move one after another but do not observe each others' moves: a sequential PGG without observation. If players are forced to move one after another with no knowledge of each other's moves and we observe more cooperation among players who move earlier in the sequence (an increase borne of the mere knowledge that others will be acting in the same game *after* them), that is interesting for two reasons. First, knowing how to induce

cooperation is useful—and doubly so in the particular case of people who are only trying to do best by themselves. Second, it gives us a window into how these decisions are being made. It could be the case that self-interested players are cooperating because they are calculating an expected payoff on the assumption that everyone moving subsequent to them will make the same move they have, providing valuable insight into possible mechanisms by which this cooperation emerges.

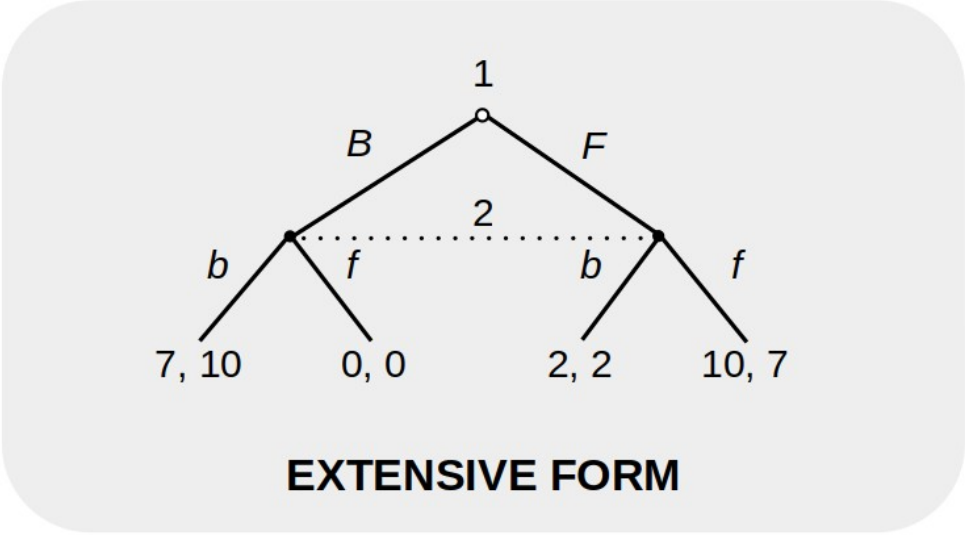
The games used here are “sequential” in that players move one after another and players’ moves are unobserved in that there is no information flow between players: any given player knows nothing about the decisions players who have already moved have made. For example, a five-person PGG is sequential with unobserved moves when the five players move one after another, but each player knows nothing about what the other players have done and also knows other players will not know his move. Traditional game theory suggests that the order in which players in the same game are moving is irrelevant as long as they don’t know anything about what moves others make, and therefore the distinction between events that have happened and have not happened—even if they are unknown—is lost.

2 Review

2.1 Incorporating order of play, social preferences, and psychology into equilibrium analyses

von Neumann and Morgenstern (1944/2004) articulate the difference between priority in chronological order of play (which they term “anteriority”) and priority in information (“preliminarity”). Preliminarity implies anteriority (if Player B has information about Player A’s moves, it is necessarily the case that Player A has moved before player B), but it is not the case that anteriority implies preliminarity (it is possible to not know about things that have already happened). von Neumann and Morgenstern then develop extensive form notation based purely on preliminarity (information), ignoring the chronological ordering of moves. Because of this, the standard extensive form representation of the simultaneous version of a game is identical to that for a sequential version today.

By the mid 1980s there was a small chorus raising the question of whether ignoring anteriority (which allows the expression of certain extensive form games in normal form with no distinction) is a good idea (Kohlberg & Mertens, 1986; Kreps, 1990; Luce, 1992). Luce (1992) mentions the lack of a time variable in extensive form games (while real life inevitably involves one), and Kreps asks explicitly: “Can we find a pair of extensive form games that give rise to the same strategic form such that, when played by a reasonable subject population, there is a statistically significant difference in how the games are played?”, setting the stage for the investigation of sequential games with and without observation.



	Prize Fight	Ballet
Prize Fight	10, 7	2, 2
Ballet	0, 0	7, 10

NORMAL FORM

Figure 1: Representations of the Battle of the Sexes *game in both Normal Form and Extensive Form.*

2.2 Sequential games with observation

There are several bodies of empirical work that investigate the effects of agents acting one after another with observability. There is a rich literature investigating team effects, in which individual agents acting as part of a team optimize for the team's success, rather than for their own best interests, under certain conditions (see Colman

& Gold, 2018 for a review). Similarly, there is substantial work investigating leader- and follower- effects in games which have some element of sequential moves. Eichenseer (2023) provides a comprehensive meta-analysis that examines the role of order of play in public goods experiments. He develops a taxonomy of PGGs: linear PGGs, threshold or step-level PGGs, PGGs with interior equilibria, field experiments, and weakest link games. Linear PGGs are the most common type of PGGs studied, and in these games Eichenseer found that leading by example significantly increased contributions. The effect was stronger when the leader was exogenously assigned rather than chosen by the group. The leader's contribution was found to be a strong predictor of the followers' contributions, indicating a degree of conditional cooperation. However, the followers typically contributed less than the leader, suggesting some degree of free-riding and raising the question of how effective sequential contribution with observation might be if one must find many leaders willing to be exploited. Tangentially related, Bohnet & Frey (1999b, 1999a) report Prisoner's Dilemmas and Dictator Games wherein they manipulate the degree to which players can communicate. They examine the effects of merely identifying the other player (but providing no information about the other player's moves), and find that mere identification results in significantly more cooperation in these games. Who another player is and how that person came to be there matters.

Herding and information cascades (Banerjee, 1992; Bikhchandani et al., 1992) may also be informative when considering sequential games with observation. The essential idea is that there are certain circumstances under which it is optimal to copy the behavior of those moving before you, namely when you estimate that the external information from the moves of those prior to you is very informative, the distinction being that in informational cascades private information is ignored, while in herding private information is taken into account and the action occurs anyway (Çelen & Kariv, 2004). An example might be selecting a restaurant in large part based on how full it is: if a lot of other people have chosen to eat there, it must be good. In the context of sequential games, herding-type behavior could explain part of the increase in cooperation following a generous first-mover especially in the case where the cost of considering other moves is large relative to the cost of following the leader.

Theoretical accounts from literatures on sequential games with observation rely on the flow of information about earlier players' moves to later players, but they also rely on earlier and later players knowing that will happen. Evidence for order effects in the absence of any information flow at all may, then, be consequential for prior work where it has not controlled for the possibility that order-based effects could be due in part to mere position in a sequence alone.

2.3 Sequential games without observation

2.3.1 Common-pool resource dilemmas

There is a small body of work investigating the effects of sequential moves without observation in common-pool resource dilemmas, which represents the first empirical work on sequential games without observation. Budescu et al. (1995) conduct an early empirical investigation of a sequential game without observation, which they refer to as the "positional order protocol". The common-pool resource games they investigate are games in which players each make a request from a common resource pool of limited size and receive nothing if too much is requested in total. They offer a theoretical account such that behavior in the sequential game without observation is intermediate between the Nash equilibrium for simultaneous play and that for sequential play with observation. This implies that, in the sequential no-observation condition, requests from the common pool will decline with increasing order (Player 1 requests more than Player 2, Player 2 more than Player 3, etc.), but slower than in the case of the sequential game with observation. They hypothesize that a given player, say Player 3, is acting as if the requests of all the players *preceding* her are known to her, and that these players are in fact playing the Nash equilibria. In some sense, norms of play from a sequential game with observation are invoked in one without. They report results from two empirical studies, each of which examined the effects of uncertainty in the size of the common pool resource as well as order effects. The first study was conducted with

45 undergraduates, and they find evidence for the decline in contributions in the sequential protocol without observation in groups of five that are weaker than those in the sequential protocol (in line with their theorizing). The second, in 180 undergraduates and groups of two or three, supports the same conclusion. A second paper, Budescu et al. (1997), reports essentially the same findings among 87 undergraduates, with the sequential game without observation showing a weaker decline in requests with increasing order than in the condition with observation. They add to previous results measures of social orientation, noting that the more self-oriented a player is the more he is likely to request, but they do not note any interaction with order effects and social orientation. Budescu & Au, 2002 extend this further, offering a formal model of behavior in these games. They conduct two more experiments, with 62 and 38 undergraduates each, replicate previous results, and conduct a variety of model-fitting exercises.

2.3.2 The role of uncertainty and causality

In a curious study, Quattrone & Tversky (1984) report evidence from two experiments for what they term “diagnostic” actions—actions that have no direct causal relationship to desirable outcomes, but which are indicative of them. In the first experiment (39 undergraduates), they report that subjects who are performing a task that involves holding an arm in circulating ice water (a painful experience) are able to hold their arms in the water *longer* when they believe this is indicative of having a strong heart, and for shorter amounts of time when that is believed to be indicative of having a bad heart. The experience of holding one’s arm in water of course has no bearing on heart type, but it does appear subjects are changing the data they themselves produce in order to receive good news in apparent disregard of the causal relationship (see Bodner & Prelec, 2003 for the development of the idea as self-signaling).

In related work, Shafir & Tversky (1992) explore nonconsequential reasoning—by which they mean reasoning that at least appears to either not produce estimates of, or which ignores, the consequences of a particular action given an information set. This

class of decisions violate the sure thing principle, which states that if X is preferred to Y under all states of the world, then X should still be preferred to Y even if the state of the world is unknown. For example, Shafir & Tversky show empirically that there are many people who would prefer to pay for a vacation to Hawaii in the event that they pass an exam *and* in the event that they fail, but who would also prefer not to buy in the case where the outcome of the exam is unknown. They refer to this pattern of events as "accept when win, accept when lose, reject when do not know" and refer to it as the "disjunction effect". They observe more cooperation in one-shot games when uncertainty about the other player's move is highest. Shafir & Tversky report results from a one-shot simultaneous prisoner's dilemma with 80 undergraduates playing 40 games each in three conditions: it is known that the other player defected, it is known that the other player cooperated, and it is not known to the player what move the other player made. The authors report 3% cooperation when the player knows his counterpart defected, 16% cooperation when he knows his counterpart cooperated, and 37% cooperation when the counterpart's move is unknown. They suggest this effect may be due to a tendency to take the perspective of the group in cases of uncertainty, perhaps not even considering the consequences of each branch of the game. It could also be due to a desire to induce cooperation in the other player. Because the other player has some matching characteristics (in this particular case, it is another student), it might be assumed that she will approach the game in the same way. In this case, each cooperates and they reap the rewards relative to the defect-defect equilibrium. They subsume these ideas in the concept of *quasi-magical thinking*, the idea that people act *as if* they can influence as-yet unresolved events even when they "know" (or will report) they cannot. It is important to note here that Shafir & Tversky do not distinguish between two different senses of uncertainty: the player experiences uncertainty in the sense that his counterpart's move is not known to *him* but may have been made and is therefore known to the counterpart and possibly the experimenters (preliminarity), and potentially also uncertainty in the sense of anteriority—the counterpart's move has yet to be made at all. It is not clear what impression the subjects had.

Hristova & Grinberg (2010) investigate two hypotheses that could explain the disjunction effect reported by Shafir & Tversky: the complexity hypothesis, which suggests that the disjunction effect is the result of it being computationally difficult to compute and reason over the multiple possibilities inherent in an uncertain situation, and quasi-magical thinking. They report that the disjunction effect in the Prisoner's Dilemma is weakened by two manipulations: making the quantities that appear in the game easier to do arithmetic with reduces cooperation under uncertainty (33 subjects), and informing participants that their computer opponent has selected moves prior to the experiment reduces cooperation under uncertainty (27 subjects). The fact that having moves selected prior to play reduces cooperation under uncertainty suggests that the disjunction effect is driven by some sort of implicit causal thinking.

Chen & Zhong (2022) find that uncertainty results in more honesty in a dice game cheating experiment, and they find that subjects are more generous under uncertainty in a dictator game experiment. They propose a model that incorporates what they call a "karmic state", where under conditions of uncertainty a player believes to some degree that moral behavior leads to better outcomes than immoral behavior. In this experiment, subjects are uncertain about which of six boxes contains a high or low reward, but are certain of which of the six boxes contains a bonus in addition to the high or low reward. After having rolled a die which picks out a certain box, they *are* certain about which one they are supposed to choose (and, therefore, whether or not it contains a bonus), but they still do not know whether it is high or low reward. In the case where there is less uncertainty about a given player's reward (e.g., 6 of 6 boxes contain the high reward), subjects are more willing to lie about their dice roll in order to receive a bonus. With more uncertainty (e.g., 3 of 6 boxes are high and 3 of 6 are low), less cheating to receive a bonus is observed. This is interesting in that it is entirely "sealed" fates except for the subject's own decision: all parameters of the game are known, if not to the subject then to the universe. The subject does have direct, obvious, causal control over the proportion of the outcome represented by the bonus, but this study still evidences more prosocial behavior under uncertainty.

2.3.3 Public Goods Games

There is a modest line of empirical work examining order effects in Prisoners Dilemmas and PGGs, both of which model the conflict between individual gain and collective benefits in a way that escapes common-pool resource dilemmas and coordination games. Social dilemmas like Prisoner's Dilemmas and PGGs are situations where members of a group are faced with tension between two choices: maximizing their own gains (defection) or maximizing their collective interests (cooperation). Abele & Ehrhart (2005), Figuières et al. (2012), and Morris et al. (1998) each provide some theorizing in addition to their empirical results. Morris et al. use a framework of heuristics, arguing that real players do not compute game-theoretic optima, and attempt to disentangle a "matching" heuristic from a "control" heuristic. In the matching heuristic, players cooperate in one-shot games because they wish to match others' acts of cooperation towards them—and the only way they can be sure of doing that is to cooperate. The control heuristic finds its origin in Shafir & Tversky's quasi-magical thinking as a theory, but Morris et al. see quasi-magical thinking as a type of illusion of control or control heuristic. The related body of work on the illusion of control begins with Langer (1975). This literature asserts that people are motivated to believe they have more control than they do over a situation, especially when the lack of control should be logical or observable. The feeling of control when there is none may be due to social motivations or to the desire to preserve self-esteem (Stefan & David, 2013 for a review). Morris et al. point out that in this literature it is commonly found that the timing of events affects behavior. For example, subjects bet more on a future roll of the dice than a past roll, suggesting the control heuristic is present more in situations with "open fates" vs. "sealed fates".

Morris et al. investigate these theories with a sequential Prisoner's Dilemma without observation among 86 students in their first experiment, and 267 MBA students in the second. Interestingly, rather than moving sequentially but directly after one another, in these experiments players in sequential conditions move on different days, up to a week apart. In the first experiment they compare three conditions, as with Shafir

& Tversky 1992: it is known that the other player defected, it is known that the other player cooperated, and it is not known to the player what move the other player made. These three conditions are played within subjects, and are crossed with timing: either the other player's move was made in the past, or has yet to be made. The "control heuristic" pattern of cooperating when the other's strategy is unknown and defecting otherwise is much more frequent in the "open fate" case where the other player's move has yet to be made. Other than this, they report results similar to Shafir and Tversky's. In their second experiment they include a simultaneous condition, and report similar results in that the control heuristic is observed more with future moves rather than past moves, but they also observe high rates of the control heuristic when the other player is making his move at the same time. The tasks used more resembled reporting strategies in response to hypothetical situations than playing games with other players in that moves, the computation of results, and payment were entirely decoupled from each other.

Abele & Ehrhart (2005) develop this work further, investigating the effects of moving sequentially with no observation in the PGG. They consider two competing theoretical accounts: first, they ask whether their "schemata activation" theory holds in these games as they assert it does in certain coordination games. The schemata activation account suggests that moving one after another—even with no observation—activates deeply held priors about how to act in social situations since social interaction is usually sequential, leading to more cooperation. Second, in a different line of reasoning specifically for the sequential PGG without observation, they suggest that moving simultaneously may activate feelings of "groupness", while moving sequentially could allow for thinking of oneself alone, leading to more cooperation in the simultaneous condition.

In their first experiment (86 students), they find that simultaneous-movers in a PGG contribute approximately double what either first- or second-movers contribute, with no difference between first- or second-movers in the sequential condition. In their second experiment (192 students), they cross the design with either a "high

expectation" (subjects are told the average contribution in the past was high) or "low expectation (subjects are told past average contribution was low) conditions. They observe no difference among simultaneous, first-, and second-movers in the low expectation condition but do observe players in the simultaneous-high condition contributing significantly more (approximately double) what first- and second-movers contribute. They interpret this as evidence for the "groupness" theory, given that elevated cooperation is observed in one of the simultaneous conditions.

Robinson et al. (2010) investigate causality with information directly (116 undergraduates) in two experiments. The first experiment is the only one to examine sequentiality with no observation, and they report no difference between simultaneous and sequential (first-mover only) behavior in one-shot Prisoner's Dilemmas.

Figuières et al., (2012) report the results of a series of simultaneous PGGs and sequential PGGs with and without observation, crossed with group sizes of four and eight players per PGG (252 undergraduates). They find that, in aggregate, contributions are higher under sequential play with observation than either in simultaneous games or sequential games without observation, and that contributions decline with increasing order in games with observation, but there is no effect of order in games without observation.

2.3.4 Theoretical approaches to the sequential PGG without observation

Masel (2007) offers an interesting theoretical account of quasi-magical thinking. Masel's model has players coming to a game already having a Bayesian prior distribution over human behavior (which, indeed, we all have). Upon observing additional information during the game, the player's prior distribution is updated in the usual fashion—one's own behavior being just another data point. A weighting function makes recent data more significant than data from earlier rounds since other players' behavior will change in response to their environment over time. The player's own move, or potential move, is an additional data point that goes in to the conditional

expected utility calculation following Jeffrey (1990). This account, however, does not distinguish between “open” and “sealed” fates, and so does not incorporate the arrow of time. Daley & Sadowski (2017) develop a similar model of magical thinking that applies to players’ preferences over actions rather than outcomes.

A related body of work examines universalization as an explanatory model for many morally-relevant behaviors. The basic idea is that, at some level, people ask themselves: What if everyone did this? Roemer (2010, 2015) develops the idea of a “Kantian equilibrium”, where each player asks: “if I deviate from my action and everyone else were to deviate in the same way, would I prefer the consequences of the new action profile versus not deviating at all?”, and Levine et al. (2020) present a computational model of universalization in moral judgment, and, significantly, refine the motivating question to, “What if everyone felt free to do that?”. Levine et al. report good evidence for universalization across a series of vignette studies across adults and children. These studies make use of threshold problems, which might be formalized as threshold PGGs.

2.4 Conclusions from prior work

Previous literature has highlighted a number of interesting questions with respect to uncertainty and causality in sequential games with no observation. However, the extant empirical literature considering sequential games without observation suffers from a few general problems. First, experiments use relatively small samples and split these small samples across many conditions. Second, these samples are almost without exception made up entirely of students. Even if students do not know each other, they know they are all part of the same rather small community. It is nearly certain that any given student is connected socially by one or two degrees of separation to any other if they aren’t connected directly, and this could engender quite different behavior than true anonymity. Third, this literature generally uses repeated measures. Even if participants for a given round were randomized and anonymous, everyone

involved is still aware that all partners are drawn from the same small pool. There may also be important differences in play that manifest after many rounds of the same game. Finally, the games reported were often not played in real time: subjects were often asked to report strategies in response to prompts on static paper cards, rather than actually playing with—and moving before or after—other players in real time, as a game is usually played. On top of this, of course, there is on average several decades' advancement in the practical application of causal inference to empirical questions in behavioral science in the form of preregistration, power analyses, etc.

Acknowledging these weaknesses, what have we learned from these literatures? Order effects are present in sequential games with observation, and in most cases we have a good theoretical handle on why this might be. Considering the effect of anteriority on play in sequential games, Kreps (1990) gets to the core of the issue: "Can we find a pair of extensive form games that give rise to the same strategic form game such that, when played by a reasonable subject population, there is a statistically significant difference in how the games are played?" The answer is unequivocally yes. There is the case of coordination games where order becomes an obvious Schelling point, and the similar case of threshold PGGs. But it is also clear that we observe differences in play between sequential games without observation and their simultaneous or sequential with observation counterparts even in games where there is no obvious use of order to facilitate coordination.

Beyond the fact that anteriority matters for behavior, what generalizations can we make for social dilemmas like the PGG? We can conclude that uncertainty matters, for one. Uncertainty seems to push people towards more prosocial actions. Exactly what kind of uncertainty is a bit hazy, though perhaps it appears that uncertainty in the sense of an "open fate" would engender the largest change in behavior relative to certainty. We can also conclude that, insofar as empirical evidence is available, in sequential PGGs without observation early-movers tend to contribute more than late movers and there is substantial heterogeneity in effects among subjects.

3 Main

In a standard PGG, *n* players are each given an endowment *e,* and are asked to decide what proportion of their endowments to contribute to the public good, from nothing to all of it. A given player's contribution to the public good is represented by *a*. The total amount from all the players that is contributed to the public good, *c, i*s then multiplied by a multiplier *m* (which must be less than the number of players), and this amount is distributed *evenly* among *all* the players—even those who chose to contribute nothing. An individual player's payoff function in a standard simultaneous-move PGG is as follows:

$$p = \frac{mc}{n} + e(1-a) \qquad \text{(Equation 1)}$$

Consequently, whenever the multiplier *m* is less than the number of players *n*, the group as a whole does better if everyone contributes their entire endowment (cooperates), but each individual player is better off if he or she contributes nothing (defects). Put another way, the total amount of money in the group is maximized if everyone cooperates, but any individual player always makes more by defecting—independent of anyone else's moves. Because other players do not know your move, they cannot change their own moves in reaction to it. If a group plays the game only once, it is impossible to build reputations, enact retribution, or to reward others for their actions.

In the sequential games described here, there is no information flow. Players are informed they will not know others' moves from the beginning (so they know they will not know), they do not find out others' moves during the game, and they do not know the size of the public good when it is their turn. But they may use their own move as a signal of what others will do. If a player believes subsequent players will make the same

move she has, the payoff-maximizing move changes from the standard Nash equilibrium (defect) to something else. We will call the player of interest in a sequence of players the “focal player”. Under most conditions, players moving earlier in a sequence will be best off if they contribute all of their endowment to the public good, and later players will be better off if they defect and contribute nothing. We would expect to see a decline in cooperation as the focal player’s position nears the end of the sequence if there is some heterogeneity among players in how strong this type of reasoning is. In this case, we would expect to see the most cooperation from Player 1 of 5, who will tend to cooperate more than Player 2, who will cooperate more than Player 3, etc.

In four studies we test whether the temporal order of unobserved moves influences decisions to contribute to a public good. The first is a three-person PGG, and the remaining studies use a five-person PGG. Study 1 verifies that there is an order effect, and how that varies with Social Value Orientation (SVO, a measure of willingness to give up gains in order to benefit others. See Murphy et al., 2011). In the case of quasi-magical thinking, players who are prosocial on the SVO measure might be expected to help others even at some cost, therefore showing no order effect, while those who are Individualistic (and therefore maximizing their own rewards) might show an order effect since the number of “open fates” varies with order. Study 2 expands this to five people to better rule out any anomalous first-mover and last-mover effects. Our chief interest is respondents who are trying to maximize their own financial rewards, but those who arrived at the experiment clearly self-interested are sufficiently rare in the study population that forming five-person groups in real time proved difficult. For this reason, Study 3 asks whether the mere instruction to try to maximize their own payouts produces the order effect. Study 4 deploys the technique from Study 3 and asks whether we would observe an order effect in the case where all players either before or after the focal player have their contribution decisions made for them by a random process. In the case where having random-movers before the focal player but regular players after results in an order effect or vice-versa, the possible mechanisms supporting the effect are constrained. This gives an indication of whether the difference

between “open” and “closed” fates matters. In all studies participants contribute three inputs: comprehension checks, game playing decisions, and predictions of the responses of other players. Apart from game compensation, participants are also paid for correct answers to comprehension checks and for accurate predictions.

4 Results

All studies are one-shot linear PGGs with a multiplier of two, and share several characteristics. First, before learning what game they are to play, players participate in a chat room with their groupmates in order to serve as a rough and ready Turing test. We hope that this chat gives the task more psychological reality than might otherwise be felt in an online task with no human interaction. Second, all games are real-time interactions between real players. When players are playing a PGG, they are playing with the groupmates they chatted with in real time. Third, all pre-play attention checks, comprehension questions, gameplay decisions, and predictions are incentive-aligned. Participants are paid more for correct answers. Fourth, all experiments have simultaneous-play PGG control conditions. Fifth, all players pass familiarization tasks and comprehension checks. Data from players who miss a single comprehension check is excluded from all analyses unless otherwise noted, but players who fail a comprehension check still complete the task. Because they are part of a group that is playing in real time their moves are necessary to calculate payoffs. All experiments share the following three up-front comprehension and attention check questions:

Q1: *Do any of the other players **know how much YOU decide to contribute?***

Q2: *Jack and Jill are playing this game together. Jack decided to **TRANSFER** and Jill decided to **KEEP**. Who will make more money, Jack or Jill?*

Q3: *What year is it?*

Participants are given one chance to get each of these questions right, and a single wrong answer results in that data being excluded. Later studies incorporate more extensive training and comprehension check regimes.

4.1 Study 1: 3-Person Sequential Public Goods Game

Participants played one round of a three-person PGG. Our primary interest was how contribution to the public good varied with order of play. Along with standard measures, we estimated participants' interpersonal utility tradeoffs using the SVO scale, which divides almost all[1] participants into two categories, Individualistic and Prosocial. Individualists are working to maximize their own outcomes and are indifferent to others' outcomes, while Prosocials care about maximizing their own outcomes but do take others' outcomes into account. Order effects should be more pronounced for participants who are primarily interested in their own payoff (Individualists). In contrast, Prosocials should be less sensitive to order of play, as altruistic motivation should not be biased toward future players.

We find some evidence for the preregistered order effect when pooling pilot data with data collected post-preregistration, due to insufficient power. First-movers contribute more than later players, though we do not resolve a difference between second- and third-movers and therefore do not meet the conditions of the preregistration. A linear regression of contribution on move order yields a significant negative slope, $\beta = -0.042$, 95% CI = [-0.081, -0.003], $F(1, 780) = 4.7$, $p = 0.03$. First-movers contribute more than second-movers ($p=.013$) and more than third-movers ($p=.031$). The difference between second- and third-movers' contributions is not significant.

[1] Nearly all subjects were SVO classified as Individualistic or Prosocial; two respondents were classified as Altruistic, and two as Competitive. These respondents' data are excluded from analyses.

We also find support for the preregistered prediction that the order effect is concentrated among participants classified as Individualistic in the SVO task. Participants classified as Prosocial exhibit no significant differences in contribution levels as function of order, while we do see a difference between the first-mover data and grouped second- and third-mover contributions ($\beta = -0.142$, 95% CI = [-0.248, -0.035], $F(1, 288) = 7.104$, $p = 0.008$). As with the aggregated data, we do not see the hypothesized difference between positions two and three among respondents SVO-classified as Individualistic.

In addition, we do not find support for the pre-registered prediction that correlations going forward in time, between a player's own move and her predictions of future players' moves, will be stronger than those going backwards in time. It is possible that the mechanism driving the order effects we observe is fundamentally sub-conscious; when forced to explicitly consider and report on their expectations of what others in the game have done, players may deploy a different strategy.

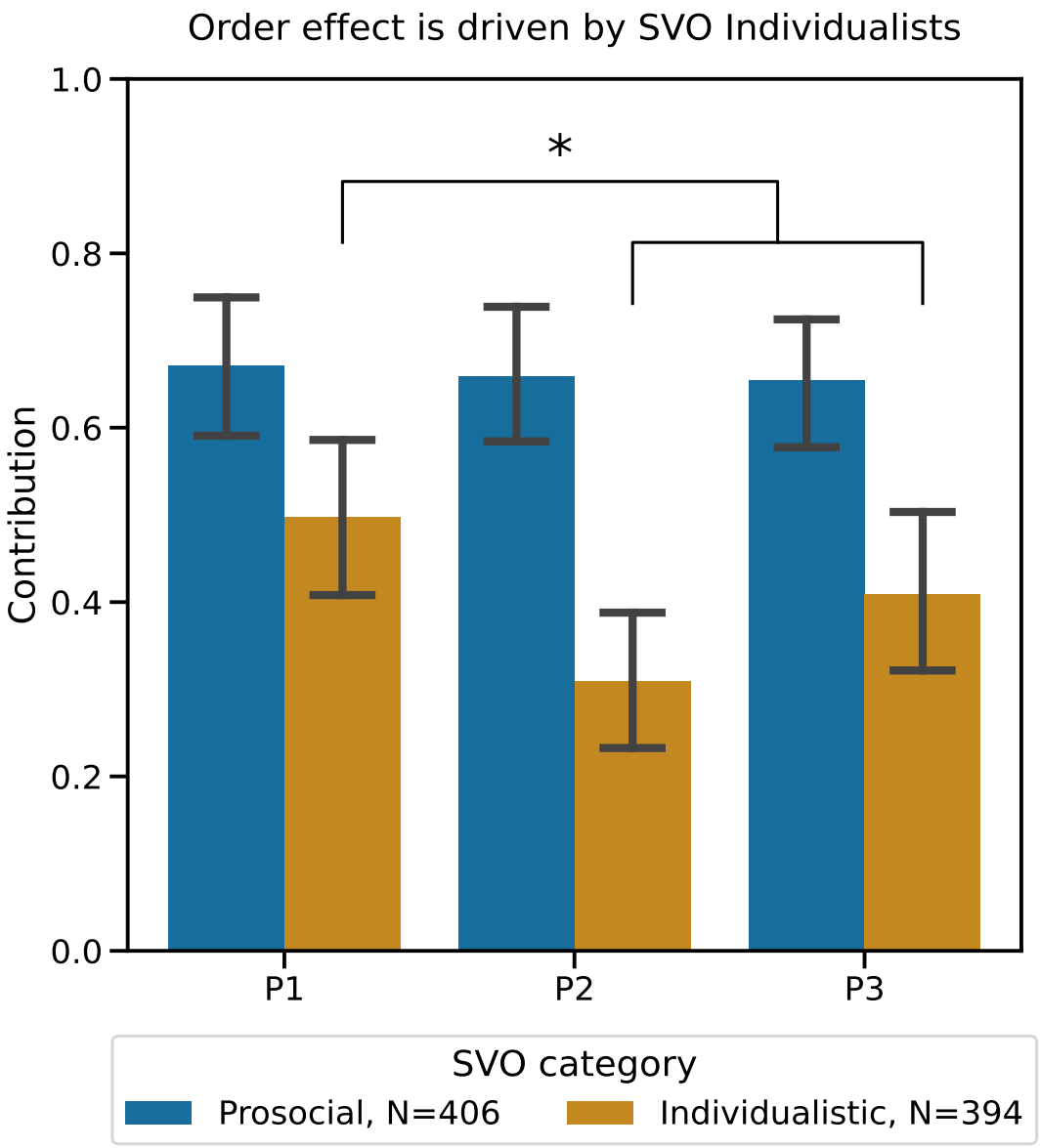


Figure 2: Change in contribution with order is driven by subjects SVO-classified as Individualistic. Subjects who passed comprehension checks. 95% CIs.

4.2 Study 2: 5-Person Sequential Public Goods Game

A five-person PGG allows for more insight into the order effect, especially given the potential for effects due to being either first in a sequence of any length (“leader effects”, e.g. Eichenseer, 2023) or last. We report results from a sequential 5-person game where respondents were classified based on an SVO task performed up-front. In Study 2 we do not meet our pre-registered threshold to detect an order effect, $\beta = 0.011$, 95% CI = [-0.032, 0.055], $F(3, 595) = 33.285$, $p = 0.6206$ for the interaction. A programming error meant that the time Player 1 and Player 2 had to make a decision was not correct, sometimes being shorter than for other players and sometimes close to zero. We do observe the predicted trend in Players 2-5.

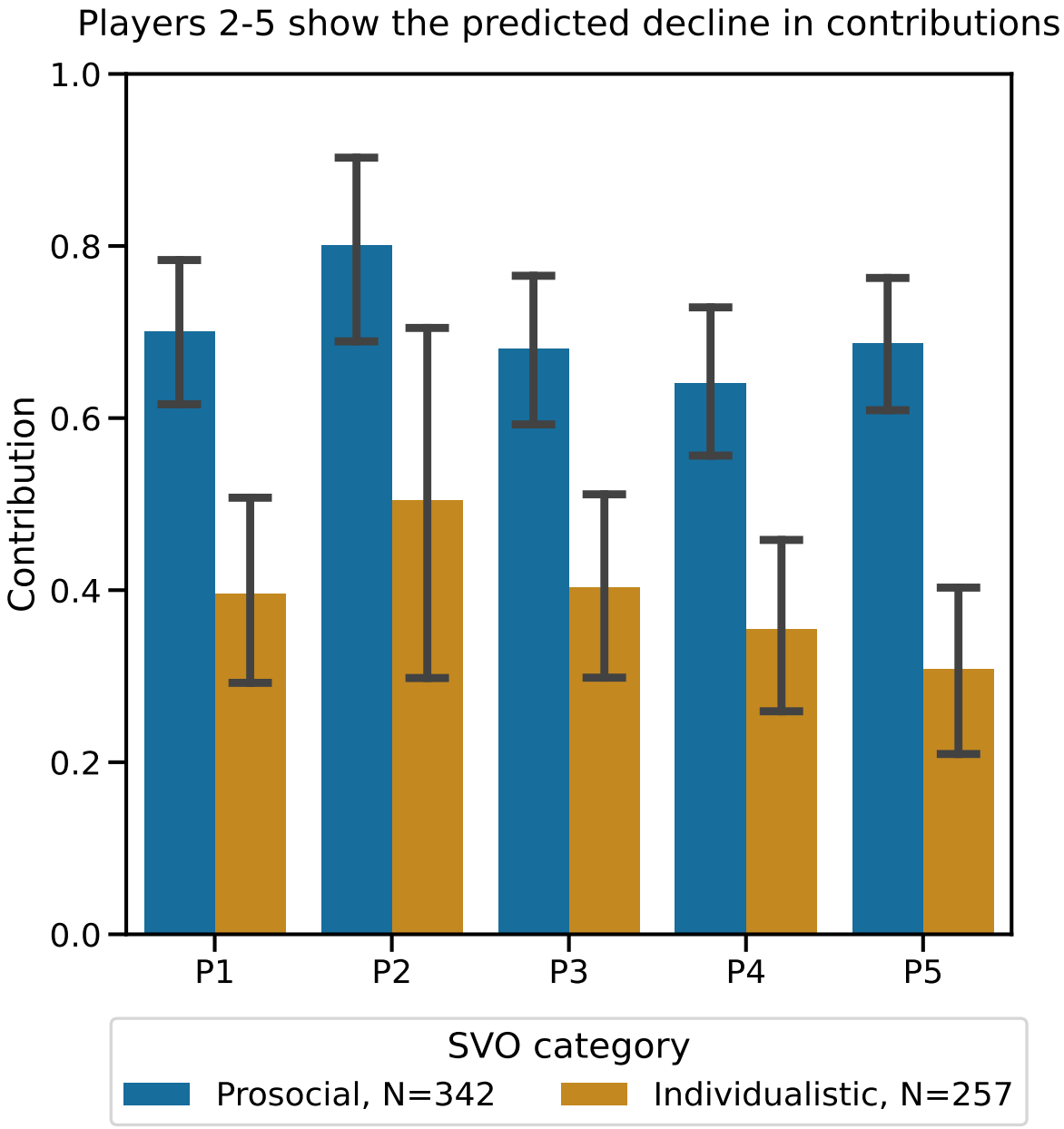


Figure 3: SVO-individualist players show a decline in contribution with increasing order from Player 2, and an anomalous result for Player 1. Player 1 was affected by a programming error that resulted in incorrect timing of stimuli presentation. Subjects who passed comprehension checks. 95% CIs.

We noted that the effect became apparent among all positions if analysis is limited to respondents who were close to 0 degrees on the SVO scale, i.e. those who were most clearly maximizing their own returns, as opposed to those who were merely classified as “Individualistic”. If we restrict our analysis to all players whose SVO degree measure was +/- 10 degrees (clustered around maximally self-interested), the predicted pattern appears despite the anomaly with position 1.

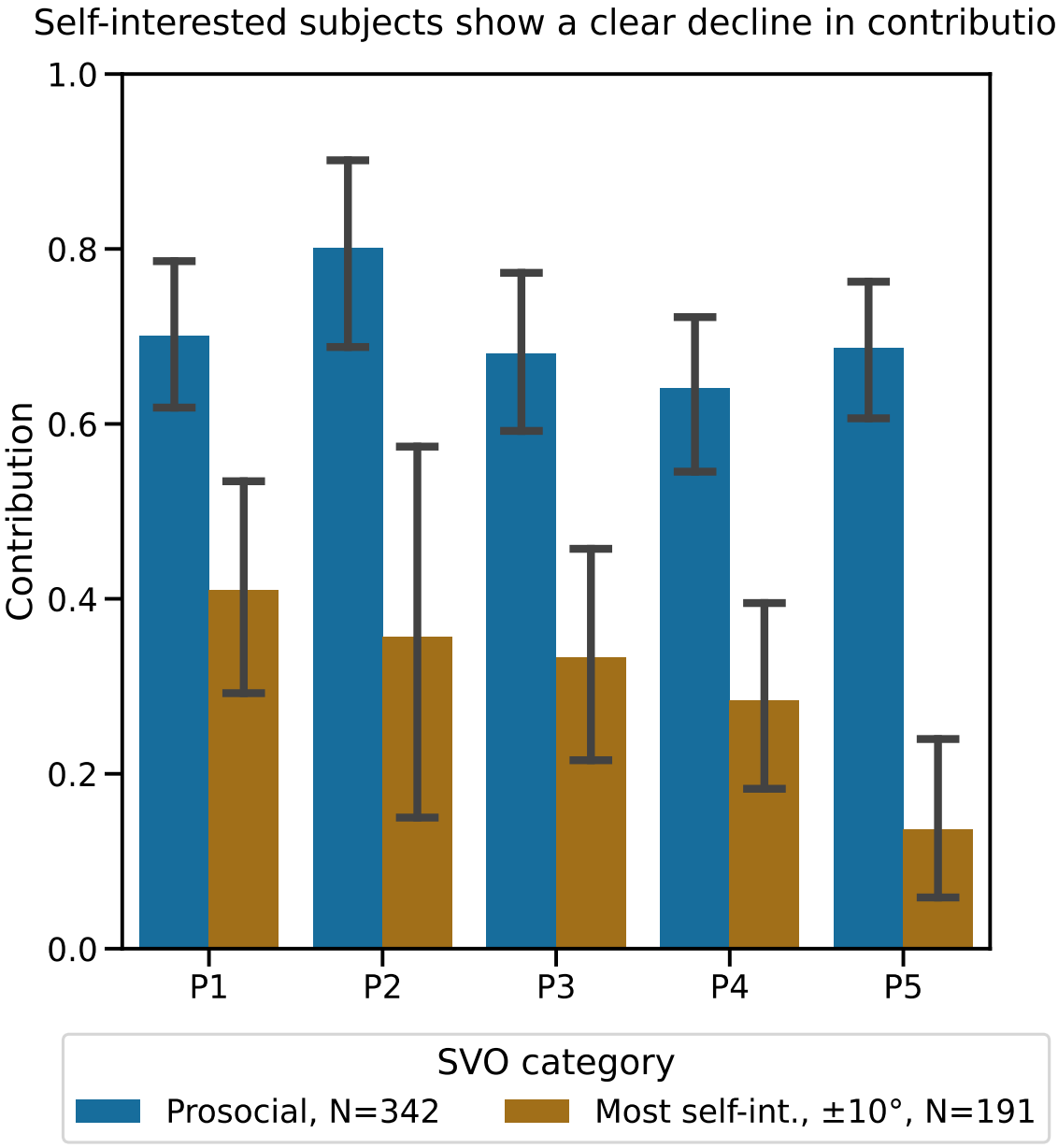


Figure 4: Players whose SVO degree measure is +/-10 degrees show the predicted decline of contribution to the public good with increasing order. All subjects. 95% CIs.

The experiment was not sufficiently powered to detect an effect among the most self-interested +/-10 degrees SVO population who passed comprehension checks, but a linear regression of contribution on order interacted with a binary self-interested / not self-interested variable using data from all respondents (including those who failed a comprehension check) does detect the effect, $\beta = -0.042$, 95% CI = [-0.084, -0.0], $F(3, 879) = 45.661$, $p = 0.043$. The preregistered tests for the partial correlation between predictions of other group members' moves and the focal player's moves being stronger going forward show no effect.

4.3 Study 3: 5-Person Sequential Public Goods Game with induced self-interest

The observation from Study 2 that the most self-interested respondents were those exhibiting the largest order effect led to the design of Study 3. Filling real-time 5-person games with enough self-interested respondents proved impractical due to the rarity of respondents who score +/- 10 degrees on the SVO battery, so Study 3 was meant to efficiently examine if a mere prompt to act in one's own interests would allow us to replicate the pattern observed in respondents who arrived at earlier studies already self-interested. The task is similar to Study 2 but has some improvements. The main difference is that respondents did not perform the SVO filtering task. Instead, respondents were randomized to a condition with no prompt, or to a condition with the prompt:

> *Please try to play this game **however you think will make you the most money.** We understand that sometimes you want to help other people, but for the purposes of this experiment we want you to try to make as much money as possible.*

In addition to the prompt, Study 3 incorporates three substantive improvements. First, Study 3 adds an additional simultaneous-play control condition that implements a delay of 80 seconds. These participants will wait about as long as sequential-condition players who are moving last (order = 5). This condition was incorporated to control for the possibility of effects dependent on time spent waiting. While waiting, respondents are shown the task's standard wait screen which incorporates the option to play a simple game to keep respondents engaged with the task. Second, Study 3 incorporates an interactive practice game after the instructions and comprehension questions. This practice game asks respondents to calculate the correct answers to questions about payoffs for hypothetical players in a PGG. Respondents are paid for correct answers, and they can make multiple attempts at any given question, limited only by time. Third, participants in Study 3 move in lock-step with one another. Each page in the study takes an allotted amount of time no matter the respondent's behavior. This is to ensure that information cannot leak to other players in one's group via response times. For instance, Player 2 might notice that Player 1 made a decision rather quickly if Player 1 is allowed to advance from the Contribution page as quickly as she likes, since Player 1's advance triggers Player 2's decision period.

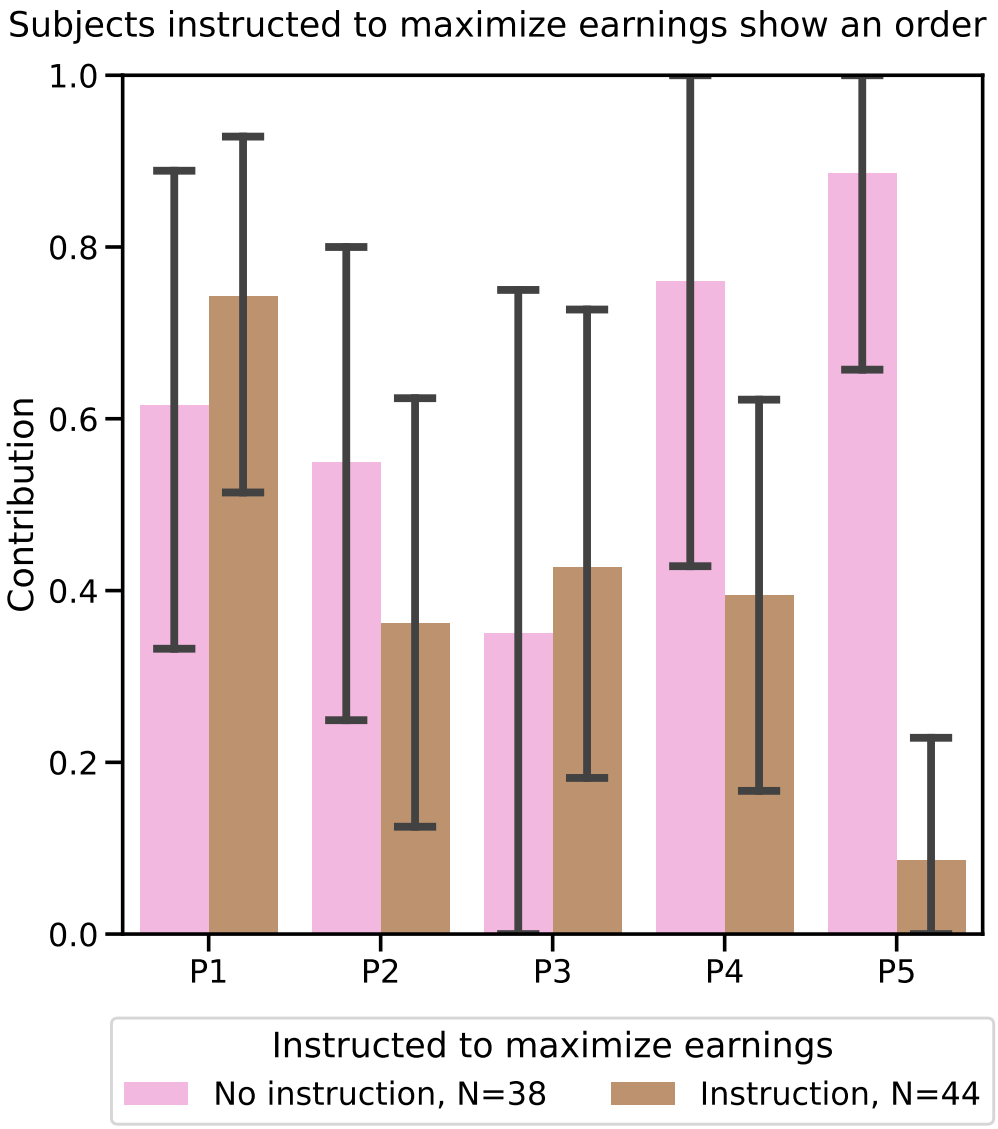


Figure 5: Study 3 shows the hypothesized decline with order among those who were instructed to be greedy. Subjects who passed comprehension checks. 95% CIs.

We observe the order effect in this non-preregistered study. A linear regression of contribution on order interacted with a binary instructed to be greedy / not instructed to be greedy variable using data from respondents who pass comprehension checks detects the interaction effect (β = -0.189, 95% CI = [-0.294, -0.074], F(3, 78) = 5.038, p = 0.006 for the interaction). Respondents receiving the prompt show a decline in

contribution with increasing order. There is substantial noise in estimates of the means, but we felt this result provided enough confidence to justify deploying this technique in the next, larger experiment.

4.4 Study 4: 5-Person Sequential Public Goods Game with random moves

Having collected some evidence suggesting that a prompt to maximize one's own payouts works as well as arriving at the experiment already wanting to do so, Study 4 extends Study 3 by applying the instruction to act to maximize one's own payouts to all participants and at larger scale, but with two new conditions: all respondents are either told that every player *before* them has his or her contribution determined randomly ("Random Before"), or that every player moving *after* them has his or her contribution determined randomly ("Random After"). This allows some insight into whether the order effect is somehow driven by the fact that other *people*, specifically, will be moving after the focal player—even though he cannot see their moves. This contrasts "open fate" vs. "closed fate" uncertainty, in that the Random Before condition probes closed fates and the Random After condition tests whether open fates are necessary for the effect, the necessity of open fates implying some causal thinking. Players are presented with a page that explains the setup, and are presented with symbols that make which players' moves were randomly decided clear. For instance, players see graphical representations similar to that shown in *Figure 6* on all pages from the point at which the concept of random moves is introduced until the end of the game. Respondents in Study 4 continue to move in lock-step, preventing the flow of information to other players in their group via response times. It may be noted that in this study Player 1 (in the Random Before condition) and Player 5 (in the Random After condition) play a standard sequential PGG in that they do not play with any players that have their contributions randomly determined at all, since there is no one before Player 1 and no one after Player 5.

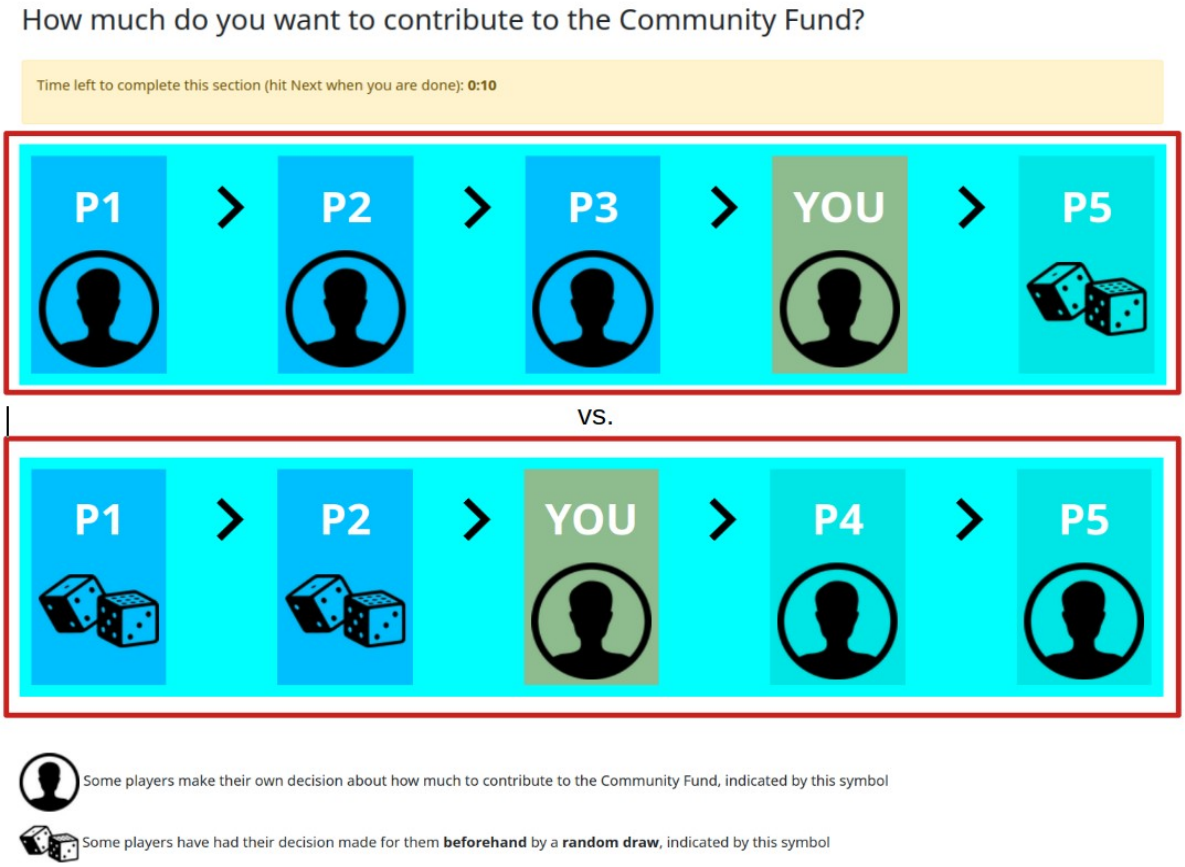


Figure 6: The stimuli on the Contribution page are shown in two conditions, in red boxes: above, Random After for Player 4 and below, Random Before for Player 3. Players see a graphical representation of their position relative to other players that clearly conveys which players are having their moves made by a random process. This is in addition to a previous screen that explains how some players are having their moves made for them by random processes.

Recall that in Study 4, all players are instructed to maximize their earnings—to act selfishly. We observe a decline in contribution with order only among those players who are told that everyone moving *before* them has his move determined randomly, while everyone moving *after* them is deciding on what move to make as usual. The preregistered linear regression contribution ~ order * random_before + wealth, differing

from previous analyses in that it controls for a measure of wealth, finds the effect. A significant regression equation was found, $\beta = -0.079$, 95% CI = [-0.134, -0.022], $F(4, 435) = 3.946$, $p = 0.0059$ for the interaction. We also find a significant equation not controlling for wealth, $\beta = -0.081$, 95% CI = [-0.136, -0.024], $F(3, 436) = 4.276$, $p = 0.005$. Among players told the opposite, that everyone moving after them has their move made randomly, we observe no order effect. 75% of players in this experiment contribute either 100% or 0% of their endowment, and the effect size and direction are preserved in this subset, $\beta = -0.096$, 95% CI = [-0.163, -0.028], $F(4, 354) = 3.849$, $p = 0.006$ for the interaction, lending credence to the idea that heterogeneity in the point at which the optimal move switches from cooperate to defect is driving the order effect. When we restrict the main analysis to only those players who passed a second set of comprehension checks at the end of the experiment (80% of players who passed the initial checks), we observe a larger effect ($\beta = -0.101$, 95% CI = [-0.158, -0.043], $F(4, 359) = 4.356$, $p = 0.0009$ for the interaction). This gives further reason to believe that the effect is present in respondents who actually understand the game. We do not observe a difference between the two simultaneous-play control conditions, one with no delay and one with a delay similar to that which Player 5 experiences before moving, which rules out the effects being due to mere time in the experiment.

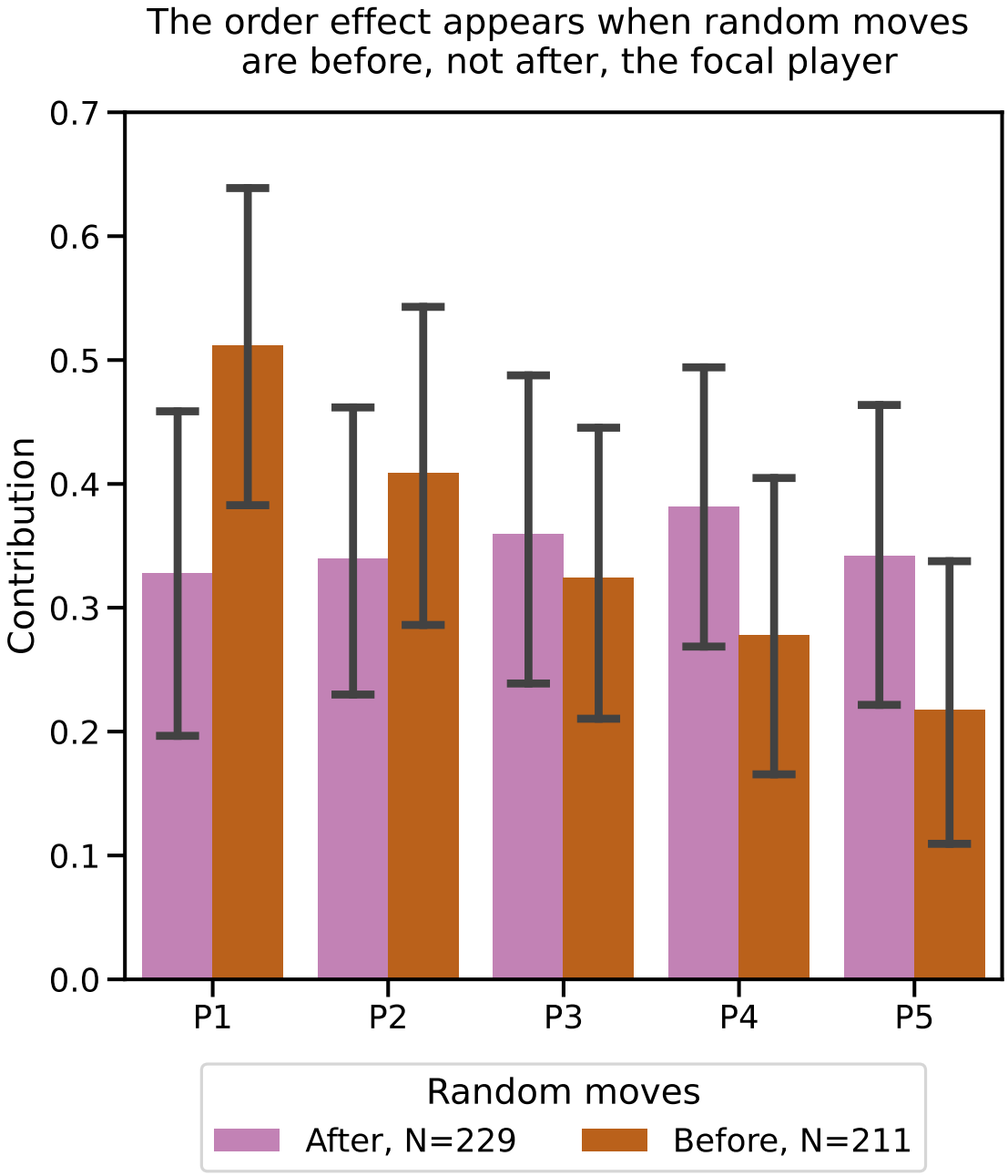


Figure 7: Study 4 shows a decline in contribution to the public good among players who are told that all players moving after them are making their own moves, and all players moving before them are having their moves made randomly. No effect is observed among players who are told that everyone moving after them has a move selected at random. Subjects who passed comprehension checks. 95% CIs.

5 Discussion

Reward-maximizing players in sequential PGGs without observation display an order effect. They cooperate more when they believe there are people moving after them, in proportion to the number of people moving after them. We have demonstrated, first, that it is a decline in contribution with increasing order; second, that the effect is present only in people trying to maximize their own rewards; third, that it is only present when the other players are making their own decisions, and fourth, that the effect goes forward in time (the presence or absence of decision-makers in the past does not matter). Four experiments support this view. Substantial training, practice games, and comprehension checks provide evidence that participants understand the game, and control conditions demonstrate the effects are not due to time spent waiting. Furthermore, when we filter based on *ex post* comprehension checks included in Study 4 in addition to the preregistered up-front checks, effect size in Study 4 increases. The fact that we observe this effect in participants who understand the game and who are trying to maximize their own rewards narrows the space of possible mechanisms: it appears that earlier movers tend to believe that contributing to the public good will maximize their payouts, and later movers believe that less contribution will maximize payouts—and so are more inclined to defect. The order effect's absence when subsequent players have their moves made randomly suggests implicit causal thinking at play: It is not just *that* I cooperate that suggests others will cooperate (in this case the effect would propagate backwards in time), but *if* I cooperate, others will cooperate—quasi-magical thinking. This makes it clear that the distinction between events that have happened (and therefore have fixed outcomes known to someone, "closed fates") and those that have not yet happened (which means they are uncertain in a deeper sense, "open fates") is important for behavior in this case. We speculate that a simple model may capture something of the process generating this behavior specifically in self-interested agents: these agents understand the rules of the game and are trying to

maximize their payouts—they just act as if their move is informative about all subsequent players' moves in a sequential game, and make the move that maximizes payouts if everyone who has not yet moved were to make the same move they do. A formalization of this model is included in the appendix.

If players evince quasi-magical thinking, if they are acting *as if*, what is driving this behavior? It may be that there is a source of information about what others might do in these games after all. In the total absence of other information, it is possible that players look to their own behavior in an attempt to learn about what others will do via social projection. If a focal player assumes some similarity between himself and the other players, it may seem reasonable to look to his own behavior as a source of information. If this is the case, there may be mechanisms by which people who are self-interested—who are trying to maximize their own payoffs independent of what is good for others—end up cooperating anyway. Projection from personal decisions to collective behavior can be rational in the sense that it can be consistent with Bayes' rule (Dawes, 1989; Hoch, 1987; Tarantola et al., 2017). Social projection could explain the sensitivity to other players making their own decisions (or not), but would not explain why the arrow of time ("closed fates" vs. "open fates") is important.

Self-signaling is another mechanism that may explain cooperation among these self-interested agents. In a self-signaling account, individuals regard their own decisions as informative about their unknown "deep" characteristics, such as morality, affection, dedication or willpower. Self-signaling implies that individuals will favor decisions that generate good news (a positive self-signal) about these characteristics, and that the effect is conditional on (a) the signal being costly (since signals that are too easy to generate are not informative) and (b) some prior uncertainty about the characteristics (since being quite sure about these types means self-signals are uninformative in comparison to what is already known) (Bernheim & Thomadsen, 2005; Bodner & Prelec, 2003; Dhar & Wertenbroch, 2012; Mijovic-Prelec & Prelec, 2010). Agents who are self-signaling are motivated to produce signals that give them good news. In the case of a PGG, self-interested players may be motivated to learn from their own

behavior that others moving after them will also contribute, thereby raising their estimate of their payoffs. Adjusting your own estimate of your future profits upwards is pleasurable, so there is utility to be gained from that adjustment (diagnostic utility) in addition to the standard utility from the payout itself (outcome utility). Crucially, from the standpoint of both theory and empirical evidence, self-signaling does not require a perceived causal link between decisions and the underlying characteristic of interest; it can influence decisions even when their causal irrelevance is made obvious by experimental design as in Quattrone & Tversky (1984). Projection from personal decisions to collective behavior, as in social projection, is consistent with Bayes' rule. With decisions, however, there is a causal component to projection. By freely choosing an action, the individual also chooses the signal about collective behavior that the action delivers. Causal power over one's expectations about others' prosocial behavior may be motivationally, if not logically, equivalent to a feeling of power over their actual behavior. However, like with social projection, the usual formulation of self-signaling does not naturally provide a direction in time for the effect: it is possible to self-signal about open and closed fates.

The idea of universalization (Levine et al., 2020) may also shed some light. While it is a mechanistic account of moral judgment rather than rational inference or decision-making revealed in behavior, the fact that asking the question "What if everyone felt free to do this?" occurs in the moral domain may imply that it is a special case of a more general strategy: considering one's own move as a signal about what others will do, and then considering the utility to be found in the circumstances that many moves like your own create: "What if everyone acted as I have"?

Self-signaling, social projection, and universalization each could lead to patterns of behavior that appear to be people acting as if their actions can influence other people without communicating, i.e., as if they had magical powers. However, maybe even magical powers have limits: they can be circumscribed by logic and commonsense metaphysics. In particular, past actions of other people may be unknown, but are not reversible. In contrast, future actions of other people are both unknown and potentially

open to influence. Miller and Gunasegaram (1990) demonstrated that, while events in the past are considered fixed, future events are treated as mutable. Moreover, future actions are perceived as more intentional and blameworthy than otherwise identical past actions (Burns et al., 2012). These facts point to deeply-held priors that direct thoughts like these towards the future, potentially making any of self-signaling, social projection, or universalization viable underlying mechanisms given a strong enough prior.

We observe an order effect, and most players contribute 0% or 100% of their endowments, but not all are at ceiling or floor. We do observe average contributions among self-interested respondents to be above floor, to be more than nothing, at the end of a sequence, and lower than ceiling, less than 100% of the endowment, at the start. There are several things that could contribute to this, including mis-classification by the up-front SVO battery, mismatch between social preferences on the SVO and social preferences in the subsequent PGG, inconsistent effects of the prompt to act selfishly, and subjects who do not understand the game making it through comprehension checks (some percentage of even random responders will make it through). It remains for future work to investigate which of these explanations contribute to the phenomenon, and aside from them why responses are not exactly floor or ceiling even among the self-interested who understand the game. Sampling-based approaches may shed some light on this feature of the data.

There is also some question about the behavior of Prosocials. Why don't we see contributions at ceiling or an order effect? Prosocials may be giving less than 100% for many of the same reasons behavior among payoff-maximizers is not optimal: mis-classification by the up-front SVO battery, mismatch between social preferences on the SVO and social preferences in the subsequent PGG, and subjects who do not understand the game making it through comprehension checks by chance. We would not predict an order effect because, on our model, the benefits from defecting in a situation where quasi-magical thinking is strong are very weak compared to those from cooperating, resulting only in edge cases where a significantly prosocial player might

defect in last position if the multiplier is small enough. But, going beyond our model, it is reasonable to think that people who arrive at the task already prosocial are likely not even doing the kind of utilitarian calculus those who are trying to maximize payouts engage in. It seems reasonable that they have decided to cooperate in a general sense in advance of the game, perhaps using some simple heuristic that cooperating in games with small stakes is always the winning move, and they stick to that heuristic—thereby avoiding the costs of carefully considering the different states the game may take, which may be large in comparison to our relatively small stakes.

Finally, whatever the mechanism, understanding a means by which self-interested actors might decide to contribute to the public good is relevant to many practical policy questions. For example, applying our findings, an agent might say to herself: if I vote then it is more likely that other people will vote; if I conserve energy, then others will conserve as well; if I contribute to a public good, so will others—and this action is actually best for me independent of what's good for everyone else. This could explain why even purely self-interested individuals might feel that their investment of time or effort for a public cause will pay off, pointing to a class of interventions that highlight that there are people deciding for themselves to contribute—or not—at a later time.

6 Methods

Ethics: All studies reported here were approved by MIT's Committee on the Use of Humans as Experimental Subjects (COUHES) and comply with all relevant ethical regulations. We obtained electronic consent from all respondents.

A convenience sample provided by Amazon Mechanical Turk (MTurk) was selected for all experiments because it is a reasonable approximation of American adults for our purposes. Studies 3 and 4 used a panel filtered by Cloud Research due to declines in the quality of unfiltered MTurk samples. This work makes the point that these effects exist in human populations, and it is left for future work to examine how they vary across ages, sexes, SES, cultures, and other covariates of interest. All experiments also involved extensive training and comprehension checks. Data from respondents who failed one or more pre-play comprehension check questions was excluded. All experiments are real-time online group tasks, where respondents interact via text chat before learning the rules of the game in order to establish some sense that they are completing the task with actual people in real time. All studies except for Study 3 were preregistered on osf.io.

6.1 Study 1

Participants. 1668 U.S.-based participants from Amazon Mechanical Turk completed the study. Median total pay per respondent (including bonuses for accurate predictions) is $3.16 (*SD* = 0.90), yielding an hourly rate o $18.46 per hour at 10 minutes' duration (SD = 8.38). Of 1668 respondents, 69% (1151) passed all of the comprehension check questions. Data from batches 1 and 8 were excluded due to technical problems resulting in server crashes during the experiment. Analysis is limited

to the 60% (1002 total; 800 sequential) responses which passed comprehension checks and were not in batches 1 or 8. To estimate the sample size required, we performed a power analysis via simulation using pilot data.

Materials and procedure. Study 1 is a one-shot sequential PGG with a multiplier of two. Three players can transfer any part of their individual $1 endowment. The total transfer amount from all participants is then doubled and distributed evenly among the players, irrespective of individual transfers. Order of play is determined randomly, with no communication among players. The only difference in information among the players is knowledge of their position in the sequence. Each participant was assigned to one of four conditions: orders 1-3 and a simultaneous-move condition. Players arrive at the experiment web page, complete a consent form, and then engage in a real effort task transcribing nonsense sentences in order to filter out bots. After this, they are placed in a chat room for 30 seconds after all players in their group have arrived to ensure participants believe the experiment is, in fact, a real game in real time with real people. After the chat, respondents are provided with an explanation of the rules of the game (which appear on every subsequent page for reference). The PGG is framed as a question of how much to contribute to a “Community Fund”. A player can “transfer” some or all of her endowment to the Community Fund, and she may “keep” some amount. Instructions include if-then statements about the consequences of certain moves to aid understanding.

Respondents are then asked three comprehension and attention check questions: (1) Do any of the other players know how much you decide to contribute? (2) No matter what the other players do, what earns you the most money? TRANSFERRING to the community fund or KEEPING your endowment? and (3) What year is it? As with the Prisoner’s Dilemma, responses to the comprehension questions are only relevant to data analysis: players continue on whether or not they have answered correctly. Since players do not interact after the initial chat, players who fail the comprehension checks can have no further influence on those that pass. Players who fail comprehension checks remain in the game because the games are real games

happening in real time, and so there moves are needed to calculate payouts without deception.

After having completed the comprehension questions, players make their move. The contribution page includes a graphic at the top highlighting their place in the sequence of moves in red (see supplemental online materials). Players in the simultaneous condition do not see any indication of sequence since they are moving simultaneously. Respondents then complete prediction questions, and then a Social Value Orientation (SVO) slider battery (Murphy et al., 2011; code based on Bakker, 2016/2019)[2]. The SVO battery measures preferences for how to allocate resources between oneself and others. The standard battery categorizes respondents into Individualistic (concerned only with what is best for self), Competitive (maximize own outcomes as with Individualistic, but also minimize the outcomes for others), Prosocial (maximize outcomes for both self and other), and Altruistic (eager to give up own gains to help others). Players then exit the experiment and are paid.

Analysis. The preregistered analysis used to investigate the impact of order on contribution is a linear regression contribution ~ order, with order treated as ordinal and backwards-difference coded. Backwards difference coding enforces a statistical significance test for each comparison, 1 vs. 2 and 2 vs. 3, enforcing a stepwise change from 1 to 2, 2 to 3, etc.

6.2 Study 2

Participants. 1089 U.S.-based participants from Amazon Mechanical Turk completed the study. Median total pay per respondent (including bonuses for accurate predictions) is $3.40 (*SD* = 0.53), yielding an hourly rate of $11.18 per hour at 18 minutes' duration (SD = 3.83). Of 1089 respondents, 66% (720 total, 599 sequential) passed all of the up-front comprehension check questions. Time on the decision-making page for players 1 and 2 was variable due to a programming error. Amazon Mechanical

[2] SVO is measured post-treatment, but we do not observe an effect of treatment on SVO.

Turk was selected because it is a reasonable approximation of a representative sample of American adults for our purposes. To estimate the sample size required, we performed a power analysis via simulation using pilot data.

Materials and procedure. Study 2 is a one-shot sequential PGG identical to Study 1, with the exception that there are five players rather than three, that the up-front chat was 90 instead of 30 seconds, and that respondents complete the SVO slider battery before the PGG.

Analysis. The preregistered analysis used to investigate the impact of order on contribution is a simple OLS linear regression contribution ~ order (excluding simultaneous participants). Backwards-difference coding (as specified in Study 1) would have required unworkably large sample sizes per bootstrapped power analyses.

6.3 Study 3

Participants. 86 U.S.-based participants from Amazon Mechanical Turk completed the study via Cloud Research. Median total pay per respondent (including bonuses for accurate predictions) is $3.99 (*SD* = 1.25), yielding an hourly rate of $15.99 per hour at 15.2 minutes duration (SD = 4.39). Of the 86 respondents who completed the task, 82 (95.0%) passed all of the up-front comprehension check questions.

Materials and procedure. Study 3 adds some features to the basic design from Study 2. In Study 3, SVO is not measured. Instead, players are randomized to a "self-interested" and a "Non-self-interested" condition. In the self-interested condition, players see a prompt:

> Please try to play this game **however you think will make you the most money.** We understand that sometimes you want to help other people, but for the purposes of this experiment we want you to try to make as much money as possible.

Players are also randomized to a delayed simultaneous condition in addition to the simultaneous condition from previous studies, to control for effects that arise merely from waiting. Respondents randomized to the delayed simultaneous condition wait for 80 seconds on the standard wait page for the task (which contains a simple game they may play if they wish). In addition, players in Study 3 move in lock-step throughout the task. Instead of being able to advance on certain pages when they feel they are ready, players move in lock-step with a certain number of seconds allotted for each page (so subsequent players cannot infer anything from how quickly those previous to them have moved). Pages on which players make their contribution or make predictions do not force a player to stay for a certain amount of time, but rather let the player move on to a wait page when the decision has been made. The wait page soaks up any remaining time.

Analysis. There was no preregistration for Study 3 since it was meant to be a simple, fast test of whether or not instruction to be self-interested would produce an order effect. The analysis used is an OLS linear regression, contribution ~ order * instruct_or_no, with instruct_or_no being a binary indicator of whether or not respondents were instructed to be self-interested.

6.4 Study 4

Participants. 539 U.S.-based participants from Amazon Mechanical Turk via Cloud Research completed the study, with 440 in sequential conditions and 99 in simultaneous conditions. Median total pay per respondent (including bonuses for accurate predictions) is $4.24 (*SD* = 1.19), yielding an hourly rate of $16.13 per hour at 16.0 minutes duration (SD = 3.97). Of 539 respondents, 440 (82%) passed all of the up-front comprehension check questions. To estimate the sample size required, we performed a power analysis via simulation using pilot data and data from previous experiments.

Materials and procedure. Study 4 is a one-shot sequential PGG identical to Study 3, with the exception that players instead of being randomized to get the instruction to maximize earnings or not, all players receive that instruction and instead they are randomized between two conditions, fully crossed with orders 1-5: players are told that everyone before them in the sequence has their decision about how much to contribute to the public good made by a random process ("Random Before"), or players are told that everyone after them has their decision made by a random process ("Random After"). As in Study 3, there are two simultaneous control conditions: one with a delay equivalent to the wait time 5th-movers experience in the sequential game, and one without which is equivalent to moving first.

Analysis. The preregistered analysis used to investigate the impact of order on contribution in Study 4 is a simple OLS linear regression that, in addition to what is used for Study 3, controls for self-reported wealth: contribution ~ order + wealth among those who are told that players before them have their moves made randomly ("Random Before"). Wealth was added to the regression given the expectation, common across economics, that players' sensitivity to payoffs is modulated by the marginal change in their wealth or similar.

7 References

Abele, S., & Ehrhart, K.-M. (2005). The timing effect in public good games. *Journal of Experimental Social Psychology*, *41*(5), 470–481. https://doi.org/10/bkgq9d

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *The Quarterly Journal of Economics*, *107*(3), 797–817. https://doi.org/10.2307/2118364

Bernheim, B. D., & Thomadsen, R. (2005). Memory and Anticipation. *The Economic Journal*, *115*(503), 271–304. https://doi.org/10.1111/j.1468-0297.2005.00989.x

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, *100*(5), 992–1026. https://doi.org/10.1086/261849

Bodner, R., & Prelec, D. (2003). Self-signaling and diagnostic utility in everyday decision making. In I. Brocas & J. D. Carrillo (Eds.), *The Psychology of Economic Decisions*. Oxford University Press.

Bohnet, I., & Frey, B. S. (1999a). Social Distance and Other-Regarding Behavior in Dictator Games: Comment. *American Economic Review*, *89*(1), 335–339. https://doi.org/10.1257/aer.89.1.335

Bohnet, I., & Frey, B. S. (1999b). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior & Organization*.

Budescu, D. V., & Au, W. T. (2002). A model of sequential effects in common pool resource dilemmas. *Journal of Behavioral Decision Making*, *15*(1), 37–63. https://doi.org/10.1002/bdm.402

Budescu, D. V., Au, W. T., & Chen, X.-P. (1997). Effects of Protocol of Play and Social Orientation on Behavior in Sequential Resource Dilemmas. *Organizational Behavior and Human Decision Processes*, *69*(3), 179–193. https://doi.org/10.1006/obhd.1997.2684

Budescu, D. V., Suleiman, R., & Rapoport, A. (1995). Positional Order and Group Size Effects in Resource Dilemmas with Uncertain Resources. *Organizational Behavior and Human Decision Processes*, *61*(3), 225–238. https://doi.org/10.1006/obhd.1995.1018

Burns, Z. C., Caruso, E. M., & Bartels, D. M. (2012). Predicting premeditation: Future behavior is seen as more intentional than past behavior. *Journal of Experimental Psychology: General*, *141*(2), 227–232. https://doi.org/10/bz6ngt

Çelen, B., & Kariv, S. (2004). Observational learning under imperfect information. *Games and Economic Behavior*, *47*(1), 72–86. https://doi.org/10.1016/S0899-8256(03)00179-9

Chen, Y., & Zhong, S. (2022). Uncertainty Motivates Morality. *SSRN Electronic Journal*.

Colman, A. M., & Gold, N. (2018). Team reasoning: Solving the puzzle of coordination. *Psychonomic Bulletin & Review*, *25*(5), 1770–1783. https://doi.org/10.3758/s13423-017-1399-0

Daley, B., & Sadowski, P. (2017). Magical thinking: A representation result. *Theoretical Economics*, *12*(2), 909–956. https://doi.org/10/f99g8r

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology*, *25*(1), 1–17. https://doi.org/10.1016/0022-1031(89)90036-X

Dhar, R., & Wertenbroch, K. (2012). Self-Signaling and the Costs and Benefits of Temptation in Consumer Choice. *Journal of Marketing Research*, *49*(1), 15–25. https://doi.org/10/bbng3z

Eichenseer, M. (2023). Leading-by-example in public goods experiments: What do we know? *The Leadership Quarterly*, 101695. https://doi.org/10.1016/j.leaqua.2023.101695

Figuières, C., Masclet, D., & Willinger, M. (2012). Vanishing Leadership and Declining Reciprocity in a Sequential Contributions Experiment. *Economic Inquiry*, *50*(3), 567–584. https://doi.org/10.1111/j.1465-7295.2011.00415.x

Henrich, J., & Muthukrishna, M. (2021). The Origins and Psychology of Human Cooperation. *Annual Review of Psychology*, *72*(1), 207–240. https://doi.org/10.1146/annurev-psych-081920-042106

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology*, *53*(2), 221–234. https://doi.org/10.1037/0022-3514.53.2.221

Hristova, E., & Grinberg, M. (2010). *Testing Two Explanations for the Disjunction Effect in Prisoner's Dilemma Games: Complexity and Quasi-Magical Thinking*. Annual Meeting of the Cognitive Science Society.

Jeffrey, R. C. (1990). *The Logic of Decision*. University of Chicago Press.

Kohlberg, E., & Mertens, J.-F. (1986). On the Strategic Stability of Equilibria. *Econometrica*, *54*(5), 1003–1037. https://doi.org/10.2307/1912320

Kreps, D. M. (1990). *Game Theory and Economic Modelling*. Oxford University Press.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. https://doi.org/10.1037/0022-3514.32.2.311

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National*

Academy of Sciences, *117*(42), 26158–26169. https://doi.org/10.1073/pnas.2014505117

Luce, D. R. (1992). Where does subjective expected utility fail descriptively? *Journal of Risk and Uncertainty*, *5*(1), 5–27. https://doi.org/10.1007/BF00208784

Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. *Journal of Economic Behavior & Organization*, *64*(2), 216–231. https://doi.org/10.1016/j.jebo.2005.07.003

Mijovic-Prelec, D., & Prelec, D. (2010). Self-deception as self-signalling: A model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 227–240. https://doi.org/10/c2tqv2

Miller, D. T., & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology*, *59*(6), 1111–1118. https://doi.org/10.1037/0022-3514.59.6.1111

Morris, M. W., Sim, D. L. H., & Girotto, V. (1998). Distinguishing Sources of Cooperation in the One-Round Prisoner's Dilemma: Evidence for Cooperative Decisions Based on the Illusion of Control. *Journal of Experimental Social Psychology*, *34*(5), 494–512. https://doi.org/10/d4cs3w

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring Social Value Orientation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1804189

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46*(2), 237–248. https://doi.org/10/dr4gxj

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425. https://doi.org/10.1016/j.tics.2013.06.003

Robinson, A. E., Sloman, S. A., Hagmayer, Y., & Hertzog, C. K. (2010). Causality in Solving Economic Problems. *The Journal of Problem Solving*, *3*(1). https://doi.org/10/ggdjxn

Roemer, J. E. (2010). Kantian Equilibrium. *The Scandinavian Journal of Economics*, *112*(1), 1–24. https://doi.org/10.1111/j.1467-9442.2009.01592.x

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, *127*, 45–57. https://doi.org/10.1016/j.jpubeco.2014.03.011

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, *24*(4), 449–474. https://doi.org/10/d6thrq

Stefan, S., & David, D. (2013). Recent developments in the experimental investigation of the illusion of control. A meta-analytic review: A meta-analysis of the illusion of control. *Journal of Applied Social Psychology*, *43*(2), 377–386. https://doi.org/10.1111/j.1559-1816.2013.01007.x

Tarantola, T., Kumaran, D., Dayan, P., & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00826-8

Von Neumann, J., & Morgenstern, O. (2004). *Theory of games and economic behavior* (60th anniversary ed). Princeton University Press.

Zelmer, J. (2003). Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics*, *6*(3), 299–310. https://doi.org/10.1023/A:1026277420119

8 Appendix

8.1 Preregistrations

Please note that these pre-registrations reference a quadratic effect, which is unrelated to the order effect and will be the subject of a separate paper.

Study 1:
OSF preregistration: https://osf.io/3vsxk

Anonymous link to registration for review: https://osf.io/3vsxk/?view_only=bf35d2d3d39d48b68869c2cf78bf8e2b

Study 2:
OSF preregistration: https://osf.io/gw8nc

Anonymous link to registration for review: https://osf.io/gw8nc/?view_only=aa0c4825dac4469a82f0156b77390e3c

Study 3:
no preregistration

Study 4:
OSF preregistration: https://osf.io/3kepm

Anonymous link to registration for review: https://osf.io/3kepm/?view_only=614de27fdf4b40a0bad48847f32c879d

8.2 Model

Here we provide a more precise statement of a model that generates the hypothesized interaction between the order effect and pro-social motivation.

8.2.1 Prosocial preferences

Consider a sequential PGG with n players endowed with 1 payoff unit each, and multiplier m, with $1 < m < n$. Players are indexed by their order of play in the sequence, $\mathrm{i}{=}1, \ldots, \mathrm{n}$. Let a_i denote the contribution of player i, $0 \leq a_i \leq 1$, and p_i the payoff to player i.

$$p_i = 1 - a_i + \frac{m}{n} \sum_{k=1}^{n} a_k \qquad \text{(Equation 2)}$$

Prosocial preferences are modeled through a prosocial parameter s_i where $s_i = 0$ indicates pure self-interest and $s_i = 1$ pure prosocial motivation. In keeping with the experimental setup, we assume that players do not learn the specific contributions of other players. The utility of player i is therefore a function of the two variables the player does or will know, namely contribution a_i and payoff p_i:

$$u_i\left(a_1, \ldots, a_n\right) = \left(1 - s_i\right) p_i + s_i m a_i \qquad \text{(Equation 3)}$$

where p_i is determined by the game formula, eq. 1 . A purely self-interested player $\left(s_i = 0\right)$ will aim to maximize own payoff, $u_i = p_i$; a purely prosocial player $s_i = 1$ will aim to maximize the impact of his contribution to the public good, $u_i = m a_i$. The prosocial motive, captured by the second term, thus reflects the impact of own contribution to the public good; other players' contributions enter the utility model only insofar they determine the first, self-interested utility term. In other words, players: (a) care how their action affects the payoffs of others, (b) care how other players' contribution affect their own payoff, but (c) do not care how other players' actions affect each others' payoffs.

8.2.2 Decision dependent expectations

We assume that players compare expected utilities conditional on contributing $(a_i = 1)$ or not contributing $(a_i = 0)$, and choose whichever expected utility is higher (we ignore here fractional contributions). The decision criterion is therefore the difference between the two expected utilities:

$$a_i = 1 \Longleftrightarrow E\left[u_i \mid a_i = 1, s_i\right] > E\left[u_i \mid a_i = 0, s_i\right](1) \quad \text{(Equation 4)}$$

A player knows the value of their prosocial parameter and hence also knows the utility function in eq. 1. If he were just a spectator, not making a decision, his expectation of the contribution a_k of another, randomly selected player k would exhibit projection, along the lines of Bayesian updating. The simplest version of such updating is linear:

$$E\left[a_k \mid s_i\right] = b + cs_i(1) \quad \text{(Equation 5)}$$

Prosocial players are more optimistic about the overall contribution level, other things equal.

The critical assumption we now make is that expectations of future players' contributions are additionally influenced by a player's own action, while expectations of prior players' contributions are not influenced. Let $a_{k<i}$ denote the contribution of any player moving before player i, and $a_{k>i}$ the contribution of any player moving after player i. We assume:

$$\begin{aligned} E\left[a_{k<i} \mid a_i, s_i\right] &= b + cs_i \\ E\left[a_{k>i} \mid a_i, s_i\right] &= b + cs_i + d\left(a_i - E\left[a_k \mid s_i\right]\right) \\ &= (b - d) + (c - d)s_i + da_i \end{aligned}$$

where $\mathrm{E}[a_k \mid s_i] = b + cs_i$ from eq. 4 is substituted in the final line.

There is no perceived causality with respect to previous players, since expectations are the same irrespective of contribution:

$$E\left[a_{k>i} \mid 1, s_i\right] - E\left[a_{k>i} \mid 0, s_i\right] = d$$

There is perceived causality with respect to future players, proportional to the 'magical influence' parameter 'd':

$$E\left[a_{k>i} \mid 1, s_i\right] - E\left[a_{k>i} \mid 0, s_i\right] = d$$

The decision criterion in eq. 3 can be expressed as:

$$E[u_i a_i = 1, s_i] - E[u_i a_i = 0, s_i] = (1 - s_i)E[p_i a_i = 1, s_i] + s_i m - (1 - s_i)E[p_i a_i = 0, s_i]$$
$$= (1 - s_i)(E[p_i a_i = 1, s_i] - E[p_i a_i = 0, s_i]) + s_i m$$
$$= (1 - s_i)(-1 + mnE[k = 1^n a_k a_i = 1, s_i] - mnE[k = 1^n a_k a_i = 0, s_i]) + s_i m$$

where the first line follows from equation 2 and the third line from equation 3.

Assuming that expectations about contributions of previous players are not affected by own contribution, the difference in expected total contribution resolves as:

$$E\left[\sum_{k=1}^{n} a_k \mid a_i = 1, s_i\right] - E\left[\sum_{k=1}^{n} a_k \mid a_i = 0, s_i\right] = 1 + E\left[\sum_{k=i+1}^{n} a_k \mid a_i = 1, s_i\right] - E\left[\sum_{k=i+1}^{n} a_k \mid a_i = 0, s_i\right]$$
$$= 1 + d(n - i)$$

Substituting into the criterion,

$$E\left[u_i \mid a_i = 1, s_i\right] - E\left[u_i \mid a_i = 0, s_i\right] = (1 - s_i)\left(-1 + \tfrac{m}{n}(1 + d(n - i)\right) + s_i m$$

(Equation 6)

For any particular value of s_i, the minimum 'magical influence' parameter $d^*(i)$ that leads to $a_i = 1$, i.e., full contribution to the Public Good, is computed as:

$$E\left[u_i \mid a_i = 1, s_i\right] - E\left[u_i \mid a_i = 0, s_i\right] = 0 \iff d^*(i) = \tfrac{-m - smn + n}{m(n-i)}$$

(Equation 7)

Note that $d^*(i)$ is increasing in i (if the expression is positive) and decreasing in s_i. The increase in i is the order effect: Players later in the sequence require a higher value of $d^*(i)$ in order to contribute. Assuming that d is an exogenous parameter with some distribution in the respondent sample, fewer players will clear the cutoff and contribute if they are later in the sequence. The decrease in s_i simply indicates that prosocial players require less magical thinking in order to contribute.

The second implication of the model is that the slope of this function with respect to i (the term in the brackets in eq. 6) is steeper if s_i is smaller, that is, if players are more self-interested. To show this, we differentiate:

$$\frac{dd^*(i)}{di} = \frac{1}{(n-i)^2}\left(\frac{n-m}{m} - \frac{s_i}{(1-s_i)}n\right)$$

which is decreasing in s_i. This is the hypothesized interaction of order and prosociality. Less prosocial players will exhibit a stronger order effect. Conversely, the order effect should disappear if s_i is sufficiently high.