# EPISTEMIC EXPERIMENTS: UTILITIES, BELIEFS, AND IRRATIONAL PLAY[†]

PAUL J. HEALY[*]

ABSTRACT. Inspired by the epistemic game theory framework, we elicit subjects' utilities, beliefs about strategies, and beliefs about beliefs in a variety of classic games. In the centipede game, subjects with selfish preferences pass in early nodes because they correctly believe that altruistic opponents will pass back to them. Cooperation in the prisoners' dilemma is largely irrational: many who cooperate do so knowing they'll receive outcomes they prefer less. In the $2 \times 2$ games where social preferences aren't observed we still observe irrational play apparently driven by factors such as loss aversion and stubbornness.

Keywords: Behavioral game theory; payoff uncertainty; rationality.

JEL Classification: C72, C90, D03, D81.

Draft: May 17, 2024

[†]This work subsumes a paper previously circulated as "Epistemic Conditions for the Failure of Nash Equilibrium". I thank seminar audiences at George Mason, U. British Columbia, Royal Holloway, U. Edinburgh, U.C. San Diego, Florida State, U. Pittsburgh, N.C. State, U. Melbourne, Monash, U. New South Wales, U.C. Davis, Heidelberg U., Karlsruhe Inst. Tech., U. Cologne, ITAM, Wisconsin, New York U., Purdue, Arizona, and Michigan State. I have benefitted greatly from conversations with Yaron Azrieli, Christoph Brunner, Evan Calford, Christopher Chambers, Brad Clark, Amanda Friedenberg, Aviad Heifetz, Kirby Nielsen, Hyoeun Park, Ryan Oprea, Antonio Penta, Ariel Rubinstein, Marciano Siniscalchi, Lise Vesterlund, Alistair Wilson, and many, many more. I am extremely grateful for the research assistance provided by Caleb Cox, Alex Gotthard-Real, Ritesh Jain, Siqi Pan, Kirby Nielsen, and Hyoeun Park.

[*]Dept. of Economics, The Ohio State University, 1945 North High street, Columbus, Ohio 43210, U.S.A.; healy.52@osu.edu.

## I. INTRODUCTION

When we observe strategic behavior that differs from theoretical equilibrium predictions, the natural next step is to understand why those deviations occurred and to construct models of bounded rationality that explain them. For example, popular theories have been built on the idea that players have incorrect beliefs (Nagel, 1995; Stahl and Wilson, 1994, 1995; Camerer et al., 2004), imperfectly best respond (McKelvey and Palfrey, 1995), have social preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), or fail to engage in contingent thinking (Eyster and Rabin, 2005). Though often bolstered by choice-process data, most of these models were initially developed from strategy choice data alone, without direct observation of the underlying causes for equilibrium to fail. This begs the question of whether there can be a more theoretically-disciplined way to identify the causes of these deviations.

Epistemic game theory provides a framework well-suited for exactly this purpose. Theorems in this literature identify which properties of beliefs, utilities, and rationality are sufficient for a given solution concept such as Nash equilibrium to occur. The contrapositive of such a theorem thus tells us, if the solution concept does fail, which properties of beliefs, utilities, and rationality could be to blame. For example, Aumann and Brandenburger (1995) prove that if Nash equilibrium fails in a two player game then it must be that players have incorrect beliefs over strategies, do not believe in the rationality of their opponent, or do not believe that their opponents have correct beliefs.[1]

To apply this framework, however, requires that we measure utilities, beliefs, and rationality directly. That is exactly the aim of this paper, and why we refer to our experiments as "epistemic experiments." We run several experiments in which subjects play a variety of classic games and, in each, we elicit the players' utilities over outcomes, belief over their opponent's utilities, first- and second-order belief over actions, and first-order belief in rationality. This procedure allows us to see the true game being played (including the players' actual utilities, not just their payoffs), and whether it is a truly a game of complete information. It also allows us to determine whether beliefs and actions are in equilibrium. And, most importantly, it gives us direct insight into why equilibrium fails when it does.

The amount of elicitation employed here is admittedly extreme, relative to past studies. We have therefore taken steps to ensure that the data are as reliable as possible. We incentivize each elicitation using the binary choice list procedure that underlies the Becker et al. (1964) value elicitation mechanism, the belief elicitation procedure of Grether (1981) and Karni (2009) which has been validated by Holt and Smith (2016), and the risk preference elicitation of Holt and Laury (2002). These procedures are all

[1]See (Dekel and Siniscalchi, 2015, Theorem 5) for an updated and precise statement of this theorem.

incentive compatible under relatively weak assumptions (Azrieli et al., 2018), and subjects are explicitly told that truth-telling is in their best interest. Finally, we run additional experiments without elicitation to check whether the presence of elicitation alters strategic behavior. In seven of the eight games we study we find that it does not. Thus we feel reasonably confident that our elicitation methods offer some potential value in understanding the reasoning behind strategic choice.

Our first finding is that some players do have non-selfish preferences, and in certain games those non-selfish preferences drastically alter the game being played even for those with selfish preferences. For example, in a centipede game with sharply increasing dollar payoffs, if we ask a subject whether they prefer to "Take" at the current node or have their opponent "Take" at the next node (which reduces their own payoff but significantly improves their opponent's), a sizeable fraction actually prefer the latter outcome. These altruistic players thus have a dominant strategy to "Pass". But then a selfish player who knows these altruists exist might be willing to Pass at early nodes because they know it is reasonably likely Pass back.[2] Thus, Passing is quite rational even for selfish players, and certainly not due to a failure of backward induction. If we flatten the payoff structure, however, the relative benefit to the opponent shrinks when they Take. This reduces the altruistic motivation, and indeed we find fewer altruistic types in this setting. Selfish types seem to know this and, as a result, Take immediately. In these games the standard backwards induction result obtains.

In the prisoners' dilemma, roughly one third of subjects exhibit non-selfish preferences, and one fifth believe their opponent is most likely to have non-selfish preferences. Thus, like the centipede game, it is a Bayesian game with multiple preference types. Once we take preferences and beliefs into account, we find that about half of those who cooperate are doing so rationally.[3] But we also find that a sizeable fraction of subjects behave in a way that is inconsistent with rationality, according to their elicited beliefs and utilities. Specifically, many subjects report utilities such that Defect is a dominant strategy, and yet they still choose to Cooperate. For these subjects their preference over strategies is nonconsequential; it is not driven solely by the outcomes it generates. Instead, they have a direct preference for the act of Cooperating, regardless of its material consequences.

We see other forms of irrationality in $2 \times 2$ games where almost all players have purely selfish preferences. In a dominance solvable game we see many subjects whose beliefs

[2]Selfish players in this game may not be universally selfish. They are identified as selfish (or, more properly, as consistent with selfishness) because their social preferences are not strong enough to change their best response function to be different from that of a truly selfish person.

[3]Following the epistemic game theory literature, "rationality" in this study is defined as maximizing subjective expected utility as measured through elicitation.

indicate that they're quite certain their opponent will follow their dominant strategy, yet they choose not to best respond to that belief. We conjecture that they do so because best responding involves the possibility of a low outcomes, and an irrational form of loss aversion may cause them to deviate from that best response. Similarly, in an asymmetric coordination game we see that players often stubbornly target their more-preferred Nash equilibrium outcome, even though they are nearly certain the opponent will target the other equilibrium. This stubbornness is similarly irrational, given their reported preferences and beliefs.

Our definition of rationality is necessarily quite narrow, assuming two parts: First, that subjects care only about payments to themselves and their opponent, but not about the strategies that led to those payments. We refer to this as *consequentialism*, and discuss in Section VIII why our methodology forces us to make this assumption. Second, it assumes expected utility preferences. Thus, we refer to rationality more properly as *consequentialist expected utility (C-EU) rationality*. The examples of irrationality we observe can therefore be due to failures of consequentialism (for example, preferring to cooperate not because of the outcomes it generates, but because cooperation is perceived as an ethical strategy) or failures of expected utility (for example, ambiguity aversion in the face of strategic uncertainty). In each game we conjecture what we believe to be the cause of the failure, and note that the cause in one game appears quite different from the cause in the next. Thus, we suggest that departures from C-EU-rationality may be highly game-specific. Indeed, we find zero correlation in C-EU-irrational behavior across games, indicating that the various phenomena we observe are not tightly linked.

Finally, we see that failures of C-EU-rationality are significantly reduced when the $2 \times 2$ games are converted to sequential-move games. For example, in the asymmetric coordination game in which each player prefers a different equilibrium, we see players in the simultaneous-move version stubbornly play their own preferred equilibrium strategy, but when the game becomes sequential the second mover almost always (rationally) follows whichever equilibrium the first mover targets. We see this conclusion across all games: the incidence of C-EU-irrationality is greatly reduced when strategic uncertainty is removed.

We discuss related literature in Section II. The analytical framework in which we operate—which motivates our choice of what to elicit and also describes how those things are elicited—appears in Section III. We present our results for the centipede game forms in Section V and the prisoners' dilemma game form in Section VI. We then describe results from two other game forms (a dominance solvable game form and an asymmetric coordination game) in Section VII. We conclude in Section IX with a synthesis of our results and directions for future work.

## II. Related Literature

Weibull (2004) notes that game theory cannot be tested without information about preferences, and is careful to distinguish between games (where payoffs are in utils) and game forms (where payoffs are in dollars).[4] We therefore view our utility elicitation as a potential away around Weibull's criticism, allowing us to observe actual utilities and test game theoretic concepts directly.[5]

To our knowledge, the only other study in which players' preferences over game outcomes are elicited is Brunner et al. (2016). They elicit each subject's ordinal ranking over outcomes and then reveal these rankings to their opponent before playing a game. They find that Nash equilibrium play increases significantly—compared to a baseline in which the ordinal rankings are not revealed—though the minmax and maxmax solution concepts are more predictive than Nash even when preferences are revealed.[6] Although focusing on ordinal preferences simplifies the elicitation procedure, it necessarily limits their analysis to pure best responses. We instead elicit cardinal utility and, as a consequence, can always identify whether or not observed play is rational.

Fischbacher et al. (2001) and Fischbacher and Gachter (2010) elicit players' best response functions in public goods settings and use this to show that the presence of non-selfish players (in particular, imperfect conditional cooperators) drives both contributions to the public good and the decline in contributions over time. Fischbacher and Gachter (2010) also elicit the mode of players' beliefs about the contributions of others, and find that the elicited best response to that modal belief does not perfectly predict play.

There are complementary approaches for studying beliefs and rationality using clever experimental designs rather than direct elicitation. For example, Kneeland (2015) develops a novel "ring game" in which one player has a dominant strategy, the next a best response, the next a best response to that, and so on. She finds that 93% of subjects are rational (they follow their dominant strategy), 71% are rational and believe in rationality, and that these percentages decline significantly for higher orders of belief in rationality. Friedenberg and Kneeland (2023) extend the ring game structure to identify players who have a limited ability to reason about opponents versus those who are able to reason iteratively but have a limited belief in rationality. Calford and Chakraborty

[4]What we call a game form he calls a "game protocol". We prefer "game form" because it is consistent with the mechanism design literature.

[5]Weibull (2004) also discusses how players' preferences might depend on their opponents' preferences (building on Levine, 1998), and players may update their preferences mid-game as they infer their opponent's preferences from their actions. We do not test for this possibility because we elicit utilities only once for each game.

[6]Revealing the elicited data to the opponent creates an incentive for subjects to misrepresent their preferences, but Brunner et al. (2016) find no evidence of such manipulations in their data.

(2022) study a sequential social dilemma and use successively "pruned" versions of the game to explore the effects of higher order beliefs. They find that deviations from subgame perfection are often due to inconsistencies between a player's belief about an opponent and what they believe others believe about that opponent. Since we study only two-player games, this possibility is absent by construction.

Our first game of interest is the centipede game, which was introduced by Rosenthal (1981), named by Binmore (1987), and first studied in the lab by McKelvey and Palfrey (1992). Our focus is on how changing the payoffs can change the preference types of the players, and how that results in drastic changes in strategic behavior among "selfish" players. Several other studies vary the payoffs in centipede games and see similar effects on strategies, though without the elicitation data used in this study. For example, Fey et al. (1996), Kawagoe and Takizawa (2012), Pulford et al. (2017), and Bornstein et al. (2004) consider constant-sum versions of the centipede game, which eliminate efficiency considerations and also feature increasing inequality across terminal nodes. All of these studies find that subjects take earlier in the constant-sum game form.

To try to eliminate any scope for altruism we design a centipede game form in which the player who chooses take wins (essentially) the entire pie, and that pie grows linearly across a player's decision nodes. The vast majority of games end with the first player taking at the first node. Mcintosh et al. (2009) and Krockow et al. (2015) also study "winner-take-all" variants of the centipede game and find that players choose take at earlier nodes, though the differences are not as large as in our experiment.[7]

García-Pola et al. (2020) also vary payments in centipede game forms and find correlations between payments and behavior that are similar to our results. They do not elicit subjects' preferences; instead, they estimate a mixture model containing a wide range of proposed behavioral types, some of which are based on social preferences. The preference-based theories they consider do not explain much of the data; instead, non-equilibrium theories (with selfish preferences) such as quantal response equilibrium and level-$k$ behavior fit better their data.

The patterns of behavior we find having striking similarities with the "gang of four" explanation, pioneered by Kreps et al. (1982) and applied to the centipede game form by McKelvey and Palfrey (1992). According to that theory, selfish players choose to pass in order to convince their opponent that they are actually altruistic, incentivizing the opponent to choose pass as well. This theory has been refuted by Kagel and McGee (2016) and Cox et al. (2015) in the related setting of finitely-repeated prisoners' dilemma games. Our elicitation data in centipede games points to a simpler explanation for early

[7]Cox and James (2012) study a winner-take-all centipede game form with private values and time pressure, meant to emulate features of a clock auction. They find unraveling in the clock format.

cooperation in these games: selfish types believe there are a significant fraction of altruist types in game forms with significant payoff growth through the tree, and therefore choose pass not to trick their opponent, but to take advantage of their altruistic preferences.

There are many other studies of the classic games we study such as the prisoners dilemma, coordination games, and dominance solvable games. For one excellent review (among many), see Camerer (2003).

## III. Analytical Framework

We describe first the framework for two-player simultaneous-move games and game forms. The two players are indexed by $i \in I = \{1,2\}$ and we use the notation $-i$ to refer to $i$'s opponent. There is a set of physical outcomes $X$ that can be paid to the subjects. These are typically dollar payments to each player, so let $X = X_1 \times X_2$, where $X_i$ is the set of possible payments to player $i$. For example, $x = (\$5, \$10)$ is the outcome in which player 1 receives \$5 and player 2 receives \$10. The experimenter chooses a *game form*, which is a tuple $\Gamma = (I, (S_i)_i, \pi)$, where each $S_i$ (for $i \in I$) is the set of strategies available to player $i$ and $\pi : S_1 \times S_2 \to X$ is the outcome function that specifies a physical outcome for each strategy profile $s \in S = S_1 \times S_2$. Let $\pi_i$ denote the projection of $\pi$ onto $X_i$.

The game form is fixed by the experimenter and publicly observable. The players' preferences, strategies, and beliefs, on the other hand, are all private information. We refer to these as the *state* of player $i$. Players form beliefs about the states of their opponent, and the experimenter can use incentive compatible elicitation techniques to elicit the state (or components of the state) from each player.

Formally, a state of player $i$ is a tuple $\omega_i = (u_i, s_i, \vec{p}_i)$. The first component is player $i$'s cardinal utility function $u_i : X \to \mathbb{R}$, defined only over physical outcomes. This is elicited via probability equivalents. Specifically, for any $x$, the value of $u_i(x)$ can be elicited by selecting "good" and "bad" outcomes $\overline{x}$ and $\underline{x}$ such that $u_i(\overline{x}) > u_i(x) > u_i(\underline{x})$ and then finding the probability $q^*$ such that player $i$ is indifferent between $x$ and the lottery $(q^*, \overline{x}; 1-q^*, \underline{x})$, which pays $\overline{x}$ with probability $q^*$ and $\underline{x}$ with probability $1-q^*$. Assuming expected utility and normalizing $u_i(\overline{x}) = 1$ and $u_i(\underline{x}) = 0$, indifference at $q^*$ means that $u_i(x) = q^* \cdot 1 + (1-q^*) \cdot 0 = q^*$. Thus, the indifference probability $q^*$ exactly identifies the cardinal utility. In the lab we elicit $u_i(x)$ for each $x$ in the range of $\pi$ (meaning, for each possible outcome of the game form).[8]

[8] To ensure $u_i(\overline{x}) > u_i(x) > u_i(\underline{x})$, we choose $\overline{x}$ and $\underline{x}$ such that $\overline{x}_i > x_i > \underline{x}_i$ for each $i$ for every $x$ in the range of $\pi$. If in fact $u_i(\overline{x}) < u_i(x)$ or $u_i(\underline{x}) > u_i(x)$ (perhaps because of inequality aversion) then we would observe $q^* = 1$ or $q^* = 0$, respectively. This occurs rarely in our data.

Recall that $X = X_1 \times X_2$, so player $i$'s utility $u_i(x_1, x_2)$ can depend on both players' payoffs. When $X_i \subseteq \mathbb{R}$, player $i$ is said to be *consistent with selfishness in* $\Gamma$ (or, simply, *selfish in* $\Gamma$) if $x_i' > x_i$ implies $u_i(x') > u_i(x)$ for all $x', x$ in the range of $\pi$. It is possible for someone to be consistent with selfishness in some games, but not others. For example, someone may be non-selfish in a public goods game where the cost of contributing is low, but consistent with selfishness when the cost of contributing is high. In other words, "selfishness" a statement about a player's preferences only in a given context, and is not an indictment of their behavior globally. Our preference elicitation exercise will allow us to measure those games in which people tend to exhibit selfishness and those in which they do not.

The second component of player $i$'s state is their pure strategy choice $s_i$. Players in this framework do not choose mixed strategies. Instead, "mixing" happens in players' uncertainty about their opponents' pure strategy choices. For example, in matching pennies Ann might believe there is a 50% chance Bob is in a state where he plays Heads, and 50% he's in a state where he plays Tails. This is the perspective of Aumann (1987), who views mixed strategy Nash equilibrium as a property of players' beliefs about each other, rather than their actual play of the game.[9]

The last component of a state identifies these beliefs over strategies, as well as beliefs over utilities, beliefs over beliefs, and so on. Let $p_i^1(u_{-i}, s_{-i})$ be player $i$'s first-order belief about $u_{-i}$ and $s_{-i}$. This belief allows for correlation between $u_{-i}$ and $s_{-i}$, which is important since player types with different utilities would rationally choose different strategies. Player $i$ also forms beliefs about their opponent's first order belief $p_{-i}^1$, so let $p_i^2(p_{-i}^1, u_{-i}, s_{-i})$ be $i$'s second-order belief.[10] An entire infinite hierarchy of beliefs $\vec{p}_i = (p_i^1, p_i^2, p_i^3, \ldots)$ can thus be constructed.

For simplicity we only elicit players' marginal beliefs. This helps simplify an already-complicated design, but at the cost of potentially limiting how much we can learn about players' belief in rationality in game forms where they believe their opponent might have more than one utility function. Let $p_i^{1s}(s_{-i})$, $p_i^{1u}(u_{-i})$, and $p_i^{2p}(p_{-i}^1)$ denote the respective marginal distributions over $s_{-i}$, $u_{-i}$, and $p_{-i}^1$.

[9] There is no loss of generality, however, if players explicitly mix. In that case, define the states of the player conditional on the realization of their mixed strategy. For example, if a player flips a coin to pick their strategy then one state would identify the player's preferences, strategy, and beliefs when the coin lands Heads, and another would identify their preferences, strategy, and beliefs when the coin lands Tails. Opponents' beliefs would presumably view these two states as equally likely.

[10] It is necessary that $p_i^2$ also include beliefs over $u_{-i}$ and $s_{-i}$—despite the redundancy with $p_i^1$—to capture any believed correlation between $s_{-i}$ and $p_{-i}^1$. This correlation is natural: player types with different beliefs would rationally choose different strategies. Normally we would assume *coherency*—meaning the marginal of each $p_i^k$ over $u_{-i}$ and $s_{-i}$ agrees with that of $p_i^{k-1}$—and common knowledge of coherency, but we do not elicit enough data to test this assumption. See Dekel and Siniscalchi (2015, pp.625–626) for details.

To elicit $p_i^{1s}(s_{-i})$ we simply find the probability $q^*$ such that $i$ is indifferent between an act which pays outcome $\overline{x}$ if $-i$ plays $s_{-i}$ (and $\underline{x}$ otherwise) and a lottery that pays $\overline{x}$ with probability $q^*$ (and $\underline{x}$ otherwise). Because $p_i^{1u}$ and $p_i^{2p}$ have much larger domains, we elicit only the mode of these distributions by having the player announce their best guess of the elicited values of $u_{-i}$ and $p_{-i}^{1s}$, respectively, and paying them $\overline{x}$ if their guess is correct.[11]

Given the data we have available, our notion of rationality must subsume two stronger concepts: consequentialism and expected utility. Consequentialism is the idea that players care about the payoffs they both receive, but not the strategy choices that led to those payoffs. And expected utility (EU) describes a player whose strategy choice is a best response to their first-order beliefs, given utility $u_i$. Formally, if $BR_i(p_i^{1s}|u_i) = \arg\max_{s_i \in S_i} \sum_{s_{-i}} p_i^{1s}(s_{-i})\, u_i(\pi(s_i, s_{-i}))$ is $i$'s EU best response correspondence given their elicited $u_i$, then player $i$ is said to be *consequentialist expected utility rational (or, C-EU-rational) at state* $\omega_i = (u_i, s_i, \vec{p}_i)$ if $s_i \in BR_i(p_i^{1s}|u_i)$. Otherwise, they are *C-EU-irrational* at $\omega_i$.

To understand the necessity of studying C-EU-rationality, consider instead if we applied the weaker notion of *rationality*, with neither consequentialism nor expected utility added. To define the concept formally, first note that choice objects are strategies $s_i \in S_i$, which can be thought of as Savage-style acts since each $s_i$ maps an unknown state $\omega_{-i} = (u_{-i}, s_{-i}, \vec{p}_{-i})$ into a consequence $\pi(s_i, s_{-i})$. If we assume the player has a complete and transitive preference over the (finite) strategy space then it can be represented by a utility function $U_i : S_i \to \mathbb{R}$. Player $i$ is *rational at* $\omega_i$ if $U_i(s_i) \geq U_i(s_i')$ for all $s_i' \in S_i$.[12] C-EU-rationality requires that, in addition, $U_i(s_i) = \sum_{s_{-i}} p_i^{1s}(s_{-i}) u_i(\pi(s_i, s_{-i}))$.

While it would be ideal to study rationality instead of C-EU-rationality, we claim that $U_i$ cannot be elicited in an incentive compatible way without deception. For example, if we elicited player $i$'s value (in terms of either dollars or probabilities) for playing $s_i$ in game $g$, and if that elicitation is incentivized, then with some chance player $i$ must be forced to play $s_i$ in game $g$ so that its outcome can be paid to the subject. But if the opponent knows this then their beliefs about $s_i$ being played will clearly be altered, as will $i$'s second-order beliefs, and so on.[13] Preferences over outcomes, however, can be elicited since the payments do not depend on any decisions made by the other player.

[11] We also elicit their belief at the mode by finding the probability $q^*$ that makes the subject indifferent between an act that pays $\overline{x}$ if their guess is correct and a lottery that pays $\overline{x}$ with probability $q^*$. Note that the expected payment of this procedure is maximized when their guess is the mode, so it remains incentive compatible for them to guess the mode.

[12] In this more general framework $U_i$ would be included as a component of $\omega_i$.

[13] Since $U_i$ is ordinal one might consider testing rationality by testing Sen's conditions $\alpha$ and $\beta$, for example. But adding or deleting strategies from a game necessarily changes the game, so such consistency should not be expected.

Thus, we are restricted to eliciting $u_i(\pi(s_i,s_{-i}))$ instead of $U_i(s_i)$, which then means that we can only test rationality under the joint hypotheses of rationality, expected utility, and consequentialism.

It is also possible to measure players' belief about the C-EU-rationality of their opponent. Letting $R_{-i}$ be the set of states $\omega_{-i}$ for which $-i$ is C-EU-rational, we can elicit player $i$'s belief that $\omega_{-i} \in R_{-i}$. This is done by finding the probability $q^*$ at which they are indifferent between an act that pays $\overline{x}$ if $\omega_{-i} \in R_{-i}$ (and $\underline{x}$ otherwise) and a lottery that pays $\overline{x}$ with probability $q^*$ (and $\underline{x}$ otherwise). For payment, the realization of whether or not $\omega_{-i} \in R_{-i}$ is taken from player $-i$'s elicitation data.[14]

At state $\omega = (\omega_i,\omega_{-i})$, the (Bayesian) game induced by $\Gamma$ is $G(\omega) = (I,S,(u_i \circ \pi)_{i\in I},(\vec{p}_i)_{i\in I})$, where $u_i(\pi(s_i,s_{-i}))$ is $i$'s utility over strategy profiles (rather than outcomes) at state $\omega_i = (u_i,s_i,\vec{p}_i)$.[15] The experimenter selects the game form $\Gamma$, and $\Gamma$ is common information among all participants, but the experimenter cannot observe the actual game $G(\omega)$ without eliciting players' utilities and beliefs.

We also study the 6-node centipede game form, so we need to generalize our framework to apply to extensive-form games. To do so, define histories of the form $h^t = (a^1,a^2,\ldots,a^{t-1})$, where $t \leq 6$ is the index of the current decision node and, for each $t' < t$, $a^{t'} \in \{T,P\}$ is the action (either $T$ake or $P$ass) chosen by the active player at decision node $t'$. Strategies map histories into actions ($T$ or $P$). Let $\{z_1,\ldots,z_7\}$ denote the 7 terminal histories of the game, where $z_1 = (T)$, $z_2 = (P,T)$, $z_3 = (P,P,T)$, and so on. Assuming consequentialism, player $i$'s utility of the game ending at $z_t$ is given by $u_i(\pi(z_t))$.[16]

In the centipede game beliefs form a conditional probability system (see Myerson, 1991), where $p_i^k(\cdot|h^t)$ denotes $i$'s $k$th order belief when the game play has reached history $h^t$. We can therefore elicit strategies (complete contingent plans), utilities, and beliefs at every realized history, including at the initial history $h^1$.[17] First-order beliefs are now over complete contingent plans, but in the centipede game these are easily elicited at any history $h^t$ by asking the player the probability with which their opponent plans to

[14]This requires that we assume that $i$ believes the elicitation procedures are incentive compatible for $-i$. The experimental instructions therefore emphasize incentive compatibility multiple times.

[15]Specifically, this is the game induced under the assumption of consequentialism. However we do not impose a common prior. A common prior would be a distribution $\hat{p}$ over $u$ and $s$ such that each $p_i^1$ is the posterior conditional on observing $u_i$ and $s_i$. Higher-order beliefs would then be defined as above. In that framework each $\vec{p}_i$ would be completely determined by $u_i$ and $s_i$, but our data is rich enough to allow this to be violated.

[16]This is slightly overloaded notation since $\pi$ was originally defined on strategy profiles, not histories. As is standard, we assume that if $s$ and $s'$ both lead to terminal history $z_t$ then $\pi(s) = \pi(s')$; this outcome is then denoted by $\pi(z_t)$ for simplicity.

[17]This allows that a subject's strategy—and therefore their rationality—might change from one history to the next. In the theoretical literature strategies are almost always assumed to be unchanging within a game.

choose Take at each remaining decision node. This belief is then compared to the actual plan reported by the opponent to determine the player's payment.

Many methodological issues arise when eliciting these variables. We defer discussion of these issues—and the detailed description of our elicitation techniques—to Section VIII, after the results.

## IV. Experimental Design

We report three experiments in which subjects play multiple game forms. In the first, which we denote by CENT, subjects play a fixed six-node centipede game form four times against different opponents and with feedback. We perform elicitation only in the last two plays of the game, after subjects have had some opportunity to learn. We report three different treatments, CENT-LO, CENT-HI, and CENT-ALL, that differ only in their payoffs. Each subject participated in only one of the three treatments.

In the second experiment, denoted by SIM, subjects play five different simultaneous $2\times 2$ game forms one time each, without feedback, and with elicitation performed in each game.

In the third experiment, denoted SEQ, a new group of subjects play sequential-move versions of SIM game forms. Specifically, the row player chooses an action in the first stage and then the column player, upon observing the row player's action, chooses an action in the second stage. Again we perform elicitation in every game.

Our elicitation procedure is the same regardless of the game form, and regardless of whether it is a simultaneous-move game form or a multistage game form. For each $i$, at the initial history $h^1$ we elicit

(1) $u_i(\pi(S))$ (cardinal utilities for all outcomes in the game form),
(2) $\arg\max_{u_{-i}} p_i^{1u}(u_{-i}|h^1)$ (the mode of $i$'s initial belief about $-i$'s utility), and
(3) $\max_{u_{-i}} p_i^{1u}(u_{-i}|h^1)$ (the density at the mode of that belief about $u_{-i}$).

At every non-terminal history $h^t$ (including the initial history) we also elicit from all players, regardless of whether they are active or not,

(4) $s_i$ ($i$'s chosen strategy, expressed as a complete contingent plan),
(5) $p_i^{1s}(s_{-i}|h^t)$ ($i$'s belief distribution over $s_{-i}$),
(6) $\arg\max_{p_{-i}^{1s}} p_i^{2p}(p_{-i}^{1s}(\cdot|h^t)|h^t)$ (the mode of $i$'s current belief about $-i$'s current belief about $s_i$),
(7) $\max_{p_{-i}^{1s}} p_i^{2p}(p_{-i}^{1s}(\cdot|h^t)|h^t)$ (the probability of that modal belief), and
(8) $i$'s current belief that $-i$ is rational.[18]

[18]Recall that in single-stage games, the only non-terminal history is the initial history. According to the framework, $s_i$ and $u_i$ should not vary across histories. We measure $s_i$ at each history to see if in fact it is stable. We measure $u_i$ only at the initial history. In the centipede game form, if player $i$'s action at $h^t$

We describe in Section VIII how we elicit each of these objects in an incentive compatible way.

Subjects in the CENT experiment (centipede game forms) interacted anonymously via a custom-built computer interface. The game tree was visible during all elicitation questions. For example, when eliciting cardinal utilities for outcomes, the subject filled in their cardinal utility for each outcome directly below that outcome on the computer screen. After entering utilities for all seven outcomes, the computer then showed a table with the outcomes ranked from best to worst according to the reported utilities, and the subject was asked to confirm that the ranking of outcomes was as they prefer.

We did not use the strategy method; subjects' elicited strategies at each node (which are complete contingent plans specifying at what node they plan on choosing Take, if ever) determined whether the game proceeded to the next node or not. Subjects learned whether their opponent chose Pass or Take after each node, but nothing else until the experiment was finished. At the end of the experiment the subjects were given 16 binary choices between gambles designed to estimate their risk and ambiguity attitudes (Holt and Laury, 2002).

The treatments CENT-LO, CENT-HI, and CENT-ALL simply vary the payoffs in the game form. Roughly, they differ in how much money is at risk (relative to how much can be gained) by choosing Pass at a given node. This measures how costly it is to be altruistic: If potential losses are small if the opponent Takes at the next node then a modestly-altruistic subject might even prefer that outcome over Taking themselves, since the opponent benefits quite a lot from this outcome. CENT-LO is a low-risk, low-cost treatment. CENT-HI and CENT-ALL drastically increase the risk and cost of altruism. Exact payoffs for each are shown in the next section. The number of subjects in each treatment were 54, 36, and 62, respectively.[19]

Subjects in the SIM and SEQ experiments ($2 \times 2$ game forms) were given a printed booklet of seven pages. Each of the first five pages showed a game form at the top, followed by the eight elicitation questions immediately below. For example, the third page of the booklet showed the Prisoners' Dilemma game form at the top and all elicitation questions below. In that sense, game play and elicitation were effectively simultaneous since subjects could work through the pages in any order they wish. The specific game forms are shown in the results section. The sixth and seventh pages contain 16 individual binary decisions intended to measure the subject's risk aversion and ambiguity

terminates the game, then player $i$ knows that they will not observe any further components of $s_{-i}$. In that case we do not elicit $p_i^{1s}(s_{-i}|h^t)$ or $i$'s belief in $R_{-i}$ since the elicitation would not be strictly incentive compatible.

[19]Two other treatments were run with moderate costs. These are described in the online appendix. As expected, results lie "between" CENT-LO and CENT-HI.

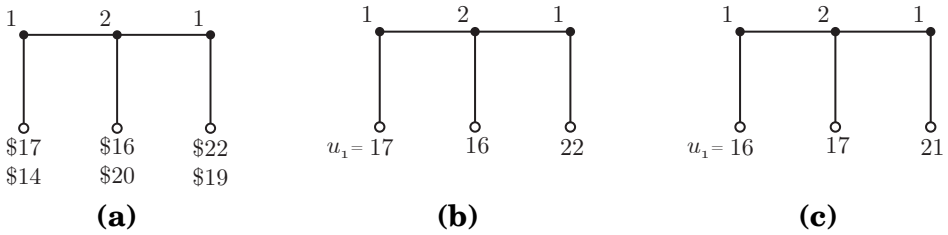



FIGURE I. (a) A three-node segment of a centipede game form. (b) The game with 'selfish' utilities. (c) The game with other-regarding preferences.

aversion. Each subject filled in their answers to all questions on all pages and turned it in to the experimenter. The experimenter then matched each booklet with that of the corresponding opponent and calculated payments.

The SEQ experiment is identical to SIM, except that row players moved first and column players second. Specifically, row players were asked to enter each strategy choice into a computer terminal immediately after making it, and the computer then transmitted the choice anonymously to the corresponding column player's computer. Column players were instructed to wait until they could see their row player's choice in each game before filling out that page of their own printed booklet.

At the end of each experiment subjects were paid for only one randomly-selected decision (Allais, 1953). This method is incentive compatible assuming subjects' preferences over gambles satisfy monotonicity, which is strictly weaker than expected utility when reduction of compound lotteries is not assumed; see Azrieli et al. (2018) for details. Subjects for all experiments were recruited via ORSEE (Greiner, 2015) from a database of potential subjects at Ohio State University. One hundred fifty subjects participated in the SIM experiment, and sixty four in SEQ. Instructions, screenshots, and booklets are all available in an online appendix.

Finally, we ran follow-up experiments to test whether the presence of elicitation affects subjects' strategy choices. The details of these experiments are relegated to Section VIII, but the general result is that elicitation had a significant impact on behavior in only one game: the asymmetric coordination game from the SIM treatment. In all other games we found no significant differences; see Figure XIV and Table XIII in Section VIII for the complete results.

## V. The Centipede Game as a Bayesian Game

### *An Illustration of the Main Result*

To understand the incentives in a centipede game form—and to preview our results—consider an arbitrary three-node segment of a centipede game form, shown in panel (a) of Figure I. A three-node segment is simply the smaller centipede game form created by taking three consecutive decision nodes from a larger centipede game and removing the option to Pass at the third node. The one in Figure I shows decision nodes three, four, and five from CENT-LO. Panel (a) shows the actual game form, with player 1 moving first and player 1's payoffs shown above player 2's payoffs at each terminal node. Panel (b) shows example utilities for a hypothetical player who is consistent with selfishness. Panel (c) shows example utilities of a (hypothetical) non-selfish player who exhibits some degree of altruism.

Now consider the incentive of player 1 to choose Pass at the root node of this segment. If player 1 has the selfish utilities from panel (b) and believes the second mover will Pass with probability $p$, then their best response is to Pass at the first node if and only if $p \geq (17-16)/(22-16) = 1/6$. Thus, their "basin of attraction" for Pass—the set of beliefs such that Passing is a best response—is the interval $[1/6, 1]$. We refer to the size of this basin as *SizeBAP*, which here equals 5/6.[20] Roughly speaking, *SizeBAP* provides a measure of how tempted a player may be to Pass.[21] If Passing is not too risky (in terms of utilities, not dollars), then the *SizeBAP* will be large.

For the altruistic player 1 in panel (c), Pass is a strictly dominant strategy. Intuitively, they are willing to sacrifice \$1 to give their opponent \$6, and so they face no temptation to choose Take. Their *SizeBAP* is therefore 1.00. Importantly, this subject is *not playing a centipede game*. Instead, they are playing a game with a dominant strategy to Pass. If we observe them choosing Pass, we should *not* conclude that backwards induction has failed. This highlights the importance of measuring preferences in these game forms.

Finally, consider again the selfish player 1 from panel (b) playing the three-node game form against a pool of subjects in which Pass is a dominant strategy for 1/3 of their opponents. Recall that they should choose Pass if $p > 1/6$. But since $p \geq 1/3$, it is rational for this selfish player to Pass at this node.

Despite being selfish, this subject is *also not playing a centipede game*. They are playing a Bayesian game with heterogeneous utilities. Consequently, their choice of Pass in early nodes should also not be viewed as a failure of backwards induction.

[20] This is inspired by a similar measure analyzed by Dal Bó and Fréchette (2011) for repeated prisoners' dilemmas.

[21] In the larger game with more than three nodes, *SizeBAP* is only an approximate measure of the true temptation to Pass.

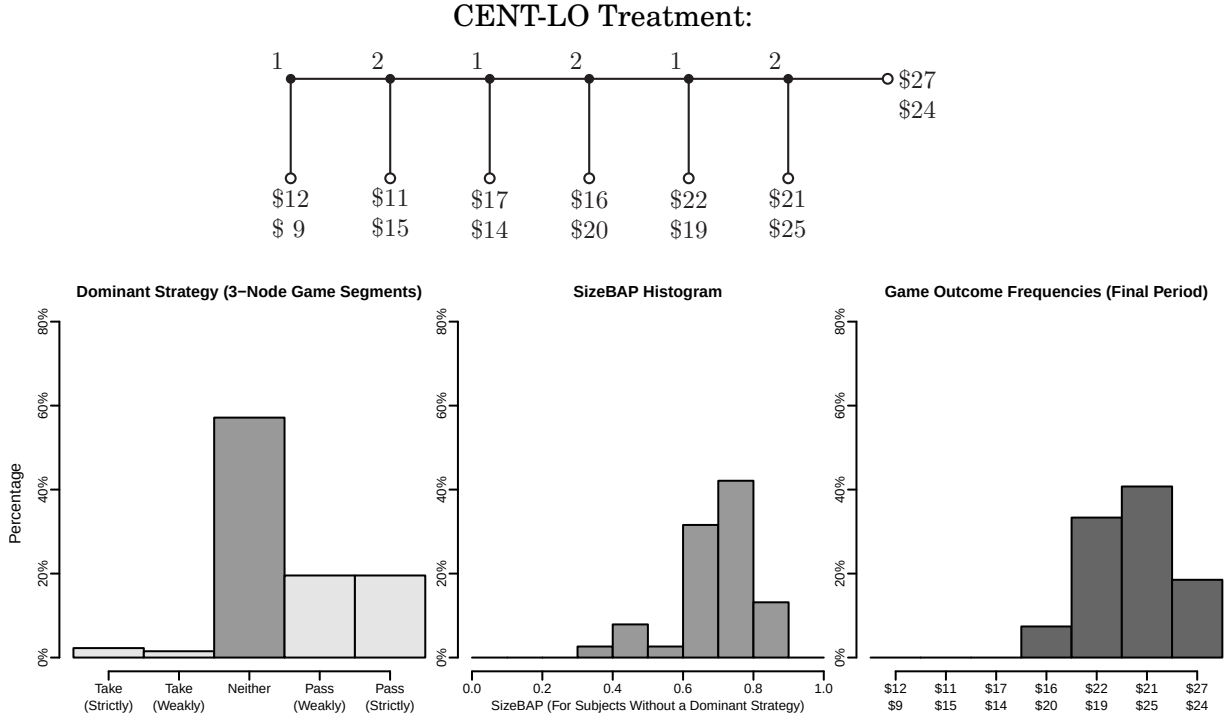



FIGURE II. The CENT-LO treatment. Top: The game form. Bottom left: Across all 3-node game segments, the percentage of subjects who have a dominant strategy to Take or Pass (or neither). Bottom middle: The temptation to Pass for subjects without a dominant strategy, as measured by *SizeBAP*. Bottom right: Actual outcome frequencies.

This simplified example illustrates our main finding: We see non-selfish preferences in much of our data, these subjects tend to choose Pass, and the presence of these non-selfish types induces selfish players to Pass as well in early nodes.

Furthermore, if we increase the risk and cost of choosing Pass (as in CENT-HI and CENT-ALL) then fewer subjects will be altruistic. Selfish subjects, knowing this, will no longer Pass at early nodes and the game will end almost immediately. This is exactly what we observe. Indeed, the results are quite consistent with backwards induction.

### *Types and Beliefs in the Three Main Treatments*

Figure II shows the game form and results for the CENT-LO treatment. The top panel displays the game form. The risk to Passing is relatively low: In any 3-node segment, choosing Pass risks \$1 (if the opponent Takes at the next node) to gain \$5 (if the opponent Passes and the player subsequently Takes). The *SizeBAP* based on selfish dollar payments is therefore $5/6 \approx 0.83$ at every node.

Fifty-four subjects participated in this treatment across three sessions, earning an average of \$19.93. We take the elicited utilities for each subject and calculate for each

3-node segment whether they have a dominant strategy to Pass, a dominant strategy to Take, or neither. We do this for all 3-node segments, but only in the last period of play. The resulting histogram (combining all 3-node segments) is shown in the bottom left panel. Over 40% of subjects have a dominant strategy, and the vast majority of those have a dominant strategy to Pass. Thus, we see a substantial incidence of altruism-like preferences in this game form.[22]

In the middle histogram we take those subjects who have no dominant strategy and calculate the *SizeBAP* measure for each 3-node segment. Again, the larger the *SizeBAP*, the more the selfish subject is willing to Pass. The *SizeBAP* here is typically large, often nearly as high as the dollar-based measure of 0.83. Given that 40% of subjects have a dominant strategy to Pass, we should expect that many of these subjects will best respond by choosing Pass as well.

The histogram of actual game outcomes (for the final period) is shown in the bottom right panel. In no case does any player Take in the first three nodes. The modal outcome is for players to Pass until the very last decision node, and then Take. This behavior is not paradoxical, however; it is easily rationalized given the preferences we observe. Again, this is a Bayesian game, and the selfish types are willing to Pass because (1) it is not very risky (the *SizeBAP* is large), and (2) they correctly believe there are non-selfish types who will Pass. Indeed, we see direct evidence of non-selfish types: conditional on the game reaching the final decision node the last mover chooses Pass 31% of the time.

Next we consider a centipede game form in which the risk to Passing is much higher. In the CENT-HI treatment (shown in Figure III) Passing risks \$2 to gain only \$1. The resulting *SizeBAP* based on dollar payoffs is 1/3. Examining elicited preferences, we now see that far fewer subjects have a dominant strategy, and among them a slight majority have a dominant strategy to play Take. Of the 71% of subjects who have no dominant strategy, most have a low *SizeBAP*. Most subjects' utilities are such that they would need at least a 50% belief that the opponent will choose Pass in order to rationalize Passing, but the distribution of actual types simply doesn't justify that belief. Indeed, the game typically ends much earlier as a result. In fact, over 2/3 of games end in the first two nodes.

Finally, we test a winner-take-all version of the centipede game form in our CENT-ALL treatment (Figure IV).[23] Here, Passing risks almost the entire payoff for a gain of \$4. For a selfish subject with dollar-based utilities the *SizeBAP* is 0.22 at their first

[22]This histogram omits segments at which players reported complete indifference (0.7%) or "reverse" preferences for which $u_i(\pi(z_t)) < u_i(\pi(z_{t+1}))$ but $u_i(\pi(z_{t+2})) < u_i(\pi(z_t))$ (0.7%). It also excludes the last decision node, which corresponds to a 2-node game segment. For that segment 74% of last movers reported selfish preferences, 15% (4 of 27) reported indifference, and 11% (3 of 27) reported strictly altruistic preferences.

[23]This type of centipede game was proposed by Reny (1993). In similar game forms Danz et al. (2016) and Krockow et al. (2015) find higher Pass rates than we do here, though Cox and James (2012, 2015) find

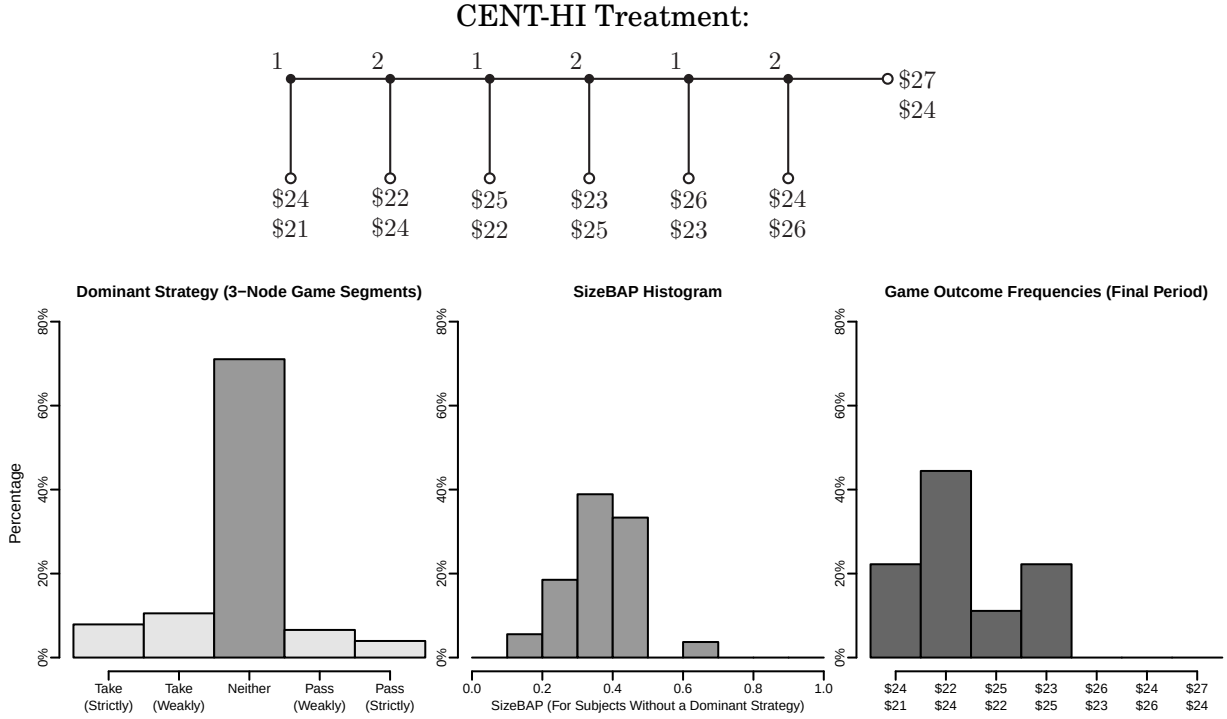



FIGURE III. The CENT-HI treatment.

decision node, 0.18 at their second, and 0.15 at their third. Looking at elicited utilities, we see that over 80% have no dominant strategy in this game form. Thus, this is reasonably close to a complete-information game with selfish preferences.[24] Even the *SizeBAP* measures are close to the selfish dollar-based values of 0.18–0.22. Arguably this game form provides the best test of a true complete-information centipede game with little to no social preferences. And the predictions of backwards induction (and extensive form rationalizability) are largely confirmed: The first mover plays Take in 68% of games, and no game proceeds beyond the third node.[25]

### *Utility, Rationality, and Accuracy of Beliefs*

In this section we carefully confirm the details of the above story by looking deeper into the elicited beliefs, rationality, and beliefs about rationality. Here we include data from both periods in which elicitation data is collected.

similar results to ours. There are multiple design differences between these studies, making it difficult to pinpoint which cause lower Pass rates.

[24]Again, "selfish" here doesn't mean the players are universally selfish. It just means that, in this particular game form, the cost of improving the opponent's payoff is too great to be worthwhile for most subjects.

[25]Recall this is for the fourth period only, but even in the first period 68% of first movers chose Take at the first node. This drops to 52% in period 2, 55% in period 3, and back to 68% in period 4. Across all four periods, 94% of games ended in the first three nodes.

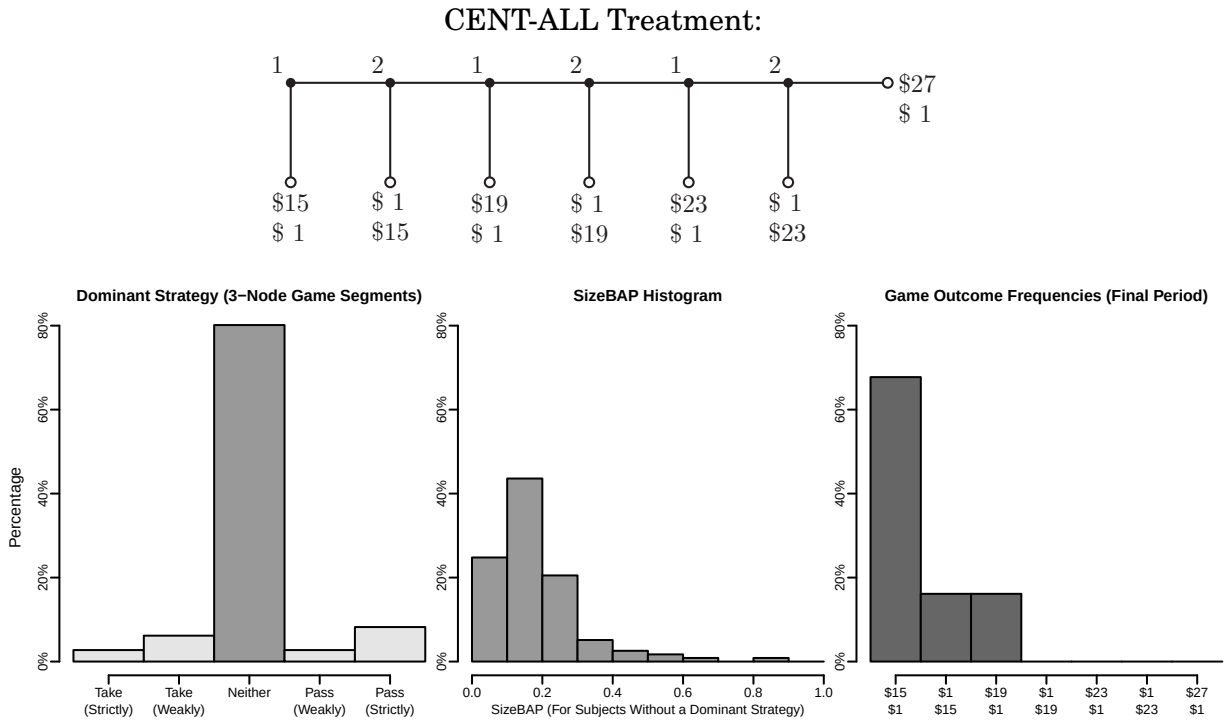


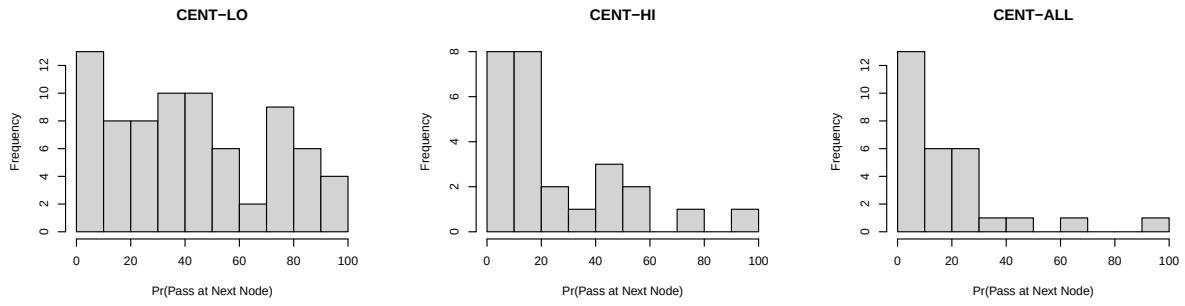

FIGURE IV. The CENT-ALL treatment.

The fist step is confirming the existence of players whose reported utility indicates a dominant strategy to pass in a given 3-node segment. Specifically, a player $i$ has a dominant strategy to pass in a 3-node segment starting at node $t \leq 5$ if $u_i(\pi(z_t)) \leq u_i(\pi(z_{t+1}))$ and $u_i(\pi(z_t)) \leq u_i(\pi(z_{t+2}))$, with one inequality strict. At node 6 (the last decision node) the criterion is simply that $u_i(\pi(z_6)) < u_i(\pi(z_7))$. For expositional simplicity we will refer to such a subject as an "altruist" at that node, noting that this is a very situation-specific definition since they may not be labeled as an altruist at other nodes or in different game forms.

In CENT-LO we observe altruist subjects in 43.7% of the 3-node segments. This drops to 8.9% in CENT-HI ($\chi^2$ $p$-value $< 0.001$) and 8.7% in CENT-ALL ($\chi^2$ $p$-value $< 0.001$, with no significant difference between CENT-HI and CENT-ALL). Thus, there is a substantial fraction of subjects who would prefer to pass altruistically when the cost of doing so is low, but not when that cost is increased.[26]

Do these altruists actually choose to pass? We look at those 3-node segments where the active player has reported altruist utilities (as defined above). In CENT-LO there

[26]Interestingly, at node 6 the percentage of altruists drops to 15% in CENT-LO. but in CENT-HI and CENT-ALL it rises to 14% and 18%, respectively. A far more stringent definition of an altruist is that "Always Pass" be a dominant strategy of the entire game. We observe such preferences in 29.6% of subjects in CENT-LO (pooled across periods 3 and 4), 4.2% of subjects in CENT-HI, and 8.9% of subjects in CENT-ALL.




(a) (b) (c)

FIGURE V. Histograms of non-altruists' belief that the opponent will Pass at the next node.

are 117 such observations, and the subject chooses Pass in 89.7% of them. Thus, their actions are largely consistent with their reported utilities. In CENT-HI and CENT-ALL there are far fewer subjects reporting such utilities. In CENT-HI we observe Passing in 6 of the 9 segments with altruist preferences, and in CENT-ALL we observe Passing in only 2 of the 12 segments with altruist preferences.[27] We conjecture that the fairly high rates of C-EU-irrationality for these altruist types is driven by the presence of a handful of noisy subjects: If every treatment has a certain percentage of subjects with noisy utility reports, but in CENT-ALL no subject is truly altruistic, then the few observations of altruistic utility reports would all be "false positives," and we might expect high rates of C-EU-irrationality from these noisy subjects.[28]

We have established that in CENT-LO there is a sizable fraction of altruists (43.7%), and most of the time they do choose to pass (89.7%). Players who recognize this should therefore expect Passing in at least 39.2% of interactions. Now we ask whether non-altruists' beliefs are consistent with this observation. In CENT-LO, when we consider non-altruists and look at their belief that the opponent will Pass at the next node, we find that 54.8% of them have a belief greater than 39.2%. As Figure V shows, however, beliefs are significantly lower in CENT-HI and CENT-ALL, consistent with the lower incidence of altruistic types.[29]

A natural question is whether this belief in Passing correlates with the player's belief that their opponent's utility is altruistic. Unfortunately, our data are not rich enough to

[27]Despite the small sample size, Passing in CENT-ALL is significantly lower than in CENT-LO, with a $\chi^2$ $p$-value of 0.038.

[28]If we define altruists as having "Always Pass" be a dominant strategy of the entire game, then we see that 56% of those announce an initial plan of Always Pass in CENT-LO, while 0% choose this plan in either CENT-HI or CENT-ALL. This is also consistent with a constant fraction of noisy subjects across treatments.

[29]Figure V looks very similar if we restrict attention to those with selfish preferences. This is because selfish types constitute 88% of all non-altruists.

address this question. The reason is that we only elicit the player's most-likely utility values for their opponent. And most players exhibit a self-similarity bias: they believe their opponent will have similar utilities to themselves. For example, those who have selfish preferences guess that their opponent has selfish preferences in 78% of 3-node segments, while those who are altruistic guess that their opponent has altruistic preferences in 67% of 3-node segments.[30] But that's only their modal vector; it could be that selfish types think the second-most-likely vector is altruistic, or the set of all altruistic vectors has fairly high probability. Since we don't observe the entire distribution of beliefs, we cannot measure whether this is the case.

Given the self-similarity result, an alternative hypothesis arises: Perhaps the selfish types in CENT-LO do *not* believe there are altruists, but instead believe their opponents are also selfish but play irrationally. To test this, consider player 1 at node 5. If they believe their opponent is a selfish type who will irrationally choose Pass then their belief in rationality should negatively correlate with their belief in Pass. Instead, we find a positive correlation coefficient of 0.41 ($p$-value of 0.11), which is inconsistent with a belief in irrational selfish types. It is consistent with a belief in altruistic types, for whom Passing is rational. The estimated correlation is insignificantly different from zero, however, so the support for a belief in altruism is present but not strong.

We can further test for a belief in altruism by moving back to player 2 at node 4. If they believe player 1 is selfish, they'll say: "If player 1 is rational and selfish and thinks I'm likely to Take at node 6, then they'll Take at node 5." If player 2 believes player 1 is altruistic and rational, however, then they'll say: "Regardless of their belief, they will Pass at node 5." To test this, we first correlate player 2's belief in rationality at node 4 with their belief that player 1 will Pass at node 5. Again we find a weak positive relationship (Pearson coefficient of 0.21 with a $p$-value of 0.21), which is suggestive of a belief in altruism: if player 1 is more likely to be rational, they're more like to Pass.

We've established that selfish types have reasonably accurate first-order beliefs about the strategies of others, and find some suggestive evidence that this may come from a belief that a fraction of others are altruistic. The last question then is whether players best respond to their first-order beliefs about others' strategies. We measure this in two ways: First, we take the more stringent approach and ask whether the player's entire strategy at a node—their complete contingent plan—is C-EU-rational, given their elicited utilities and current beliefs. Second, we look at 3-node segments and ask whether the current action of Take or Pass is optimal within that 3-node segment. Note that under either measure we cannot observe whether choosing Take at the first node is rational

[30]Here we define someone to be selfish or altruistic if they have that preference in at least two of their three 3-node segments. If we restrict to those that have selfish or altruistic preferences in all 3-node segments then the self-similarity percentages are 91% for selfish types and 56% for altruistic types.

or not, because we cannot elicit beliefs about the continuation game from a subject who has just chosen to end the game at the first node. A similar problem arises if a subject chooses Take at a later node; in those cases we use the subject's reported belief from the previous node as a proxy for their current belief.[31]

Looking at rationality of the entire strategy, we find that subjects are C-EU-rational in 58.0% of observations in CENT-LO, 49.4% in CENT-HI, and 48.1% in CENT-ALL ($\chi^2$ $p$-values are 0.048 for CENT-LO vs. CENT-HI, 0.027 for CENT-LO vs. CENT-ALL, and 0.811 for CENT-HI vs. CENT-ALL).[32] We do find that the altruists are generally more C-EU-rational than non-altruists in CENT-LO (64.1% vs. 46.3%; $\chi^2$ $p$-value 0.004). In the other two treatments altruists are too rare to find significant differences across groups.

If we focus only on 3-node segments, C-EU-rationality increases to 83.8% for CENT-LO, 58.5% for CENT-HI, and 38.3% for CENT-ALL (pairwise $\chi^2$ test $p$-values are $<$ 0.001, $<$ 0.001, and 0.017). However, the sample size for measuring rationality in CENT-ALL is quite small because most subjects Take at the first node: Only 60 3-node segments originate beyond the first node, compared to 266 in CENT-LO. Thus, the observations of irrationality we observe in CENT-ALL may be driven by some fixed percentage of subjects whose reports and decisions are especially noisy, regardless of treatment. Indeed, if we look at the 3-node rationality of those who Pass, it is quite high in CENT-LO, where there are many observations (211 of 244, or 86.5%), but low in CENT-HI (23 of 54 or 42.6%) and CENT-ALL (8 of 36, or 22.2%), where there are few observations.[33]

In summary, our elicitation data suggest that most centipede game forms induce Bayesian games. There are subjects whose preferences are consistent with selfishness, but also those with altruistic preferences. A larger presence of altruistic types gives a clear incentive for selfish types to choose Pass early in the game. But, if the payoffs are changed so that the riskiness of Passing is increased then more players become consistent with selfishness, and so the selfish types no longer have an incentive to Pass.[34] Consequently, players Take earlier.

[31]Beliefs from the previous node correlate with the current node with a correlation coefficient of 0.722. The average belief change is only -1.34 percentage points, though the average absolute change is 9.67 percentage points. In other words, beliefs do fluctuate to some degree, but the fluctuations are roughly mean-zero.

[32]The number of observations is significantly lower in CENT-HI and CENT-ALL since many subjects choose Take at the first node, in which case rationality cannot be measured.

[33]Similarly, altruists are rational very frequently in CENT-LO, best-responding in 92.3% of 3-node segments. This is compared to 77.2% for non-altruists. Altruists are so infrequent in the other two treatments (three 3-node segments in CENT-HI and two 3-node segments in CENT-ALL) that their rationality is not meaningfully measured.

[34]Recall that becoming consistent with selfishness does not necessarily mean their preferences have changed. It just means that, with these payoffs, their social preferences are not strong enough to push their best responses away from the selfish best response.

Much of the existing literature analyzes centipede game forms as though they induce complete-information centipede games. These results clearly demonstrate that this is not always the case. Arguably, CENT-ALL is the closest to a complete-information centipede game to have been studied in the laboratory. For that game form the predictions of backwards induction perform quite well. The question of what other game forms also induce a complete-information centipede game is an interesting open question.[35]

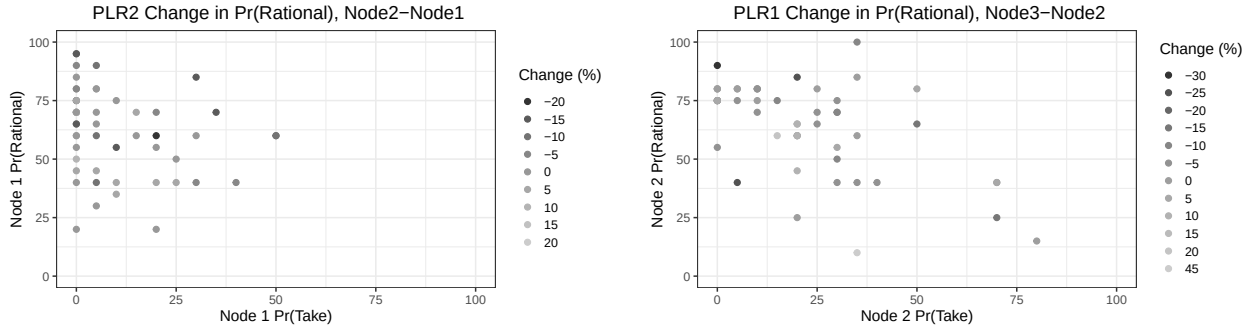
### *Initial vs. Strong Belief in Rationality*

In the theoretical literature on backwards induction there is a very important distinction between initial belief in rationality and "strong" belief in rationality, which requires that the initial belief in rationality be maintained even after zero-probability events. As Reny (1993) points out, initial common belief in rationality is not sufficient for backwards induction. His argument is as follows: Suppose that at the initial node player 2 believes in higher-order rationality and that this implies they must believe player 1 will play Take at every node. If player 1 instead chooses Pass at the first node then this is a probability-zero event for player 2, so their updated belief is unconstrained. If they now believe that player 1 is irrational and will Pass again at the third node then player 2 will rationally choose to Pass at the second node to take advantage of player 1's irrationality. But if a rational player 1 anticipates this then they actually should choose Pass at the first node because it will "trick" player 2 into choosing Pass. Thus, a common initial belief in rationality does not necessarily imply the backwards induction outcome.

Only if player 2 staunchly maintains a belief in rationality—even after zero-probability events—is the backwards induction outcome necessarily predicted. Thus, an important theoretical question is how beliefs in rationality change after surprise events. In particular, do players who believe in rationality maintain that belief even when their opponent plays Pass?

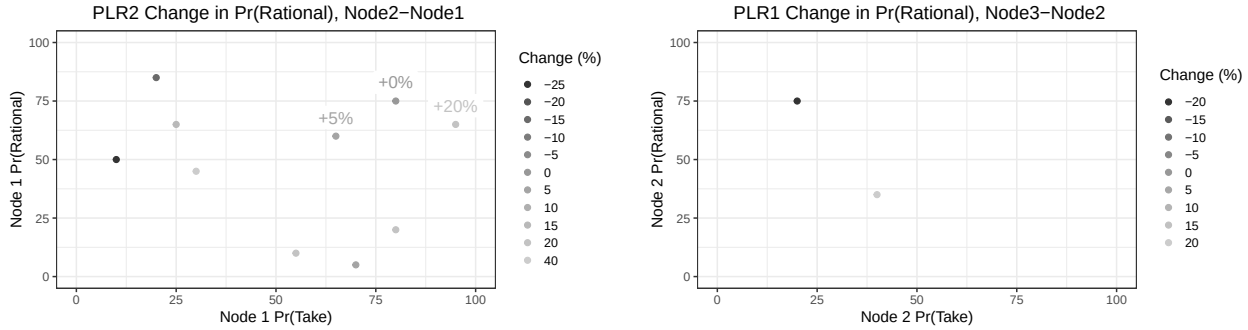
To analyze this, we first consider exactly the situation described above: a player 2 at node 1 who believes player 1 is rational and will choose Take, but player 1 actually chooses Pass. How does player 2's belief in rationality change in response? First consider CENT-LO. Panel (a) of Figure VI takes all games where player 1 initially chose Pass and shows the change in player 2's belief in rationality from node 1 to node 2. Darker points indicate a larger drop in that belief. To examine strong belief in rationality we focus

[35]In the appendix we report two additional treatments. CENT-MID is between CENT-LO and CENT-HI in terms of payoffs, though more similar to CENT-LO. Indeed, the results are similar to CENT-LO. The other treatment, CENT-CONST, tests a constant-sum game form with only four decision nodes. The *SizeBAP* is high for the first node, but drops quickly in subsequent nodes. We see the vast majority of subjects choosing Take in the first node, with the remaining games ending at the second node. Thus, the results are similar to CENT-ALL.




(a) (b)

FIGURE VI. Change in belief in rationality in CENT-LO after observing the opponent choose Pass.




(a) (b)

FIGURE VII. Change in belief in rationality in CENT-ALL after observing the opponent choose Pass.

on subjects who had high initial beliefs in both Take and rationality, which appear to the northeast in the graph. Unfortunately, no subjects have such beliefs, which is sensible since Passing is very common in CENT-LO. Thus, we cannot observe how subjects update when they're surprised because no subjects are surprised.

Panel (b) shows the same thing when moving from node 2 to node 3, where player 1 must now update because player 2 chose Pass. Again, no subjects have a high belief in both Take and rationality, and so no subjects are surprised. Thus, the critical distinction between initial belief in rationality and strong belief in rationality is moot in CENT-LO since the presence of altruist types means no subject believes both in rationality and in their opponent choosing Take at early nodes.

One might hope that we could identify surprises in CENT-ALL, where Taking is very common. But this means there are very few observations of Pass, which are needed to observe updating. This is apparent by the small number of data points in Figure VII. Recall that we cannot observe player 2's beliefs if they choose Take at node 2, so in panel (a) we are restricted to the ten observations where player 2 also chose Pass at

| | C | D |
|---|---|---|
| C | \$10,\$10 | \$ 1,\$15 |
| D | \$15,\$ 1 | \$ 5,\$ 5 |

FIGURE VIII. The Prisoners' Dilemma Game Form

node 2. And in panel (b), the two observations where Player 1 chose Pass again at node 3. The few observations we have with a high Pr(Take) and Pr(Rational) from panel (a) show that the belief in rationality actually increases, not decreases, when the opponent passes.[36] Thus, we don't see evidence that subjects switch from believing in rationality to believing in irrationality after observing Pass, but with so few observations the data are far from conclusive.

## VI. THE PRISONERS' DILEMMA

In the SIM experiment subjects play five $2 \times 2$ simultaneous-move game forms without feedback. The third game form subjects face in that experiment is the classic Prisoners' Dilemma, shown in Figure VIII.[37] In our data, 30.4% of subjects choose cooperation ($C$). This is in line with previous experiments using similar payoffs but no elicitation (see Mengel, 2018 for a meta-study). Using our elicitation data we can ask whether this cooperation is rationalized by non-selfish preferences, or whether it violates C-EU-rationality.

To that end we classify subjects into four possible types based on the best response functions implied by their elicited utilities. Selfish types are defined as those whose elicited utility indicates a (weak or strict) dominant strategy to defect. Conditional Cooperators prefer $D$ in response to $D$, and $C$ in response to $C$. Reverse types prefer $C$ in response to $D$, and $D$ in response to $C$. Finally, Unconditional Cooperators have a dominant strategy to cooperate.[38] These are summarized in the first three columns of Table I. The fourth column shows the frequency of each of these types among the 147 subjects in this experiment for whom we have valid preference and belief data.[39] The remaining columns show how many subjects made each strategy choice, broken down by whether their C-EU best response to their stated belief is $C$ (columns 5 and 6) or $D$ (columns 7 and 8).

[36] In CENT-HI there are also three observations of player 2 having Pr(Take) and Pr(Rational) both greater than 50%. The belief changes for these are −10%, −5%, and 0%. For player 1 there are two such observations with belief changes of 0% and +20%.

[37] In the actual experiment the strategies were labeled "A" and "B".

[38] For the three non-selfish types, any ranking that differs from the selfish type must be strict. For example, Conditional Cooperators are defined by $u_i(\pi(D,D)) \geq u_i(\pi(C,D))$ and $u_i(\pi(D,C)) < u_i(\pi(C,C))$.

[39] Two of the 150 subjects did not make a strategy choice for this game form and one did not fill in their first-order beliefs. Additionally, three selfish types reported utilities and beliefs such that their expected utility of $C$ and $D$ are identical. We exclude them from this table, though all three chose $s_i = D$.

| Pref. Type | $BR_i(C)$ | $BR_i(D)$ | % Subj. | $BR_i(p_i^{1s}\|u_i)=C$ $s_i=C$ | $BR_i(p_i^{1s}\|u_i)=C$ $s_i=D$ | $BR_i(p_i^{1s}\|u_i)=D$ $s_i=C$ | $BR_i(p_i^{1s}\|u_i)=D$ $s_i=D$ |
|---|---|---|---|---|---|---|---|
| Selfish | $D$ | $D$ | 68.0% | – | – | 18 | 79 |
| Cond. Coop. | $C$ | $D$ | 19.7% | 15 | 5 | 3 | 6 |
| Reverse | $D$ | $C$ | 2.7% | 1 | 2 | 0 | 1 |
| Uncond. Coop. | $C$ | $C$ | 9.5% | 8 | 6 | – | – |

TABLE I. The strategy choices of the four types in the Prisoners' Dilemma game form, broken down by whether their best response to their beliefs is $C$ or $D$.

Our main result is that, although a slight majority of subjects report selfish preferences and choose $D$ (79 out of 144), there is substantial heterogeneity in both preferences and actions. This game form induces a Bayesian game of incomplete information in which roughly one third of subjects report non-selfish preferences. The question then is whether the cooperation we observe comes entirely from these non-selfish subjects maximizing their expected utility. The results are mixed: Of the 45 subjects who choose $C$ (columns 5 and 7), only 24 (53%) do so rationally (column 5). The remaining 21 violate C-EU-rationality (column 7), and the vast majority of those (18 of 21) come from subjects who reported a dominant strategy to defect. In these cases the violation of C-EU-rationality cannot be explained as a failure of expected utility, but instead must be either a failure of consequentialism or a failure of dominance. If we assume players don't violate dominance then it must be that these subjects have a preference for cooperation that depends on more than just the outcomes it generates. This is line with the findings of Shafir and Tversky (1992), who write "evidently, some people are willing to forego some gains in order to make the cooperative, ethical decision."[40]

These results add nuance to the received wisdom regarding cooperation in the prisoners' dilemma. Much of the extant literature views cooperation as a rational response to social preferences defined narrowly over the four outcomes of the game (Mengel, 2018; Gächter et al., 2024, , *e.g.*). While that does explain 53% of the cooperation we observe, we also find that 47% cooperate even though their stated preferences and beliefs indicate

[40]In the sequential-move Prisoners' Dilemma (treatment SEQ), 62% of first-movers choose $D$, after which *all* second-movers respond with $D$. In the 12 cases where the first mover chooses $C$, eight second-movers reciprocate with $C$ and four choose $D$. This is very similar to the results of Shafir and Tversky (1992). Based on elicited utilities, 93% of second-movers choose rationally. Unfortunately, the elicited utility types for second movers are significantly more selfish when the first-mover chooses $D$ (Wilcoxon $p$-value 0.004), indicating that many second-movers may have waited to see the first-mover's choice before reporting their utilities. This places a significant caveat on any interpretation of the high level of rationality we observe, though does provide an interesting observation of non-consequentialism. We do not observe this problem in any of the other game forms in the SEQ treatment. For more on the sequential prisoners' dilemma, see Ross et al. (1977), Clark and Sefton (2001), Altmann et al. (2008), Blanco et al. (2011), Gächter et al. (2012), Blanco et al. (2014), Rubinstein and Salant (2016), and Miettinen et al. (2020), among others.

| | L | R |
|---|---|---|
| U | \$10,\$ 5 | \$15,\$15 |
| D | \$ 5,\$10 | \$ 1,\$ 1 |

FIGURE IX. The Dominance Solvable Game Form

| Column Players' Types | | | | $BR_i(p_i^{1s}\|u_i)=L$ | | $BR_i(p_i^{1s}\|u_i)=R$ | |
|---|---|---|---|---|---|---|---|
| Pref. Type | $BR_i(U)$ | $BR_i(D)$ | % Subj. | $s_i=L$ | $s_i=R$ | $s_i=L$ | $s_i=R$ |
| Selfish | $R$ | $L$ | 91.9% | 0 | 0 | 14 | 53 |
| DomStrat $L$ | $L$ | $L$ | 5.4% | 3 | 1 | – | – |
| DomStrat $R$ | $R$ | $R$ | 2.7% | – | – | 1 | 1 |
| Reversed | $L$ | $R$ | 0% | 0 | 0 | 0 | 0 |

TABLE II. The strategy choices of the four types of column player in the dominance solvable game form, broken down by whether their best response to their beliefs is $L$ or $R$.

that they should have defected. It is thus their revealed preferences over *strategies* that deviate from the selfish prediction, but not necessarily their preferences over outcomes.

## VII. GAMES WITHOUT SOCIAL PREFERENCES

Finally, we highlight two game forms in the SIM and SEQ treatments for which elicited preferences are almost all consistent with selfishness. Therefore, deviations from the selfish equilibrium predictions are driven almost entirely by strategic uncertainty rather than preference uncertainty.[41]

### *A Dominance Solvable Game Form*

The first of five game forms that subjects face in the SIM experiment is the dominance solvable game form shown in Figure IX. In terms of monetary payoffs, the row player has a strict dominant strategy to play $U$. Anticipating this, a money-maximizing column player should respond with $R$, even though it does expose them to the possibility of the outcome (\$1,\$1) should the row player tremble.

In our data, 100% of row players play $U$. Looking at the elicitation data, 71 of the 75 row players report utilities consistent with selfish preferences, and the remaining four do not have a dominant strategy but do have beliefs such that $U$ is their best response. Thus, 100% of the row players are C-EU-rational.

Table II shows that 91.9% of column players also report selfish preferences.[42] Yet 25% of column players violate iterated dominance (in terms of payoffs) by choosing $L$. The

[41]For brevity, the remaining two game forms are relegated to the online appendix.

[42]One column player left one of the utility elicitation question blank and another reported exact indifference between $L$ and $R$. Both are omitted from these analyses.

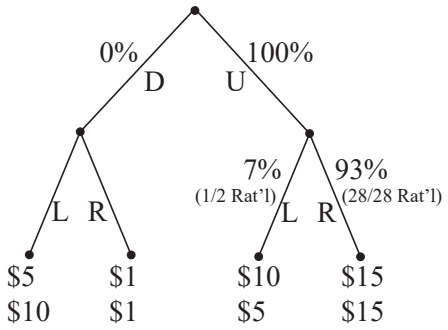



FIGURE X. Frequencies of choices and fractions of subjects who are C-EU-Rational in the sequential-move Dominance Solvable game form.

question then is whether these players have incorrect beliefs, or whether they violate C-EU-rationality.

We can see from Table II that all 67 of those with selfish preferences report a belief $p_2^{1s}(U)$ high enough such that their best response is to play $R$. Thus, the 14 (21%) who play $L$ are violating C-EU-rationality. Pooling across all types, a total of 18 subjects (25%) play $L$ but only three do so rationally.[43] Thus, the failure of iterated dominance here is not due to beliefs, but instead represents a failure of either consequentialism or expected utility preferences.

Our conjecture is that many of the 15 column players who irrationally play Left do so to avoid the ($1,$1) outcome, even among players who reported a 100% belief that the row player would play $U$. But this distaste for the ($1,$1) outcome is not captured by their elicited utilities, for if it were then $L$ would be a C-EU-rational response. Instead, the avoidance of ($1,$1) occurs in the face of strategic uncertainty within the context of the game. It could be explained by a form of loss aversion or ambiguity aversion that violates expected utility. If this is the case then their preference over strategies in the game may differ from their elicited cardinal preferences over the outcomes those strategies generate. And this difference may push them to play $L$ even though $R$ gives the better expected utility of outcomes.[44]

To further understand the role of strategic uncertainty, consider the sequential version of the game in which row players move first. The data from the SEQ experiment are shown in Figure X.[45] As in the SIM treatment, all row players choose Up, and all do

[43]These three subjects report preferences such that Left is a dominant strategy, which means $u_2(\$10,\$5) > u_2(\$15,\$15)$. We view such preferences as implausible and mostly likely the result of noise or mistakes.

[44]Ambiguity aversion alone seems unlikely to explain the data here. For example, if a subject satisfies maxmin-EU (Gilboa and Schmeidler, 1989) then, given that we only elicit $p_2^1(U)$ and assume $p_2^1(D) = 1 - p_2^1(U)$, our assumption of expected utility would result in a correct estimate of $U_2(R)$ but an overestimate of $U_2(L)$. Thus, their actual best response is even more likely to be $R$.

[45]Two column players are omitted because action choices of the row players were not recorded.

| | L | R |
|---|---|---|
| U | \$15,\$ 5 | \$ 2,\$ 1 |
| D | \$ 1,\$ 2 | \$ 5,\$10 |

FIGURE XI. The Asymmetric Coordination Game Form

| Row's Type | $BR_1(L)$ | $BR_1(R)$ | % Subj. | $BR_1(p_1^{1s}\|u_1)=U$ $s_1=U$ | $BR_1(p_1^{1s}\|u_1)=U$ $s_1=D$ | $BR_1(p_1^{1s}\|u_1)=D$ $s_1=U$ | $BR_1(p_1^{1s}\|u_1)=D$ $s_1=D$ |
|---|---|---|---|---|---|---|---|
| Selfish | $U$ | $D$ | 95.8% | 43 | 2 | 20 | 3 |
| DomStrat $U$ | $U$ | $U$ | 4.2% | 3 | 0 | – | – |

| Col's Type | $BR_2(U)$ | $BR_2(D)$ | % Subj. | $BR_2(p_2^{1s}\|u_2)=L$ $s_2=L$ | $BR_2(p_2^{1s}\|u_2)=L$ $s_2=R$ | $BR_2(p_2^{1s}\|u_2)=R$ $s_2=L$ | $BR_2(p_2^{1s}\|u_2)=R$ $s_2=R$ |
|---|---|---|---|---|---|---|---|
| Selfish | $L$ | $R$ | 93.0% | 26 | 27 | 5 | 8 |
| DomStrat $L$ | $L$ | $L$ | 7.0% | 5 | 0 | – | – |

TABLE III. The strategy choices of the two types of row and column players observed in the asymmetric coordination game, broken down by the best responses to their beliefs.

so rationally because all have preferences consistent with selfishness. But, unlike the SIM treatment, 28 out of 30 column players choose Right in response. And they do so rationally because all 28 report preferences such that $u_2(\$15,\$15) \geq u_2(\$10,\$5)$. Only one of the 30 column players violates C-EU-rationality in this game.

Comparing the simultaneous-move to the sequential-move version of the game, we find there is not a significant difference in preference types of column players (92% selfish up to 94% selfish, giving a Wilcoxon $p$-value of 0.746), but there is a significant reduction in their incidence of irrationality (23% irrational down to 3% irrational, giving a Wilcoxon $p$-value of 0.013). Thus, removing strategic uncertainty nearly wipes out violations of C-EU-rationality.

### *An Asymmetric Coordination Game Form*

The fifth game form subjects encounter in the SIM experiment is the asymmetric coordination game form shown in Figure XI. Both $(U,L)$ and $(D,R)$ are pure-strategy equilibria (in terms of dollar payoffs), but coordination is difficult because the two players prefer different equilibrium outcomes. In addition, the row player gets a higher payoff in their more-preferred equilibrium than the column player gets in theirs.

In this game form 95% of subjects report preferences that are consistent with selfishness.[46] Of the eight that don't report selfish preferences, all report that they have a dominant strategy (either $U$ or $L$) and all rationally play that strategy. But of those

[46]Four row players and one column player failed to make an action choice. Three of these four row players reported selfish preferences, one reported a dominant strategy of $U$, and the one column player reported a

with selfish preferences only 62% are C-EU-rational. By far the most common source of irrationality is those subjects whose reported beliefs indicate that they should acquiesce and follow their opponent's preferred equilibrium, yet they deviate and choose the strategy consistent with their own preferred equilibrium. For example, of the 53 selfish column players for whom $BR_2(p_2^{1s}|u_2) = L$, the median belief is $p_2^{1s}(U) = 0.90$, so they are quite sure the row player will choose $U$. Yet a slight majority of them still choose $R$, presumably targeting their own preferred equilibrium. For row players this tendency is even more extreme: 87% of those for whom $BR_1(p_1^{1s}|u_1) = D$ instead play $U$.[47] Thus, players appear to target their preferred equilibrium irrationally, and do so more when their preferred equilibrium offers a relatively higher payoff.

It is hard to argue that irrational subjects are loss averse in this case, since both strategies expose them to the possibility of either ($1,$2) or ($2,$1). And most subjects assign similar utilities to these two outcomes: The average reported utility difference between them is only 4.3, with 41% of subjects reporting no difference at all.[48] Thus, we conclude that failures of C-EU-rationality are due either to a form of optimism that is not reflected in elicited beliefs, or a form of stubbornness—or even spite—in their preferences over strategies.

The one caveat with this game is that, when elicitation is removed, we see both players shifting play more towards their opponent's preferred equilibrium (see Figure XIV and Table IV in Section VIII for details). Although this could be evidence that elicitation actually increases stubbornness, we cannot draw definitive conclusions since it's possible that beliefs also changed when elicitation was removed. Regardless, the fact that we observe this sort of stubbornness in our elicitation experiment suggests that we should at least be concerned that it may exist in other settings.

The data from the SEQ experiment are summarized in Figure XII. As we saw in the previous games, irrationality essentially disappears once strategic uncertainty is removed. The vast majority of first movers target their preferred equilibrium by choosing Up, after which 26 out of 28 second movers acquiesce and choose Left. Thus, negative reciprocity (or, spite) is rare in the sequential-move version, suggesting that the irrationality in the SIM treatment may stem from a form of stubbornness that is not reflected in their beliefs.[49]

---

dominant strategy of $R$. Additionally, three column players reported indifference because $p_2^{1s}(U) = 1$ and $u_2(\pi((U,L)) = u_2(\pi(U,R))$, and then chose $R$. All eight subjects are excluded from the analysis.

[47]The median belief for this group is $p_1^{1s}(R) = 0.80$.

[48]Following footnote 44, ambiguity aversion also doesn't seem to explain the results for this game form. By assuming expected utility we overestimate $U_2(R)$, meaning ambiguity averse column players are more likely to have a best response of $L$. So this doesn't rationalize why we see players choosing $R$.

[49]The three first movers who chose Down may have done so out of fear of such spiteful behavior. Unfortunately we cannot test this because we cannot incentivize these players' beliefs in the counterfactual event in which they choose Up.

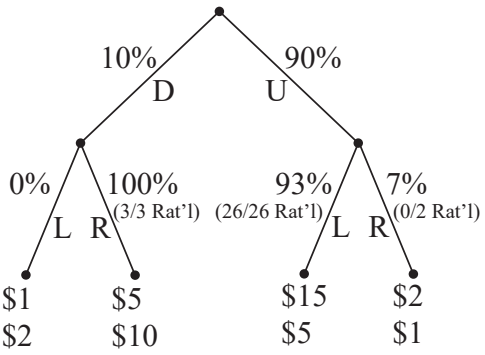



FIGURE XII. Frequencies of choices and fractions of subjects who are C-EU-Rational in the sequential-move Asymmetric Coordination game form.

### *Correlation Across Games*

Finally, we ask whether C-EU-irrationality is a persistent trait of an individual across game forms, or whether different people are C-EU-irrational in different game forms. We perform this test only on the SIM experiment, since the CENT experiments are all between-subject. For each pair of game forms in SIM we test for correlation in rationality between those two game forms using Kendall's $\tau$ statistic.

For all ten pairs of games we cannot reject the null hypothesis of zero correlation, with about half of the point estimates indicating negative (but insignificant) correlation. This suggests that not only are different phenomena driving C-EU-rationality in different game forms, but that there's little predictive power between them at the individual level. For example, subjects who are stubborn in the asymmetric coordination game form are not more likely to be irrational cooperators in the prisoners' dilemma.

## VIII. ELICITATION METHODOLOGY

In this section we provide an extensive discussion of our elicitation methodology, including incentive compatibility and order effects. We also discuss the issues of contamination (showing that the presence of elicitation generally doesn't affect play) and consequentialism (which cannot be tested).

### *Eliciting Cardinal Utilities*

For each cell of the game form's matrix we elicit the subject's cardinal utility for that outcome. Each question is described to the subject as a 101-item binary choice list. For example, if the top-left cell of the game has payoffs ($\$5, \$10$), then to elicit the subject's cardinal utility for that cell's outcome the subject is asked to consider the list shown in

| **Option A** | or | **Option B** |
|---|---|---|
| You get \$5 and they get \$10 | or | 0% chance you both get \$20 |
| You get \$5 and they get \$10 | or | 1% chance you both get \$20 |
| ⋮ | ⋮ | ⋮ |
| You get \$5 and they get \$10 | or | 100% chance you both get \$20 |

FIGURE XIII. The choice list used in cardinal utility elicitation.

Figure XIII. The subject is asked at what probability they would switch from choosing Option A to choosing Option B. One row is then randomly selected, and the subject's choice for that row is paid. In practice, the entire list is not shown while making actual decisions; rather, the subject is shown an example list in the instructions and then simply asked for their switch point when making actual elicitation decisions.

If we normalize $u(\$0,\$0) = 0$ and $u(\$20,\$20) = 1$, then a switch point of $q^*$ represents the subject's cardinal utility for $(\$5,\$10)$, since indifference implies $u(\$5,\$10) = q^*u(\$20,\$20) + (1-q^*)u(\$0,\$0) = q^*$. Notice that the elicited cardinal utility captures both risk aversion and social preferences. If this question is chosen for payment, only one of the two subjects (selected at random) is paid. This ensures that $i$'s choice truly determines the final payment of both $i$ and $-i$, and that the subjects receive no other payments.

All randomness in the experiment is resolved using die rolls or draws from a Bingo cage performed in front of the subjects at the conclusion of the experiment.[50]

We assume subjects have a unique switch point. Reporting that switch point generates a two-stage gamble in which the question number is first chosen, and then the selected alternative for that question is paid out. (If Option A is selected then the second stage is degenerate.) As long as preferences over these two-stage gambles respects statewise monotonicity, the procedure is incentive compatible; see Azrieli et al. (2018) for details.

### *Eliciting Beliefs*

We elicit subjects' beliefs about a variety events, including which strategy their opponent plays, whether their opponent's elicitation data is consistent with C-EU-rationality, whether their guess of their opponent's utility is correct, and so on. In general, eliciting beliefs about any event $E$ is done by asking the subject to consider their switch point in a 101-question list similar to Figure XIII, except Option A is now "You get \$20 if $E$ occurs" and Option B is "You get \$20 with probability $p$", with $p$ varying from 0% to 100%.

[50]In the domain of individual decision-making, Oechssler et al. (2019) and Baillon et al. (2022) show that the timing of the resolution of this uncertainty does not affect choices.

If this elicitation question is chosen for payment, both players are paid independently based on their own choice from one randomly-selected row.

Again, assuming a unique switch point, this procedure is incentive compatible as long as players' preferences over two-stage gambles respects statewise monotonicity. Holt and Smith (2016) verify that this mechanism performs well when reporting beliefs, and Healy and Kagel (2023) (following lessons from Danz et al., 2022) show that it works especially well when the list is presented and the incentives are explained in the instructions, but not shown on the screen at the time of decision.

### *Eliciting Modes of Others' Beliefs*

Belief distributions that live in spaces with more than one dimension (such as beliefs over the opponent's four utility values) are difficult to elicit in practice. Instead, we ask subjects to report their two most likely guesses of the realized value. For example, we first ask the subject to report their best guess of the four utility values submitted by their opponent. We then ask their belief that their guess is correct. This is done using the procedure described above, where event $E$ is now the event that "your guess is correct".

Given a fixed guess, it is clear that reporting the true belief for that guess is optimal under eventwise monotonicity. When choosing which guess to report, it is clear that reporting the guess that has the highest probability (the subject's "best guess") is also optimal. Doing so maximizes both the value of Option A in the list, and the number of rows for which Option A is paid. A similar insight appears in Möbius et al. (2022).

To elicit the subject's second-most-likely value, we simply ask them to report a second guess—which must be different from the first—and the probability this guess is correct. Again they are paid via the above mechanism. Given that they cannot report their most likely guess, their optimal report is their second-most-likely guess.[51]

### *C-EU-Rationality*

It may seem that $i$'s belief in C-EU-rationality does not need to be elicited separately, as it could be inferred from their beliefs about opponent's strategies, beliefs, and utilities. But since we only elicit marginal beliefs, this is not the case. For example, suppose $i$ believes $-i$ plays two possible strategies $\hat{s}_{-i}$ and $\tilde{s}_{-i}$ and has two possible beliefs $\hat{p}^{1s}_{-i}$ and $\tilde{p}^{1s}_{-i}$. Utility is known to be $\hat{u}_{-i}$. Suppose player $-i$ is rational at $(\hat{s}_{-i}, \hat{p}^{1s}_{-i})$ and $(\tilde{s}_{-i}, \tilde{p}^{1s}_{-i})$, but not at $(\hat{s}_{-i}, \tilde{p}^{1s}_{-i})$ or $(\tilde{s}_{-i}, \hat{p}^{1s}_{-i})$. If $i$ reports that both strategies are equally likely and

[51]The subject could switch which report is their first- and second-best, but their probabilities of each guess being correct would then reveal which is truly their best guess. When this occurs we switch which is labeled as their highest-probability guess when analyzing the data.

both beliefs are equally likely then their belief in rationality cannot be determined: This report could come from a belief that the two are perfectly correlated, so that $i$ assigns probability one to $-i$ being C-EU-rational. It could come from a belief that the two are independently drawn, so that $i$ assigns probability 1/2 to $-i$ being C-EU-rational. Or it could come from a belief that they are perfectly negatively correlated, putting zero probability on C-EU-rationality. Without information about the joint distribution, we cannot infer anything about $i$'s belief in C-EU-rationality from their marginal beliefs alone. Rather than eliciting a joint distribution, we elicit beliefs about C-EU-rationality directly in our experiment.

In order to elicit beliefs about C-EU-rationality we must first explain C-EU-rationality to subjects. To do this, we teach subjects simple expected utility calculations, and say that their opponent is "consistent" (rather than "rational") if their action choice gives a higher expected utility than the unchosen action. We then ask the subject's belief that their opponent's action choice is "consistent." We can then incentivize this belief elicitation—using the 101-item list described above—because we can observe whether the opponent's choices actually are "consistent" or not.

### *Issue #1: Contamination*

The elicitation of beliefs and utilities may alter game play. And playing the game may alter the beliefs and utilities that subjects report. We refer to both of these possibilities as examples of "contamination" that may be present in our experiment. And both are possible in our experiment because the strategy choices and elicitation questions were intermingled.[52] Furthermore, the instructions make it clear that both strategy choices and elicitation questions will be given. Thus, we view this as a fully contaminated experiment in which game play is contaminated by elicitation, and elicitation is contaminated by game play.

Having the experiment be fully contaminated was a conscious design choice. We are not aware of any way to remove contamination through the experimental design, so instead we embrace it. One argument is that an experiment with elicitation will increase observed rates of C-EU-rationality, since subjects are more likely to think carefully about the choices of others and to submit consistent responses. Thus, our results would provide a *lower* bound on the levels of C-EU-irrationality one might observe in typical experiments. And yet we still see significant levels of C-EU-irrationality across the games we study.

[52]Recall that in the SIM and SEQ experiments, strategy choices and elicitation questions are all presented on the same page, so there was no forced ordering in which they were answered. In the CENT treatments subjects did choose actions first, but action choices in the fourth period did follow elicitation questions from the third.

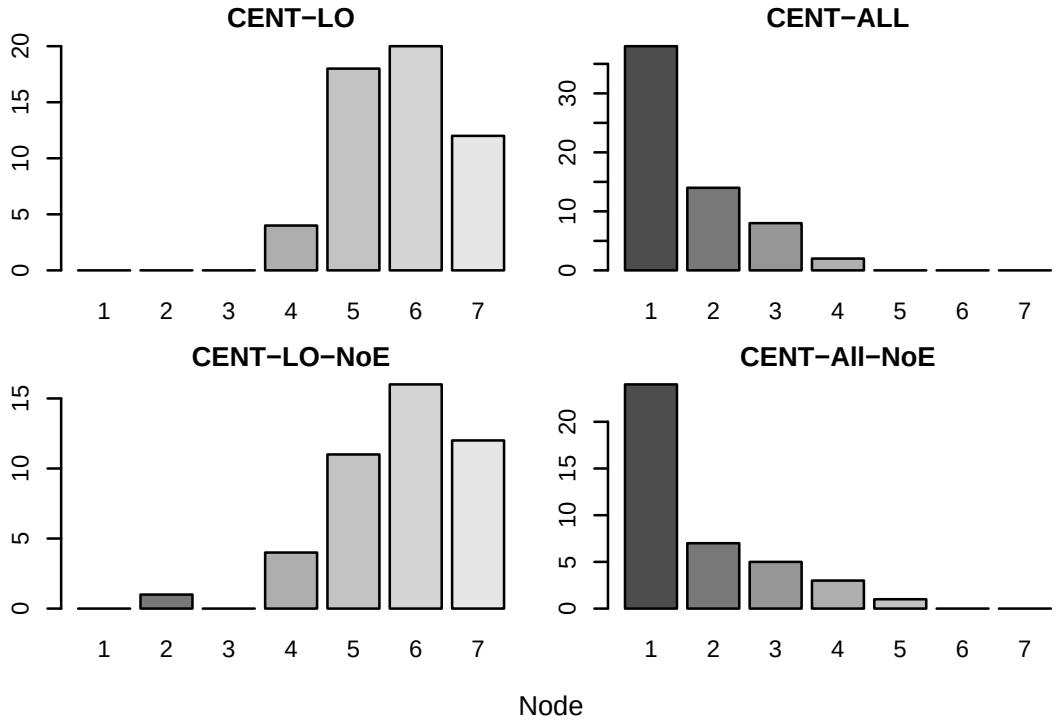



FIGURE XIV. Histograms of how frequently each terminal node was reached in each treatment. NoE refers to a "No Elicitation" treatment. Periods 3 and 4 are pooled.

However, we can also test for contamination by re-running all of these game forms without elicitation. To that end, we run the CENT-LO-NoE and CENT-ALL-NoE treatments, which replicate the CENT-LO and CENT-ALL treatments exactly, except with all elicitation questions removed.[53] We also run a SIM-NoE that is identical to the SIM experiment, but with no elicitation questions. Subjects simply make choices in the five games (again, with one game per page in the booklet) and are paid for one randomly-chosen game. The CENT-LO-NoE treatment had 44 subjects, CENT-ALL-NoE had 40, and SIM-NoE had 60.[54]

Figure XIV shows histograms of how frequently different terminal nodes were reached in each of the centipede treatments. We pool periods 3 and 4 since their distributions of terminal nodes are not significantly different (Fisher's exact test p-values of 0.93 for CENT-LO and 0.31 for CENT-ALL). Comparing CENT-LO and CENT-LO-NoE using a

[53]NoE is mnemonic for "no elicitation." Subjects in CENT-LO-NoE and CENT-ALL-NoE do not submit entire strategies; instead, they play out the extensive form by choosing $T$ or $P$ at each node reached.

[54]It is tempting to test the other direction of contamination by running treatments with elicitation but without strategy choices. But this would require that subjects become inactive third parties, stating beliefs about a game played between two others. This might drastically change how they view and analyze the game, and so it's not clear that we should expect those beliefs to be comparable to our current data.

| | Game: | Dom. Solvable | | Common Interest† | | Prisoners' Dilemma | | Asymm. M.P.† | | Asymm. Coord. | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Treatment | Strategy | Row | Col | Row | Col | Row | Col | Row | Col | Row | Col |
| SIM | $U$ or $L$ | 75 | 19 | 72 | 74 | 19 | 26 | 64 | 32 | 66 | 36 |
| | $D$ or $R$ | 0 | 56 | 3 | 1 | 55 | 48 | 9 | 41 | 5 | 38 |
| SIM-NoE | $U$ or $L$ | 30 | 2 | 28 | 30 | 11 | 9 | 24 | 7 | 19 | 23 |
| | $D$ or $R$ | 0 | 28 | 2 | 0 | 19 | 21 | 6 | 23 | 11 | 7 |
| $\chi^2$ test $p$-value: | | 1.00 | 0.03 | 0.56 | 0.53 | 0.26 | 0.62 | 0.32 | 0.051 | 0.0002* | 0.009 |

TABLE IV. Strategy choices with elicitation (SIM) and without (SIM-NoE) for each player role in each game, with chi-squared test $p$-values. †Elicitation data reported in the appendix. *Rejection of equality between treatments at the 5% level with either a Bonferroni or Holm-Bonferroni correction.

Fisher's exact test yields no significant difference (p-value of 0.777). Similarly, no significant difference is found between CENT-ALL and CENT-ALL-NoE (p-value of 0.612). We conclude that elicitation does not significantly affect game play in these centipede game treatments, suggesting that contamination may actually not be a large concern here.

Results for the SIM and SIM-NoE treatments are shown in Table IV. The frequency of each strategy choice for each player role in each game is shown. The bottom row shows $p$-values from $\chi^2$ tests of equality between treatments. Given that we report ten separate tests, we drop the threshold for significance accordingly (using either the Bonferroni method or Holm-Bonferroni methods) and find only one significant difference: row players in the asymmetric coordination game appear to be sensitive to the presence of elicitation. Specifically, it appears that elicitation may have increased the "stubborn" behavior of row players in favoring their preferred equilibrium, with row players choosing $U$. The difference for the column player is marginally insignificant after correction, but directionally is consistent with them favoring their preferred equilibrium by playing $R$. There is also suggestive evidence that, in the dominance solvable game form, elicitation caused the column players to choose $L$ more often, perhaps because elicitation made the cost of the ($1,$1) outcome more salient.

Overall, evidence for contamination is limited. In most games we see no difference in behavior when elicitation is removed. And, under our assumption that elicitation only increases C-EU-rationality, we view our results as providing a lower bound on the levels of C-EU-irrationality present in typical experiments without elicitation.

| | L | R |
|---|---|---|
| U | \$5,\$5 | \$5,\$5 |
| D | \$100,\$5 | \$5,\$5 |

FIGURE XV. A game form in which consequentialism is likely to fail.

*Issue #2: Consequentialism*

Traditional game theory takes as primitive cardinal payoffs of the form $u_i(s_i, s_{-i})$, which allows the possibility that two strategy profiles which lead to the same outcome are evaluated differently. To illustrate, consider the game form shown in Figure XV. For the row player it is plausible that $u_i(U,L) \neq u_i(D,R)$. Profile $(U,L)$ generates a payoff of (\$5,\$5) (instead of (\$100,\$5)) because the row player chose to play $U$. Intuitively, it is their own fault they did not get \$100. The profile $(D,R)$ also generates a payoff of (\$5,\$5), but in this case the "fault" belongs to the opponent. It seems entirely plausible that, for the row player, $u_i(U,L) \neq u_i(D,R)$ even though $\pi(U,L) = \pi(D,R)$. Thus, consequentialism (which assumes $\pi(s) = \pi(s') \Rightarrow u_i(s) = u_i(s')$) is violated.

Ideally, we would elicit $u_i(s_i, s_{-i})$ instead of $u_i(\pi(s_i, s_{-i}))$, avoiding the need to assume consequentialism. To our knowledge, however, it has not been demonstrated that $u_i(s_i, s_{-i})$ is an elicitable quantity. And we conjecture that it cannot be elicited in an incentive-compatible way. In particular, the row player in Figure XV who knows they will select $s_i = D$ cannot be paid in the counterfactual event that $s_i = U$. Thus, $u_i(U,L)$ cannot be incentivized in the play of this game. One might construct related decision problems or games that might be used to infer $u_i(U,L)$, but by changing the game or decision problem we change the embedded meaning of the strategies and therefore cannot be sure that they are interpreted by the subject as truly identical to $(U,L)$. Given this difficulty, we are forced to assume consequentialism and concede that all apparent failures of rationality could in fact be failures of consequentialism. Indeed, for the prisoners' dilemma, this is our preferred interpretation.

## IX. DISCUSSION

What have we learned from these epistemic experiments? If we were to construct a *post hoc* model to try to organize these results, what elements would it contain? Certainly it would need to respect the fact that preferences are uncertain, and this uncertainty can drastically alter behavior as in the centipede game form. It would model games as Bayesian games, rather than complete-information games. But it would also need to feature certain types of C-EU-irrationality to capture the irrational cooperation we see in the prisoners' dilemma, or the stubbornness we see in the asymmetric coordination game.

For better or worse, the phenomena we observe are quite specific to their game forms, making any search for a unifying model appear to be hopeless. This mirrors a frequent complaint of game theory: Its predictions are often very sensitive to the details and mechanics of the game. While one model of oligopoly predicts stark competition, a perturbed version leads to successful collusion. Knowing which prediction to apply therefore requires extensive knowledge of the underlying interaction. But, as many have pointed out (including Kreps, 1990), this criticism can be turned on its head: Game theory should instead be lauded for its ability to capture the importance of such fine details of an interaction. In the same way, behavioral phenomena appear to be quite game-specific, and maintaining a healthy respect for that sensitivity both helps us to understand the role of the game form on human behavior and prevents us from writing overly-simplified behavioral models that miss this important heterogeneity.

Although not pursued here, elicitation data can also be used to provide stronger tests of existing theories. For example, both the level-$k$ theories (Nagel, 1995; Stahl and Wilson, 1994, 1995; Camerer et al., 2004) and quantal response equilibrium (McKelvey and Palfrey, 1995) could be used to explain the strategy choice data from the asymmetric coordination game, but our elicited belief data do not line up with the underlying assumptions about beliefs from either model.

Similarly, one could use elicited utilities to test various models of social preferences. Such models are often tested using choice data alone, but in principle these tests could be augmented with direct measurement of preferences over outcomes. Our measurement shows substantial heterogeneity in the exact shape of players' social preferences.[55]

One important area for future work is to understand better the role of noise or stochastic choice in elicitation. First, to what degree are elicited quantities stochastic? Is that stochasticity intentional or better modelled as noise? Second, if responses are stochastic, how does that affect our conclusions? Will it generate systematic biases? For example, Collins and James (2015) show how noise can generate a bias: the preference reversal phenomenon can largely be explained by stochastic choice in the Becker et al. (1964) elicitation method. On the other hand, McGranaghan et al. (2024) show how the choice list method we apply can be more robust to noise. They prove that, when studying common ratio effects with stochastic choice, eliciting lottery values via choice lists removes the effect of noise, whereas binary choices can lead to systematic differences depending on how close to indifferent are the two options.

Obviously, our elicitation methodology can be applied wholesale to any games of interest. Given the findings of **?**, it might be interesting to run epistemic experiments on games with more than two players to explore whether players believe others have the

[55]Details of this, and of the tests of level-$k$ and quantal response equilibrium beliefs, are available upon request.

same beliefs as themselves. Another open question is whether there are other quantities that would be valuable to elicit. Those chosen here were based on the epistemic game theory framework, but in practice other quantities may be important in actual decision-making. And which quantities are important may also depend on the game form. Along these lines, we view elicitation as complementary to other choice-process data, such as eye tracking or response times, all of which are used to augment strategy choice data to help understand and model the underpinnings of strategic choice.

## REFERENCES

Allais, M., 1953. Le Comportement de L'Homme Rationnel Devant le Risque: Critique des Postulats et Axiomes de L'Ecole Americaine. Econometrica 21, 503–546.

Altmann, S., Dohmen, T., Wibral, M., 2008. Do the reciprocal trust less? Economics Letters 99, 454–457. doi:10.1016/j.econlet.2007.09.012.

Aumann, R., 1987. Correlated Equilibrium as an Expression of Bayesian Rationality. Econometrica 55, 1–18.

Aumann, R., Brandenburger, A., 1995. Epistemic Conditions for Nash Equilibrium. Econometrica 63, 1161–1180.

Azrieli, Y., Chambers, C.P., Healy, P.J., 2018. Incentives in Experiments: A Theoretical Analysis. Journal of Political Economy 126, 1472–1503. doi:10.1086/698136.

Baillon, A., Halevy, Y., Li, C., 2022. Randomize at Your Own Risk: On the Observability of Ambiguity Aversion. Econometrica 90, 1085–1107. doi:10.3982/ECTA18137.

Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring Utility by a Single-Response Sequential Method. Behavioral Science 9, 226–232.

Binmore, K., 1987. Modeling Rational Players: Part I. Economics & Philosophy 3, 179–214. doi:10.1017/S0266267100002893.

Blanco, M., Engelmann, D., Koch, A.K., Normann, H.T., 2014. Preferences and beliefs in a sequential social dilemma: A within-subjects analysis. Games and Economic Behavior 87, 122–135. doi:10.1016/j.geb.2014.05.005.

Blanco, M., Engelmann, D., Normann, H.T., 2011. A within-subject analysis of other-regarding preferences. Games and Economic Behavior 72, 321–338. doi:10.1016/j.geb.2010.09.008.

Bolton, G.E., Ockenfels, A., 2000. ERC: A Theory of Equity, Reciprocity, and Competition. American Economic Review 90, 166–193.

Bornstein, G., Kugler, T., Ziegelmeyer, A., 2004. Individual and group decisions in the centipede game: Are groups more "rational" players? Journal of Experimental Social Psychology 40, 599–605. doi:10.1016/j.jesp.2003.11.003.

Brunner, C., Kauffeldt, T.F., Rau, H., 2016. Mutual knowledge of preferences and equilibrium play: Experimental evidence.

Calford, E.M., Chakraborty, A., 2022. Higher-Order Beliefs in a Sequential Social Dilemma .

Camerer, C.F., 2003. Behavioral Game Theory. Princeton University Press, Princeton, NJ.

Camerer, C.F., Ho, T.H., Chong, J.K., 2004. A Cognitive Heirarchy Model of Games. Quarterly Journal of Economics 119, 861–898.

Clark, K., Sefton, M., 2001. The Sequential Prisoner's Dilemma: Evidence on Reciprocation. Economic Journal 111, 51–68.

Collins, S.M., James, D., 2015. Response mode and stochastic choice together explain preference reversals. Quantitative Economics 6, 825–856.

Cox, C.A., Jones, M.T., Pflum, K.E., Healy, P.J., 2015. Revealed reputations in the finitely repeated prisoners' dilemma. Economic Theory 58, 441–484. doi:10.1007/s00199-015-0863-1.

Cox, J.C., James, D., 2012. Clocks and Trees: Isomorphic Dutch Auctions and Centipede Games. Econometrica 80, 883–903. doi:10.3982/ECTA9589.

Cox, J.C., James, D., 2015. On Replication and Perturbation of the McKelvey and Palfrey Centipede Game Experiment, in: Replication in Experimental Economics. Emerald Group Publishing Limited. volume 18 of *Research in Experimental Economics*, pp. 53–94. doi:10.1108/S0193-230620150000018003.

Dal Bó, P., Fréchette, G.R., 2011. The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. American Economic Review 101, 411–429.

Danz, D., Huck, S., Jehiel, P., 2016. Public Statistics and Private Experience: Varying Feedback Information in a Take-or-Pass Game. German Economic Review 17, 359–377. doi:10.1111/geer.12098.

Danz, D., Vesterlund, L., Wilson, A.J., 2022. Belief Elicitation and Behavioral Incentive Compatibility. American Economic Review 112, 2851–2883. doi:10.1257/aer.20201248.

Dekel, E., Siniscalchi, M., 2015. Epistemic Game Theory, in: Young, H.P., Zamir, S. (Eds.), Handbook of Game Theory. North Holland, Oxford. volume 4, pp. 619–702.

Eyster, E., Rabin, M., 2005. Cursed Equilibrium. Econometrica 73, 1623–1672. doi:10.1111/j.1468-0262.2005.00631.x.

Fehr, E., Schmidt, K.M., 1999. A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics 114, 817–868.

Fey, M., McKelvey, R.D., Palfrey, T., 1996. An experimental study of constant-sum centipede games. International Journal of Game Theory 25, 269–287.

Fischbacher, U., Gachter, S., 2010. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. American Economic Review 100, 514–556.

Fischbacher, U., Gachter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 394–404.

Friedenberg, A., Kneeland, T., 2023. Is Bounded Reasoning about Rationality Driven by Limited Ability? .

Gächter, S., Lee, K., Sefton, M., Weber, T.O., 2024. The Role of Payoff Parameters for Cooperation in the One-Shot Prisoner's Dilemma. European Economic Review , 104753doi:10.1016/j.euroecorev.2024.104753.

Gächter, S., Nosenzo, D., Renner, E., Sefton, M., 2012. Who Makes a Good Leader? Cooperativeness, Optimism, and Leading-by-Example. Economic Inquiry 50, 953–967. doi:10.1111/j.1465-7295.2010.00295.x.

García-Pola, B., Iriberri, N., Kovářík, J., 2020. Non-equilibrium play in centipede games. Games and Economic Behavior 120, 391–433. doi:10.1016/j.geb.2020.01.007.

Gilboa, I., Schmeidler, D., 1989. Maxmin Expected Utility with Non-Unique Prior. Journal of Mathematical Economics vol. 18, issue 2, pp. 141–153. doi:10.1016/0304-4068(89)90018-9.

Greiner, B., 2015. Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE. Journal of the Economic Science Association 1, 114–125.

Grether, D., 1981. Financial incentive effects and individual decision-making.

Healy, P.J., Kagel, J., 2023. Testing Elicitation Mechanisms Via Team Chat.

Holt, C.A., Laury, S.K., 2002. Risk Aversion and Incentive Effects. American Economic Review 92, 1644–1655. doi:10.1257/000282802762024700.

Holt, C.A., Smith, A.M., 2016. Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes. American Economic Journal: Microeconomics 8, 110–139. doi:10.1257/mic.20130274.

Kagel, J.H., McGee, P., 2016. Team versus Individual Play in Finitely Repeated Prisoner Dilemma Games. American Economic Journal: Microeconomics 8, 253–276. doi:10.1257/mic.20140068.

Karni, E., 2009. A Mechanism for Eliciting Probabilities. Econometrica 77, 603–606.

Kawagoe, T., Takizawa, H., 2012. Level-$k$ analysis of experimental centipede games. Journal of Economic Behavior and Organization 82, 548–566.

Kneeland, T., 2015. Identifying Higher-Order Rationality. Econometrica 83, 2065–2079. doi:10.3982/ECTA11983.

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

Kreps, D.M., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. Journal of Economic Theory 27, 245–252.

Krockow, E.M., Pulford, B.D., Colman, A.M., 2015. Competitive Centipede Games: Zero-End Payoffs and Payoff Inequality Deter Reciprocal Cooperation. Games 6, 262–272. doi:10.3390/g6030262.

Levine, D., 1998. Modelling Altruism and Spitefulness in Experiments. Review of Economic Dynamics 1, 593–622.

McGranaghan, C., Nielsen, K., O'Donoghue, T., Somerville, J., Sprenger, C.D., 2024. Distinguishing Common Ratio Preferences from Common Ratio Effects Using Paired Valuation Tasks. American Economic Review 114, 307–347. doi:10.1257/aer.20221535.

Mcintosh, C.R., Shogren, J.F., Moravec, A.J., 2009. Can tournaments induce rational play in the Centipede game? Exploring dominance vs. strategic uncertainty. Economics Bulletin 29, 2018–2024.

McKelvey, R.D., Palfrey, T.R., 1992. An Experimental Study of the Centipede Game. Econometrica 60, 803–836.

McKelvey, R.D., Palfrey, T.R., 1995. Quantal Response Equilibria for Normal Form Games. Games and Economic Behavior 10, 6–38.

Mengel, F., 2018. Risk and Temptation: A Meta-study on Prisoner's Dilemma Games. The Economic Journal 128, 3182–3209. doi:10.1111/ecoj.12548.

Miettinen, T., Kosfeld, M., Fehr, E., Weibull, J., 2020. Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. Journal of Economic Behavior & Organization 173, 1–25. doi:10.1016/j.jebo.2020.02.018.

Möbius, M.M., Niederle, M., Niehaus, P., Rosenblat, T.S., 2022. Managing self-confidence: Theory and experimental evidence. Management Science .

Myerson, R.B., 1991. Game Theory: Analysis of Conflict. Harvard University Press, Cambridge, MA.

Nagel, R.C., 1995. Unraveling in Guessing Games: An Experimental Study. American Economic Review 85, 1313–1326.

Oechssler, J., Rau, H., Roomets, A., 2019. Hedging, ambiguity, and the reversal of order axiom. Games and Economic Behavior 117, 380–387.

Pulford, B.D., Colman, A.M., Lawrence, C.L., Krockow, E.M., 2017. Reasons for cooperating in repeated interactions: Social value orientations, fuzzy traces, reciprocity, and activity bias. Decision 4, 102–122. doi:10.1037/dec0000057.

Reny, P., 1993. Common belief and the theory of games with perfect information. Journal of Economic Theory 59, 257–274.

Rosenthal, R.W., 1981. Games of Perfect Information, Predatory Pricing, and the Chain-Store Paradox. Journal of Economic Theory 25, 92–100.

Ross, L., Greene, D., House, P., 1977. The “false consensus effect”: An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology 13, 279–301. doi:10.1016/0022-1031(77)90049-X.

Rubinstein, A., Salant, Y., 2016. “Isn’t everyone like me?”: On the presence of self-similarity in strategic interactions. Judgement and Decision Making 11, 168–173.

Shafir, E., Tversky, A., 1992. Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive psychology 24, 449–474.

Stahl, D.O., Wilson, P.O., 1994. Experimental Evidence on Players’ Models of Other Players. Journal of Economic Behavior and Organization 25, 309–327.

Stahl, D.O., Wilson, P.W., 1995. On Players’ Models of Other Players: Theory and Experimental Evidence. Games and Economic Behavior 10, 218–254.

Weibull, J.W., 2004. Testing Game Theory, in: Huck, S. (Ed.), Advances in Understanding Strategic Behavior; Game Theory, Experiments and Bounded Rationality. Essays in Honour of Werner Guth. Palgrave Macmillan, pp. 85–104.