# Nothing to See, Nothing to Say, and Noting How Much

## Three Essays on Information and Behavior

By

Matthew Cashman

A.B. Chemistry & Philosophy
Hamilton College, 2008

S.M. Management Research
Massachusetts Institute of Technology, 2020

Submitted to the Department of Management in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023



Authored by: Matthew Cashman
Department of Management
10th of August 2023

Certified by: Drazen Prelec
Department of Management
Thesis supervisor

Accepted by: Eric So
Sloan Distinguished Professor of Financial Economics
Professor, Accounting and Finance
Faculty Chair, MIT Sloan PhD Program

# Nothing to See, Nothing to Say, and Noting How Much

## Three Essays on Information and Behavior

by

Matthew Cashman

Submitted to the Department of Management on the 10th of August 2023 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Management

## ABSTRACT

I present three essays that examine information flows and behavior. The first examines the effect of sequential play in Public Goods Games in cases where players move one after another but do not see each others' moves. Even with no information flow—when there is nothing to see of others' decisions—order of play affects contributions to the public good. The second essay considers pre-play socializing and its effects on coordination games and hold-up games. Pre-play small-talk results in better outcomes even when players talk before they know they will be playing a game—before they have anything of strategic relevance to say. The third essay presents a novel quantitative, empirical means of measuring the flow of memes through minds. Most ways of learning what other people know rely on strong commitments to what the right question to ask is. Using cloze completion tasks I outline a principled, content-agnostic method of estimating how much information from a given text is stored in a reader's mind.

Thesis supervisor: Drazen Prelec

Title: Digital Equipment Corp. Leaders for Global Operations Professor of Management
Professor of Management Science
Professor of Economics
Professor of Brain and Cognitive Sciences

# Acknowledgments

I want to thank my academic families—Fiery Cushman's lab at Harvard Psychology, Joe Henrich's lab at Harvard Human Evolutionary Biology, and Dave Rand's lab at MIT Sloan—for making earning a Ph.D. a pleasure in addition to a lot of hard work. Many of you have become friends, and I wouldn't have done it without you. Or, more to the point, I did it largely because of you. I have to thank Fiery for giving me a chance in academia (and a home to the World Crokinole Championship team), Joe for making me think about evolution, which has become such an important part of how I see the world, and Dave for making our department a community instead of just a place to work.

Finally, I'd like to thank my advisor Drazen for six years of wisdom, carefully applied, which I hope has made me something of a scientist.

This work is dedicated to Alex, who helped me forward every step of the way, and to Cat & O.C., who held me back.

# Table of Contents

# Table of Figures

# Index of Tables

# 1 Playing blind: Absent other information, self-interested players act as if others will mirror their moves

*In collaboration with Drazen Prelec*


## 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.

## 1.1 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.

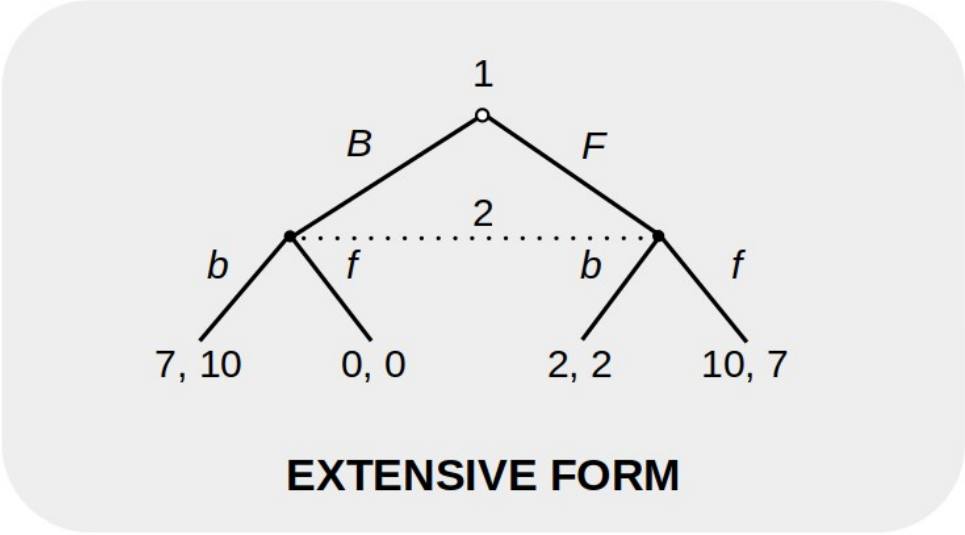
The games used here are “sequential” in that players move one after another and players’ moves are unobserved in that there is no information flow between players: any given player knows nothing about the decisions players who have already moved have made. For example, a five-person PGG is sequential 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.

## 1.2 Review

### 1.2.1 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.*

The work on equilibria in sequential games with observation has progressed a great deal since von Neumann and Morgenstern, and it may be instructive to highlight a few theoretical touchstones. Building on von Neumann and Morgenstern's foundation as well as prior work outlining Nash equilibria (Nash, 1951), subgame perfect equilibria (Selten, 1965), and related equilibrium refinements, Kreps and Wilson (1982) introduced the concept of sequential equilibria. In this formulation, the strategy of each player should be consistent with their beliefs about other players' strategies. The idea is that a

player at some stage of a sequential game develops beliefs in the form of probability distributions over the information available about each node of the extensive-form game. These beliefs can be about unobserved moves that have happened and moves that are yet to happen. Given this set of beliefs, it is possible to maximize expected value by developing a set of strategies and beliefs Kreps and Wilson refer to as an assessment. This formulation incorporates the sequential nature of games naturally, allowing for incomplete or imperfect information.

The advent of psychological game theory in the late 1980s further enriched the modeling of sequential games. Geanakoplos et al. (1989) introduced psychological considerations into game theory by treating players' beliefs as an integral part of the game structure rather than as independent entities. In sequential games in particular, these psychological considerations allow accounting for players' evolving beliefs over time and their subsequent impact on their decision-making process. The authors point out that a player's utility function takes into account his own beliefs, beliefs about others' beliefs, etc. These beliefs are not limited to probability distributions over possible information sets of the game, as in Kreps and Wilson, but encompass things like being indignant when held up, or the sheer joy of retaliation even when it is costly. It may be the case that some of these phenomena might be captured by sufficiently complex utility functions (and so the usual machinery might be used), but Geanakoplos et al. make the point that these psychological states are endogenous and are dependent on previous psychological states. Given this relationship between psychological states and the objective payouts of any given game, the standard game-theoretic machinery cannot effectively model the state of the game: backwards induction no longer leads to unique strategies. The authors describe the expanded problem space and develop analogs of Nash equilibria and subgame perfect equilibria that take into account these psychological properties.

Rabin (1993) introduces a game-theoretic framework that incorporates fairness into a broad range of economic models. In a very general sense, Rabin introduces the idea that a player's utility function can be a function of other players' utility functions,

with the relationship between the two reflecting something like a simple normative ethics. Rather than psychological properties being endogenous to the game but the exact machinery generating them being somewhat mysterious, Rabin provides a stripped-down normative ethics that makes exactly which fairness-related emotions occur when and where intelligible based on the material aspects of the game. Rabin's framework is based on three stylized facts:

1. People are willing to sacrifice their own material well-being to help those who are being kind.
2. People are willing to sacrifice their own material well-being to punish those who are being unkind.
3. Both motivations (helping the kind and punishing the unkind) have a greater effect on behavior as the material cost of sacrificing becomes smaller.

He develops a game-theoretic solution concept, the "fairness equilibrium", that incorporates these stylized facts. Fairness equilibria do not in general constitute either a subset or a superset of Nash equilibria; that is, incorporating fairness considerations can both add new predictions to economic models and eliminate conventional predictions. Among other results, he demonstrates the special role of "mutual-max" outcomes (in which, given the other person's behavior, each person maximizes the other's material payoffs) and "mutual-min" outcomes (in which, given the other person's behavior, each person minimizes the other's material payoffs). Rabin’s results in this paper are, however, limited to normal-form games as he acknowledges. The additional structure in extensive-form games, in particular representations of sequentiality, remain unexplored in this work. Still, Rabin’s contribution is an important addition to the theoretical toolkit necessary for understanding behavior in sequential games.

Fehr and Schmidt (1999) extended Rabin’s theme and introducing a formal model for inequality aversion. They argued that individuals might not purely maximize their own material payoffs, but also consider the allocation of payoffs to other players, reflecting a preference for fairness. This preference for fairness, they argue, explains a wider range of empirical phenomena than does Rabin’s framework based on three

stylized facts. The authors propose a utility function that captures a player's disutility from inequitable outcomes. This utility function includes two parameters: one that captures the player's aversion to disadvantageous inequity (i.e., receiving less than others) and another that captures the player's aversion to advantageous inequity (i.e., receiving more than others).

The authors argue their model explains behavior observed empirically in a variety of games, including ultimatum games, PGGs, and market games. In particular, the model predicts that players will cooperate in one-shot prisoner's dilemma games and contribute to public goods, even in the absence of repeated interaction or reputation effects. The model also predicts that competition can lead to equitable outcomes, even in the absence of perfect information or large numbers of competitors.

In the context of sequential games, the authors' model of inequity aversion can explain why players might choose to cooperate in the first stage of a game. If there is a group of “conditionally cooperative enforcers” who care about inequality and are willing to punish defectors, then full cooperation can be sustained as an equilibrium outcome. This is because these enforcers are happy to cooperate if all others cooperate as well, and they can credibly threaten to punish a defector if they care sufficiently about inequality.

Dufwenberg & Kirchsteiger (2004) add to our understanding of extensive-form games, where a player's decision at any stage depends not only on his rational understanding of the game but also on his belief about the fairness of previous actions and his inclination to reciprocate accordingly. In comparing their model to Rabin (1993), the authors note that the novelty of their approach is that it takes into account changes in strategic choices and reciprocity as new subgames occur. Their sequential reciprocity equilibrium (SRE) framework models players' expectations and intentions over the course of play, allowing for a richer understanding of strategic interactions that begins to model the game-theoretic underpinnings of normative ethics. The authors propose a utility function that includes both a player's material payoff and a reciprocity payoff, which depends on the player's beliefs about the kindness of other players' actions. The

kindness of an action is determined by comparing the material payoffs of the action to the material payoffs of other possible actions, given the player's beliefs about the other players' strategies. Applying their model to several well-known games, including the ultimatum game, the trust game, and the prisoner's dilemma, they show that their model can explain why players might choose to cooperate in these games even in the absence of repeated interaction or reputation effects.

### 1.2.2 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. In threshold or step-level PGGs, a public good is only provided when a minimum level of contributions (often referred to as a "provision point") is met; the meta-analysis found that sequential play was more

efficient in providing the public good compared to simultaneous play. However, not all subjects followed the game theoretic prediction of exploiting later movers.

Eichenseer also investigated weakest-link PGGs. In these games, the minimum contribution of the players involved determines the total amount of public good available to all. He found that minimum choices in the weakest-link PGG were significantly higher in the fully sequential treatment compared to the simultaneous treatment. In a more classical leader-follower relationship, groups in the leadership treatments did better in coordination in that minimum contributions were larger and subjects earned more indicating an increase in efficiency.

Finally, field experiments reveal something about sequentiality with observation in real life. In some cases, leadership was effective in improving public good provision, for instance in one particular example when the leader was a democratically elected local authority. However, in other cases, leadership did not improve cooperation, particularly in culturally heterogeneous groups. It appears that, while sequential contribution *can* result in social benefits, the selection process for leaders (and the subsequent players' beliefs about those selection processes) are very relevant for success. 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.

### 1.2.3 Sequential games without observation

#### 1.2.3.1 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.

#### 1.2.3.2 Coordination games

Rapoport (1997) Is the genesis of a related, modest body of work looking at order effects in coordination games. Rapoport suggests that a given player, in the absence of information except for her position, will assume that prior players take advantage of their position. In three experiments, a common-pool resource dilemma (45 undergraduates), a step-level binary PGG (70 students), and a coordination game (36 students). Rapoport detects some evidence of an order effect without observation. In the common-pool resource dilemma, requests decline with increasing order; in the binary step-level PGG, players 3, 4, and 5 show more cooperation than players 1 and 2, where three players' contributions are necessary to produce the public good. In the coordination

game, following Budescu et al., Rapoport observes players using the order of play as a coordination mechanism. These experiments also involved within-subject treatments and several trials per subject, making for important differences from canonical one-shot games.

Building on work from Cooper et al., (1993), Güth et al., (1998) study the order effect with no observation specifically in coordination games such as Battle of the Sexes, games that have a first-mover advantage. In such games, it is possible that the mere decision to play a game that is sequential with no observation leads to both players coordinating on the first-mover's preferred outcome, analogous to a Schelling point. Player 2 knows that Player 1 decided to play the game and that Player 1 knows she is moving first, so Player 2 defers to Player 1's preferred outcome. They report experimental results from a Battle of the Sexes game (254 students), and an Independent Moves game (170 students). In both tasks they find support for order effects without observation, and specifically they argue that when there are two pure strategies and symmetric equilibria, as in Battle of the Sexes, order of play serves as a coordination device which makes the equilibrium outcome most favored by the first mover a Schelling point. Weber et al., (2004) further investigate these phenomena under a theory of "virtual observability" introduced by Amershi et al. (1989), where players play as if they could observe the moves made *prior* to their own move. They find evidence for this theoretical account in ultimatum bargaining and weak-link games, but see Li (2007) for a discussion of possible design issues.

#### 1.2.3.3 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.

#### 1.2.3.4 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.

#### 1.2.3.5 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.

### 1.2.4 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.

## 1.3 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:

$$p = \frac{mc}{n} + e(1-a) \qquad \text{(Equation 1.1)}$$

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.

## 1.4 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.

### 1.4.1 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.

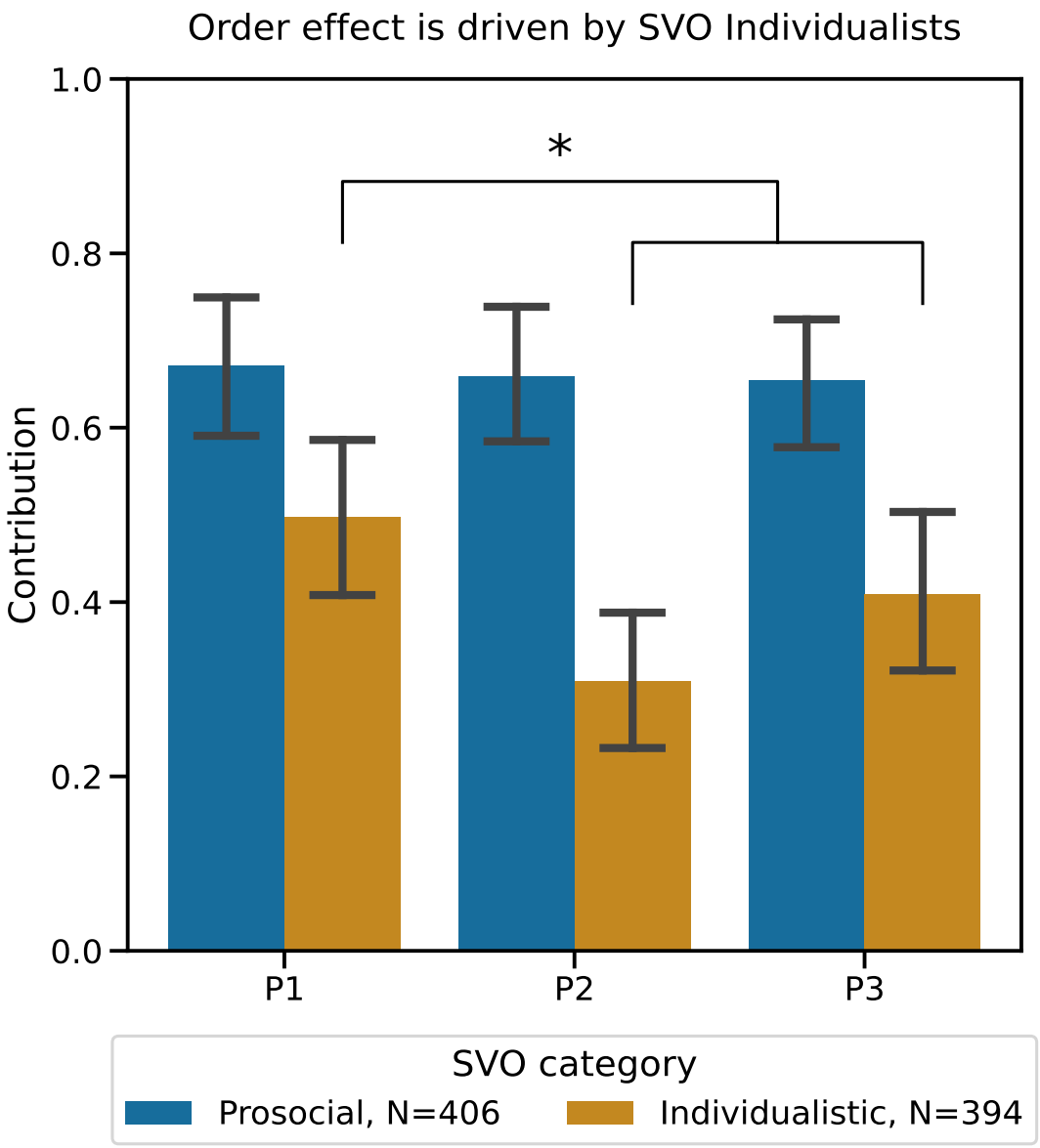



*Figure 2: Change in contribution with order is driven by subjects SVO-classified as Individualistic. Subjects who passed comprehension checks. 95% CIs.*

### 1.4.2 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.

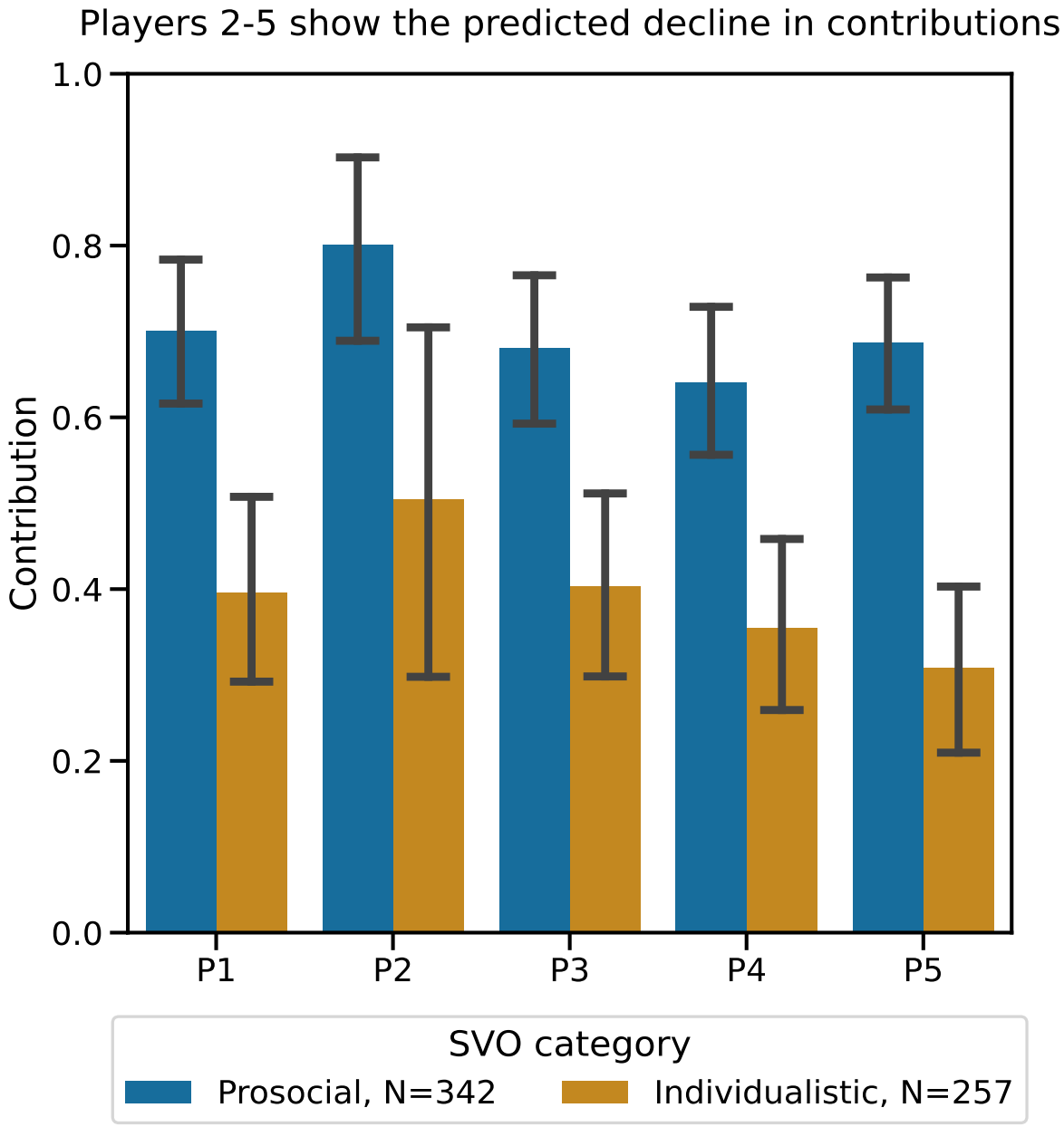



*Figure 3: SVO-individualist players show a decline in contribution with increasing order from Player 2, and an anomalous result for Player 1. Player 1 was affected by a programming error that resulted in incorrect timing of stimuli presentation. Subjects who passed comprehension checks. 95% CIs.*

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.

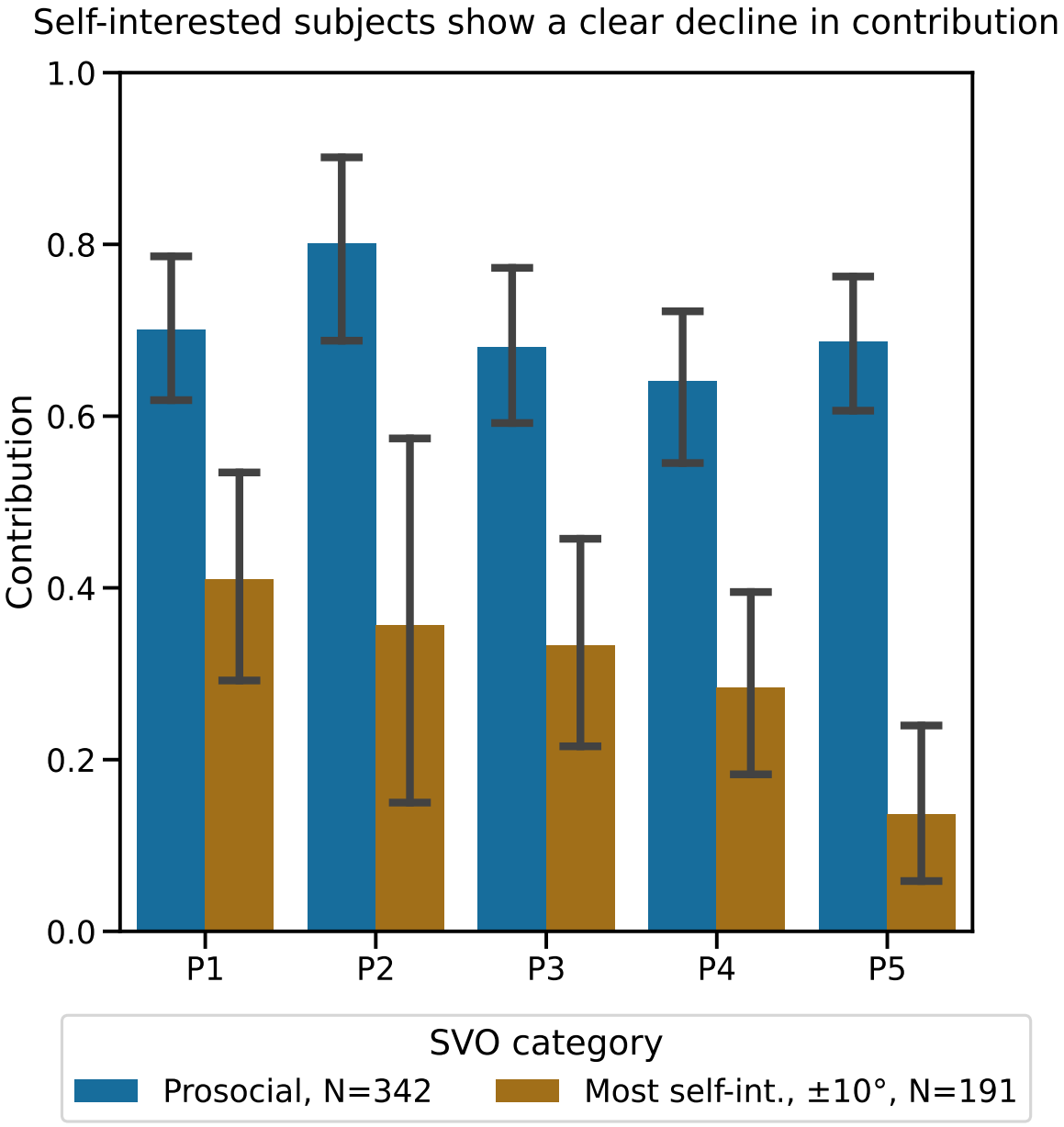



*Figure 4: Players whose SVO degree measure is +/-10 degrees show the predicted decline of contribution to the public good with increasing order. All subjects. 95% CIs.*

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.

### 1.4.3 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.

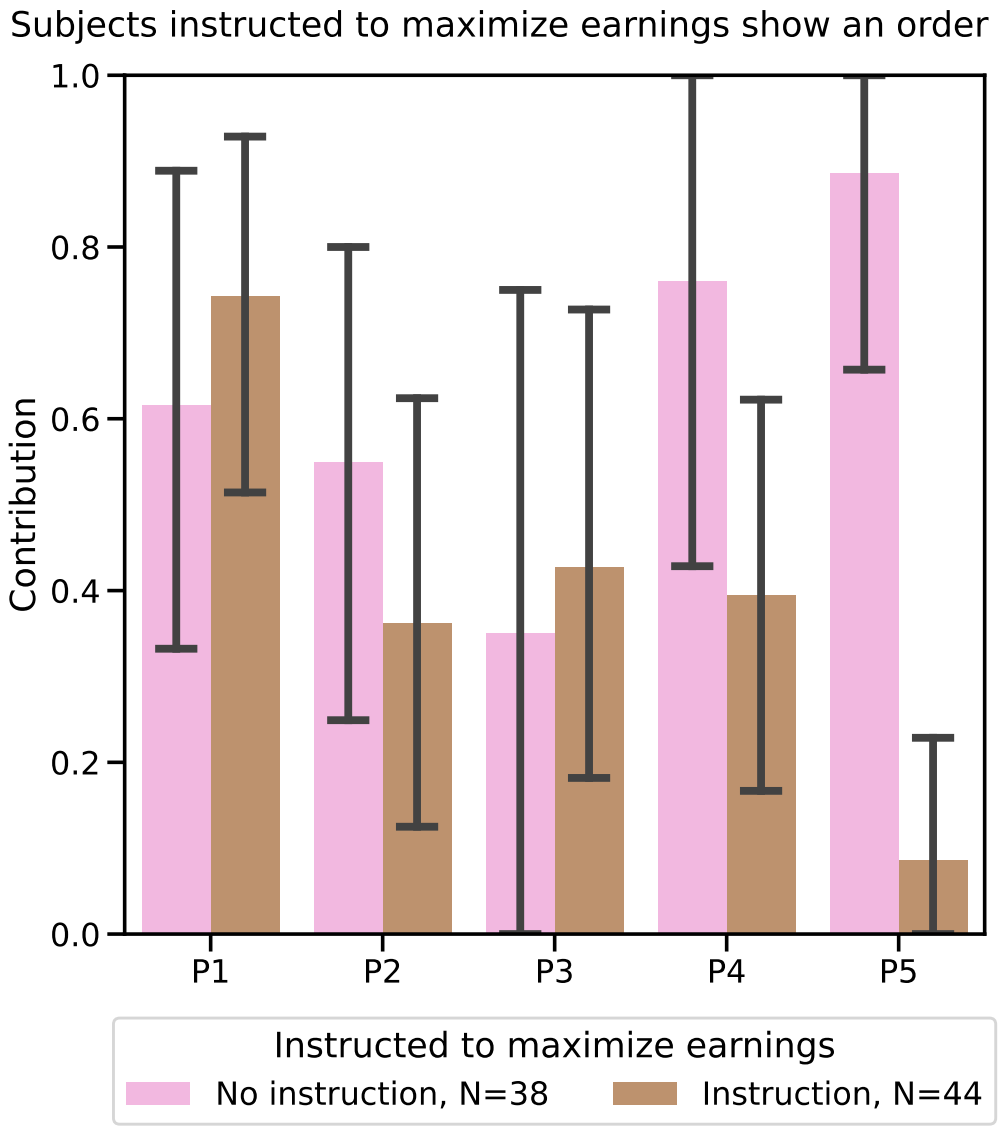



*Figure 5: Study 3 shows the hypothesized decline with order among those who were instructed to be greedy. Subjects who passed comprehension checks. 95% CIs.*

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.

### 1.4.4 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.

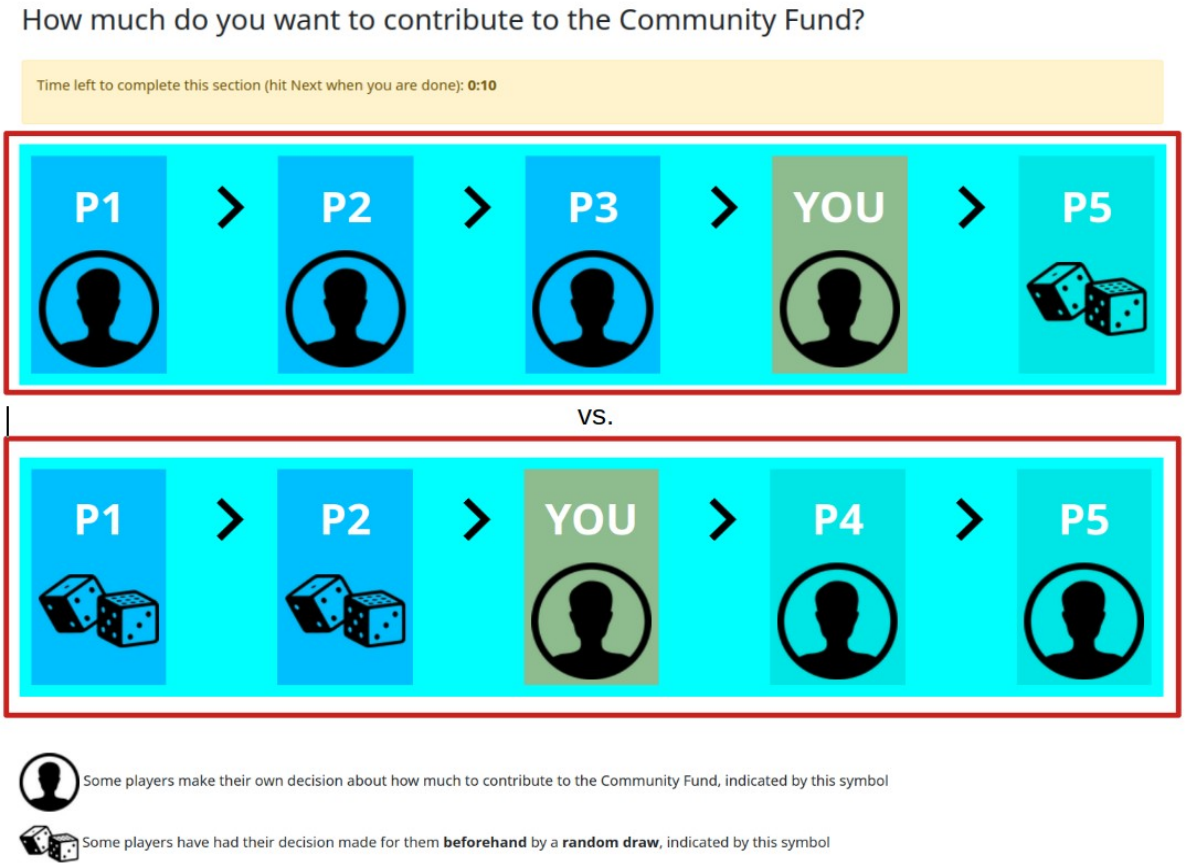



*Figure 6: The stimuli on the Contribution page are shown in two conditions, in red boxes: above, Random After for Player 4 and below, Random Before for Player 3. Players see a graphical representation of their position relative to other players that clearly conveys which players are having their moves 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.

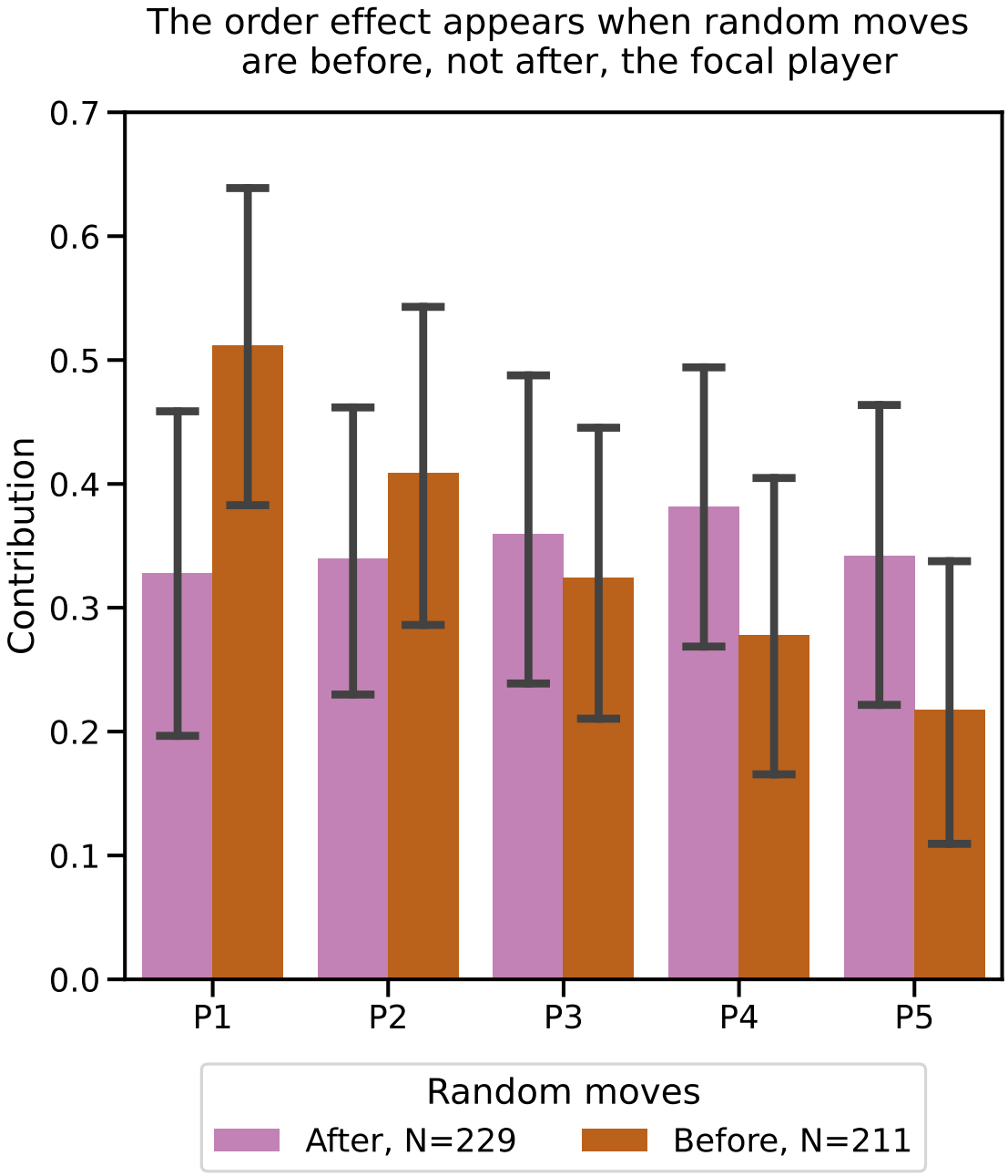



*Figure 7: Study 4 shows a decline in contribution to the public good among players who are told that all players moving after them are making their own moves, and all players moving before them are having their moves made randomly. No effect is observed among players who are told that everyone moving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

## 1.5 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.

## 1.6 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.

### 1.6.1 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.

### 1.6.2 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.

### 1.6.3 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.

### 1.6.4 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.

## 1.7 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

Amershi, A. H., Sadanand, A. B., & Sadanand, V. (1989). Manipulated Nash Equilibria - I: Forward Induction And Thought Process Dynamics In Extensive Form. *Working Papers*, Article 1989–4. https://ideas.repec.org//p/gue/guelph/1989-4.html

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

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1993). Forward Induction in the Battle-of-the-Sexes Games. *The American Economic Review*, *83*(5), 1303–1316.

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

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298. https://doi.org/10.1016/j.geb.2003.06.003

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

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.

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

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79. https://doi.org/10.1016/0899-8256(89)90005-5

Güth, W., Huck, S., & Rapoport, A. (1998). The limitations of the positional order effect: Can it support silent threats and non-equilibrium behavior? *Journal of Economic Behavior & Organization*, *34*(2), 313–325. https://doi.org/10.1016/S0167-2681(97)00057-7

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.

Kreps, D. M., & Wilson, R. (1982). Sequential Equilibria. *Econometrica*, *50*(4), 863. https://doi.org/10.2307/1912767

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

Li, T. (2007). Are there timing effects in coordination game experiments? *Economics Bulletin*, *3*(13), 1–9.

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

Nash, J. (1951). Non-Cooperative Games. *Annals of Mathematics*, *54*(2), 286–295. https://doi.org/10.2307/1969529

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

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

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

Rapoport, A. (1997). Order of play in strategically equivalent games in extensive form. *International Journal of Game Theory*, *26*(1), 113–136. https://doi.org/10.1007/BF01262516

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

Selten, R. (1965). Spieltheoretische Behandlung Eines Oligopolmodells Mit Nachfrageträgheit: Teil I: Bestimmung Des Dynamischen Preisgleichgewichts. *Zeitschrift Für Die Gesamte Staatswissenschaft / Journal of Institutional and Theoretical Economics*, *121*(2), 301–324.

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.

Weber, R. A., Camerer, C. F., & Knez, M. (2004). Timing and Virtual Observability in Ultimatum Bargaining and “Weak Link” Coordination Games. *Experimental Economics*, *7*, 25–48.

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

## 1.8 Appendix

### 1.8.1 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

### 1.8.2 Model

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

#### 1.8.2.1 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 $\mathrm{i}$, $0 \leq a_i \leq 1$, and $p_i$ the payoff to player $i$.

$$p_i = 1 - a_i + \tfrac{m}{n} \textstyle\sum_{k=1}^{n} a_k \qquad \text{(Equation 1.2)}$$

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 $p_i$:

$$u_i\left(a_1, \ldots, a_n\right) = \left(1 - s_i\right) p_i + s_i m a_i \qquad \text{(Equation 1.3)}$$

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.

#### 1.8.2.2 Decision dependent expectations

We assume that players compare expected utilities conditional on contributing $\left(a_i = 1\right)$ or not contributing $\left(a_i = 0\right)$, and choose whichever expected utility is higher (we ignore

here fractional contributions). The decision criterion is therefore the difference between the two expected utilities:

$$a_i = 1 \iff E\left[u_i \mid a_i = 1, s_i\right] > E\left[u_i \mid a_i = 0, s_i\right] (1)$$ (Equation 1.4)

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:

$$E\left[a_k \mid s_i\right] = b + cs_i (1)$$ (Equation 1.5)

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 1 and the third line from equation 2.

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,

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

(Equation 1.6)

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 1.7)

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:

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

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.

# 2 Small talk as a contracting device: Trust, cooperative norms, and changing equilibria

*In collaboration with Birger Wernerfelt and Boris Maciejovsky*


## Abstract

We show experimentally that a very brief face-to-face talk with a potential trading partner may have a contracting function by enhancing trust and strengthening cooperative norms. Specifically, subjects engage in three-minute video calls with no agenda prior to playing Hold Up and Stag Hunt games. In spite of the fact that the players had no advance knowledge of the games, the call had large effects on trust, cooperation, and efficiency: There was more investment and less stealing in Hold Up games and twice-repeated Stag Hunt games much more frequently ended up in the efficient equilibrium. Beyond suggesting that small talk can alleviate contractual incompleteness, the results also explain several other phenomena.

## 2.1 Introduction

This is a paper about contracts that are incomplete in the sense that unforeseen, and therefore not-contracted-on, contingencies are likely to have a significant effect on the payoff implications of different actions. The parties to such a contract should only have incentives to communicate about the foreseen contingencies, since nothing relevant can be said about anything else. And yet, parties to such contracts often incur costs to engage in “small talk” (here defined as a face-to-face meeting in which no issues with payoff relevance are discussed) with potential future trading partners. We will rationalize these meetings by showing, in a tightly controlled experimental setting, that they nurture trust and cooperative norms, thus compensating for contractual incompleteness. However, the same effect can also help explain a number of other widespread behaviors.[3] One example is networking: This very common practice seems to be motivated by the belief that the other party, should you ever want to contact them, will be more receptive if the two of you have met - even if very briefly. (A variant of this is the perceived advantages of “knowing” your boss). A final and quite different example are corporate team-building exercises: These are generally seen as attempts to change the organizational equilibrium to a more efficient one. The prevalence of these and many other examples raise the question: does small talk make any difference?

To start thinking about the contracting function of small talk, it is helpful to review some stylized facts about when it is and is not demanded. First, it is not deemed necessary in settings such as grocery stores, online retailing, or stock markets, where simple formal contracts cover all relevant contingencies. Second, other informal

[3] It should be acknowledged that we appear to be more willing to break these informal contracts, violate these norms, and doubt this trust, when the economic gains from doing so are greater, although the paper by Frydlinger and Hart (forthcoming) suggests that similar effects exist even when very large sums are involved. In general though, if a formal contract is possible, it is more likely to be used when more is at stake, for example if you are buying real estate. However, small talk is typically cheaper and results in an “agreement” that is less incomplete than formal contracts. It arguably shares these advantages with relational contracting but does not depend on repeated play. (Gibbons et al., 2021, look at incomplete relational contracts).

contracts, such as handshakes or verbal promises, are used when the agreement involves a small number of well-understood ways to defect. Examples include “I will do the job to a reasonable standard, and you will then pay me $X”, “Once this foal is weaned, I will sell it to you for $Y”, and “If you agree to bring me the money tomorrow, I will not sell the car to anyone else in the meantime”. Third, small talk is used when there is not a complete list of potential conflicts you can talk about ex-ante such that a complete contract is unattainable. One class of examples are cases in which you select a partner for a complex trade or service (preferred supplier, kitchen renovator, exclusive retailer,..). In such situations it is very likely that conflicts will present themselves but neither party knows what they all may be. So the best one can do is to try to establish norms of cooperation and hope to enhance trust. There is a widespread belief that this can be accomplished through small talk. In particular, the popular management literature is full of assertions to that effect. For example, the Wikijob Team (2021) claims that “Small talk […] helps to form social cohesion that […] builds trust”, and Jeevan Sivasubramanian (2021) writes that “Small talk helps to establish trust”.[4]

We report on two experiments that throw light on some novel effects of small talk. In both cases the subjects did not know each other and had no prospect of even meeting again. The first experiment is based on a simultaneous move “Hold Up” game: One player, the “investor”, decides whether to invest and if they do, another player, the “operator”, chooses between theft and cooperation. So the operator’s choice reflects the power of cooperative norms, whereas the investor’s decision is an indicator of their trust that the operator will adhere to these norms. While the efficient outcome is not an equilibrium in the standard sense, our main hypothesis is that investor-operator pairs are more likely to reach it if they have a chance to engage in small talk before the game. We represent small talk by letting two opposing players spend three minutes together (on a video call), knowing only that they are about to play some sort of a game for money (such that they cannot make promises or agreements about any specific moves). The results of these pairs are then contrasted with those obtained by a control

[4] In the academic literature, Morris et al. (2002) and Mislin, Campagna, and Bottom (1999) show that trusting behavior and efficiency are enhanced by communication prior to playing a known game, and Bickmore and Cassell (1999), even propose developing computerized agents capable of simulating small talk.

group in which the players never meet, and we find that the three minutes of small talk almost doubles the fraction of games that achieve the efficient outcome.[5]

We thus find that small talk increases the investors' willingness to trust their operators, and in turn makes the latter more trustworthy (more likely to follow cooperative norms). The precise underlying mechanism is hard to pin down, but the result is consistent with the idea that small talk develops trust and strengthens cooperative norms.

In the second experiment we look at a twice-repeated Stag Hunt game and show that pairs who engage in small talk between rounds are much more likely to play the efficient, but risk dominated equilibrium in the second game.[6] In fact, these pairs are 150% more likely to play the efficient equilibrium than those in the control group. Beyond supporting the findings from the first experiment, this may explain how small talk not only establishes rapport and develops trust, but also leads to downstream consequences for future social interactions. In the business context, the widespread use of "team-building exercises" in which groups of employees from the same company are put through a number of activities that, among other things, require them to communicate might be an example of such social interactions.[7]

We discuss related literature in Section 2, derive our hypotheses in Section 3, and present the experiments and the results in Section 4. Section 5 concludes with a brief discussion.

[5] We admit that there, at least anecdotally, are cases in which small talk leads to a complete break-down of relations, in stark contrast with our hypothesis. However, these are presumably very rare cases, and our data only allows us to look at mean effects. (In our second experiment, there is not a single occasion in which small talk caused the parties to move from a good equilibrium to a less good one.)

[6] In our study, subjects do not know that they are to play the same game again after the small talk is over and are thus very unlikely to spend the time making promises about it.

[7] Buller and Bell (1986) remark that "one of the most popular intervention techniques in organizational development (OD) is teambuilding".

## 2.2 Related Literature

Our first experiment is motivated by the “guiding principles” described by Frydlinger and Hart (forthcoming). They describe a range of situations in which executives from firms about to enter into trading relationships have extended meetings in which they agree to follow certain “guiding principles”. These principles suggest, among other things, that each of them will try to see things from the perspective of the other, take the other’s payoffs into account, and behave cooperatively whenever foreseen and unforeseen circumstances afford one of them the ability to hold up the other. The authors report that the practice has been adopted by several businesses and that it seems to be successful. The process described by them can be interpreted as taking advantage of the mechanisms studied here. The fact that their subjects are real executives engaged in actual contracts with large sums at stake is a major strength of their paper. However, three key differences are that we run subjects through identical controlled experiments with objective measures of success, that our sample arguably is subject to fewer selection effects, and that our setting eliminates any fear of retaliation or reputation effects.

Another closely related paper is by Chen and Chen (2011). They ask some pairs of subjects to engage in electronic communication prior to playing a minimum effort game and show that those who communicated selected more cooperative equilibria. As in our experiment, the subjects communicated without knowing about the game they would play afterwards. Our experiments differ in three ways: We measure both trust and the strength of cooperative norms, our subjects communicate face-to-face, and we measure (in our second experiment) changes in the equilibria played.

The observation that subjects are nicer to those they know better has been explored in several studies in the behavioral economics literature on fairness (Kahneman, Knetch, and Thaler, 1986; Camerer and Thaler, 1995; Fehr and Schmidt, 1999). For example, Bohnet and Frey (1999) show that players are more generous in dictator games when they have a chance to see their opponents prior to playing, and Brooks, Dai, and

Sweitzer (2013) show that subjects are more trusting of opponents who start an interaction by making an irrelevant apology for the weather. A second related branch of the economics literature is concerned with betrayal, guilt, and aversion to lying (Frank, 1987; Gneezy, 2005; Mazar, Amir, and Ariely, 2008; Lundquist, et al., 2009; Belot, Bhasar, and van de Ven, 2010)[8] and a third branch is looking at the effects of cheap talk Tingley and Walter, 2011). However, except for the above-mentioned paper by Chen and Chen (2011), the economics literature on pregame communication has invariably assumed that players know which game they are about to play. So, while these studies show an effect of communication, they do not throw light on the incomplete contracting angle pursued in the present paper.

There is finally a large literature in social psychology on the beneficial effects of pre-game communication, going back to at least Deutsch (1958) and including Bouas and Komorita (1996), Bicchieri and Lev-on (2007), and Baillet (2010). As far as we know, all experiments described in this literature also involve situations in which subjects are informed about the game prior to communicating.

There is less literature that explicitly addresses the change of equilibrium observed in our second experiment. Not surprisingly, it is very hard to find any economic literature on changing equilibria – such an observation almost runs counter to the definition. However, as mentioned in the Introduction, there is a lot of management literature on the ability of team-building exercises to change an organization's "culture" - which again could be interpreted as changing its equilibrium.[9]

## 2.3 Theory and Research Questions

We first look at a Hold Up game that is very similar to that used in Charness and Dufwenberg, 2006. Two players, the investor (he) and the operator (she), make simultaneous moves; the investor decides between IN ("invest") and OUT ("outside

[8] This has been taken up in recent theoretical research assuming that lying imposes a private cost on senders (Kartik, 2009; Gneezy, Kajackeite, and Sobel, 2018).

[9] See Klein, DiazGranados, Sales, and Le (2009) for a meta-analysis of this literature.

option"), and the operator between KEEP ("hold up") and ROLL ("implement the proposed venture"). If the investor selects OUT, both parties get *1* no matter what the operator chooses. However, if the investor selects IN, payoffs do depend on the operator*'s* choice: When they pick KEEP, the operator gets $\kappa$ and the investor gets *0.* When the operator picks ROLL, she gets $\sigma$ while the investor gets *0* with probability $q$ and $\pi$ with probability $1 - q$.[10] *Figure 8* gives the game matrix.

*Figure 8: Basic Investor-Operator Game*

| Investor, Operator expected payoffs | KEEP | ROLL |
|---|---|---|
| OUT | *1, 1* | *1, 1* |
| IN | *0, κ* | *q0 + (1 – q)π, σ* |

If $\kappa > \sigma$ the investor plays OUT in all Nash equilibria and if $(1 - q)\pi + \sigma > Max\{2, \kappa\}$, (IN, ROLL) is first best.

Contrary to the above analysis, experiments on many similar one-shot games have shown that some pairs manage to end up in the first best outcome (Johnson and Mislin, 2011). This is often thought of as the result of players anticipating feeling guilty if they violate cooperative norms and play the Nash moves (Attanasi, Battigalli, and Manzoni, 2016). Equivalently, we can imagine that operators feel bad if they betray trust or that they to some extent are altruistic. We could model cooperative norms in all three ways but will illustrate the point by using the latter. If both weigh the opponent's payoffs by $w$ the game changes to that in *Figure 9*.

[10] A common problem in experiments on cooperation is that subjects are "too cooperative" in the control condition such that a ceiling effect reduces statistical power. We use this construction (with $q > 0$) because it enables the operator to play KEEP without the investor knowing for sure that she did so. While the operators "shouldn't" worry about this when the players have no common acquaintances and will not meet again, the construction did in fact result in more of the operators playing KEEP in our pilot studies. To further strengthen the effect, we explicitly pointed this out to them.

*Figure 9: Investor-Operator Game with Cooperative Norms (as Altruism)*

| Investor, Operator expected payoffs | KEEP | ROLL |
|---|---|---|
| OUT | $1 + w, 1 + w$ | $1 + w, 1 + w$ |
| IN | $w\kappa, \kappa$ | $(1 - q)\pi + w\sigma , \sigma + w(1 - q)\pi$ |

As can be seen, if $\sigma > 1$ the efficient (IN, ROLL) is a Nash equilibrium for sufficiently large *w.* So we can think of operators playing ROLL when they place a high value on cooperative norms and investors playing IN when they know this and therefore have a high level of trust in their operators. Our main hypothesis is that the players, if they spend time together prior to playing the game, could develop an element of trust and cooperative norms, thereby growing the values of *w.*

We can investigate the size and nature of the small talk effect by comparing games with and without small talk in the following ways: (i) Do more games end in (IN, ROLL) after small talk? (ii) Do more investors trust their operators after small talk and therefore play IN? (iii) Do more operators play ROLL after small talk, thereby rewarding the trust placed in them by the investors? And (iv) Does small talk allow investors to identify more trustworthy operators? If so, in the treatment with small talk, investors who play IN have a better chance of their opponent playing ROLL than investors who play OUT.

In the second experiment we look at a twice repeated Stag Hunt game (though the players do not know that their second activity will be the same game). In the STAG, STAG outcome, the players share *s*, and in the HARE, HARE outcome, they both get *1.* If they fail to coordinate, the STAG hunter gets *0* while the player going for a HARE gets $1 + c$.[11] The game matrix is given in *Figure 10*.

[11] $c \geq 0$ reflects the fact that it is easier to catch a hare when nobody else is hunting them.,

*Figure 10: Stag Hunt Game*

| Row, Column Hunter payoffs | STAG | HARE |
|---|---|---|
| STAG | *s/2, s/2* | *0, 1 + c* |
| HARE | *1 + c, 0* | *1, 1* |

If we assume that $s > 2 > s/2 - c > 1 > 1 - c,$ there are two equilibria and the risk-dominant, but inefficient (HARE, HARE) equilibrium is often played because players are uncertain about each other. Since we represented cooperative norms as altruism in *Figure 10*, we now use guilt to change the stage game to that depicted in *Figure 11*.

*Figure 11: Stag Hunt Game with Cooperative Norms (as Guilt)*

| Row, Column Hunter payoffs | STAG | HARE |
|---|---|---|
| STAG | *s/2, s/2* | *0, 1 + c – g* |
| HARE | *1 + c - g, 0* | *1 - g, 1 – g* |

So (STAG, STAG) is the only Nash equilibrium if $g > 1.$

We can test this by asking the following questions: (v) Do more games end in (STAG, STAG) after small talk? (vi) Conversely, do fewer games end in (HARE, HARE) after small talk? (vii) Do more games change from the inefficient to the efficient equilibrium after small talk? (viii) Do fewer games change from a non-equilibrium outcome to the inefficient equilibrium after small talk? (ix) Do more games change from a non-equilibrium to the efficient equilibrium after small talk?

## 2.4 Experiments and Results

All studies used US residents aged 26 and up (to help ensure that they share similar norms and had some first-hand experience with the economy) and were run on Amazon Mechanical Turk. Subjects were paid their winnings. The exact procedures and instructions are reproduced in the Appendix.

### 2.4.1 Experiment 1: Small talk increases trust and cooperation in a one-shot game.

Pairs of subjects engage in simultaneous move Investor-Operator games with the following payoff matrix:

*Figure 12: Investor-Operator Game with Dollar Parameter Values Used in Experiment 1*

| Investor, Operator expected payoffs | KEEP | ROLL |
|---|---|---|
| OUT | *3.5, 3.5* | *3.5, 3.5* |
| IN | *0, 9* | *(2/3) x 7.5, 5* |

We compare the outcomes of this game in two different treatments:

-Treatment 1: Players are informed about, and play, the game. They do not meet or see each other.

-Treatment 2: Opponents spend 3 minutes together on a video call.[12] After the video call, they are informed about, and play, the game.

The number of agents choosing each action are shown in Table 1 below.

[12] In an extensive pilot study, we seeded the conversations in three different ways. Some pairs are encouraged to use the time to identify the two most interesting things they have in common. If they independently report the same two things afterwards, they get a reward. Other pairs answer ten binary lifestyle questions (rural/urban, tacos/sushi, beach/mountain, etc.). Each pair is then told, prior to engaging in the 3-minute video conversation, on which of the ten questions they agree. Finally, the last group were not given any instructions. All three groups performed identically. In particular, the number of questions on which the players agree does not correlate with their actions.

*Table 1: Treatment 1: No Contact*

| Investors | OUT | IN | Totals |
|---|---|---|---|
| | 58 | 40 | 98 |
| Operators | KEEP | ROLL | |
| | 58 | 40 | 98 |

*Table 1: Treatment 2: Small Talk*

| Pairs | KEEP | ROLL | Totals |
|---|---|---|---|
| OUT | 22 | 25 | 47 |
| IN | 22 | 31** | 53* |
| Totals | 44 | 56** | 100 |

Significantly different from the proportion in Treatment 1,* $p < .1$, ** $p < .05$, Chi Square-test

We will now turn to answer questions (i) – (iv) from Section 3.

(i) Since the subjects did not interact in Treatment 1, they did not play against specific opponents. However, the expected fraction of games ending in (IN, ROLL) was 0.17, while it was 0.31 ($p = .02$) in Treatment 2. The difference between Treatments 1 and 2 is consistent with our main hypothesis, that more games end in (IN, ROLL) after small talk.

(ii) The fractions of investors playing IN was 0.41 in Treatment 1 and 0.53 ($p = .09$) in Treatment 2. So investors appear to be more willing to trust operators after small talk.

(iii) Similarly, the fraction of operators who played ROLL was 0.41 in Treatment 1 and 0.56 *( $p = .03$)* in Treatment 2. The result is consistent with the operators anticipating feeling guilty after playing KEEP and violating cooperative norms.

(iv) If an investor plays IN (OUT), the chance that his opponent plays ROLL is 0.58 (0.53) in Treatment 2. Since these are not significantly different, we cannot conclude that agents after small talk can tell whether their opponent is more trustworthy.

While experiment 1 was concerned with the effect of small talk on trust and cooperation, experiment 2 is focused on cooperative norms. However, it also tackles the question of whether small talk might help move players from inefficient outcomes and equilibria to more efficient ones.

### 2.4.2 Experiment 2: Small talk can allow players in a repeated game to move from one stage game equilibrium to another.

Pairs of subjects engage in two Stag Hunt games with the following payoff matrix:[13]

*Figure 13: Stag Hunt Game with Dollar Payoff Values Used in Experiment 2*

| Hunter payoffs | STAG | HARE |
|---|---|---|
| STAG | 4, 4 | 1, 3 |
| HARE | 3, 1 | 3, 3 |

None of the players know their opponents prior to the first round. Half the pairs play the second game immediately after the first, but the other half have a three-minute face-to-face meeting between the two games (and thus meet).[14] Both groups knew that they were to engage in a second "task" after the first game but did not know that it turned out to be the same game.

We ran the experiment with 55 pairs that did not engage in small talk between games and 60 pairs that did. Looking first at the condition with no small talk between

[13] Dal Bo, Frechette, and Kim (2021) look at the relationship between payoff matrices and equilibrium selection in stag hunt games. Our findings are consistent with theirs.
[14] We did not seed these conversations, but it is possible that they discussed the game.

games, (HARE, HARE) was played by 18 pairs and (STAG, STAG) was played by 10. All of these played the same equilibrium on the second game. Of the 27 pairs who did not play an equilibrium in the first game, 18 went to (HARE, HARE), only one went to (STAG, STAG), and eight again failed to find an equilibrium. So in the second game, a total of 36/55 = 0.65 of the pairs played (HARE, HARE) while only 11/55 = 0.2 played (STAG, STAG).

In the condition with small talk between games, we ran 60 pairs and 28 played (HARE, HARE) in the first game. Four of these switched to (STAG, STAG) in the second, while 23 continued to play (HARE, HARE). In the same condition, six pairs started with (STAG, STAG) and all of these played the same equilibrium in the second game. Of the 26 pairs who did not find an equilibrium in the first game, 20 went to (STAG, STAG) and four ended up playing (HARE, HARE). So in the second game, 27/60 = 0.45 of the pairs played (HARE, HARE) while 30/60 = 0.50 played (STAG, STAG), many more than without small talk. The data in Table 3 summarizes the higher efficiency in the condition with small talk.

*Table 2: Increased Efficiency Following Small Talk*

| Fraction of pairs | No small talk | Small talk between games |
|---|---|---|
| Playing efficient equilibrium in second game | 11/55 | 30/60**** |
| Playing inefficient equilibrium in second game | 36/55 | 27/60** |
| Switching from inefficient to efficient equilibrium | 0/18 | 4/28 |
| Switching from non-equilibrium to inefficient equilibrium | 18/27 | 4/26**** |
| Switching from non-equilibrium to efficient equilibrium | 1/27 | 20/26**** |

Significantly different from the results in column 1, **** $p < .001$,** $p < .05$.

(v) As hypothesized, the fraction of pairs who play (STAG, STAG) in the second game is significantly higher after small talk ($p = 0.0008$, Chi square test).

(vi) The fraction of pairs who play (HARE, HARE) in the second game is significantly smaller ($p = 0.028$, Chi square test).

(vii) Four games do change from the inefficient equilibrium all the way to the efficient equilibrium after small talk, but the effect is not significant ($p = .14$, Fisher test).

(viii) Fewer games change from a non-equilibrium outcome to the inefficient equilibrium after small talk ($p = .0002$, Fisher test).

(ix) More games change from a non-equilibrium outcome to the efficient equilibrium after small talk ($p = .0000$, Fisher test).

Taken together, the results strongly suggest that the players follow cooperative norms more closely after small talk. In addition, they show that our simple intervention can help migrate a finitely repeated game to a more efficient equilibrium.

## 2.5 Further questions suggested by our results.

We show that a very limited amount of small talk can cause people to trust and cooperate with strangers. Small talk overcomes contractual incompleteness by covering a broad range of contingencies, including some that are truly unforeseen (e.g., our subjects socialize before they know that they are to play a game, much less which game). We also show that small talk can be used effectively to change a finitely repeated game from a less efficient equilibrium to a more efficient one.

The results provide one explanation why people appear eager to “get to know” potential trading partners, as well as the popularity of networking. Our results also apply to “acquaintanceship corruption” (cronyism, nepotism, patronage, or clientelism) in which employees make discretionary decisions on behalf of firms or governments with no immediate quid pro quo (so it is different from regular corruption). Since the employee has to trust that some sort of payback eventually will materialize, we conjecture that this behavior more important and more common in societies where the rule of law is weaker, and trust is higher.[15] A similar but different phenomenon is the widely held belief that “knowing your boss” confers advantages in situations where discretionary decisions are made. We conjecture that this is more important in societies and industries with less efficient labor markets. Also these conjectures seem eminently testable.

More generally, it would be interesting to look at small talk between more than two people. At what point does it cease to be effective? Along similar lines, what happens if people are put through a large number of brief encounters? Is there a scale

[15] Kosse et al (2020) show that prosocial norms are shaped by social environments.

at which small talk no longer works? And could intensive exposure eventually inoculate participants against its effects?

## 2.6 What is going on?

There could be a mechanical explanation for our findings. We do not record what the subjects are talking about. It is possible that they guess what is about to happen and agree to "cooperate" (whatever that means).

A different class of possibilities, which in our view are more likely, is that a face-to-face meeting makes use of psychological traits that evolved to stabilize cooperation among group members. There are many versions of this: It could be that subjects cooperate simply because they instantly find the opponent "reasonable" based on their experiences with similar looking people (They may look like a former neighbor, be physically attractive, or share race or gender with the subject)[16]. This could then reduce strategic uncertainty and make reliance on the opponent's behavior feel less risky. Consistent with the idea that the effect has a social origin, Roth (1995, p. 295) summarizes part of the experimental literature on bargaining by saying that "Face-to-face interactions call into play all the social training we are endowed with".

These traits could have originated because we originally only communicated face-to-face with members of our own tribe and that small talk causes subjects to, unconsciously, impute in-group membership to their opponents. The existence of these norms and the fact that they affect play in unrelated games can presumably be traced very far back, and one could conjecture that they at some point were supported by community enforcement (Coleman, 1955; Kandori, 1992).[17]

The idea that people have a tendency to favor other members of groups to which they belong, has a long history in the literature on tribalism.[18] It has been studied in a

[16] See Vogt, Efferson, and Fehr (2013)

[17] This would be consistent with the widespread practice in which strangers, when they first meet, try to find a social connection ("So you are a doctor from Cleveland. Do you know Lisa Smith?").

[18] A representative early statement is due to Taylor and Doria (1981).

large number of experiments (Goette, Huffman, and Meier, 2006) and field studies (Ert, Fleischer, and Magen, 2016; Karlsson, Kemperman, and Dolnicar, 2017; Edelman, Luca, and Svirsky, 2017), some of which suggest that group membership can change within relatively short periods (Efferson, Lalive, and Fehr, 2008; Rand et al., 2009).[19]

It is an important goal of future research to try to disentangle some of these mechanisms.

[19] It is interesting, though perhaps a coincidence, (a) that you often see a person's in-group defined as the set of people whose welfare matters in their utility function (Dawes, Van De Kragt, and Orbell, 1988), and (b) that one of the things participants in the Frydlinger-Hart process promise is to take each others' payoffs into account.

## 2.7 References


Attanasi, Giuseppe, Pierpaolo Battigalli, and Elena Manzoni, “Incomplete-Information Models of Guilt Aversion in the Trust Game,” *Management Science*, 62, pp. 648-667, 2016.

Balliet, Daniel, “Communication and Cooperation in Social Dilemmas: A Meta-Analytic Review”, *Journal of Conflict Resolution,* 52(1), pp. 39-57, 2010.

Belot, Michele, V. Bhasar, and Jeroen van de Ven, ”Promises and Cooperation: Evidence from a TV Game Show”, *Journal of Economic Behavior and Organization,* 73, pp.396-405, 2010.

Bicchieri, Christina, and Azi Lev-On, “Computer Mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis,” *Politics, Philosophy, and Economics*, 6, pp. 139-168, 2007.

Bickmore, Timothy, and Justine Cassell, “Small Talk and Conversational Storytelling in Embodied Conversational Interface Agents”, pp. 23 – 54, *The Association for the Advancement of Artificial Intelligence* (*AAAI) Fall Symposium on Narrative Intelligence*, 1999.

Bohnet, Iris, and Bruno Frey, “Social Distance and Other-Regarding Behavior in the Dictator Games: Comment”, *American Economic Review,* 89, pp. 335-39, 1999.

Bouas, Kelly S., and Samuel S. Komorita, “Group Discussion and Cooperation in Social Dilemmas”, *Personality and Social Psychology Bulletin,* 22, pp. 1144-50, 1996.

Brooks, Allison Wood, Hengchen Dai, and Maurice E Schweitzer, “I am sorry about the rain! Superfluous Apologies Demonstrate Empathic Concern and Increase Trust”, *Social Psychological and Personality Science,*5, pp. 467-74, 2013.

Buller, Paul, and Cecil Bell, “Effects of Team Building and Goal Setting on Productivity: A Field Experiment“, *Academy of Management Journal,* 29, pp. 305-28, 1986.

Camerer, Colin, and Richard Thaler, “Anomalies: Ultimatums, Dictators, and Manners”, *Journal of Economic perspectives,* 9, pp.209-19, 1995.

Charness, Gary, and Martin Dufwenberg, “Promises and Partnership”, *Econometrica,* 74, pp. 1579-1601, 2006.

Chen, Roy, and Yan Chen, "The Potential of Social Identity for Equilibrium Selection." *American Economic Review*, 101, pp. 2562-89, 2011

Coleman, James, “Social Capital in the Creation of Human Capital”, *American Journal of Sociology,* 94, pp. S95-S120, 1988.

Dal Bo, Pedro, Guillaume R. Frechette, and Jeongbin Kim “The Determinants of Efficient Behavior in Coordination Games”, *Games and Economic Behavior,* 130, pp. 352-68, 2021.

Dawes, Robin, Alphons Van De Kragt, and John M. Orbell, “Not Me or Thee, but We: The Importance of Group Identity in Eliciting Cooperation in Dilemma Situations: Experimental Manipulations”, *Acta Psychologica,* 68, pp. 83-97, 1988.

Deutsch, Morton, “Trust and Suspicion”, *Journal of Conflict Resolution*, 2, pp. 256-79, 1958.

Edelman, Benjamin, Michael Luca, and Dan Svirsky,” Racial Discrimination in The Sharing Economy: Evidence from a Field Experiment”, *American Economic Journal: Applied Economics,* 9, no. 2, pp. 1 – 22, 2017.

Efferson, Charles, Raphael Lalive, and Ernst Fehr, “The Coevolution of Social Groups and Ingroup Favoritism”, *Science,* 321 (5897), pp. 1844-49, 2008.

Ert, Eyal, Aliza Fleischer, and Nathan Magen, “Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb”, *Tourism Management* 55, pp. 62-73, 2016.

Fehr, Ernst, and Klaus Schmidt, “A Theory of Fairness, Competition, and Cooperation”, *Quarterly Journal of Economics,* 114, pp. 817-68.

Fehr,Ernst, Helen Bernhard, and Bettina Rockenbach, “Egalitarianism in Young Children”, *Nature,* 454, pp. 1079-83, 2008.

Frank, Robert H., “If Homo Economicus Could Choose his Own Utility Function, Would He Want One with A Conscience?”, *American Economic Review,* 77, pp. 593-604, 1987.

Frydlinger, David, and Oliver Hart, “Overcoming Contractual Incompleteness: The Role of Guiding Principles”, *Journal of Law, Economics, and Organization,* forthcoming.

Gibbons, Robert S., Manuel Greider, Holger Herz, and Christian Zehnder, “Building an Equilibrium: Rules versus Principles in Relational Contracts”, *Organization Science,* forthcoming, 2021.

Goette, Lorenz, David Huffman, and Stephan Meier, “The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups”, *American Economic Review,* 96, no. 2 (May), pp. 212-16, 2006.

Gneezy, Uri, “Deception: The Role of Consequences”, *American Economic Review,* 95, pp. 384-94, 2005.

Gneezy, Uri, Agne Kajckaite, and Joel Sobel, “Lying Aversion and the Size of the Lie”, *American Economic Review,* 108, pp. 419-53, 2018.

Johnson, Noel D., and Alexandra A. Mislin, “Trust Games: A Meta Analysis”, *Journal of Economic Psychology*, 32, pp. 865-889.

Kahneman, Daniel, Jack Knetch, and Richard Thaler, “Fairness as a Constraint on Profit Seeking: Entitlements in the Market”, *American Economic Review,* 76, 728-41, 1986

Kandori, Michihiro, “Social Norms and Community Enforcement”, *Review of Economic Studies,* 59, pp. 63-80.

Karlsson, Logi, Astrid Kemperman, and Sara Dolnicar, “ May I sleep in Your Bed? Getting Permission to Book”, *Annals of Tourism Research,* 62, pp, 1 – 12, 2017.

Kartik, Navin, “Strategic Communication with Lying Costs”, *Review of Economic Studies,* 76, pp.1359-95, 2009.

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Horisch, and Armin Falk, “The Formation of Prosociality: Causal Evidence on the Role of the Environment”, *Journal of Political Economy,* 128, pp. 434-67, 2020.

Klein, Cameron, Deborah DiazGranados, Eduardo Sales, and Huy Le, “Does Team Building Work?”, *Small Group Research*, 40, pp 181-220. 2009.

Lundquist, Tobias, Tore Ellingsen, Erik Magnus Johannesson, “The Aversion to Lying”, *Journal of Economic Behavior and Organization,* 70, pp, 81-92, 2009.

Mazar, Nina, On Amir, and Dan Ariely, “The Dis-honesty of Honest People: A Theory of Self-concept Maintenance”, *Journal of Marketing Research,* 45, pp. 633-44, 2008.

Mislin, Alexandra A., Rachel L. Campagna, and William P. Bottom, “After the Deal: Talk, Trust Building and the Implementation of Negotiated Agreements”, *Organizational Behavior and Human Decision Processes*, 115(1), pp. 55-68, 2011.

Morris, Michael, Janice Nadler, Terri Kurtzberg, and Leigh Thompson, “Schmooze or Lose: Social Friction and Lubrication in E-Mail Negotiations”, *Group Dynamics: Theory, Research, and Practice*, 6(1), pp. 89-100, 2002.

Rand, David, Thomas Pheffer, Anna Dreber, Rachel Sheketoff, Nils Wernerfelt, and Yochai Benkler, “Dynamic Remodeling of In-group Bias During the 2008 Presidential Election”, *Proceedings of the National Academy of Sciences,* 106 (15), pp. 6187-91, 2009.

Roth, Alvin E., “Bargaining Experiments”, in *Handbook of Experimental Economics.* John H. Kagel and Alvin E. Roth, Eds*.,* Princeton, NJ: Princeton University Press, 1995.

Subramaniam, Jeevan, “Small Talk Helps to Establish Trust”, ideas.bkconnection.com/5-ways-small-talk-serves-a-big-purpose , accessed 5/11/2012.

Taylor, Donald, and Janet Doria, “Self-serving and group-serving Bias in Attribution”, *Journal of Social Psychology,* 113 (2) pp. 201-11, 1981.

Tingley, Dustin, and Barbara Walter, “Does Cheap Talk Matter? An Experimental Analysis”, *Journal of Conflict Resolution,* 55, pp. 996-1020, 2011

Vogt, Sonja, Charles Efferson, and Ernst Fehr, “Can we see Inside? Predicting Strategic Behavior Given Limited Information”, *Evolution and Human Behavior,* 34, pp. 258-64, 2013.

Wikijob Team, “The Best Ways to Make Business Small Talk”, https://www.wikijob.co.uk/content/features/useful-resources/best-ways-make-business-small-talk, accessed 5/11/2012.

Wilson, Edward O., *The Social Conquest of Earth,* New York, NY: W. W. Norton, 2012.

## 2.8 Appendix

### 2.8.1 Procedures and instructions for the two experiments

#### 2.8.1.1 Experiment 1: Investor-Operator game

1. Subjects are recruited via Amazon Mechanical Turk. They report being resident in the United States and being 26 years old or older.
2. Subjects enter the experiment and are consented.
3. Subjects proceed to a screening task. In this task, subjects transcribe some nonsense text according to some rules that are given (e.g., “Only transcribe the first and fourth sentences. Make sure each sentence you transcribe has an exclamation point at the end”). This task is easy for native speakers and is meant to screen out subjects who do not speak English well.
4. Subjects proceed to the main task. Some answer a series of lifestyle questions before proceeding (rural/urban, tacos/sushi, beach/mountain, etc.). Others are asked to try to find the two most interesting things they have in common.
5. Subjects enter a waiting room where they are given the opportunity to play a game while they wait for a partner. Once a suitable partner has entered the waiting room, the two are paired and the game proceeds. There is a maximum wait time of 10 minutes, for which they are paid.

   (i) In the case of Treatment 1, the two simply proceed to the next step

   (ii) In the case of Treatment 2, the subjects have a video chat for three minutes with the instruction to find the most interesting thing they have in common. Subjects often fail to get their video equipment working, so if a subject reports his partner has not been able to video chat for more than a minute that subject tries a new partner.

   (iii) The pairs who answered the ten lifestyle questions are shown what answers

they had in common but are given no instruction other than to chat with their partner.

6. Subjects are then instructed in the rules of the Investor / Operator game. The rules to the game are then reproduced at the bottom of subsequent pages. Subjects must spend two minutes on this page.
7. Subjects are given a series of comprehension questions and are not allowed to proceed until they get them right.
8. Subjects play a practice game.
9. Subjects are notified that on the next page they will play the game for real with their partner.
10. Subjects then play the Investor / Operator game for real.
11. Subjects wait a few seconds to make sure their partner has moved.
12. Subjects then answer a variety of demographic questions.
13. Subjects are told the results of the game, are debriefed, and paid.

#### 2.8.1.2 Experiment 2: Twice-repeated Stag Hunt

Subjects are recruited via Amazon Mechanical Turk. They report being resident in the United States and being 26 years old or older.

Subjects enter the experiment and are consented.

Subjects proceed to a screening task.

(i) In the treatment with no small talk, subjects transcribe some nonsense text according to some rules that are given (e.g., “Only transcribe the first and fourth sentences. Make sure each sentence you transcribe has an exclamation point at the end”). This task is easy for native speakers and is meant to screen out subjects who do not speak English well.

(ii) In the treatment with small talk, subjects give a code word to an experimenter via video, and the experimenter gives a corresponding code word which allows

the subject to proceed. This verifies that the subject can speak English and that the subject has working video equipment.

Subjects enter a wait room where they wait to be paired with a partner. They are able to play a game while they wait if they wish. There is a maximum wait time of 10 minutes, for which they are paid.

Subjects proceed to the main task, which begins with an explanation of the rules to the Stag Hunt game (cast as Rabbit / Buffalo due to higher comprehension). The rules are reproduced at the bottom of subsequent pages. Subjects must spend two minutes on this page.

Subjects are given a series of comprehension questions and are not allowed to proceed until they get them right.

Subjects play a practice game.

Subjects are notified that on the next page they will play the game for real with their partner.

Subjects then play the Stag Hunt game for real.

Subjects wait a few seconds to make sure their partner has moved.

Subjects either video chat or proceed.

(i) In the treatment with no small talk, subjects are told the result of the first game and proceed.

(ii) In the treatment with small talk, subjects learn the result of the game and are told, “You have finished this game and will now video chat for three minutes with the person you just played with before moving on to the next task” in order to make it non-obvious that they will be playing the exact same game again. Subjects then talk with their partner for three minutes. They are told, “You will talk with your partner from the last game for 3 minutes before we move on to the next phase of the task.”. They must exchange a code word with each other in order to move on, verifying that the video chat happened.

Subjects are then told that they will play the same game again with the same person.

Subjects make their decision for the second Stag Hunt game.

Subjects wait a few seconds to make sure their partner has moved.

Subjects are told the results of the second game.

Subjects then answer a variety of demographic questions.

Subjects are told their earnings breakdown, are debriefed, and paid.

# 3 An Information-Theoretic Measure of Cultural Success


## Abstract

Models of genetic evolution are tested empirically by counting alleles: a good model of genetic evolution successfully predicts which genes will be found where. Culture is changing humanity at an astounding rate, but at present but we lack a means for measuring the flow of memes through minds in a quantitative, content-agnostic way analogous to counting alleles. I develop a method for measuring the information from a written work that is retained in the minds of those exposed to it, and which is therefore capable of influencing behavior. I estimate the entropy of samples from a target written work using a cloze-completion tasks in a treatment group (those that have read a target work) and a control group (those who have not read the target work). In doing this, we use human minds as encoders-decoders in Shannon's communication model. Difference measures taken between the entropy estimated with the treatment group and that taken with the control quantifies the information that the treatment group already knows relative to the control group, in bits. This method can control for shared cultural inheritance naturally, and it is content-agnostic—it does not require strong commitments to what information from the target work is important, nor commitments to what questions are important to ask. The technique can be extended to a variety of domains including evolutionary theory, methods of teaching, and the study of music.

## 3.1 Introduction

Natural selection provides us with a success criterion: success is replication. While the process of genetic evolution has remained relatively stable over human history thus far, there has been a remarkable increase in the bandwidth available for transmission of culture in the past few centuries. This increase is due to genetic adaptations for cultural transmission (as in evolved social learning psychology), cultural adaptations increasing transmission bandwidth and fidelity (such as schools) and most recently cultural artifacts that facilitate information storage, transmission, and interpretation (such as books, libraries, telephones, automatic translators, and the internet).

Dual Inheritance Theory, also referred to as gene-culture coevolution (Boyd & Richerson 1985), describes two mutually influential routes by which variations, either cultural or genetic, can be selected among and passed on—or transmitted to—subsequent generations. The gene-culture feedback loop has been present for as long as there has been transmissible culture, beginning with behaviorally modern humans approximately 80,000 years ago (Fisher & Ridley 2013). In brief, dual-inheritance theory describes the way that cultural traits (defined as socially learned information stored in human brains and capable of affecting behavior) change the environment under which genetic selection operates (Richerson & Boyd, 2005). The canonical example of culture influencing genes is lactase persistence: three groups of humans have separately evolved the ability to produce lactase, an enzyme that enables digesting dairy foods, into adulthood. This happened as a direct result of the rise of dairying, and hence the availability of milk products as food, in these cultures (O'Brien & Laland, 2012). Dairying is culturally-transmitted, and this cultural information has resulted in the genetic adaptation of lactase persistence. Causation in the other direction, genetic influence on

the ability to transmit culture, is also widespread, manifesting as psychological adaptations for social learning.

Evolution via natural selection at the genetic level is well-understood. Theories of genetic evolution model empirical data, the data being, at its core, simply counts of alleles: we are interested in the frequency of a particular variation of a gene found in some population being measured. Our models of genetic evolution are grounded in this data, and we evaluate models based on their fit with it.

The study of cultural evolution is a burgeoning field, with many important advances made over the last few decades: scholars have developed mathematical models of cultural evolution starting in the 1970s, (e.g. Cavalli-Sforza & Feldman 1981, Boyd & Richerson 1985, Boyd & Richerson 2005, McElreath & Henrich 2007). However, theories of cultural evolution have thus far lacked empirical grounding analogous to counting alleles. What these theories require is a content-agnostic means of measuring the flow of ideas from one human mind to the next, and for measuring the frequency of these ideas in populations of interest.[20] Up to now we have had limited means for empirically verifying models of cultural evolution. Empirical investigation has thus far relied on qualitative techniques, or on quantitative techniques that require two things: first, strong commitments to what pieces of culture are important and second, measurement of behaviors induced outside of natural settings, as in a lab.

Substantive questions in the study of cultural evolution are often motivated by the flow of whole packages of ideas that are meaningfully connected. Consider, for instance, religion and religious texts: the Bible, the Koran, the Vedas, the Torah, the Pāli Canon, etc. Each is a rich tapestry of interwoven ideas. Probing the minds of informants with specific questions about e.g. the Bible may give some insight into their familiarity with the text or the religion, but it does require a commitment to knowing the right questions to ask beforehand. For instance, we might wish to investigate how Christianity has shaped the mind and behavior of the people making up a small-scale society, and

[20] See Chvaja (2020) for a discussion of Memetics, a largely qualitative and theoretical effort to articulate how ideas might be quantized and and their flow modeled, around the turn of the century.

so we might ask something like: “What is the name of the mountain on which Moses received the Ten Commandments?”. And answers to this question will probably produce some signal of Christian-ness. But there are many ways that ideas, beliefs, and practices might flow from the Bible and in to our subjects’ minds while leaving this particular piece of information behind—or at least making it difficult to recall. Even a whole set of exam-type questions makes for a very strong commitment to a particular kind of interaction with the text being the *right* kind of interaction. People in this hypothetical society may, for instance, have focused on the New Testament, or they may have received the substantive content but not the specific labels associated with the Ten Commandments episode. They may not even know these ideas come from the Bible, and so could be completely unable to answer. In either case, having strong commitments to what questions are the right questions means we will have missed a substantial information flow from the text into their minds.

We have seen increasing empirical investigation with laboratory experiments (see Mesoudi 2016 for a review) that clarifies, among other things, the mechanisms by which we decide what information to retain (e.g., prestige), the demographic structure that is required in a society to maintain transmission (e.g. Henrich 2004, Muthukrishna et al. 2014), and the mechanisms by which the information transmitted changes (e.g. Derex et al. 2019). But empirical investigations of the transmission information via culture have thus far involved artificial settings and deliberately induced transmission, as in laboratory experiments, and limitation to semantically simple ideas, and strong commitments as to what the right questions to ask are in all cases. In the wild, the behaviors that result in the transmission of information via culture happen across many different contexts in response to rich sets of cues about what to transmit, what to receive, and from whom. The ideas that flow from one mind to the next are semantically rich and difficult to bound, making laboratory measurement illuminating but necessarily limited. In addition to working outside natural settings, lab experiments must use very particular, necessarily idiosyncratic ideas: a specific drawing task, transmission of a given sentence, certain software, etc.

These two classes of limitations highlight the unmet need and potential opportunity developing the empirical underpinnings of theories of cultural evolution. The method set out in this paper is a step towards the kinds of measurement necessary to make theories of cultural evolution as empirically quantifiable as those of genes. It is hoped that it is a step towards their further integration as well.

## 3.2 Aims

The value inherent in modern models of genetic evolution via natural selection is twofold: first, they can predict change in a measurable quantity related to traits (*e.g.* $\Delta z$ in the Price equation), and second, their form reflects the underlying data-generating process (albeit greatly simplified). I outline a method for measuring the amount of information from a particular written cultural artifact that is actually present in a human mind, which I will call *retained novel information* (RNI). RNI could allow for a principled measurement of an analog to $\Delta z$ for culture, as well as allowing for the comparison of information flow via various teaching methods, the mapping of story archetypes across the world, and many other uses.

The method derives from Claude Shannon's 1951 procedure for measuring per-character entropy of written English. Shannon designed this procedure using human minds to predict encoded message content (i.e., written English) because he did not have access to large digitized corpora and because his computers used slide rules rather than microchips. We all do, in fact, walk around with a detailed knowledge of the statistics of the natural languages we speak held in our minds, and Shannon cleverly took advantage of this fact. Modern computational techniques and corpora remove the need for such a procedure if the quantity of interest is merely the entropy of written language—but what if we were to use human minds to measure entropy *precisely because* we are interested in their properties as codecs (encoders-decoders) when applied to the written word?

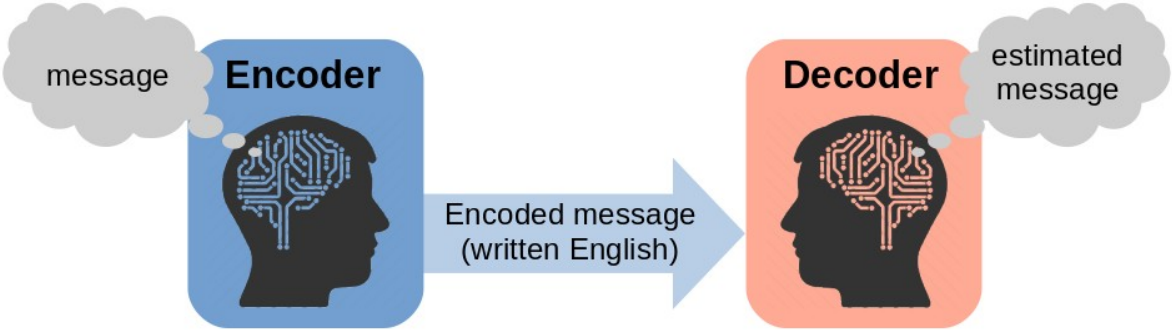



*Figure 14: A diagram showing the communication system under study. The message begins in the mind as thoughts, and those thoughts are formed into language. The language is written, then read, and becomes thoughts again. The mind of each speaker of a natural language has within it a sophisticated model that allows prediction of the encoded message to varying degrees, depending on familiarity with the work and the extent to which the information in that work is inherited via culture even without exposure via direct experience.*

The procedure, then, is very simple: define a treatment group (those familiar with a target cultural artifact, perhaps a book they have read) and a control group (matched to the treatment group, but not exposed to the target artifact). Then, each group completes tasks designed to measure the entropy of written language, in our case English, but only with samples from the target artifact. We then compute a difference measure between these two entropy estimates; Kullback-Leiber divergence gives us a measure of the number of additional bits it would take to encode the text with the worse (controls') language model, as an example. The entropy measurement from the treatment group is a property of the coded sequence (in the case of a book, the text) itself *given a particular decoder*, and RNI is the quantity of information from the target that is revealed to be in the decoder—in the mind. This is the information that is both novel relative to the treatment group and the control group's shared experiences and also retained. It is not the case that all novel information from the target artifact is

retained, of course, as memory has evolved to be highly selective (Kuhl & Chun 2014). We can quantify RNI by multiplying the result of the difference measure by the length of the text, denominated in the relevant unit of analysis (character, word, sentence, etc.). This will yield an estimate of the amount of information in a given target artifact for a particular reader.

## 3.3 Background: information theory

The theory underpinning RNI measurements stems ultimately from *A Mathematical Theory of Communication* (Shannon 1948). In brief, Shannon outlines a measure that describes the information entropy associated with a discrete random variable *X*:

$$\mathrm{H(X)} = \text{-}\sum_{i=1}^{n} \mathrm{P}(x_i) \log_2 \mathrm{P}(x_i) \qquad \text{(Equation 3.1)}$$

Where H is information entropy and P($x_i$) is the probability of possible value $x_i$. As H*(X)* increases, the amount of information a given datum communicates increases. We can take the example of a fair coin and calculate that each flip of that fair coin is one bit:

$$\mathrm{H}(X) = -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} \qquad \mathrm{H}(X) = -\sum_{i=1}^{2} \frac{1}{2} \cdot (-1) = 1 \qquad \text{(Equation 3.2)}$$

The intuition is that any flip of a fair coin carries the same amount of information—you are no more surprised if it comes up Heads than if it comes up Tails. However, in a situation where all outcomes of an information-generating process are *not* of equal probability, we are more "surprised" by some outcomes than others. Suprisal is formalized as the reciprocal of the probability, $1/P_x$, and is perhaps the more intuitive quantity. In a situation with an unfair coin where that has a 99% probability of producing

Heads and 1% for tails, we are very surprised when Tails occurs—its probability was low.

$$\mathrm{H}(X) = -(\frac{99}{100}\log_2\frac{99}{100} + \frac{1}{100}\log_2\frac{1}{100}) = 0.08 \qquad \text{(Equation 3.3)}$$

In this case, each flip of the unfair coin delivers, *on average,* only 0.08 bits of information. This is because we are quite sure that any given flip will result in a Heads. Each flip that results in a Heads delivers much less information than a flip that results in a Tails. A coin with Heads on both sides delivers exactly zero information with each flip; you are completely sure the outcome will be heads. As with a fair coin, a string of letters drawn randomly from the 26 in the English alphabet has the highest possible entropy for that encoding scheme—there is no way to predict what the next character will be on the basis of previous characters. In an optimally compressed data stream it becomes impossible to predict the next datum based on previous data.

This sort of measurement can be applied to any encoding scheme: the bits that are ubiquitous in digital information storage and transmission are a base-2 encoding scheme in that they can take two states (one and zero), language (particularly written language) is amenable to treatment, DNA uses a base-4 rather than base-2 encoding scheme, and many others.

## 3.4 Measuring entropy in language

We now have the necessary concepts, but we need a way to estimate how much information is transmitted via a particular artifact to a defined set of human minds. Language is made up of words which can be expressed in writing with characters, and characters are limited in number: there are only 26 letters in the English alphabet. Characters as they occur in writing can then be understood as a random variable. If we are using human minds as decoders, entropy per character can be calculated by

asking, “To what degree is a person able to predict the next letter given a sample string?” An optimal encoding scheme for transmission without noise would answer “not at all” (and hence bits per character would be high and uniform). But people are quite able to predict the next letter given the previous *n-1* letters of an n-gram in any natural language. This is due to redundancies built in to language to deal with noise in the transmission channel, as well as higher-level prior expectations about content (as in cultural inheritance; much of what is said is already known to the receiver with high probability). Here we will consider English, but it is expected that all natural languages have very similar properties in this regard (Behr 2006).

In 1951 Shannon published a method for empirically estimating the upper bound of entropy of written English, known as the “Shannon game”. The experiment asked subjects to guess the letters making up a sentence, where the set of symbols is the 26-letter English alphabet plus space, for 27 total symbols. Starting at the first letter, the subject makes a guess. The subject is given feedback on whether her guess is correct, and if it is correct she writes it in the blank and moves on to the next test item. If her guess is wrong, she tries again until it is right. The number of tries it takes to guess correctly is recorded.

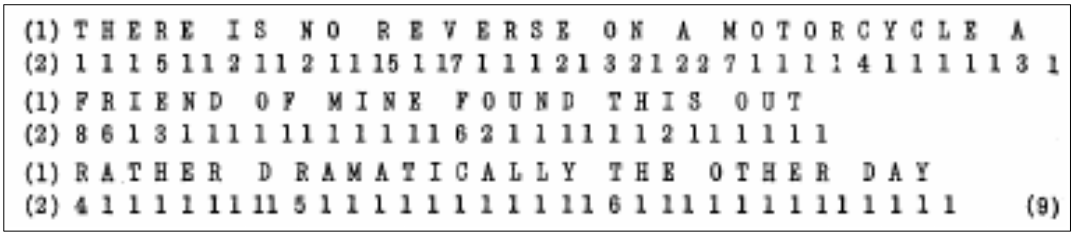



*Figure 15: Example data from Shannon (1951). Rows labeled (1) are the target sentence, and rows labeled (2) show the number of guesses necessary to hit upon the correct letter for a given subject. Note the preponderance of ones.*

This data allows the estimation of entropy per character of written English conditional on the model used (*i.e.,* on the particular reader’s mind). Shannon estimates $\hat{q}_i^N$, the probability that the subject requires *i* guesses to land upon the correct letter following a

sequence of *N-1* symbols, and shows the following is an upper bound for the entropy of printed English:

$$H(X) \leq -\sum_{i=1}^{27} \hat{q}_i^N \log(\hat{q}_i^N) \qquad \text{(Equation 3.4)}$$

More recent work has improved upon that technique. Using a paradigm that involved asking subjects to bet on subsequent characters, Cover and King (1978) estimated about 1.3 bits per character and subsequent estimates have coalesced around approximately 1-1.3 bits per character (Takahira 2016, Ren et al. 2019). An improved estimator such as Cover and King's together with an incentive-compatible task would be used for empirical measurements today.

One may rightly worry that a character-level model may not capture the deeper semantics underpinning our interest in it as a medium for the transmission of culture. After all, interest in the transmission of culture is generally rooted in the ideas that are transmitted[21]. For the sake of clarity I have used examples from character-level models here, though this general technique can be extended to semantically richer word-level models and even further to sentence fragments. There are a variety of enhancements that could allow for accurate estimation of entropy moving up through words and sentence fragments while keeping the task tractable for subjects, such as using language models to eliminate trivial questions (questions where the probability any given respondent is able to get the right answer in one try is quite high), and questions the subject has no hope of getting right. In addition, the sampling of n-grams from the target work would benefit from importance sampling. The specifics remain to be developed in future work.

[21] There is some indication that character-level models with long enough n-grams capture higher-level semantic features to some extent. CharRNN character-level models trained on a large English corpus vs. those trained on a large English corpus plus a particular target work do evince a difference in entropy measurements of ~0.2-0.4 bits per character when tested on n-grams (n=64 or 100) sampled from the target work. That these relatively crude models are able to discern a difference is somewhat surprising in itself (Cashman 2018, unpublished).

## 3.5 Applying entropy measurements to information flows through culture

If H(P(*x*)) is the entropy measured with the treatment group (those who have read the target work) and H(Q(*x*)) is the entropy measured with the control group (matched subjects who have not read the target work), then a measure of the distance between these two probability distributions is Kullback-Leiber divergence,

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right) \qquad \text{(Equation 3.5)}$$

which can also be, perhaps more intuitively, expressed as the entropy given P(x) minus the cross-entropy:

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) && (1) \\ &= \sum_{x \in \mathcal{X}} p(x) \log p(x) - p(x) \log q(x) && (2) \\ &= H(P, Q) - H(P) && (3) \end{aligned} \qquad \text{(Equation 3.6)}$$

The KL divergence reflects the additional number of bits required to encode *X,* a random variable representing a stream of written English, when using an inferior model, here *q(x),* the minds of the non-readers, rather than *p(x),* the minds of the readers. This, when multiplied by the length of the target artifact, is a measure of RNI—the novel information the subjects retained from the target artifact. The RNI a representation of everything subjects in the control group failed to predict but which subjects in the treatment group successfully predicted, so it is a measure of the information from the target work that has been retained in the minds of the treatment group and is therefore

capable of influencing behavior. The proportion of RNI will vary among different target artifacts, giving us a window into the extent to which a particular artifact is predictable by anyone regardless of exposure. For example, upon hearing a story that begins with a local youth leaving for parts unknown, many would guess that she will undergo trials and return wiser and stronger. Only a small proportion of a given written work is likely to be novel in the sense of being unpredictable by the naïve, while the remainder can be thought of as shared cultural inheritance because it can be predicted by any member of the culture defined in the experiment by the control group.

To illustrate with a character-level example, we might consider a culture that has been successful in recent centuries: Mormons. Given a control group (Americans), a treatment group (Mormon Americans), and a target work only the treatment group (presumably) has been directly exposed to (the Book of Mormon), we can then estimate entropy of the target work in each group. The difference in entropy measurements between the two groups, the RNI, is driven by their minds' differing properties as decoders—differences in their language models themselves. RNI gives us a window into how much information from the target work has actually been transmitted to the treatment group's minds and is capable of influencing behavior; it is the number of additional bits that non-readers require to encode the sample, on average. If we measure 1.3 bits per character in the control group (non-Mormons) and 1.0 bits per character in the treatment group (Mormons), we might see a difference measurement of ~0.3 bits per character. This, multiplied by the number of characters in the Book of Mormon, is the number of additional bits required to decode the Book of Mormon using a non-optimized model (about 500Kb). The minimal task is very simple: subjects are given randomly selected n-grams from the target work with the last item covered, and the instruction to guess the missing character. The number of guesses necessary to get the right letter for the blank is recorded, and entropy is calculated from that data.

However, without randomizing assignment to treatment or not the experiment would be confounded by subjects' differing abilities at cloze completion tasks, among other things. A difference-in-difference design would allow for meaningful

measurements in the field without random assignment. If all subjects complete both items from the target work and items from a huge, diverse corpus we can implement subject-level controls for the influence on RNI of factors other than exposure to the target work. The procedure is simple: at the subject level, calculate H for the target work and H for the large corpus control. Take the difference between these two measures to arrive at a measurement of entropy in the target work controlling for the factors that would affect entropy measurements in the large corpus control but which would *not* affect entropy measurements in the target work. The set of things that may influence entropy measurements of the target work but *not* the control are quite small, and probably limited to exposure to the target work itself. Once this is done, one may calculate the difference between the treatment and control groups' entropy estimates to arrive at an RNI measure that controls for ability on cloze tasks, among other things.

## 3.6 Discussion

I have outlined a general procedure for measuring retained novel information given a control group who are not familiar with a particular book, and a treatment group who are. The ability to perform precise estimations of just how much information from a particular written cultural artifact is present in a living human mind may be novel, but what, really, are we measuring? And how can we use it?

We can consider several different scenarios in order to develop intuitions about what RNI is measuring. First, consider the situation where the target artifact is a book full of truly random gibberish. Entropy of the work is very high, but very little is going to be retained in the minds of humans who read it. If we were to print hundreds of copies and bury them in the desert, we would have a very high-entropy cache of cultural artifacts. Nobody would ever read those books though. For the purposes of investigating cultural evolution, we really only care about information that is in *living* minds and therefore capable of influencing behavior. Random Gibberish, Vols. 1-10 is very high entropy but very low RNI. We can compare this to the Book of Mormon, a religious text

written by Joseph Smith in the 1820s, which is noted for its reuse of Abrahamic religious tropes and approximation of the style of the King James Bible. We might expect the Book of Mormon to have approximately average entropy per character for English, while having RNI lower than average. The Pāli Canon, a foundational Theravada Buddhist text, reflects its origins in an oral tradition by being extremely repetitive. It is also, perhaps necessarily, extremely long—it takes up approximately one bookcase. We would expect that a control group, non-readers, would measure below-average entropy (the text is very predictable, perhaps on a character-level model people are guessing whole words or sentences), and that the relative proportion of RNI given the treatment group would be small (there is little novel information to retain per symbol). Contrast these two books with something like a technical manual that outlines how to machine a specific part for an airplane. Though there are the usual regularities, redundancies, and other error-correcting properties of natural language, we would expect that a control group would result in approximately average entropy measurements (perhaps a little high), while the treatment group would result in an RNI measurement considerably below average for a treatment group: much of the information in such an artifact is very predictable once you have been exposed to it (indeed, you have a model of the machined part in your mind). However, this information would be very difficult to predict *a priori*. There is also a class of target artifacts for which RNI as measured will appear to be very large, but in fact it is merely a few bits. Consider the example of a simple pseudorandom number generating algorithm: such an algorithm could be used to generate a sequence of letters, words, or even sentences with a regularity that only becomes clear once the reader has finished the work. If the reader has inferred the correct algorithm, then prediction of the next symbol from an n-gram becomes trivial—we would observe a very high entropy measurement in the control group and a very low measurement in the treatment group (perhaps zero), but we know that the information that was learned can in fact be expressed with relatively few bits. This sort of problem is not likely to arise in natural language texts, but there may be certain classes of ideas, or groups of ideas, that have something of this property.

### 3.6.1 Limitations

Perhaps the most important limitation to the method described is its reliance on the written word. Much of culture is transmitted outside of written language, for example via speech, facial expression, changes made to an environment, and observing then imitating others' actions. The proportion of information transmitted via culture that is transmitted via writing has increased exponentially in recent centuries, to be sure, but it is still only a fraction (and perhaps not the most important fraction) of the total flow of information via culture.

Another concern might be that the method outlined is completely agnostic to content; it does not reflect *what*, though it does accurately reflect *how much.* However, it does depend on the selection of a target artifact from which samples can be drawn for testing. If interested in measuring the flow of novel information from that artifact, then there is no problem. But there may be instances where the ideal measurement is only approximated by a particular work. One of the advantages of using this method to track information flows is that there is no need to make a definitive claim about what is important. More conventional methods might ask questions to test subjects' knowledge of a particular concept, event, vignette, or similar. This requires making a strong claim about what knowledge is important with respect to the subject being researched, namely choosing the questions. In some circumstances we may just trading this problem for the problem of choosing the target artifact well.

Finally, though I give a method for measuring how much information from a given target artifact is present in a living mind, we do not have the theoretical machinery necessary to link this to behavior. There may be a certain number of bits stored in a mind and capable of influencing behavior, but it is far from clear when it would be influential or how influential it would be. We might gain some traction from the assumption that information that is retained has passed through an algorithm optimized to only retain information that increases inclusive fitness. If we make this assumption

then perhaps we can make the general claim that more information means more influence on behavior, though it would remain unclear exactly what and when.

### 3.6.2 Future directions

*Evolutionary theory*

This sort of procedure might also eventually be put to work bringing together models of genetic and cultural evolution. Measuring gene flow, selection among genes, etc. is relatively straightforward with modern technology, but models of cultural evolution have thus far lacked a principled, general-purpose means of empirically measuring the flow of ideas through culture. Information-theoretic measures of cultural success might provide the basis for that interaction, helping open the way to a more general dual-inheritance theory. Krakauer et al. (2020) describe a method of defining individuals on which selection happens, historically a key puzzle for those attempting to analogize culture to genes (see Chvaja 2020 for a discussion). This sort of principled definition of an individual might allow for the definition of units of culture on which selection is happening. Though it is not clear why one bit of DNA might be comparable to one bit of RNI, it's possible that being able to treat both culture and DNA (a base-4 encoding scheme) in terms of information may be useful in the future as well. Finally, we are on the cusp of having the practical ability to edit our genomes directly. This will allow the transmission of genetic information via cultural means. At first bandwidth will be very small, but it is reasonable to expect the technological capacity will grow. Popular genes may spread like popular books; selection mechanisms and network effects that were once the sole province of catchy ideas will soon be applicable to our DNA, and there may be a role for the sort of analysis presented here in modeling these changes.

*Measurements across multiple different works*

We need not limit ourselves to a single target work, a treatment, and a control. We can design measurement schemes that look at two or more works. Consider investigating the relationship between two target works, A and B. The treatment group has been exposed to A, while we have taken our test set sampled from B. By testing subjects on material from B after having been exposed to A, in the case of the treatment group, or not, in the case of controls, we can investigate the extent to which A and B are similar relative to each other. Treatment groups could be exposed to a wide variety of target works and then all test with samples from B, or we could vary the test sets to come from a wide variety of works.

*Flow through social networks*

We might also deploy this procedure while measuring information flowing from a certain target work through social networks. This would allowing the measurement of information flow in a way that is agnostic to content. Attempting to measure the flow of information, beliefs, etc. at present requires knowing which questions to ask. If we can be satisfied with a single work as a target artifact, then we can deploy the framework outlined to trace the information from that work as it is transmitted from person to person via a series of measurements across time and distributed in social networks. This sort of procedure would be amenable to measuring variation within levels of familiarity beyond just a binary familiar/not familiar measure.

*Education*

Principled methods of measuring information flow also lend themselves to the study of teaching and learning. In educational settings, we might compare different teaching methods or materials via measuring RNI pre- and post-treatment, given a target artifact of educational value. For instance, three matched groups of students might be taught using the same history textbook, but each group’s teacher is assigned a

different teaching method. We can then measure RNI from the history textbook at the end of the class to assess how much information was retained. This measurement scheme could shed light on how much pupils are learning from a target work, and critically without the assumptions about what is important inherent in traditional testing regimes: a student may not be able to recall the year of the Battle of Hastings, but the same student may have in fact learned many other things from the book and instruction.

*Charting cultural tropes and archetypes*

We might also chart how story archetypes are distributed across the world. One way this could be accomplished is to carefully select target stories that are then coded as particular archetypes and used the standard procedure, testing for RNI within many different cultures relative to a single control. Another related technique might use specific, well-known works. We might investigate how the predictability of the Bible or the Koran varies across cultures, giving us a new window into the influence the information in those books has across the globe. Given that these are long and varied texts, it would also be possible to partition them and look at the RNI for subsections such as Old Testament versus the New Testament.

*Music*

The method outlined above for written language can be applied without change to music as well. Music can be coded using a set of symbols as well (notes on a staff or scientific pitch notation), and we can measure expectation in exactly the same way. There is a small literature on expectation in music (see Large & Kim 2019 for a review) including some that incorporates Shannon's methods of estimating entropy (Manzara et al. 1992, Pearce & Wiggins 2012). We could use a very similar technique as for language to investigate the distribution and spread of musical archetypes and tropes throughout the world.

## 3.7 References


Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.

Boyd, R., & Richerson, P. J. (2005). *The origin and evolution of cultures*. Oxford University Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.

Chvaja, R. (2020). Why Did Memetics Fail? Comparative Case Study. *Perspectives on Science*, *28*(4), 542–570. https://doi.org/10.1162/posc_a_00350

Cover, T., & King, R. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, *24*(4), 413–421. https://doi.org/10/bfhsp5

Derex, M., Bonnefon, J.-F., Boyd, R., & Mesoudi, A. (2019). Causal understanding is not necessary for the improvement of culturally evolving technology. *Nature Human Behaviour*, *3*(5), 446–452. https://doi.org/10.1038/s41562-019-0567-9

Fisher, S. E., & Ridley, M. (2013). Culture, Genes, and the Human Revolution. *Science*, *340*(6135), 929–930. https://doi.org/10.1126/science.1236171

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Henrich, J. (2004). Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case. *American Antiquity*, *69*(2), 197. https://doi.org/10.2307/4128416

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., & Ay, N. (2020). The information theory of individuality. *Theory in Biosciences*, *139*(2), 209–223. https://doi.org/10.1007/s12064-020-00313-7

Kuhl, B. A., & Chun, M. (2014). *Memory and Attention* (A. C. (Kia) Nobre & S. Kastner,

Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199675111.013.034

Large, E. W., & Kim, J. C. (2019). Musical expectancy. In *Foundations in music psychology: Theory and research* (pp. 221–263). The MIT Press.
Manzara, L. C., Witten, I. H., & James, M. (1992). On the Entropy of Music: An Experiment with Bach Chorale Melodies. *Leonardo Music Journal*, *2*(1), 81–88.

McElreath, R., & Henrich, J. (2007). *Modelling cultural evolution*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198568308.013.0039

Mesoudi, A. (2016). Cultural Evolution: A Review of Theory, Findings and Controversies. *Evolutionary Biology*, *43*(4), 481–497. https://doi.org/10.1007/s11692-015-9320-0

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1774), 20132511. https://doi.org/10.1098/rspb.2013.2511

O'Brien, M. J., & Laland, K. N. (2012). Genes, Culture, and Agriculture: An Example of Human Niche Construction. *Current Anthropology*, *53*(4), 434–470. https://doi.org/10.1086/666585

Pearce, M. T., & Wiggins, G. A. (2012). Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science*, *4*(4), 625–652. https://doi.org/10.1111/j.1756-8765.2012.01214.x

Ren, G., Takahashi, S., & Tanaka-Ishii, K. (2019). Entropy Rate Estimation for English via a Large Cognitive Experiment Using Mechanical Turk. *Entropy*, *21*(12), 1201. https://doi.org/10.3390/e21121201

Shannon, C. (1951). Prediction and Entropy of Printed English. *The Bell System Technical Journal*, 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Takahira, R., & Tanaka-Ishii, K. (2016). Upper Bound of Entropy Rate Revisited; A New Extrapolation of Compressed Large-Scale Corpora. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, *Osaka, Japan*, 213–221.