Cooperation by Chance: Noise, Repetition, and Rationalization*

Aidin Hajikhameneh† **Laurence R. Iannaccone**‡

Abstract

Can noise sustain cooperation even when it reveals nothing about anyone's behavior? We study how superstition emerges from strategically irrelevant payoff noise, focusing on the relationship between repetition and causal rationalization. In an indefinitely repeated prisoner's dilemma, we compare a no-noise baseline with two noise treatments generated by the same adjustment algorithm. In one, the randomness of the adjustments is disclosed. In the other, their source is left unexplained. This design allows us to ask whether behavioral responses to favorable random outcomes appear to require a corresponding causal interpretation. Payoff noise raises direct cooperation by about eight percentage points relative to the no-noise baseline. Leaving its source unexplained does not increase average cooperation further, but it shifts strategic plans away from unconditional defection and toward more cooperative plans. It also changes how participants respond to the adjustments they receive. Under unexplained noise, cooperation followed by a favorable adjustment is more likely to continue, especially within cooperative streaks, whereas this response is substantially weaker when the randomness is disclosed. Yet few participants report that cooperation produced favorable adjustments, and such causal accounts are no more common under unexplained noise. The joint pattern is more consistent with a repetition-first process in which random reinforcement can shape behavior before any corresponding causal rationalization becomes observable.

JEL classifications: *C91, C92, Z12, Z13*
Keywords: *superstition, cooperation, random reinforcement, causal beliefs, environmental uncertainty, repeated games*

*Figures were created with R, and experiments were programmed using z-Tree (Fischbacher, 2007).

†Department of Economics, San Jose State University, Dudley Moorhead Hall, 1 Washington Square, San Jose, CA 95112, USA, email: aidin.hajikhameneh@sjsu.edu

‡Institute for the Study of Religion, Economics, and Society, Chapman University, 338 N. Glassell, Orange, CA 92866, USA, email: iannacco@chapman.edu

1 Introduction

People often make decisions under uncertainty, when the causes of observed outcomes are difficult to identify. In such environments, random reinforcement can make causally irrelevant actions appear consequential. An action that happens to be followed by a favorable outcome may become more likely to be repeated, may be interpreted as having caused the outcome, or both. These responses need not arise in the same order. Repetition may begin before any causal explanation develops, or a causal interpretation may arise first and then sustain the behavior. These responses are central to theories of superstition, which explain how random reinforcement can generate persistent behavior and causal beliefs about otherwise unrelated sequences of action and outcome (Malinowski, 1918; Skinner, 1948; Foster and Kokko, 2009; Cushman, 2020). A growing economic literature links environmental uncertainty to stronger religiosity, the apparent effectiveness of religious practices, and greater prosocial behavior (Sinding Bentzen, 2019; Espín-Sánchez *et al.*, 2026; Hajikhameneh and Iannaccone, 2023, 2025; Hajikhameneh, 2026). Yet this literature leaves open a central question for strategic interaction: can payoff shocks that change earnings but not the strategic incentives of the game alter cooperation, and when they do, can uncertainty about how those shocks are generated lead people to interpret favorable adjustments as caused by cooperation?

Skinner's (1948) seminal experiment showed how behavior can be reinforced by outcomes that are unrelated to it. Pigeons receiving food at predetermined intervals developed idiosyncratic routines, behaving as though their actions produced rewards that were in fact independent of them. Human experiments document related patterns. Noncontingent reinforcement can sustain repeated responding and create a sense of control even when outcomes do not depend on behavior (Catania and Cutts, 1963; Ono, 1987; Matute, 1994; Aeschleman *et al.*, 2003).[1] Such behavioral persistence does not require a fully articulated causal belief. People may continue to act on a superstitious intuition even while recognizing that the implied causal connection is irrational, a form of acqui-

[1] In consumer settings, accidental associations between products and favorable outcomes can also shape preferences, particularly when the desire for control is strong and perceived control is weak (Hamerman and Johar, 2013).

escence analyzed by Risen (2016). Malinowski's (1918) account emphasizes another response to uncertainty. When reliable knowledge and effective control are limited, people may assign causal meaning to otherwise unexplained outcomes.[2] Research on causal illusion similarly shows that repeated co-occurrence can produce perceived causal relationships even when the underlying contingency is absent (Alloy and Abramson, 1979; Matute *et al.*, 2011; Blanco *et al.*, 2013; Matute *et al.*, 2015). Together, these literatures identify two closely related manifestations of superstition: the persistence of behavior and the construction of causal meaning. They do not, however, determine how the two become connected.

We address these questions in an indefinitely repeated prisoner's dilemma with strategically irrelevant payoff noise in the form of adjustments to realized earnings.[3] A large experimental literature on uncertainty in indefinitely repeated prisoner's dilemmas focuses instead on imperfect monitoring, in which players receive noisy information about actions or outcomes. In such environments, uncertainty is strategically relevant because an adverse signal may reflect either intentional defection or chance. Cooperation may therefore rely on lenient or forgiving strategies that delay punishment or restore cooperation after adverse signals (Dreber *et al.*, 2008; Dal Bó and Fréchette, 2011; Rojas, 2012; Fudenberg *et al.*, 2012; Rand *et al.*, 2015; Aoyagi *et al.*, 2019). See Dal Bó and Fréchette (2018) for a comprehensive review. Our setting isolates a different channel. Adjustments alter earnings, but actions and stage-game outcomes remain perfectly observed and the adjustments convey no information about behavior.[4] This setting allows us to separate the effect of strategically irrelevant payoff adjustments from the effect of uncertainty about how those adjustments are generated.

[2]Malinowski based this argument on the fishing practices of the Trobriand Islanders. Fishing in the protected lagoon relied on knowledge and skill and involved little magic, whereas hazardous and unpredictable open-sea fishing was accompanied by extensive ritual. His point was not that ritual displaced rational reasoning, but that it emerged alongside it when ordinary knowledge and control reached their limits (Malinowski, 1918).

[3]Repeated interaction, often modeled as a repeated prisoner's dilemma, has long been central to explanations of cooperation in economics and the evolutionary sciences (Trivers, 1971; Axelrod and Hamilton, 1981; Axelrod, 1984; Boyd and Richerson, 1988). Folk-theorem results characterize conditions under which cooperation can be sustained (Fudenberg and Maskin, 1986), while evolutionary models study the emergence and stability of cooperative strategies (Smith, 1982; Nowak and Sigmund, 1993).

[4]Unlike implementation errors or noisy signals, the adjustments change neither the action selected nor the action observed, and they leave the stage-game outcome intact.

The experiment consists of three prisoner's dilemma treatments: Baseline (no noise), Random (disclosed random noise), and Unknown (opaque noise). In Baseline, earnings are determined solely by stage-game payoffs. In Random, each pair receives payoff adjustments that are explicitly described as random and unrelated to behavior. In Unknown, adjustments are generated by the same random mechanism, but participants receive no information about how they are generated, leaving their source open to interpretation. Actions and stage-game outcomes are perfectly observed in every treatment. The comparison between Baseline and Random identifies the effect of payoff noise when its randomness is disclosed, while the comparison between Random and Unknown isolates the effect of leaving the source of otherwise identical noise unexplained.

To distinguish ex ante strategic plans from behavior in the moment, the experiment uses two decision modes within subjects. At the start of each supergame, participants select one of 20 predefined contingent strategies. A computerized coin flip then determines whether the selected strategy is implemented automatically for the entire supergame or whether participants make decisions round by round. This repeated random assignment permits within-subject comparisons of the strategies participants choose, the actions they take in direct play, and the consistency between the two. After all supergames, participants in the Random and Unknown treatments answer open-ended questions about how they interpreted the payoff adjustments. The design therefore provides three complementary views of behavior and interpretation: the strategies participants intend to follow, the actions they take as outcomes unfold, and the causal explanations they articulate after play.

The results reveal three main patterns. First, both noise treatments increase cooperation in round-by-round play relative to the no-noise Baseline, and average direct-decision cooperation does not differ between the Random and Unknown treatments. Second, participants in the Unknown treatment move away from Always Defect and toward more cooperative strategies relative to participants in either the Baseline or Random treatment. Third, in the Unknown treatment, favorable adjustments following cooperation are associated with a greater likelihood that cooperation continues, especially within cooperative streaks. This response is substantially weaker in Random, when participants know that the adjustments are random. Yet few participants explicitly report that

cooperation produced favorable adjustments, and such causal accounts are no more common in Unknown than in Random. The behavioral response to favorable noise therefore emerges without a corresponding increase in articulated causal beliefs.

This divergence between behavior and articulated beliefs helps clarify how superstition may take shape. In the Unknown treatment, favorable adjustments following cooperation make cooperation more likely to be repeated, yet participants are no more likely than those in Random to report that cooperation caused favorable adjustments. The joint pattern is therefore more consistent with a repetition-first process: behavior responds to favorable coincidences without a corresponding increase in observable causal rationalization.[5]

Our findings contribute to three related literatures. First, they extend research on cooperation in repeated games by showing that strategically irrelevant payoff noise can increase cooperation even when actions and stage-game outcomes are perfectly observed and the noise conveys no information about players' behavior. This complements work on imperfect monitoring, where uncertainty is strategically relevant because it affects how players interpret and respond to one another's actions (Dal Bó and Fréchette, 2011; Fudenberg *et al.*, 2012; Cason and Mui, 2019; Aoyagi *et al.*, 2019). Second, the results contribute to the economics of superstition and religion by showing that repetition and causal can diverge. Favorable random outcomes shape subsequent cooperation, particularly when their source is left unexplained, without producing a corresponding increase in articulated causal beliefs. The joint pattern is more consistent with a repetition-first process in which behavior can respond to random reinforcement before any causal rationalization becomes observable. Third, the experiment contributes to research on strategy choice in repeated games by showing that opacity can reshape ex ante contingent plans even when it does not further increase average direct cooperation. Participants facing opaque noise move away from unconditional defection and toward more cooperative plans.[6]

[5]This interpretation does not imply a stationary reinforcement-learning process. Responses to the largest favorable adjustment fluctuate across the experiment rather than following a stable temporal pattern.

[6]The design builds on strategy-method approaches that elicit contingent plans in repeated games (Dal Bó and Fréchette, 2019; Romero and Rosokha, 2018, 2019; Cason and Mui, 2019). Appendix D uses our within-subject

The remainder of the paper proceeds as follows. Section 2 describes the experimental design and belief elicitation. Section 3 develops the conceptual framework and predictions. Section 4 presents the empirical approach and results on plan selection, direct cooperation, responses to realized adjustments, the persistence of cooperative streaks, and participants' articulated interpretations of the adjustment process. Section 5 concludes. Appendix D compares selected plans with direct play and examines the role of plan complexity.

2 Experimental Design

The experiment examines the role of noise in shaping cooperative behavior by embedding stochastic payoff adjustments into an indefinitely repeated prisoner's dilemma. Participants were assigned to one of three treatments: Baseline (no noise), Random (disclosed random noise), and Unknown (opaque noise). Baseline included no payoff adjustments. In Random, adjustments were explicitly described as random and unrelated to behavior. In Unknown, adjustments were generated by the same random mechanism as in Random, but participants received no information about how they were generated. Actions and stage-game outcomes remained perfectly observed in all treatments. At the start of each supergame, participants selected a plan before a virtual coin flip determined whether the plan was implemented automatically or participants chose actions round by round.

2.1 The Stage Game and Continuation Rule

Participants played a sequence of indefinitely repeated prisoner's dilemmas. In each round, the two matched participants simultaneously chose between actions A and B, the neutral labels used in the experimental instructions. Their joint action determined the stage-game payoffs shown in Table 1. We refer to A as cooperation and B as defection.

After each round, the supergame continued with probability 0.75, regardless of how many rounds had already been played (Roth and Murnighan, 1978; Murnighan and Roth, 1983). This continuation probability implies an expected supergame length of four rounds. The sequence of supergame

design to compare selected plans with round-by-round behavior and examine how plan complexity affects implementation consistency (Oprea, 2020; Guan and Oprea, Working paper).

Table 1: Payoff Matrix Used in All Treatments

	Other Participant's Choice: A	**Other Participant's Choice: B**
Your Choice: A	40¢, 40¢	12¢, 65¢
Your Choice: B	65¢, 12¢	25¢, 25¢

lengths was generated before the experiment using this continuation rule. The same sequence of 15 lengths was used in every session and treatment. Participants were randomly rematched with an anonymous partner at the beginning of each incentivized supergame.

For these payoffs and continuation probability, mutual cooperation can be sustained by the grim-trigger strategy. Grim trigger is also risk dominant relative to unconditional defection (Blonski and Spagnolo, 2015).

2.2 Decision Modes: Strategy Method and Direct Response

At the beginning of each supergame, participants selected one of 20 predefined plans adapted from Fudenberg *et al.* (2012). Each plan prescribed action A or B based on the actions observed in previous rounds. The menu included familiar strategies such as Tit-for-Tat, Grim Trigger, and Win-Stay-Lose-Shift, along with more lenient, forgiving, and exploitative variants. Table 2 lists the plans and the categories used in the analysis. Detailed descriptions of their decision rules appear in Table A1.

After all participants selected a plan, a virtual coin flip implemented by the computer determined the decision mode for that supergame. The same draw applied to all participants in the session. Under the *plan procedure*, the computer implemented each participant's selected plan automatically throughout the supergame. This procedure corresponds to the strategy method and captures participants' cold decisions. Under the *round-by-round procedure*, the selected plans were not implemented, and participants chose action A or B at the beginning of each round. These choices constitute direct, or hot, decisions.

Because the decision mode was determined only after participants selected their plans, a plan choice was elicited in every supergame regardless of whether it was implemented. The design

Table 2: **The Strategy List**. Following Fudenberg *et al.* (2012), all plans other than ALLD and D-TFT are classified as cooperative. The forgiving and lenient classifications also follow their definitions. "Always Defect" identifies ALLD, and "Starts with Cooperation" identifies plans 1–13. See Table A1 for detailed descriptions.

Plan	Name	Abbreviation	Category
1	Always Cooperate	ALLC	Cooperative, Starts with Cooperation
2	Tit-for-Tat	TFT	Cooperative, Forgiving, Starts with Cooperation
3	Tit-for-2-Tats	TF2T	Cooperative, Forgiving, Lenient, Starts with Cooperation
4	Tit-for-3-Tats	TF3T	Cooperative, Forgiving, Lenient, Starts with Cooperation
5	2-Tits-for-Tat	2TFT	Cooperative, Forgiving, Starts with Cooperation
6	2-Tits-for-2-Tats	2TF2T	Cooperative, Forgiving, Lenient, Starts with Cooperation
7	T2	T2	Cooperative, Starts with Cooperation
8	Win-Stay-Lose-Shift	PTFT	Cooperative, Starts with Cooperation
9	Win-Stay-Lose-Shift with Two-Round Punishment	2PTFT	Cooperative, Starts with Cooperation
10	Grim Trigger	Grim	Cooperative, Starts with Cooperation
11	Lenient Grim	Grim2	Cooperative, Lenient, Starts with Cooperation
12	More Lenient Grim	Grim3	Cooperative, Lenient, Starts with Cooperation
13	False Cooperator	C-to-ALLD	Cooperative, Starts with Cooperation
14	Always Defect	ALLD	Always Defect, Noncooperative
15	Exploitative Tit-for-Tat	D-TFT	Noncooperative
16	Exploitative Tit-for-2-Tats	D-TF2T	Cooperative
17	Exploitative Tit-for-3-Tats	D-TF3T	Cooperative
18	Exploitative Grim2	D-Grim2	Cooperative
19	Exploitative Grim3	D-Grim3	Cooperative
20	Alternator	DC-Alt	Cooperative

therefore allows us to compare the plans participants selected with the actions they chose during direct, round-by-round play.

2.3 Treatments

Sessions were randomly assigned to one of three between-subjects treatments, summarized in Table 3. The treatments varied whether participants received payoff adjustments and whether the source of those adjustments was explained.

- **Baseline:** Earnings were determined solely by the stage-game payoffs shown in Table 1. No additional adjustments were made.

- **Random:** After the stage-game payoffs were determined in each round, the computer independently drew an adjustment from {-10, -5, 0, +5, +10} cents, with each value occurring with equal probability. The same adjustment was applied to both members of a pair. Partic-

ipants were explicitly informed that the adjustments were randomly generated and unrelated to their decisions.

- **Unknown:** Adjustments were generated using the same random mechanism and distribution as in the Random treatment. Participants were told only that "the computer will choose this adjustment based on a complex sequence of calculations."[7] The instructions did not state that the adjustments were random or explain how they were generated.

In both noise treatments, adjustments were generated separately across pairs and rounds. They affected participants' final earnings but did not alter the actions selected, the stage-game outcomes, or the information participants received about one another's behavior.

Table 3: Treatment Overview

Treatment	Payoff Adjustments	Adjustment Process	Information Provided
Baseline	None	–	–
Random	$-10, -5, 0, +5, +10$	Random	Explicitly described as random
Unknown	$-10, -5, 0, +5, +10$	Random	Generation process left unexplained

Practice Phase

Each session began with an unpaid practice phase designed to ensure that participants understood the repeated game and the 20 available plans. Before making any practice decisions, participants completed a comprehension quiz. For each plan, they were shown a three-round history of play and asked which action the plan would prescribe in the fourth round. The plans were presented sequentially on separate screens. Participants received immediate feedback and could revise their responses until all questions were answered correctly.

Participants then completed three unpaid practice supergames against a computerized opponent. The length of each practice supergame was randomly determined in advance using the same

[7] As detailed in Appendix B, the adjustment algorithm is based on a multistage calculation and is parameterized so that aggregate cooperation can alter the relative probabilities of favorable and unfavorable adjustments. In each round, the program classified the session according to whether a strict majority cooperated, generated an independent random draw for each pair, and mapped that draw into one of five possible adjustments using parameters α and β. In the sessions reported in this paper, we set $\alpha = \beta = 0.5$, so each adjustment occurred with probability $1/5$ under either cooperation classification. Aggregate cooperation therefore entered the calculation without affecting the distribution of adjustments.

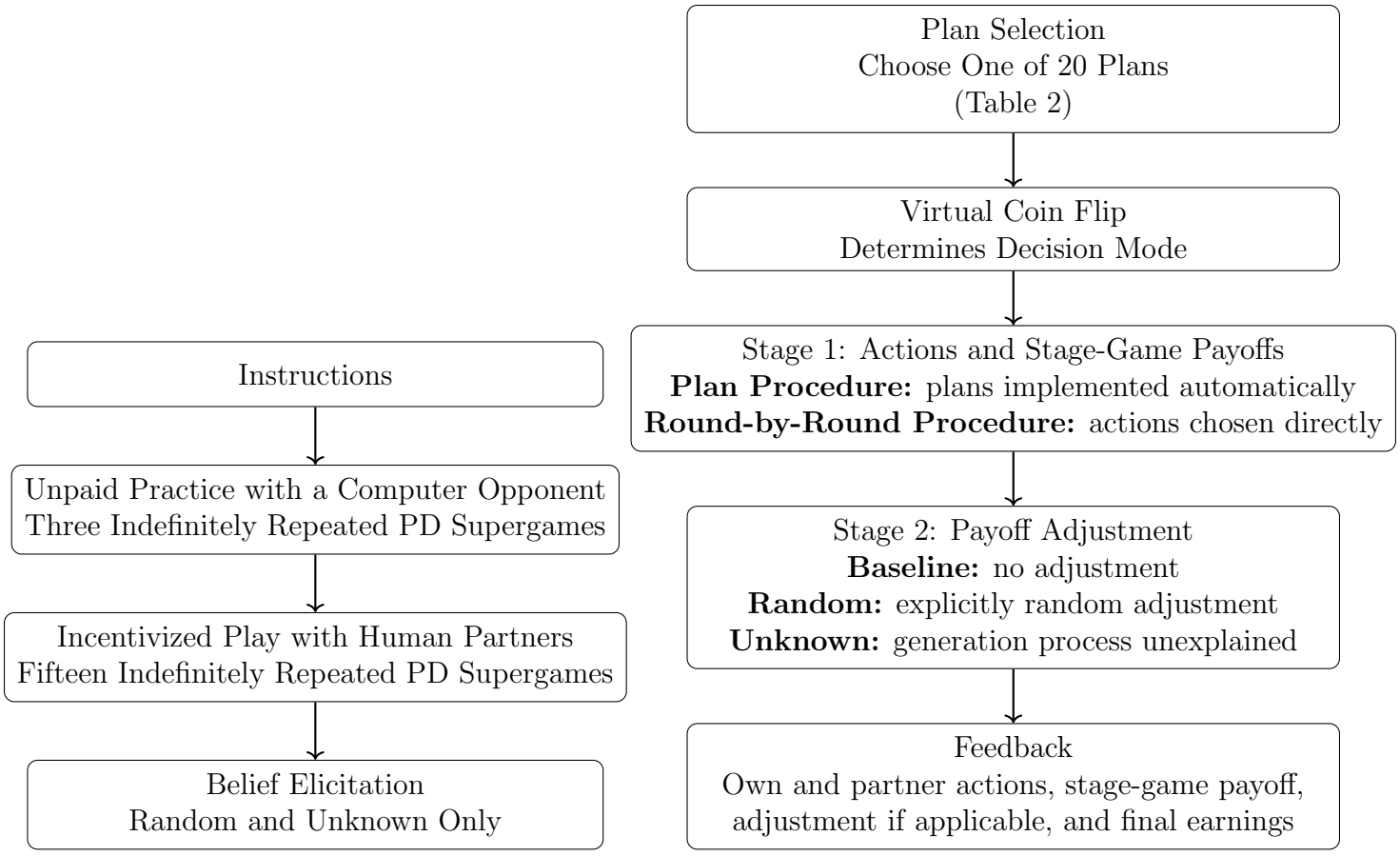
continuation rule as in the incentivized phase. The practice supergames mirrored the participant's assigned treatment. Participants in Baseline received no payoff adjustments, while participants in Random and Unknown received adjustments under the same treatment-specific procedures used in the incentivized phase. At the beginning of each practice supergame, the computer randomly selected one of the 20 plans and followed it throughout the supergame. The practice supergames also used the same two decision procedures. After selecting a plan, a computerized 50/50 draw determined whether the plan was implemented automatically or whether the participant chose actions directly, round by round. The practice phase therefore familiarized participants with the continuation rule, the plan menu, the decision procedures, and the payoff environment they would face during incentivized play.

Incentivized Phase and Session Procedures

Following the practice phase, participants completed 15 incentivized supergames with human partners. At the beginning of each supergame, participants were randomly rematched with an anonymous partner and selected one of the 20 available plans. A virtual coin flip implemented by the computer then determined the decision mode for all participants in the session. The supergame proceeded under either the plan procedure or the round-by-round procedure, as described in Section 2.2, and followed the continuation rule described in Section 2.1.

In all treatments, the first stage of each round consisted of action selection and the determination of stage-game payoffs according to Table 1. In the Random and Unknown treatments, a second stage followed in which the computer applied a payoff adjustment to both members of the pair. At the end of each round, participants observed their own action, their partner's action, and their stage-game payoff. Participants in the noise treatments also observed the adjustment and their final earnings for the round. Figure 1 summarizes the overall experimental procedure and the sequence of events within each incentivized supergame.

We conducted 12 laboratory sessions with 16 participants per session, for a total of 192 participants. Four sessions were assigned to each treatment. Participants were recruited from a mid-sized



(a) **Session Overview** (b) **Incentivized Supergame Structure**

Figure 1: **Experimental Procedure and Supergame Structure. Panel A** summarizes the sequence of activities within a session. **Panel B** summarizes the structure of each incentivized supergame. Participants first selected a plan, after which a computerized coin flip determined whether the plan procedure or the round-by-round procedure was used. In the Random and Unknown treatments, the stage-game payoff was followed by a payoff adjustment.

U.S. university, and the experiment was programmed and conducted using z-Tree (Fischbacher, 2007). At the beginning of each session, participants received printed instructions and watched a short video reviewing the instructions and procedures. An experimenter remained present throughout the session and answered clarification questions privately.

Sessions lasted approximately 65 minutes. Participants received a $7 show-up fee in addition to their cumulative earnings from the 15 incentivized supergames. Final payments ranged from $22 to $36 and averaged $28.

2.4 Belief Elicitation

Participants in the Random and Unknown treatments completed a belief-elicitation task after the 15 incentivized supergames. Participants in the Baseline treatment did not complete this task.

In the Unknown treatment, participants were asked: (1) "Did you notice any patterns in the computer's adjustments?", (2) "Did you notice any relationship between the contributions of you or your group and the computer's adjustments?", and (3) "Do you have any advice for future participants in this experiment?" The first two questions included a yes-or-no response followed by an open-ended explanation. The advice question was open-ended. Participants in the Random treatment received the second and third questions but not the general pattern question, since they had already been informed that the adjustments were randomly generated.

To encourage thoughtful responses, participants were told that their complete message could be selected and shown to a future group of participants. If selected, they would receive an additional payment equal to that group's average earnings.

The questions were open-ended so that participants could describe the adjustment process in their own terms rather than select from explanations supplied by the researchers. Eliciting these responses only after incentivized play also avoided directing participants' attention toward possible patterns or causal relationships during the experiment. The responses therefore capture participants' articulated interpretations after play. Appendix C.2 describes the coding of the responses.

3 Conceptual Framework and Predictions

Random reinforcement can give rise to superstition in two observable forms. The simplest is repetition. As in Skinner's classic experiment, an action that happens to be followed by a favorable but unrelated outcome may become more likely to recur even though the action did not cause the outcome (Skinner, 1948). We refer to this pattern as *Skinnerian repetition*. Superstition can also involve causal interpretation. A coincidence between an action and an outcome may be given meaning, so that the individual comes to believe that the action influenced the outcome. We refer to this

articulated causal account as *Malinowskian rationalization* (Malinowski, 1918). Skinnerian repetition and Malinowskian rationalization describe different manifestations of superstition. Neither, however, determines how the two become connected.

There are two natural possibilities. Repetition may come first. A favorable coincidence makes the preceding action more likely to recur, and repeated coincidences may later acquire a causal explanation. Rationalization is possible along this route, but it is not necessary for repetition to begin. Alternatively, causal interpretation may come first. The favorable coincidence is understood as evidence that the action affects the outcome, and that belief then supports repetition. Figure 2 summarizes these two possibilities.

The treatment design gives us leverage on this distinction. Baseline contains no payoff adjustments. Random and Unknown contain the same adjustment process, but participants in Random are explicitly told that the adjustments are random and unrelated to their decisions, while participants in Unknown are not told how the adjustments are generated. The Baseline–Random comparison therefore asks whether strategically irrelevant noise can affect cooperation even when its randomness is known. The Random–Unknown comparison asks what changes when the source of the same noise is left unexplained.

Behavior is observed throughout the experiment, while participants' interpretations of the adjustments are elicited after play. We therefore do not observe the precise moment at which a causal belief may arise. The design instead allows us to assess whether the behavioral response to random reinforcement is accompanied by, and appears to require, a corresponding causal interpretation. If favorable adjustments affect repetition without a corresponding increase in articulated causal beliefs, then behavior can respond before any observable articulated rationalization appears. If the stronger behavioral response under unexplained noise is accompanied by greater causal attribution, the evidence is instead consistent with an interpretation-first process.

This framework yields three predictions.

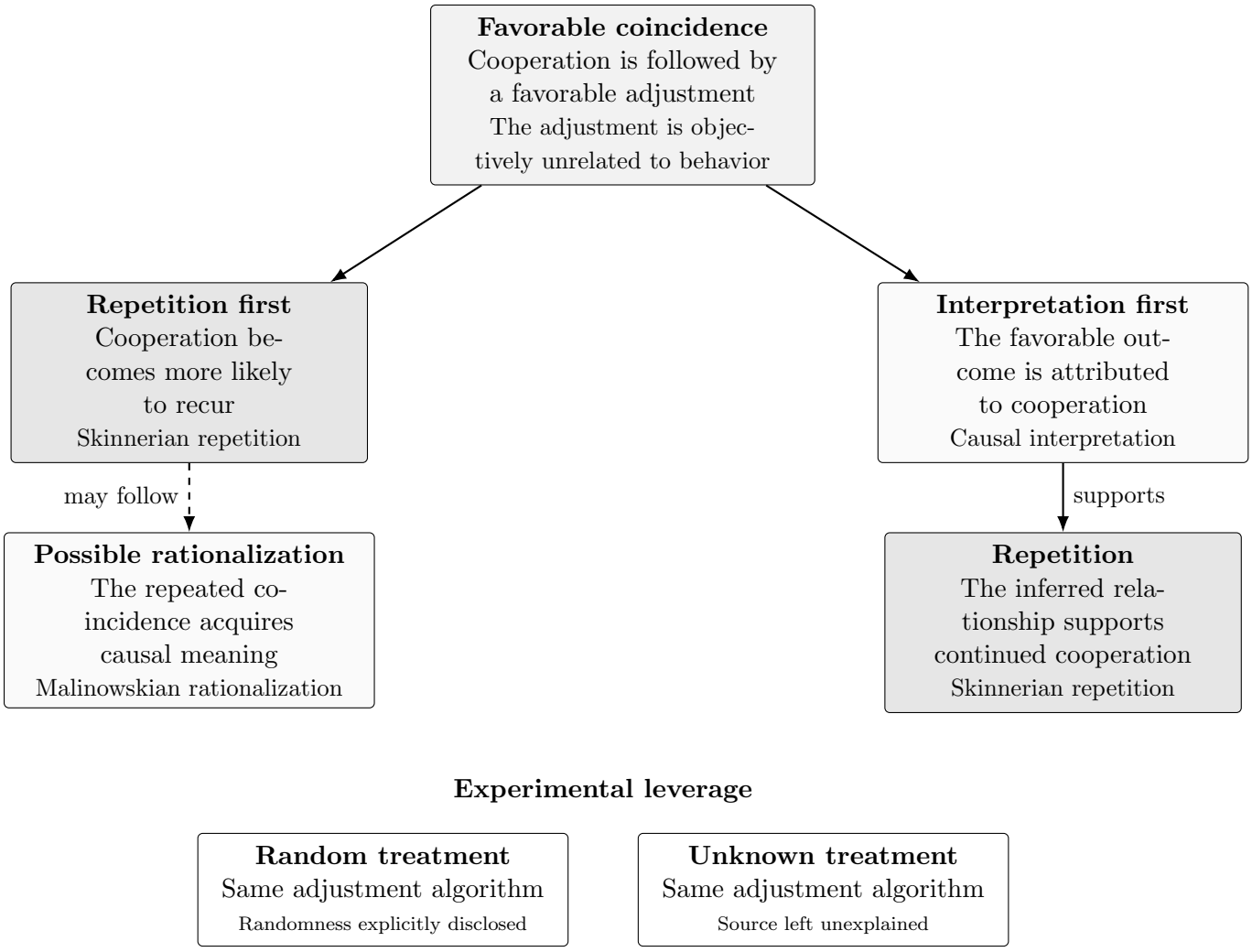


Figure 2: **Two Possible Routes from Favorable Coincidence to Superstition.** A favorable adjustment that happens to follow cooperation may first make cooperation more likely to be repeated, with causal rationalization developing later or not at all. Alternatively, the coincidence may first be interpreted causally, with that interpretation then supporting repetition. Skinnerian repetition and Malinowskian rationalization identify the observed behavioral and belief manifestations of these processes. The Random and Unknown treatments use the same adjustment algorithm but differ in whether its randomness is disclosed.

1. **Noise and cooperation.** Building on our previous findings, we expected environmental uncertainty generated by strategically irrelevant payoff shocks to increase cooperation. Cooperation should therefore be higher in both the Random and Unknown treatments than in the no-noise Baseline (Hajikhameneh and Iannaccone, 2023, 2025).

2. **Repetition following favorable noise.** A favorable adjustment following cooperation should increase the likelihood that cooperation is repeated in the next round. Because the source of the adjustment is left unexplained in Unknown, Skinnerian repetition should be stronger in Unknown than in Random.

3. **Causal rationalization.** If unexplained favorable coincidences lead participants to assign causal meaning to the adjustment process, participants in Unknown should be more likely than those in Random to report that cooperation or group behavior influenced the adjustments. Such responses constitute Malinowskian rationalization.

Predictions 2 and 3 are most informative when considered together. If stronger repetition following favorable adjustments in Unknown is accompanied by greater causal attribution, the evidence is consistent with an interpretation-first process in which causal beliefs support subsequent repetition. If repetition is stronger under unexplained noise without a corresponding increase in articulated causal beliefs, the evidence instead favors a repetition-first process in which random reinforcement can shape behavior without requiring causal rationalization. The question is therefore not whether Skinnerian repetition or Malinowskian rationalization is the "correct" form of superstition. It is whether repetition appears to require causal interpretation, or whether it can arise before any such interpretation becomes observable.

4 Results

4.1 Empirical Approach

We organize the results in the order in which decisions and outcomes unfolded. We begin with the plans participants selected at the start of each supergame, before the decision procedure was determined. We then examine choices made under the round-by-round procedure, first comparing average cooperation across treatments and then asking how realized adjustments shaped subsequent cooperation. Within the latter analysis, we test for Skinnerian repetition by asking whether cooperation that is followed by a favorable adjustment becomes more likely to be repeated. We

conclude by examining whether participants articulated causal explanations linking their behavior to the adjustments. The combination of adjustment-contingent behavior and post-play interpretations allows us to assess whether repetition under random reinforcement is accompanied by, and appears to require, a corresponding causal interpretation.

The comparison between Baseline and Random isolates the effect of payoff noise when its randomness is disclosed. The comparison between Random and Unknown isolates the effect of leaving the source of the same random adjustment process unexplained. The comparison between Baseline and Unknown captures the combined effect of payoff noise and opacity about its source.

All analyses exclude the three unpaid practice supergames and use observations from the 15 incentivized supergames. Participants selected one plan at the beginning of every supergame, regardless of whether the plan or round-by-round procedure was subsequently implemented. Direct decisions are observed only in supergames assigned to the round-by-round procedure. Analyses of lagged adjustments are restricted to Random and Unknown and exclude the first round of each supergame. The cooperative-streak analysis further restricts the sample to observations following cooperative histories of one or more consecutive actions.

For the binary choice outcomes, we estimate mixed-effects linear probability models with supergame fixed effects and random intercepts for sessions, participants, and interactions. Because treatments were assigned at the session level, standard errors are clustered by session.[8]

4.2 Plan Selection: How Noise Shapes Strategy Choice

We begin with the plans participants selected before learning which decision procedure would govern the supergame. Because participants selected a plan in every incentivized supergame, the analysis captures how they intended to approach the repeated interaction.

[8]We also conduct null-imposed score-bootstrap-t inference using 9,999 Webb-weight replications, with session-level clustering in both the original and bootstrap samples. The main results are robust. In particular, Unknown continues to exhibit significantly higher direct cooperation and less frequent selection of Always Defect than Baseline.

Following Fudenberg *et al.* (2012), we examine five overlapping plan classifications. Always Defect identifies ALLD. Cooperative includes all plans other than ALLD and D-TFT. Forgiving and Lenient identify overlapping subsets of cooperative plans, while Starts with Cooperation includes plans 1–13.

Table 4 reports plan selection in the first incentivized supergame. Participants made these choices after treatment-specific practice but before observing the action of a human partner in incentivized play. Always Defect was selected by 28.1 percent of participants in Unknown, compared with 31.3 percent in Random and 32.8 percent in Baseline. Cooperative plans were selected by 65.6 percent of participants in Unknown and 60.9 percent in each of the other treatments. The largest descriptive difference appears among Forgiving plans, which were selected by 31.3 percent of participants in Unknown, 15.6 percent in Random, and 9.4 percent in Baseline.

Table 4: **Plan Selection in the First Incentivized Supergame**

	Baseline			**Random**			**Unknown**		
Plan Category	Mean	SE	N	Mean	SE	N	Mean	SE	N
Always Defect	0.328	0.059	64	0.313	0.058	64	0.281	0.057	64
Cooperative	0.609	0.061	64	0.609	0.061	64	0.656	0.059	64
Forgiving	0.094	0.037	64	0.156	0.046	64	0.313	0.058	64
Lenient	0.109	0.039	64	0.141	0.044	64	0.156	0.046	64
Starts with Cooperation	0.469	0.063	64	0.531	0.063	64	0.547	0.063	64

Notes: The table reports the proportion of participants selecting each plan category in the first incentivized supergame. Plan classifications follow Fudenberg *et al.* (2012) and Cason and Mui (2019). The categories overlap.

Table 5 extends the analysis to all 2,880 plans selected during the incentivized supergames. Relative to Unknown, participants in Baseline were 9.3 percentage points more likely to select Always Defect and 14.1 percentage points less likely to select a Cooperative plan. Participants in Random were 7.1 percentage points more likely to select Always Defect and 6.6 percentage points less likely to select a Cooperative plan. Baseline participants were also 4.2 percentage points less likely to select a Lenient plan and 12.9 percentage points less likely to select a plan that Starts

with Cooperation. The estimated differences for Forgiving plans are negative in both Baseline and Random but are not statistically distinguishable from zero.

Across the incentivized supergames, the clearest shift in strategy choice occurs under opaque noise. Participants in Unknown move away from unconditional defection and toward more cooperative plans.[9]

	(1) Always Defect	(2) Cooperative	(3) Forgiving	(4) Lenient	(5) Starts with C
Baseline	0.093***	-0.141***	-0.083	-0.042**	-0.129***
	(0.035)	(0.048)	(0.055)	(0.017)	(0.044)
Random	0.071**	-0.066*	-0.079	-0.008	-0.053
	(0.031)	(0.035)	(0.050)	(0.022)	(0.035)
Intercept	0.253***	0.723***	0.242***	0.152***	0.576***
	(0.034)	(0.036)	(0.043)	(0.021)	(0.028)
Supergame Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	2,880	2,880	2,880	2,880	2,880

Standard errors clustered at the session level are reported in parentheses.
Random intercepts are included for sessions, participants, and interactions.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Unknown is the omitted treatment.

Table 5: **Treatment Effects on Plan Selection: All Incentivized Supergames**

4.3 Direct Decisions: Treatment Effects on Cooperation

We next turn from selected plans to the choices participants made under the round-by-round procedure. Using all direct decisions from the incentivized supergames, we estimate a mixed-effects linear probability model in which the dependent variable equals one when a participant cooperates.

The estimated cooperation rate is 22.5 percent in Baseline, compared with 30.3 percent in Random and 30.8 percent in Unknown. Cooperation is 7.8 percentage points higher in Random than in Baseline ($SE = 0.042$, $p = 0.063$) and 8.3 percentage points higher in Unknown than in

[9] Appendix D examines how selected plans translated into direct play. Direct decisions matched the selected plan in 79.4 percent of cases. Implementation declined with plan complexity in all three treatments, but the complexity gradient did not differ significantly across treatments. These results provide little evidence that payoff noise broadly impaired participants' ability to carry out their plans.

Baseline ($SE = 0.031$, $p = 0.008$). The difference between Unknown and Random is 0.5 percentage points ($SE = 0.039$, $p = 0.892$).

Strategically irrelevant payoff noise raises direct cooperation by approximately eight percentage points relative to the no-noise Baseline. Leaving the adjustment process unexplained does not further increase average cooperation.

4.4 Noise and Skinnerian Repetition

Average treatment differences do not reveal how participants respond to the adjustments they receive. Prediction 2 concerns Skinnerian repetition: whether cooperation that happens to be followed by a favorable but unrelated adjustment becomes more likely to recur. We examine this pattern in two steps. First, we estimate how the previous round's adjustment affects subsequent cooperation across all eligible rounds. We then condition on cooperative histories to test whether favorable adjustments make cooperation more likely to continue once it has begun.

4.4.1 Lagged Adjustments and Subsequent Cooperation

The analysis is restricted to Random and Unknown, the two treatments in which adjustments were present. Column (1) compares cooperation across the treatments using all direct decisions. Columns (2) and (3) exclude the first round of each supergame and control for the partner's previous action and the adjustment received in the previous round. Column (3) interacts each adjustment with Random, allowing its effect on subsequent cooperation to differ depending on whether participants knew that the adjustment was random. Unknown with a lagged adjustment of -10 is the omitted category.

Column (1) confirms the close average cooperation rates in Random and Unknown. Once previous-round information is included, the partner's previous action is the strongest predictor of subsequent cooperation. A partner's defection reduces the probability of cooperating in the next round by approximately 25 percentage points.

Table 6: **Lagged Adjustments and Subsequent Cooperation**

	(1) Cooperation All Rounds	(2) Cooperation Rounds > 1	(3) Cooperation Rounds > 1
Random Treatment	-0.003	-0.014	0.019
	(0.038)	(0.024)	(0.046)
Lagged Other Defect		-0.252***	-0.251***
		(0.026)	(0.026)
Lagged Adjustment = −5		-0.013	-0.010
		(0.023)	(0.038)
Lagged Adjustment = 0		0.015	0.051**
		(0.020)	(0.022)
Lagged Adjustment = 5		-0.007	0.006
		(0.025)	(0.045)
Lagged Adjustment = 10		0.017	0.052*
		(0.023)	(0.027)
Random × Adj. = −5			-0.005
			(0.046)
Random × Adj. = 0			-0.068**
			(0.030)
Random × Adj. = 5			-0.025
			(0.052)
Random × Adj. = 10			-0.070*
			(0.038)
Intercept	0.486***	0.604***	0.590***
	(0.035)	(0.035)	(0.036)
Supergame Fixed Effects	Yes	Yes	Yes
Observations	3,824	2,912	2,912

Standard errors clustered at the session level are reported in parentheses.
Random intercepts are included for sessions, participants, and supergames.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Unknown with a lagged adjustment of −10 is the omitted category.

Column (3) shows that adjustments affect subsequent cooperation differently across the two treatments. In Unknown, a zero adjustment raises cooperation in the next round by 5.1 percentage points relative to a −10 adjustment ($p < 0.05$), while a +10 adjustment raises it by 5.2 percentage points ($p < 0.10$). The corresponding estimates in Random are approximately −1.7 and −1.8

percentage points. The interaction terms show that the effects of zero and +10 adjustments are 6.8 and 7.0 percentage points lower in Random, respectively.[10]

This analysis includes observations following both cooperation and defection. It therefore captures the overall effect of lagged adjustments on subsequent cooperation, but does not isolate whether a favorable adjustment makes cooperation itself more likely to be repeated. We turn to that question next.

4.4.2 Direct Decisions: Noise and Cooperative Streaks

We isolate Skinnerian repetition by conditioning on participants' own cooperative histories. The dependent variable equals one when the participant cooperates in the current round. Columns (1), (2), and (3) restrict the sample to observations following one, two, and three consecutive cooperative actions, respectively. Each specification uses a lagged adjustment of −10 in Unknown as the omitted category.

Following one cooperative action, a +10 adjustment raises the probability of cooperating again by 10.3 percentage points in Unknown relative to a −10 adjustment. The corresponding estimate in Random is 2.8 percentage points, although the difference between the treatments is not statistically significant in this specification.

The contrast becomes sharper after two consecutive cooperative actions. In Unknown, +5 and +10 adjustments raise the probability of continuing cooperation by 11.3 and 11.4 percentage points, respectively. The interaction between Random and the +10 adjustment is −11.8 percentage points and statistically significant. It almost fully offsets the effect observed in Unknown, leaving a combined estimate of approximately zero in Random.[11] After three consecutive cooperative actions,

[10] Dividing the 15 incentivized supergames into five three-supergame intervals reveals no systematic time pattern. Treatment differences in cooperation and responses to lagged adjustments fluctuate across intervals rather than consistently strengthening or fading.

[11] A purely mechanical income effect cannot explain this treatment difference, nor can a mood or valence response that operates identically across the two noise treatments. The treatments used the same adjustment values, probabilities, timing, and pair-level application. Yet, after two consecutive cooperative actions, a +10 adjustment increased the probability of continued cooperation by 11.4 percentage points in the Unknown treatment, while the

Table 7: **Effect of Lagged Adjustments on the Continuation of Cooperative Streaks**

	(1) One Prior Cooperative Action	(2) Two Prior Cooperative Actions	(3) Three Prior Cooperative Actions
Random Treatment	0.069	0.068	0.100**
	(0.052)	(0.051)	(0.047)
Lagged Other Defect	-0.541***	-0.601***	-0.503***
	(0.046)	(0.037)	(0.075)
Lagged Adjustment = −5	0.027	0.028	0.005
	(0.070)	(0.081)	(0.104)
Lagged Adjustment = 0	0.044	0.003	0.022
	(0.069)	(0.077)	(0.091)
Lagged Adjustment = 5	0.092	0.113***	0.041
	(0.071)	(0.031)	(0.075)
Lagged Adjustment = 10	0.103**	0.114***	0.101
	(0.045)	(0.025)	(0.104)
Random × Adj. = −5	-0.042	-0.109	-0.081
	(0.077)	(0.111)	(0.117)
Random × Adj. = 0	-0.048	0.026	-0.017
	(0.078)	(0.093)	(0.103)
Random × Adj. = 5	-0.007	-0.057	-0.042
	(0.071)	(0.036)	(0.077)
Random × Adj. = 10	-0.075	-0.118***	-0.159
	(0.054)	(0.027)	(0.114)
Intercept	0.742***	0.714***	0.718***
	(0.089)	(0.062)	(0.113)
Supergame Fixed Effects	Yes	Yes	Yes
Observations	937	439	240

The dependent variable equals one if the participant cooperates in the current round.
Columns (1)–(3) condition on one, two, and three prior consecutive cooperative actions.
Standard errors clustered at the session level are reported in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Unknown treatment and a lagged adjustment of −10 are the omitted categories.

the estimated effect of a +10 adjustment in Unknown remains close to 10 percentage points, but the smaller sample produces a much less precise estimate.

Favorable adjustments under opaque noise therefore make an ongoing cooperative pattern more likely to persist rather than merely shifting average behavior. The clearest contrast appears after

Random × (+10) interaction was −11.8 percentage points, leaving an implied effect near zero when participants were told that the adjustment was randomly generated. Thus, the same favorable adjustment, drawn from the same distribution, had different behavioral consequences depending on whether its randomness was disclosed.

two consecutive cooperative actions, when the largest favorable adjustment increases the likelihood of continued cooperation in Unknown but has essentially no effect in Random.[12] This pattern corresponds to the Skinnerian repetition described in Section 3.

4.5 Noise and Malinowskian Rationalization

The preceding analysis documents Skinnerian repetition, with favorable adjustments following cooperation making cooperation more likely to persist, particularly when the source of the adjustments is unexplained. We now ask whether this behavioral response is accompanied by a corresponding causal interpretation. Malinowskian rationalization arises when participants form and articulate a causal explanation linking their behavior to an outcome that was, in fact, randomly generated. We therefore examine whether leaving the adjustment process unexplained made participants more likely to report that cooperation influenced the adjustments.

After the incentivized supergames, participants in Random and Unknown were asked whether they perceived a relationship between the adjustments and the actions taken by their pair. Participants answered yes or no and then explained their reasoning in an open-ended response.[13] Because the same relationship question was administered in both treatments, it provides the primary test of whether opacity increased Malinowskian rationalization. Participants in Unknown were also asked a broader question about any pattern they had noticed in the adjustments. These responses reveal the explanations that arose when participants were invited to identify a pattern without being directed toward a relationship with behavior.

We developed the coding categories, definitions, and decision rules based on the study's conceptual framework. A generative AI model then applied the resulting codebook to each anonymized open-ended response. The model was not provided with the paper's predictions or with any in-

[12] Parallel specifications conditioning on prior defection do not show the same pattern in reverse. A (-10) adjustment does not systematically reduce the likelihood that defection continues, nor do the estimates reveal a consistent difference between Unknown and Random.

[13] The belief elicitation was incentivized. Participants were told that, if their complete response block was selected and shown to a future group of participants, they would receive a payment equal to that group's average earnings.

formation about the expected relationship between treatment and the response categories. This division of labor leaves the definition of the substantive constructs with the researchers while using AI to apply the coding rules systematically and consistently. Large language models are particularly useful for analyzing and classifying open-ended survey responses, which can otherwise require substantial time and resources (Korinek, 2023; Chang *et al.*, 2026). Appendix C reports the complete codebooks, classification procedure, full response distributions, and exploratory analyses relating post-experiment beliefs to earlier behavior.

For the relationship question, responses were assigned to five mutually exclusive categories. Positive Relationship identifies explanations in which greater cooperation or more cooperative group behavior produced more favorable adjustments. Negative Relationship captures explanations in which cooperation was associated with less favorable adjustments. Random or No Relationship includes responses explicitly describing the adjustments as random or unrelated to behavior. Unclear identifies responses that did not provide an interpretable account. Other or Complex includes explanations proposing a different pattern or a more elaborate mechanism that did not imply a clear positive or negative relationship between cooperation and the adjustments.

Positive causal accounts were uncommon in both treatments. Four of 57 respondents in Random, or 7.0 percent, described a positive relationship between cooperative behavior and favorable adjustments. Four of 62 respondents in Unknown, or 6.5 percent, did so. The difference is effectively zero and is not statistically significant. One participant in Unknown described a negative relationship, while no participant in Random did so. Opacity therefore did not increase the frequency of explicit prosocial causal attributions.

Most responses fell outside a simple positive or negative account. In Random, 33.3 percent described the adjustments as random or unrelated to behavior, compared with 29.0 percent in Unknown. Other or complex explanations accounted for 50.9 percent of responses in Random and 43.5 percent in Unknown. The clearest directional difference appears in the Unclear category, which rises from 8.8 percent in Random to 19.4 percent in Unknown. Rather than indicating the emergence

Table 8: **AI-Assisted Content Analysis of Beliefs About the Adjustments**

	Random	Unknown
Panel A: Perceived Relationship Between Actions and Adjustments		
Positive Relationship	4 (7.0%)	4 (6.5%)
Negative Relationship	0 (0.0%)	1 (1.6%)
Random or No Relationship	19 (33.3%)	18 (29.0%)
Unclear	5 (8.8%)	12 (19.4%)
Other or Complex	29 (50.9%)	27 (43.5%)
Total	57 (100%)	62 (100%)
Panel B: General Pattern Reported in Unknown		
Prosocial Causal Explanation	–	3 (4.8%)
Random	–	20 (32.3%)
Uncertain	–	7 (11.3%)
Other or Complex	–	32 (51.6%)
Total	–	62 (100%)

Notes: Entries report the number of responses, with percentages among nonmissing responses in parentheses. Panel A uses the relationship question administered in both Random and Unknown. Panel B uses the general pattern question administered only in Unknown. The researchers developed the categories and coding rules, and a generative AI model applied the resulting codebooks to the anonymized responses. The model was not provided with the paper's predictions.

of a shared causal explanation, this pattern is more consistent with greater difficulty interpreting the opaque adjustment process.

The broader pattern question administered only in Unknown points in the same direction. Three of the 62 respondents, or 4.8 percent, articulated a prosocial causal explanation. Twenty described the adjustments as random, seven remained uncertain, and 32 proposed another or more complex pattern. Even when participants were invited to search for structure, few concluded that cooperative behavior produced favorable adjustments.

The advice responses reported in Appendix C contain a somewhat different directional signal. Trust or cooperation was recommended by ten of 61 respondents in Unknown, or 16.4 percent, compared with four of 58 respondents in Random, or 6.9 percent. Self-focused advice was nearly identical across treatments, at 26.2 percent in Unknown and 25.9 percent in Random. The greater frequency of cooperative recommendations in Unknown leaves open the possibility that opacity

shaped how participants advised others even when it did not generate an explicit belief that cooperation caused favorable adjustments.

Overall, Malinowskian rationalization remained rare and was no more prevalent when the adjustment process was opaque. This contrasts with the behavioral evidence. Favorable adjustments helped sustain cooperative streaks in Unknown, yet participants were no more likely than those in Random to report that cooperation produced favorable adjustments. The joint pattern is more consistent with a repetition-first process than with an interpretation-first process in which the behavioral response requires an articulated causal belief. Because interpretations were elicited after play, these results do not rule out causal beliefs that arose during play but were not subsequently articulated.

5 Conclusion

We study whether payoff noise that changes earnings but not the strategic incentives of the game affects cooperation in an indefinitely repeated prisoner's dilemma, and whether its influence depends on what participants know about its source. In Baseline, participants receive no payoff adjustments. In Random, adjustments are explicitly described as random and unrelated to behavior. In Unknown, the same random process generates the adjustments, but participants are not told how they are determined. Actions and stage-game outcomes are perfectly observed throughout. The comparison between Baseline and Random therefore isolates the effect of payoff noise when its randomness is disclosed, while the comparison between Random and Unknown isolates the effect of leaving the source of otherwise identical noise unexplained.

Payoff noise increases cooperation. Direct cooperation is approximately eight percentage points higher in both Random and Unknown than in Baseline. The nearly identical cooperation rates in the two noise treatments show that opacity does not produce an additional increase in average direct cooperation. It does, however, change how participants plan to approach the repeated interaction. In Unknown, participants move away from unconditional defection and toward more

cooperative plans. The plan-to-play analysis provides little evidence that these differences reflect greater difficulty implementing selected plans under payoff noise.

Average cooperation rates do not capture how participants respond to the adjustments they experience. In Unknown, favorable adjustments following cooperation make cooperation more likely to continue, particularly within cooperative streaks. This response is substantially weaker in Random, where participants know that the adjustments are random and unrelated to their behavior. The pattern corresponds to Skinnerian repetition, in which an action that happens to precede a favorable but unrelated outcome becomes more likely to recur.

The belief responses provide a different pattern. Few participants report that cooperation produced favorable adjustments, and such explanations are no more common in Unknown than in Random. Malinowskian rationalization therefore remains rare even when the adjustment process is left unexplained. The combination of stronger repetition under opaque noise and no corresponding increase in articulated causal beliefs is more consistent with a repetition-first process than with an interpretation-first process in which the behavioral response requires an articulated causal belief. Because participants' interpretations were elicited after play, we cannot rule out causal beliefs that arose during the experiment but were not subsequently articulated.

Viewed alongside our related experiments, these findings suggest a broader progression in the development of superstition, summarized in Appendix E and Figure E1. Environmental uncertainty creates conditions in which favorable coincidences can acquire behavioral significance. Random reinforcement can first make an action more likely to recur, and repeated behavior may subsequently acquire a causal explanation. Once formed, private rationalizations can spread through communication and cultural transmission, becoming shared beliefs that support coordinated norms and practices. Our earlier experiments further suggest that such shared beliefs can interact with costly commitment, sacrifice, and endogenous group formation in ways that reinforce cooperation (Hajikhameneh and Iannaccone, 2023, 2025; Hajikhameneh, 2026). Not every superstition will follow this progression, and socially reinforced beliefs may be functional or counterproductive. The broader

pattern is nevertheless consistent with superstition developing from behavior outward, from an individual's response to a favorable coincidence to causal rationalization, shared belief, and eventually institutional reinforcement.

These findings show that environmental uncertainty need not undermine cooperation. Payoff shocks that convey no information about the partner or the strategic environment can increase cooperation, while uncertainty about their source can affect which strategies people choose and how they respond to favorable coincidences. More broadly, the results point to a behavioral route through which superstition may emerge in social interaction. Actions that happen to precede favorable outcomes can become more likely to be repeated before any corresponding causal rationalization becomes observable.

References

Aeschleman, S. R., Rosen, C. C. and Williams, M. R. (2003). The effect of non-contingent negative and positive reinforcement operations on the acquisition of superstitious behaviors. *Behavioural Processes*, **61** (1-2), 37–45.

Alloy, L. B. and Abramson, L. Y. (1979). Judgment of contingency in depressed and non-depressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, **108** (4), 441–485.

Aoyagi, M., Bhaskar, V. and Fréchette, G. R. (2019). The impact of monitoring in infinitely repeated games: Perfect, public, and private. *American Economic Journal: Microeconomics*, **11** (1), 1–43.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

— and Hamilton, W. D. (1981). The evolution of cooperation. *science*, **211** (4489), 1390–1396.

Blanco, F., Matute, H. and Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior*, **41** (4), 333–340.

Blonski, M. and Spagnolo, G. (2015). Prisoners' other dilemma. *International Journal of Game Theory*, **44**, 61–81.

Boyd, R. and Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of theoretical Biology*, **132** (3), 337–356.

Cason, T. N. and Mui, V.-L. (2019). Individual versus group choices of repeated game strategies: A strategy method approach. *Games and Economic Behavior*, **114**, 128–145.

Catania, A. C. and Cutts, D. (1963). Experimental control of superstitious responding in humans. *Journal of the Experimental Analysis of Behavior*, **6** (2), 203–208.

Chang, S., Kennedy, A., Leonard, A. and List, J. A. (2026). 12 best practices for leveraging generative ai in experimental research. *Journal of Economic Behavior & Organization*, **246**, 107564.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, **43**, e28.

Dal Bó, P. and Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, **101** (1), 411–429.

— and Fréchette, G. R. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, **56** (1), 60–114.

— and Fréchette, G. R. (2019). Strategy choice in the infinitely repeated prisoner's dilemma. *American Economic Review*, **109** (11), 3929–3952.

Dreber, A., Rand, D. G., Fudenberg, D. and Nowak, M. A. (2008). Winners don't punish. *Nature*, **452** (7185), 348–351.

Espín-Sánchez, J.-A., Gil-Guirado, S. and Ryan, N. (2026). Praying for rain. *The Quarterly Journal of Economics*, **141** (3), 2363–2422.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, **10** (2), 171–178.

Foster, K. R. and Kokko, H. (2009). The evolution of superstitious and superstition-like behaviour. *Proceedings of the Royal Society B: Biological Sciences*, **276** (1654), 31–37.

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, **54** (3), 533–554.

—, Rand, D. G. and Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, **102** (2), 720–749.

Guan, M. and Oprea, R. (Working paper). Three faces of complexity in strategic choice.

Hajikhameneh, A. (2026). When superstition outperforms truth: Belief transmission, sacrifice, and cooperation under environmental uncertainty. *Evolution and Human Behavior*, **47** (3), 106831.

— and Iannaccone, L. R. (2023). God games: An experimental study of uncertainty, superstition, and cooperation. *Games and Economic Behavior*, **139**, 88–116.

— and — (2025). From shocks to solidarity and superstition: Exploring the foundations of faith. *Journal of Economic Behavior & Organization*, **229**, 106775.

Hamerman, E. J. and Johar, G. V. (2013). Conditioned superstition: Desire for control and consumer brand preferences. *Journal of Consumer Research*, **40** (3), 428–443.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, **61** (4), 1281–1317.

Malinowski, B. (1918). 53. fishing in the trobriand islands. *Man*, **18**, 87–92.

Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement in humans. *Learning and Motivation*, **25** (2), 216–232.

—, Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A. and Barberia, I. (2015). Illusions of causality: How they bias our everyday thinking and how they could be reduced. *Frontiers in psychology*, **6**.

—, Yarritu, I. and Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*, **102** (3), 392–405.

Murnighan, J. K. and Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games: A test of several models. *Journal of conflict resolution*, **27** (2), 279–300.

Nowak, M. and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, **364** (6432), 56–58.

Ono, K. (1987). Superstitious behavior in humans. *Journal of the experimental analysis of behavior*, **47** (3), 261–271.

Oprea, R. (2020). What makes a rule complex? *American economic review*, **110** (12), 3913–3951.

Rand, D. G., Fudenberg, D. and Dreber, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization*, **116**, 481–499.

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological review*, **123** (2), 182–207.

Rojas, C. (2012). The role of demand information and monitoring in tacit collusion. *The RAND Journal of Economics*, **43** (1), 78–109.

Romero, J. and Rosokha, Y. (2018). Constructing strategies in the indefinitely repeated prisoner's dilemma game. *European Economic Review*, **104**, 185–219.

— and — (2019). The evolution of cooperation: The role of costly strategy adjustments. *American Economic Journal: Microeconomics*, **11**, 299–328.

Roth, A. E. and Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical psychology*, **17** (2), 189–198.

Sinding Bentzen, J. (2019). Acts of god? religiosity and natural disasters across subnational world districts. *The Economic Journal*, **129** (622), 2295–2321.

Skinner, B. F. (1948). " superstition" in the pigeon. *Journal of Experimental Psychology*, **38**, 168–172.

Smith, J. M. (1982). Evolution and the theory of games. In *Did Darwin get it right? Essays on games, sex and evolution*, Springer, pp. 202–215.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, **46** (1), 35–57.

Appendices

A Experiment Instructions

This is an experiment in the economics of strategic decision making. If you follow the instructions and make appropriate decisions, you can earn an appreciable amount of money.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand but do not say anything. An experimenter will come to you to answer your question or provide assistance in private. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your cooperation.

The experiment is divided into three Phases.

Phase 1: Unpaid Practice Interactions with the Computer

Phase 1 consists of a series of practice *interactions* between you and the computer playing the role of another participant. Each interaction includes a random number of *rounds*. At the start of each round, you and the computer separately choose an *action*, either A or B. The combined actions of you and the computer then determine your payoff for the round. When the rounds of one interaction end, you will move on to another interaction, which also includes a random number of rounds. Do *not* assume that computer will behave the same way in every interaction.

The payoffs:

The following table lists the payoffs associated with different combinations of actions. The first number in each cell is *your* payoff for the round, and the second number is the computer's payoff. So, for example, if in a particular round you choose action A and the computer chooses action B, your payoff is 12 cents and the computer's is 65 cents. If both you and the computer had chosen A, your payoffs would have been 40 cents each. These payoffs would not affect your actual earnings because Phase 1 is for practice. But in Phase 2, when you interact with human participants, your payoff in each round is added to the cash payment you will receive when the experiment ends.

	The Computer's Choice	
Your Choice	A	B
A	**40¢**, 40¢	**12¢**, 65¢
B	**65¢**, 12¢	**25¢**, 25¢

The number of rounds in each interaction:

After each round, there is a 3/4 chance that the interaction will continue for another round, and a 1/4 chance that the interaction will end. It's as if a four-sided die is rolled after each round and the interaction

goes on for another round if the die comes up 1, 2, or 3 but ends if the die comes up 4. So, no matter how many rounds have already occurred, there's always 75% chance that the interaction will go on to another round.

The two decision-making procedures:

There are two possible decision-making procedures for each interaction: *round-by-round* and *plan*. In the *round-by-round* procedure, you yourself choose action A or B at the start of each round. In the *plan* procedure, a decision plan chosen by you determines your A or B actions for each round. In other words, the *plan* procedure automates A or B choices throughout the interaction. At the beginning of each interaction, you and every other participant in the experiment will choose one of 20 possible plans. After everyone has made his or her plan choice, there is a 50% chance that the *plan* procedure will be implemented automatically to determine everyone's actions (and resulting payoffs) for all the rounds in the interaction. There is also a 50% chance that everyone's plans will not be implemented, in which case every participant must choose A or B at the start of each round. For each of the practice interactions of Phase 1, the computer will randomly select one of the 20 plans and follow it for all rounds of the interaction. You will not be told what plans the computer selects in Phase 1, nor will you be told what plans other participants select in Phase 2.

The 20 different plans:

Twenty different plans are listed in the following table. Most plans specify different actions based on the outcomes of previous rounds. Plans 1-13 start in round 1 by choosing action A. Plans 2-9 start with A, but switch to action B based on actions in one or more previous rounds. These plans indicate scenarios in which a participant would switch back to action A. Plans 10-13 also start with A, but when they switch to B, they remain with B for the remainder of the interaction. Plans 14-20 start in round 1 by choosing action B. Note that whenever a plan prescribes an action other than A, it implements action B (since A and B are the only possible actions).

Once you choose a plan, a confirmation screen will display the plan and let you confirm it or choose another plan. *Once you have confirm a plan choice for the interaction, that plan cannot be changed in later rounds of the interaction.* You can only choose a new plan at the beginning of the next interaction. You will be shown the outcome of each round immediately after the round ends.

Recall that for each interaction, there is a 50% chance that the *plan* procedure will actually be implemented for each round and a 50% chance that the *round-by-round* procedure will be implemented. In both cases, you will be shown your action, the computer's action, and your earnings for each round of the interaction.

Table A1: Plans for Interaction

Plan Number	Description
1	Always choose A in all rounds
2	Start by choosing A, then always choose A *unless* the other participant's action is B in the previous round.
3	Start by choosing A, then always choose A *unless* the other participant's action is B in the *two* previous rounds.
4	Start by choosing A, then always choose A *unless* the other participant's action is B in the *three* previous rounds.
5	Start by choosing A, then always choose A *unless* the other participant's action is B in *either* of the *two* previous rounds. If your choice is B because the other participant's action was B previously, then always choose two consecutive rounds of B; but switch back to A if, and only if, the other participant's actions are two consecutive rounds of A.
6	Start by choosing A, then always choose A *unless* the other participant's action is B in *two* out of the *three* previous rounds. If your choice is B because the other participant's actions were two consecutive B actions, then always choose two consecutive rounds of B before switching back to choose A.
7	Start by choosing A, then continue choosing A until *either* your action *or* the other participant's action is B in the previous round. If this occurs, then choose B twice before switching back to choose A.
8	Start by choosing A, and choose A whenever both participants' actions match (A-A or B-B) in the previous round; otherwise choose B.
9	Start by choosing A, and choose A whenever both participants' actions match (A-A or B-B) for *two* consecutive previous rounds; otherwise choose B.
10	Start by choosing A, and continue to choose A until *either* participant's action is B in the previous round. If either participant's previous action is B, then choose B for every remaining round of the interaction.
11	Start by choosing A, and continue to choose A until *either* participant's action is B for *two* consecutive previous rounds. If this occurs, then choose B for every remaining round of the interaction.
12	Start by choosing A, and continue to choose A until *either* participant's action is B for *three* consecutive previous rounds. If this occurs, then choose B for every remaining round of the interaction.
13	Start by choosing A, then choose B for every remaining round of the interaction.
14	Always choose B in all rounds.
15	Start by choosing B, then always choose A *unless* the other participant's action is B in the previous round.
16	Start by choosing B, then always choose A *unless* the other participant's action is B in the *two* previous round.
17	Start by choosing B, then always choose A *unless* the other participant's action is B in the *three* previous round.
18	Start by choosing B, then switch to choose A until either participant's action is B for *two* consecutive previous rounds. If this occurs, then choose B for for every remaining round of the interaction.
19	Start by choosing B, then switch to choose A until either participant's action is B for *three* consecutive previous rounds. If this occurs, then choose B for for every remaining round of the interaction.
20	Start by choosing B, then switch to A, then switch to B, etc., alternating between A and B for every round of the interaction, regardless of what the other participant's actions are in previous rounds.

Summary

Phase 1 consists of interactions, each of which involves a randomly determined number of rounds. Prior to each interaction, you will choose one of 20 plans that specify how actions will be implemented in the rounds of that interaction. After you choose your plan and the computer randomly chooses its plan, there is a 50% chance that the *plan* procedure will actually be implemented and a 50% chance that the *round-by-round* procedure will be implemented.

If the *plan* procedure is implemented, then your chosen plan and the computer's randomly chosen plan will determine your actions and those of the computer in every round of the interaction. You will be shown the outcome of each round immediately after the round ends.

If the *round-by-round* procedure is implemented, your plan will not be implemented. Instead, you yourself will choose action A or B at the start of each round and the computer will choose A or B based on its randomly chosen plan. You will be shown the outcome of each round immediately after the round ends.

After each round, there is a 75% chance the interaction will continue and a 25% chance the interaction will end. After an interaction ends, you will start a new interaction.

Phase 1 of the experiment is strictly for practice. None of the "payoffs" will count toward your earnings. In Phase 2, when you will be interacting with other participants, your decisions will affect your earnings. Your payoffs in every round – both those that occur during *round-by-round* interactions and those that occur during *plan* interactions – will be added directly to the cash earnings you will be paid via PayPal.

Review Questions Before making any Phase 1 decisions, you will answer some questions to ensure you understand the actions made by different plans. Phase 1 will begin once everyone has correctly answered the questions.

Phase 2 [Unknown Treatment]: Paid Interactions with Other Participants

Phase 2 of the experiment differs from Phase 1 in three important ways: (1) for each interaction you will be randomly paired with another participant rather than the computer; (2) each round has *two stages*; and (3) your payoffs in each round will count toward your actual earnings.

In all other respects, Phase 2 is nearly identical to Phase 1. In particular, ...

- The phase consists of multiple interactions.
- Each interaction includes one or more rounds.
- At the start of the interaction, you select one of the 20 plans.
- There's a 50% chance that, for any given interaction, the *plan* procedure will be implemented (and all actions will be plan-based) and a 50% chance that the *round-by-round* procedure will implemented (and the participants will make their A or B choices at the start of each round).
- The chance of an interaction continuing after any given round is always 75%.

The two stages of each round

In addition to being paired with human participants and earning actual money, Phase 2 differs from Phase 1 because each round of every interaction has *two* stages.

Stage 1: The first stage of each Phase 2 round is essentially the same as a Phase 1 round. You and the participant you've been paired with each choose action A or B (or your plans choose for you), and you each earn payoffs based on the same payoff matrix as before.

	Other Participant's Choice	
Your Choice	A	B
A	**40¢**, 40¢	**12¢**, 65¢
B	**65¢**, 12¢	**25¢**, 25¢

Stage 2: After you and the other participant have chosen your actions (or your plans have chosen for you), **the computer will make an adjustment to your payoffs**. The adjustment will be either +10, +5, 0, -5, or -10 cents. The *same* payoff adjustment will be made to you and the participant you've been paired with. The computer will choose this adjustment based on a complex sequence of calculations. Every round in Phase 2 includes stage 1 and stage 2, so the computer will end up making many adjustments to your payoffs based on numerous calculations.

You will be shown the computer's adjustment as well as your action, the other participant's action, your stage 1 payoff, and your total earnings for the round at the end of each round. In each round, the computer will also calculate separate adjustments for all the other pairs of participants.

Examples: Suppose that you're in a *round-by-round* interaction, and in stage 1 of a round, you choose action A and the participant you're paired with chooses action B. Based on the table above, your stage 1 payoff is 12 and the other participant's payoff is 65. Suppose that in stage 2, the computer chooses a -5 adjustment. Then your total earnings for the round equal 7 cents (because 12 - 5 = 7), and the other participant's total earnings equal 60 cents (because 65 - 5 = 60). Or suppose that you're in *plan* interaction, and in stage 1 of a round, the plan you selected chooses action B and the other participant's plan chooses action B. Then, the payoffs to you and the other participant are both 25. If in stage 2, the computer chooses a +10 adjustment, your total earnings for the round equal 35 cents, and the other participant's total earnings equal 35 cents.
You can refer back to these instructions at any time. Note that the same payoffs and the same twenty plans apply to both Phase 1 and Phase 2.

Phase 3: Your Feedback

After Phase 1 and Phase 2, you will be asked to reflect on the experiment. You will be asked if you observed any patterns in the computer's adjustments. You will also be asked if you observed any relationships between people's decisions and the computer's adjustments. *Your feedback is a key part of this experiment.* We will carefully review everyone's comments, and we will share a number of the comments with future participants in this experiment. If we include some of your comments, you will receive an *additional* payment equal to the average earnings of the future participants. We are more likely to include your comments if they are clear and specific, avoid abbreviations, and use complete sentences with proper punctuation.

Phase 2 [Random Treatment]: Paid Interactions with Other Participants

Phase 2 of the experiment differs from Phase 1 in three important ways: (1) for each interaction you will be randomly paired with another participant rather than the computer; (2) each round has *two stages*; and (3) your payoffs in each round will count toward your actual earnings.

In all other respects, Phase 2 is nearly identical to Phase 1. In particular, ...

- The phase consists of multiple interactions.
- Each interaction includes one or more rounds.
- At the start of the interaction, you select one of the 20 plans.

- There's a 50% chance that, for any given interaction, the *plan* procedure will be implemented (and all actions will be plan-based) and a 50% chance that the *round-by-round* procedure will implemented (and the participants will make their A or B choices at the start of each round).
- The chance of an interaction continuing after any given round is always 75%.

The two stages of each Phase 2 round

In addition to being paired with human participants and earning actual money, Phase 2 differs from Phase 1 because each round of every interaction has *two* stages.

Stage 1: The first stage of each Phase 2 round is essentially the same as a Phase 1 round. You and the participant you've been paired with each choose action A or B (or your plans choose for you), and you each earn payoffs based on the same payoff matrix as before.

	Other Participant's Choice	
Your Choice	A	B
A	**40¢**, 40¢	**12¢**, 65¢
B	**65¢**, 12¢	**25¢**, 25¢

Stage 2: In the second stage of each round, the computer will make adjustments to the earnings of all the participants in the experiment. The adjustments will be either +10, +5, 0, -5, or -10 cents. The same adjustment will be made to the earnings of you and the other participant, but the other pairs may receive different adjustments. The computer will choose the adjustment for your group based on a random number calculation, so that there will be a 20% chance that 10 cents are added to the earnings of everyone in your group, a 20% chance that 5 cents are added, a 20% chance that 0 cents is added, a 20% chance that 5 cents are subtracted, and a 20% chance that 10 cents are subtracted from every member's earnings.

You will be shown the computer's adjustment as well as your action, the other participant's action, your stage 1 payoff, and your total earnings for the round at the end of each round. In each round, the computer will also calculate separate adjustments for all the other pairs of participants.

Examples: Suppose that you're in a *round-by-round* interaction, and in stage 1 of a round, you choose action A and the participant you're paired with chooses action B. Based on the table above, your stage 1 payoff is 12 and the other participant's payoff is 65. Suppose that in stage 2, the computer randomly chooses a -5 adjustment. Then your total earnings for the round equal 7 cents (because 12 - 5 = 7), and the other participant's total earnings equal 60 cents (because 65 - 5 = 60). Or suppose that you're in *plan* interaction, and in stage 1 of a round, the plan you selected chooses action B and the other participant's plan chooses action B. Then, the payoffs to you and the other participant are both 25. If in stage 2, the

computer randomly chooses a +10 adjustment, your total earnings for the round equal 35 cents, and the other participant's total earnings equal 35 cents.
You can refer back to these instructions at any time. Note that the same payoffs and the same twenty plans apply to both Phase 1 and Phase 2.

Phase 3: Your Feedback

After Phase 1 and Phase 2, you will be asked to reflect on the experiment. You will be asked if you observed any relationships between people's decisions and the computer's adjustments.
Your feedback is a key part of this experiment. We will carefully review everyone's comments, and we will share a number of the comments with future participants in this experiment. *If we include some of your comments, you will receive an additional payment equal to the average earnings of the future participants.* We are more likely to include your comments if they are clear and specific, avoid abbreviations, and use complete sentences with proper punctuation.

Phase 2 [Baseline Treatment]: Paid Interactions with Other Participants

Phase 2 of the experiment differs from Phase 1 in one important way: for each interaction you will be randomly paired with another participant rather than the computer.
In all other respects, Phase 2 is nearly identical to Phase 1. In particular, ...

- The phase consists of multiple interactions.
- Each interaction includes one or more rounds.
- At the start of the interaction, you select one of the 20 plans.
- There's a 50% chance that, for any given interaction, the *plan* procedure will be implemented (and all actions will be plan-based) and a 50% chance that the *round-by-round* procedure will implemented (and the participants will make their A or B choices at the start of each round).
- The chance of an interaction continuing after any given round is always 75%.

The earnings of the actions (in cents)

	Other Participant's Choice	
Your Choice	A	B
A	**40¢**, 40¢	**12¢**, 65¢
B	**65¢**, 12¢	**25¢**, 25¢

Example: Suppose that in Stage 1 of a round, you choose action A and the participant you're paired with chooses action B. Based on the table above, your stage 1 earnings equal 12 and the other participant's earnings equal 65.

B Adjustment Algorithm

Let C_t denote the number of participants who cooperated in round t. The computer first classified the session according to whether cooperation constituted a strict majority:

$$r_t \equiv \begin{cases} 1 & \text{if } C_t > 8, \\ 2 & \text{if } C_t \leq 8. \end{cases} \tag{B1}$$

For each pair g, the computer then generated an independent random draw $u_{gt} \sim U[0,1)$. Draws were independent across pairs and rounds. Let $\alpha, \beta \in [0,1]$ govern the relative weights assigned to positive and negative adjustments under the two cooperation classifications.

When a strict majority of participants cooperated, the adjustment was

$$A_{gt} \equiv \begin{cases} -10 & \text{if } 0 \leq u_{gt} < \dfrac{2\alpha}{5}, \\ -5 & \text{if } \dfrac{2\alpha}{5} \leq u_{gt} < \dfrac{4\alpha}{5}, \\ 0 & \text{if } \dfrac{4\alpha}{5} \leq u_{gt} < \dfrac{4\alpha+1}{5}, \\ 5 & \text{if } \dfrac{4\alpha+1}{5} \leq u_{gt} < \dfrac{2\alpha+3}{5}, \\ 10 & \text{if } \dfrac{2\alpha+3}{5} \leq u_{gt} < 1. \end{cases} \qquad \text{if } r_t = 1. \tag{B2}$$

When eight or fewer participants cooperated, the ordering of the adjustments was reversed:

$$A_{gt} \equiv \begin{cases} 10 & \text{if } 0 \leq u_{gt} < \dfrac{2\beta}{5}, \\ 5 & \text{if } \dfrac{2\beta}{5} \leq u_{gt} < \dfrac{4\beta}{5}, \\ 0 & \text{if } \dfrac{4\beta}{5} \leq u_{gt} < \dfrac{4\beta+1}{5}, \\ -5 & \text{if } \dfrac{4\beta+1}{5} \leq u_{gt} < \dfrac{2\beta+3}{5}, \\ -10 & \text{if } \dfrac{2\beta+3}{5} \leq u_{gt} < 1. \end{cases} \qquad \text{if } r_t = 2. \tag{B3}$$

The same adjustment A_{gt} was applied to both members of pair g. Conditional on $r_t = 1$, each negative adjustment occurred with probability $2\alpha/5$, each positive adjustment with probability $2(1-\alpha)/5$, and

the zero adjustment with probability 1/5. Conditional on $r_t = 2$, each positive adjustment occurred with probability $2\beta/5$, each negative adjustment with probability $2(1-\beta)/5$, and the zero adjustment with probability 1/5. Thus, $1-\alpha > \beta$ would make favorable adjustments more likely when a strict majority cooperated, whereas $1-\alpha < \beta$ would favor the lower-cooperation classification.

For the sessions analyzed in this paper, we set $\alpha = \beta = 0.5$. Under these parameter values, the relevant cutoffs reduce to 0.2, 0.4, 0.6, and 0.8. The two cooperation classifications therefore differed only in whether the five adjustment values were assigned to these intervals in ascending or descending order. Because u_{gt} was uniformly distributed,

$$\Pr(A_{gt} = a \mid C_t) = \frac{1}{5} \qquad \text{for every } a \in \{-10, -5, 0, 5, 10\}. \tag{B4}$$

Consequently, $\mathbb{E}[A_{gt} \mid C_t] = 0$. Although aggregate cooperation determined which mirror-image mapping was applied, it did not affect the probability of receiving any particular adjustment. The resulting adjustment algorithm was therefore fully random with respect to aggregate cooperation.

C Content Analysis of Belief Elicitation

C.1 Classification Procedure

We conducted a content analysis of the open-ended belief responses. We developed the coding categories, category definitions, and decision rules based on the conceptual distinction between Skinnerian repeat behavior and Malinowskian rationalization. A generative AI model then applied the resulting codebook to the anonymized responses. Large language models are well suited to the systematic classification of open-ended survey responses, which can otherwise require substantial time and resources (Korinek, 2023; Chang *et al.*, 2026).

Before classification, all participant identifiers were removed from the responses. The model received the relevant survey question, the anonymized response, and the applicable codebook. It did not receive information about the participant's identity.[14] The same classification instructions were used for every response. The codebooks are reported below.

C.2 Coding Categories

General Pattern

"Did you notice any patterns in the computer's adjustments?" (Unknown treatment only)

[14]The model was not provided with the paper's predictions or with any information about the expected relationship between treatment and the response categories.

Table C.21: **Coding Categories for the General Pattern Question**

Code	Label	Description
P	Prosocial Causal Explanation	The response links cooperative actions or cooperative group behavior to more favorable adjustments. This includes claims that the computer rewarded cooperation, fairness, or other prosocial behavior.
R	Random	The response describes the adjustments as random or states that no discernible pattern was present.
U	Uncertain	The response expresses doubt, confusion, or an inability to identify a pattern.
X	Other or Complex	The response proposes another pattern, an inconsistent or highly specific mechanism, or an explanation that cannot be classified as prosocial, random, or uncertain.

Relationship Between Actions and Adjustments

"Did you notice any relationship between the contributions of you or your group and the computer's adjustments?"

Table C.22: **Coding Categories for the Relationship Question**

Code	Label	Description
PR	Positive Relationship	The response states that cooperation or cooperative group behavior led to more favorable adjustments.
NR	Negative Relationship	The response states that cooperation or cooperative group behavior led to less favorable adjustments.
IR	Random or No Relationship	The response describes the adjustments as random or explicitly rejects a relationship between actions and adjustments.
UR	Unclear	The response does not provide an interpretable conclusion about the relationship.
XR	Other or Complex	The response proposes a shifting, mixed, or more elaborate relationship that does not imply a clear positive or negative link.

Advice to Future Participants

"Do you have any advice for future participants in this experiment?"

Table C.23: **Coding Categories for Advice to Future Participants**

Code	Label	Description
T	Trust or Cooperate	Recommends choosing A, cooperating, or working together.
W	Watch for Patterns	Recommends observing, testing, or adapting to perceived patterns in the adjustments.
S	Self-Focus	Recommends maximizing one's own payoff, often by choosing B.
N	No Advice or Unclear	Provides no usable advice or does not express an interpretable recommendation.
X	Other	Provides advice that does not fit the other categories.
M	Match the Other Participant	Recommends conditioning one's action on the other participant's behavior.
MS	Mixed or Adaptive Strategy	Recommends combining approaches or changing behavior across rounds.

C.3 Results

Table C.31 reports the distribution of the coded responses. The relationship question provides the direct comparison between Random and Unknown because it was administered in both treatments. Four of 57 respondents in Random, or 7.0 percent, described a positive relationship between cooperative behavior and favorable adjustments. Four of 62 respondents in Unknown, or 6.5 percent, did so. One respondent in Unknown described a negative relationship, while no respondent in Random did so. Thus, leaving the adjustment process unexplained did not increase the prevalence of explicit prosocial causal attributions.

Responses rejecting a relationship were also common in both treatments. Nineteen respondents in Random and 18 in Unknown described the adjustments as random or unrelated to participants' actions. Other or complex explanations accounted for 29 responses in Random and 27 in Unknown. Unclear responses were more frequent in Unknown, with 12 responses compared with five in Random. This difference is more consistent with greater difficulty interpreting the opaque adjustment process than with the emergence of a shared alternative explanation.

The general pattern question was administered only in Unknown. Three of the 62 respondents articulated a prosocial causal explanation. Twenty described the adjustments as random, seven remained uncertain, and 32 proposed another or more complex pattern. Thus, even when participants were explicitly asked to consider possible patterns, few concluded that cooperative behavior produced favorable adjustments.

The advice responses show somewhat larger descriptive differences. Self-focused advice was the most common category in both treatments, appearing in 15 of 58 responses in Random and 16 of 61 responses in Unknown. Trust or cooperation was recommended by four respondents in Random and ten in Unknown.

Table C.31: **Summary Statistics of Belief and Advice Types by Treatment**

Category	Label	Random, N (%)	Unknown, N (%)
Pattern Belief (`pattern_code`)			
	Prosocial Causal Explanation	–	3 (4.8%)
	Random	–	20 (32.3%)
	Uncertain	–	7 (11.3%)
	Other or Complex	–	32 (51.6%)
	Total	–	**62 (100%)**
Relationship Belief (`relationship_code`)			
	Positive Relationship	4 (7.0%)	4 (6.5%)
	Negative Relationship	0 (0.0%)	1 (1.6%)
	Random or No Relationship	19 (33.3%)	18 (29.0%)
	Unclear	5 (8.8%)	12 (19.4%)
	Other or Complex	29 (50.9%)	27 (43.5%)
	Total	**57 (100%)**	**62 (100%)**
Advice Type (`advice_code`)			
	Trust or Cooperate	4 (6.9%)	10 (16.4%)
	Watch for Patterns	3 (5.2%)	5 (8.2%)
	Self-Focus	15 (25.9%)	16 (26.2%)
	No Advice or Unclear	10 (17.2%)	12 (19.7%)
	Other	13 (22.4%)	7 (11.5%)
	Match the Other Participant	4 (6.9%)	6 (9.8%)
	Mixed or Adaptive Strategy	9 (15.5%)	5 (8.2%)
	Total	**58 (100%)**	**61 (100%)**

Notes: Entries report the number of responses, with percentages among nonmissing responses to the corresponding question in parentheses. The general pattern question was administered only in Unknown. Participants in Random were told that the adjustments were random, while the adjustment process remained unexplained in Unknown.

Matching the other participant or adopting a mixed or adaptive strategy accounted for 13 responses in Random and 11 in Unknown. Cooperative advice was therefore more common in Unknown, but such advice need not reflect a causal belief about the adjustments.

Table C.32 examines whether participants who reported a positive relationship between cooperation and favorable adjustments had previously cooperated more frequently. Positive Relationship equals one for the eight participants who articulated such a relationship, four in each treatment. All other coded responses form the comparison category. The analysis is restricted to direct decisions after the first round in Random and Unknown and includes the same lagged controls and multilevel structure used in the analysis of adjustment-contingent behavior.

Column (1) shows no overall association between reporting a positive relationship and cooperation. The estimated difference is 4.8 percentage points and is not statistically significant. Column (2) allows this

Table C.32: **Association Between Post-Experiment Beliefs and Cooperation (Rounds > 1, Direct Decisions, Noise Treatments)**

	(1) Cooperation Pooled Association	(2) Cooperation Treatment Interaction
Positive Relationship	0.048	-0.040
	(0.061)	(0.047)
Random × Positive Relationship		0.180**
		(0.071)
Intercept	0.588***	0.593***
	(0.040)	(0.042)
Supergame Fixed Effects	Yes	Yes
Lagged Partner Action	Yes	Yes
Treatment × Adjustment	Yes	Yes
Random Intercepts	Yes	Yes
Observations	2,710	2,710

Standard errors clustered at the session level are reported in parentheses.
Random intercepts are included for sessions, participants, and supergames.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
The dependent variable equals one for cooperation. Positive Relationship identifies participants who reported that cooperation or cooperative group behavior produced more favorable adjustments. All other coded relationship responses form the omitted belief category. Unknown is the omitted treatment. Treatment main effects, lagged-adjustment indicators, and treatment-by-adjustment interactions are included but not reported.

association to differ across treatments. In Unknown, participants reporting a positive relationship cooperated 4.0 percentage points less than other participants, a difference that is not statistically significant. The association is 18.0 percentage points more positive in Random than in Unknown. The implied association in Random is therefore approximately 14.0 percentage points.

These treatment-specific estimates should be interpreted cautiously because they are based on only four positive-relationship respondents in each treatment. Taken together, the belief responses provide little evidence that opacity generated Malinowskian rationalization. Explicit causal explanations linking cooperation to favorable adjustments remained rare, and those explanations were not consistently associated with participants' earlier cooperative behavior.

D From Plans to Play

The main analysis establishes how payoff noise shapes plan selection, direct cooperation, and responses to realized adjustments. We now use the experiment's two decision procedures to examine how participants translated their selected plans into round-by-round behavior. We first measure how closely direct deci-

sions followed the actions prescribed by the selected plan. We then examine whether this correspondence depended on plan complexity and whether that relationship varied across treatments.

D.1 Mapping Strategies to Direct Decisions

Participants selected a plan at the beginning of every supergame. When the round-by-round procedure was implemented, however, they remained free to choose differently. This allows us to compare the plan selected before play with the actions taken as the interaction unfolded. We begin by asking whether each direct decision matched the action the selected plan would have prescribed had it governed the participant's play throughout, given the partner's observed actions. We then use the Strategy Frequency Estimation Method (SFEM) of Dal Bó and Fréchette (2011) to identify the strategies that best describe participants' observed sequences of play.

Figure D.11 compares selected plans with direct decisions. Panel A reports the share of decisions that followed the selected plan, separately by plan and treatment. Panel B reports the corresponding number of consistent decisions. The second panel helps distinguish plans that were followed reliably from plans that account for a large share of observed play because they were selected frequently.

Across treatments and plans, 79.4 percent of direct decisions matched the action prescribed by the selected plan. This level of consistency is in line with earlier evidence that participants can carry out explicitly selected strategies with considerable accuracy in repeated games (Dal Bó and Fréchette, 2019; Romero and Rosokha, 2018).

Consistency varies across plans. ALLD has the highest overall consistency rate, at 91.9 percent. D-TFT, TF2T, and TFT follow at 80.6, 80.5, and 79.1 percent. Several other conditional plans are also implemented with substantial accuracy. The corresponding rates are 78.3 percent for 2TF2T, 77.2 percent for Grim, and 77.0 percent for C-to-ALLD. The large number of consistent ALLD decisions in Panel B reflects both its high implementation rate and its frequent selection.

We next use SFEM to characterize the strategies reflected in direct play. Unlike the preceding comparison, SFEM does not use the plan selected at the beginning of the supergame. Instead, it estimates the population shares of the fixed strategies that make the observed action sequences most likely, while allowing for decision error.

Figure D.12 presents the estimated strategy frequencies by treatment, while Table D.11 reports the four most common strategies in each treatment together with bootstrapped standard errors. In a separate SFEM estimation pooling observations across treatments, C-to-ALLD, ALLD, and D-TFT have the largest estimated shares, at 23.1, 22.9, and 20.1 percent, respectively.

The clearest difference between selected plans and inferred play appears in Baseline. SFEM estimates a 57 percent share for C-to-ALLD in that treatment, even though only 4 percent of participants selected it as their plan. This gap need not imply that participants concealed their intended strategy. The selected plan records an explicit choice made before play, while SFEM identifies the strategy that best fits the

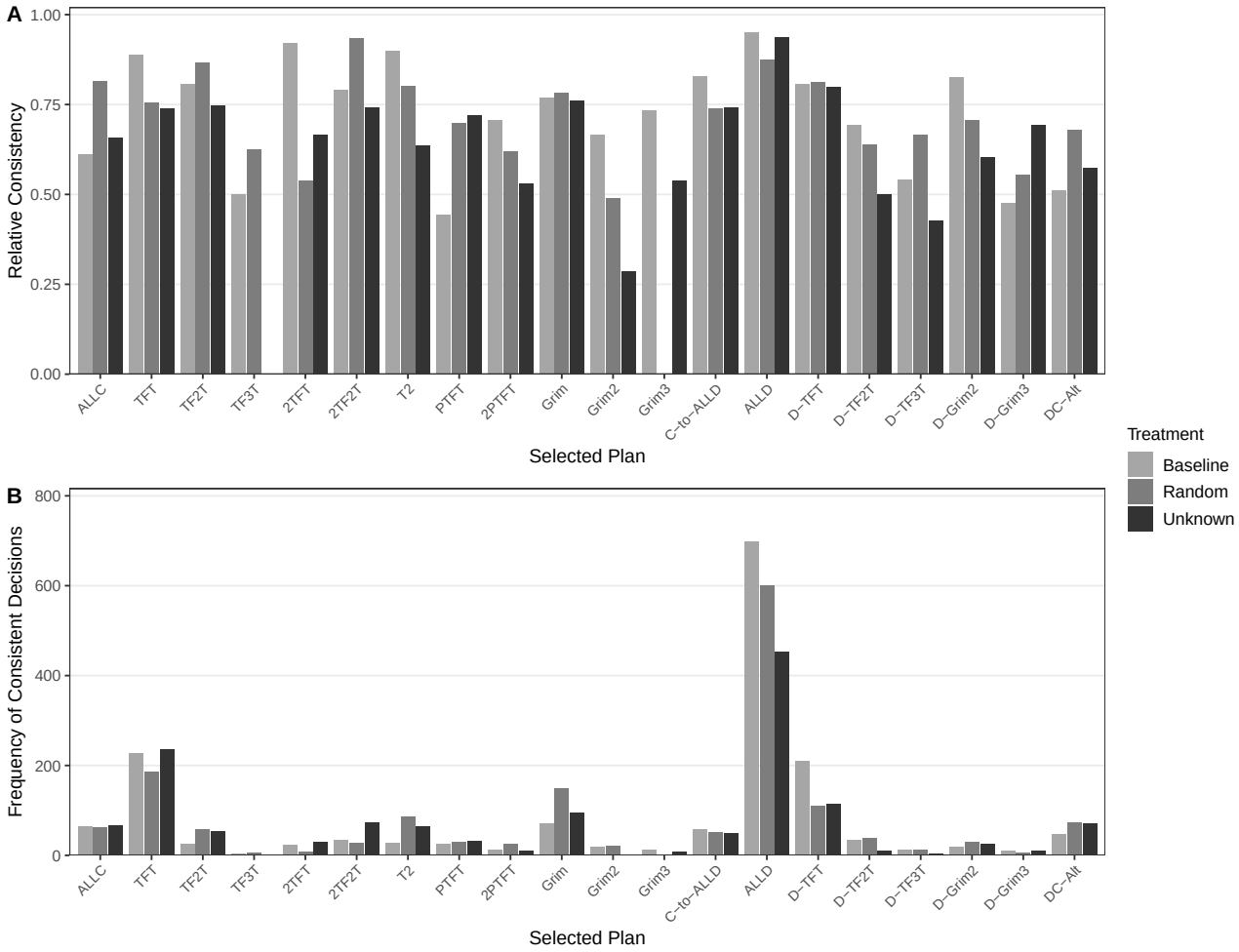


Figure D.11: **Consistency Between Selected Plans and Direct Decisions.** Panel A reports the proportion of direct decisions that match the action prescribed by the selected plan, separately by plan and treatment. Panel B reports the number of plan-consistent direct decisions.

actions observed afterward. In short supergames, several strategies may prescribe the same actions along the realized path, and even a small number of deviations can make observed play resemble a different strategy.

The two approaches show substantial continuity between plans and play. Participants followed their selected plan in roughly four out of five direct decisions, particularly when the plan was simple and commonly used. This high rate of implementation provides little evidence that participants were generally confused or overwhelmed by the decision environment. The remaining differences between selected and inferred strategies may partly reflect the complexity of some plans. We examine the role of plan complexity next.

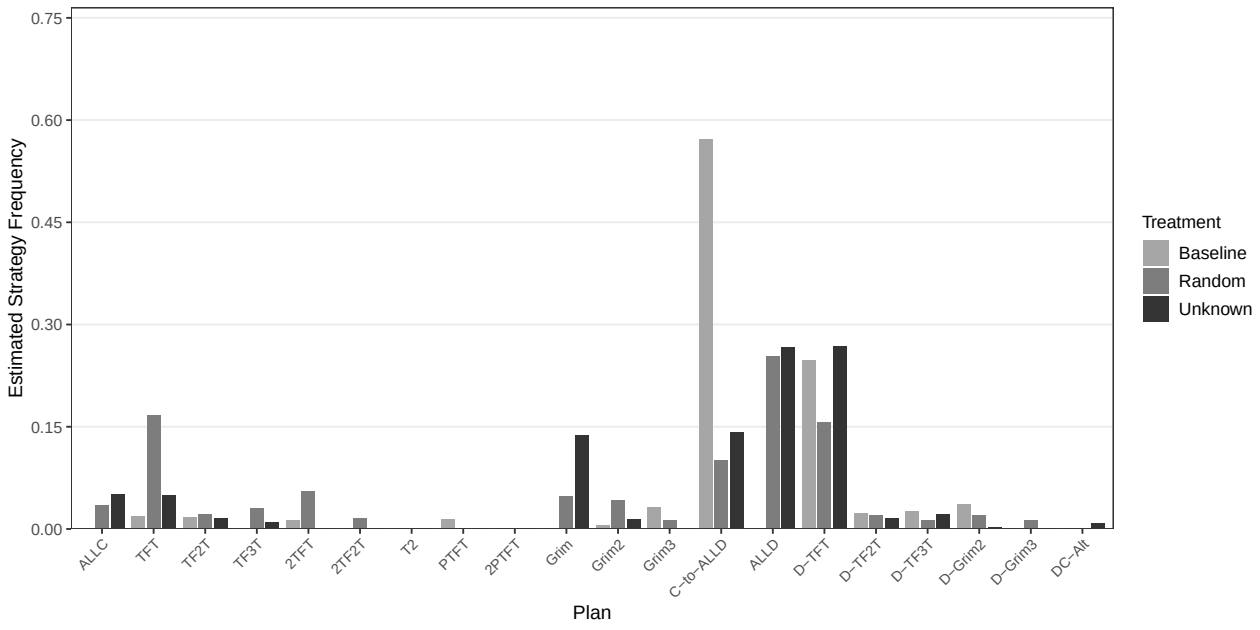


Figure D.12: **Estimated Strategy Frequencies During Direct Play.** The figure reports SFEM estimates based on choices made under the round-by-round procedure, separately by treatment.

Table D.11: **Estimated Strategy Frequencies by Treatment**

Treatment	**Four Most Frequent Estimated Strategies**				β	**Participants**
	C-to-ALLD	D-TFT	D-Grim2	Grim3		
Baseline	0.57	0.25	0.04	0.03	0.86	64
	(0.13)	(0.05)	(0.03)	(0.03)	(0.14)	
	ALLD	TFT	D-TFT	C-to-ALLD		
Random	0.25	0.17	0.16	0.10	0.85	64
	(0.06)	(0.05)	(0.05)	(0.06)	(0.09)	
	D-TFT	ALLD	C-to-ALLD	Grim		
Unknown	0.27	0.27	0.14	0.14	0.83	64
	(0.07)	(0.08)	(0.08)	(0.05)	(0.13)	

Notes: The table reports the four largest SFEM strategy frequencies in each treatment. Bootstrapped standard errors are reported in parentheses. The final two columns report the estimated β parameter and the number of participants.

D.2 Plan Complexity and Implementation

Why do participants sometimes depart from the plans they selected? One possibility is implementation difficulty. Plans with more states or effective transitions require participants to track more contingencies as play unfolds. Following Oprea (2020) and Guan and Oprea (Working paper), we assign each plan a complexity score based on its automaton states and effective transitions. We then collapse the direct-decision data to the participant–interaction level and estimate mixed-effects linear models relating the

resulting plan–play consistency rate to complexity. The models include supergame fixed effects and random intercepts for sessions, participants, and supergames. Standard errors are clustered by session.

Table D.21 shows a clear negative relationship between complexity and implementation. In Column (1), each additional unit of complexity reduces the plan–play consistency rate by 2.6 percentage points ($p < 0.001$). Column (2) allows this relationship to vary across treatments. The estimated reduction is 2.7 percentage points in Baseline ($p < 0.001$), 2.0 percentage points in Random ($p = 0.004$), and 3.2 percentage points in Unknown ($p < 0.001$). Neither the Baseline nor the Random interaction with complexity is statistically significant, leaving no evidence that the complexity penalty differs across treatments.

Table D.21: **Plan Complexity and Implementation Consistency**

	(1)	(2)
	Plan–Play Consistency	
Complexity	-0.026***	-0.032***
	(0.0039)	(0.0064)
Baseline		0.039
		(0.025)
Random		0.003
		(0.032)
Baseline × Complexity		0.005
		(0.0073)
Random × Complexity		0.012
		(0.0092)
Constant	0.842***	0.826***
	(0.059)	(0.056)
Supergame Fixed Effects	Yes	Yes
Random Intercepts	Yes	Yes
Observations	1,376	1,376

Standard errors clustered at the session level are reported in parentheses.
Random intercepts are included for sessions, participants, and supergames.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Unknown is the omitted treatment in Column (2).

Plan complexity therefore helps explain part of the remaining gap between plans and play. Participants implement their selected plans with considerable accuracy overall, but consistency declines as a plan requires them to track more contingencies. Because complexity reduces consistency in all three treatments and the estimated slopes do not differ significantly across treatments, departures from selected plans are unlikely to reflect broad confusion specific to the payoff-noise treatments. Noise mainly shapes which plans participants choose, while complexity shapes how faithfully they carry them out.

E From Favorable Coincidence to Shared Norms and Institutions

The experiments in this paper focus on the earliest stages of superstition. Viewed together with our related work, however, they suggest a broader sequence through which an initially individual response to a random outcome can develop into a socially shared and institutionally reinforced practice (Hajikhameneh and Iannaccone, 2023, 2025; Hajikhameneh, 2026). Figure E1 summarizes this progression.

The sequence begins in an environment of uncertainty, where important outcomes are observable, but the forces producing them are difficult to identify or control. Such environments create fertile conditions for accidental relationships between actions and outcomes to acquire significance. A favorable coincidence provides the initiating event. An individual takes an action and is subsequently exposed to a salient favorable outcome whose occurrence is objectively unrelated to that action. Random reinforcement can then make the preceding action more likely to recur. The present experiment provides particularly direct evidence for this first transition. Favorable adjustments following cooperation increase the likelihood that cooperation continues, especially when the source of the adjustment is unexplained, even though participants are no more likely to articulate a causal relationship between cooperation and favorable adjustments. Our earlier public-goods experiments document a closely related pattern, with positive adjustments increasing the probability that participants repeat their previous contributions (Hajikhameneh and Iannaccone, 2025).

For some individuals, repeated behavior can subsequently acquire a causal explanation. The favorable outcome is interpreted as having resulted from the preceding action, transforming a behavioral response into an articulated superstition. We refer to this individual-level process as rationalization. The evidence across our experiments suggests that this causal account need not initiate the behavioral response. Rather, repetition can emerge before any corresponding rationalization becomes observable. Rationalization is therefore a subsequent and contingent stage in the progression. Many randomly reinforced behaviors may never develop into articulated causal beliefs.

Once a causal account has formed, social interaction can transform a private interpretation into a shared one. A private rationalization can be communicated, evaluated, modified, and reinforced by others. Our communication experiments provide evidence for this transition. Group members facing the same uncertain environment developed shared interpretations of random outcomes and coordinated their actions around those interpretations (Hajikhameneh and Iannaccone, 2023). In some cases, communication strengthened cooperation. In others, groups collectively searched for winning numbers and maintained inefficient practices despite repeated evidence against the presumed relationship. Social interaction can therefore transform an individual superstition into a shared rationalization without guaranteeing that the resulting belief is accurate or socially beneficial.

Shared rationalizations can persist beyond the individuals who first developed them. Our intergenerational experiments show that explanations formed by one cohort can be transmitted to another and influence the beliefs adopted by later participants (Hajikhameneh, 2026). Transmission also changes the

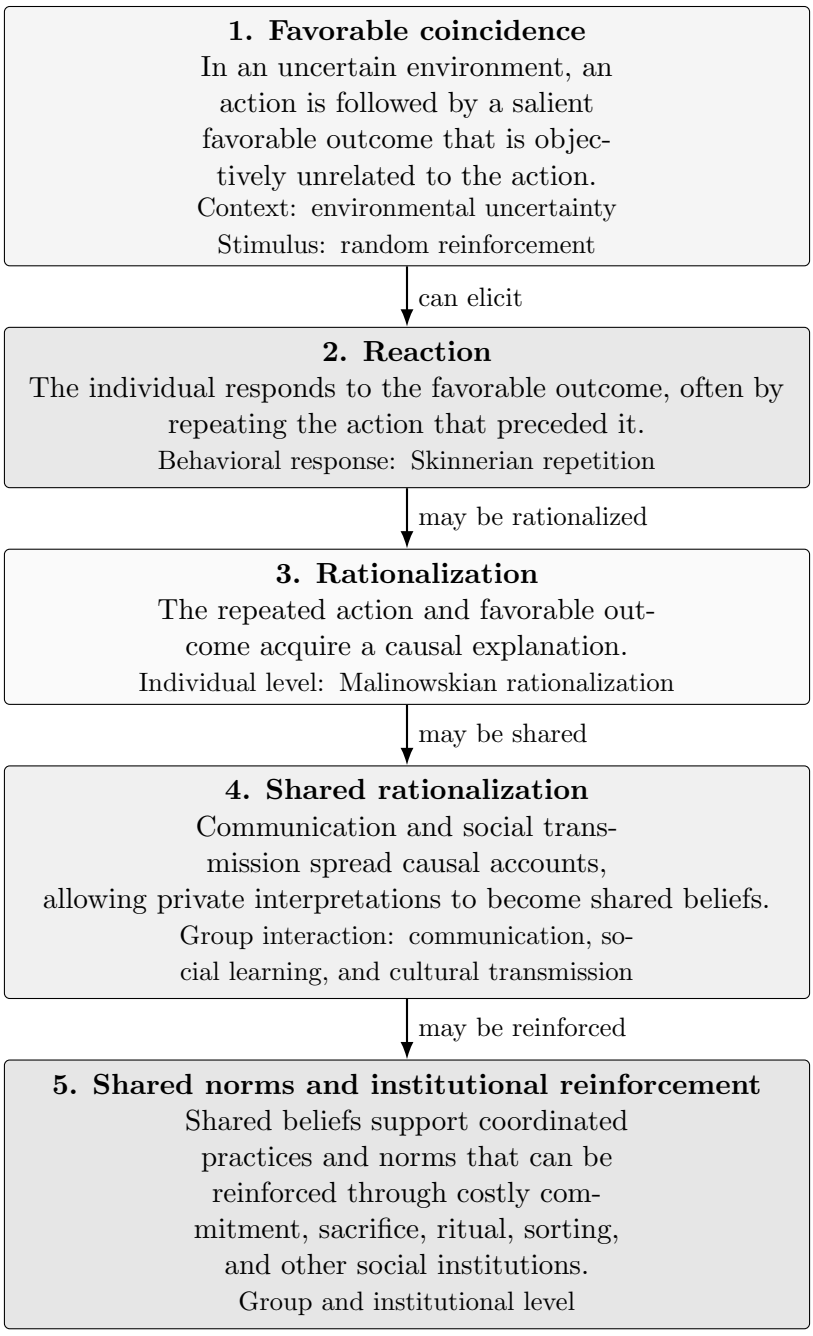


Figure E1: **A Possible Progression from Random Reinforcement to Shared Norms and Institutions.** The accumulated evidence from our experiments suggests a sequential process in which a favorable coincidence first produces a behavioral response and may subsequently acquire a causal rationalization. Once articulated, private causal accounts can spread through communication and cultural transmission, become shared within groups, and interact with norms and institutions that reinforce coordinated behavior. The arrows beyond the initial behavioral response are contingent rather than inevitable. Many reinforced behaviors may never be rationalized, and socially shared superstitions need not improve welfare.

strategic significance of belief. A private causal interpretation provides little information about what others will do, whereas a commonly transmitted narrative can create shared expectations about appropriate behavior. In this sense, social transmission provides a bridge from individual rationalization to a shared norm.

The final stage concerns institutional reinforcement. Shared beliefs and norms can interact with practices that make commitment observable or costly. In our intergenerational experiment, participants who adopted superstitious–prosocial beliefs were more willing to sacrifice private returns, and voluntary sacrifice enabled similarly committed participants to sort into highly cooperative groups (Hajikhameneh, 2026). In naturally occurring settings, analogous mechanisms may include rituals, sacrifices, taboos, costly observances, and other practices that express and reinforce shared beliefs. Such institutions can stabilize behavior by coordinating expectations, identifying committed members, and transmitting both beliefs and practices over time.

Figure E1 places these findings within a broader progression. The accumulated evidence is consistent with superstition developing from behavior outward. Environmental uncertainty creates scope for favorable coincidences to influence behavior through random reinforcement. Repeated behavior may subsequently acquire a causal explanation. Private rationalizations can become shared through communication and transmission, and shared beliefs can support norms and institutions that reinforce the associated behavior. Not every superstition will complete this progression, and the resulting practices may be either functional or counterproductive. The broader pattern suggests how environmental uncertainty can set in motion a process through which an accidental relationship between action and outcome acquires behavioral, cognitive, social, and eventually institutional significance.