

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 2×2 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.

[†]This work subsumes a paper previously circulated as “Epistemic Foundations for the Failure of Nash Equilibrium”. I am especially grateful for the many detailed comments, support, and endless encouragement I received from Amanda Friedenberg, Kirby Nielsen, and Ryan Oprea. I thank many seminar and conference audiences for valuable feedback. I have also benefited from helpful conversations with Yaron Azrieli, Christoph Brunner, Evan Calford, Christopher Chambers, Brad Clark, Aviad Heifetz, The Kool-Aid Man (who was overwhelmingly supportive), Antonio Penta, Ariel Rubinstein, Marciano Siniscalchi, Lise Vesterlund, Alistair Wilson, and many, many more. Research assistance was provided by Caleb Cox, Alex Gotthard-Real, Ritesh Jain, Siqi Pan, Kirby Nielsen, and Hyoeun Park. This research was approved by the Ohio State University Institutional Review Board (protocol #2014B0270) and partially funded by NSF grants #SES-0847406 and #SES-1426967.

*Dept. of Economics, The Ohio State University; healy.52@osu.edu.

I. INTRODUCTION

When we observe strategic behavior that differs from standard equilibrium predictions, the natural next step is to understand why those deviations occurred and to construct models of non-standard preferences or bounded rationality to 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 developed from strategy choice data alone, without direct observation of the underlying causes for equilibrium to fail. The challenge for experimentalists is then to design clever treatments that help identify the causes of these failures. For example, by having subjects play against computer opponents we can control both their beliefs and social preferences, eliminating those factors from any model that might explain the deviations that remain.

In this paper, instead of relying on clever designs, we tackle the problem much more directly by using elicitation of both beliefs and preferences. In other words, rather than trying to control for beliefs or preferences, we simply measure them directly. But then what exactly should we elicit? And is there a theoretically-disciplined way to organize elicitation data to ensure that it provides meaningful insights regardless of the game being studied?

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 then tells us that 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 With enough elicitation data we can see directly which of these are true, and thus why Nash equilibrium fails in a given context.

Following this framework, it becomes clear that the data needed are not just strategy choices, but also players' utilities over outcomes, their beliefs about their opponent's utilities, their first- and second-order beliefs over strategies, and their first-order beliefs about their opponent's rationality. We refer to experiments that elicit this data as "epistemic experiments" and use this type of experiment to re-analyze several classic games in the behavioral game theory literature. This procedure allows us to see the true game being played (including the players' actual utilities, not just their payoffs), and whether or not players view it as

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

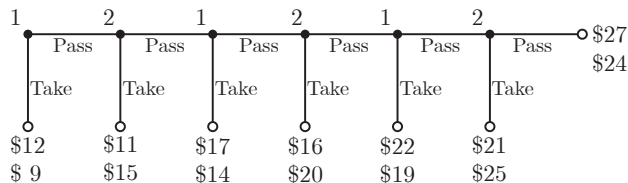


FIGURE I. A centipede game form.

a game of complete information. It 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 an appropriately-designed binary choice list (or “multiple price list”) procedure that underlies the Becker et al. (1964) value elicitation mechanism.² We describe how exactly these lists are constructed in Section VIII. These procedures are all incentive compatible under relatively weak assumptions (Azrieli et al., 2018), and subjects are given extensive training in which they are told that truth-telling is in their best interest. Importantly, we also run treatments 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 believe that our elicitation methods offer potential value in understanding the reasoning behind strategic choice.

Our first application of an epistemic experiment is to study the classic centipede game, a version of which is shown in Figure I. Although “always Take” is the predicted strategy under any standard game theoretic solution concept, experiments show that players choose “Pass” often, particularly when payoffs grow substantially across terminal nodes. Almost all of the existing literature develops explanations for this pattern using only strategy choice data, with no elicitation. This has led to a variety of different explanations. For example, McKelvey and Palfrey (1992) fit a “gang-of-four” model (à la Kreps et al., 1982) in which players believe some fraction of opponents are altruistic, and rational (selfish) players then try to build a false reputation for being altruistic to take advantage of their opponent.³ Fudenberg and Levine (1997) argue that choosing Pass (except at the final node) can be roughly consistent with a self-confirming equilibrium with heterogeneous beliefs. Reny (1993) shows how it could arise under common knowledge of rationality, as long as players abandon that belief

²See Grether (1981), Holt and Smith (2016), and Holt and Laury (2002) for prominent applications of this procedure.

³To fit the data better, McKelvey and Palfrey (1992) augment this model by allowing for trembles in actions and heterogeneous, randomly-drawn beliefs about the fraction of altruists.

upon observing Pass.⁴ Fey et al. (1996) argue that observed play fits a version of quantal response equilibrium (QRE; McKelvey and Palfrey, 1992, 1998), while Kawagoe and Takizawa (2012) claim that the Level- k model (Nagel, 1995; Stahl and Wilson, 1995) organizes the data well.

All of these models can fit the observed strategies, but which is actually the most descriptive?⁵ Or is it something else altogether? Using the elicitation data we can provide a more complete picture of players' reasoning in the centipede game. First, we clearly observe social preferences: a fair number of subjects in the Player 1 role would actually prefer the (\$11, \$15) outcome at the second terminal node over the (\$12, \$9) outcome at the first (effectively giving up \$1 to give the other player \$6), and thus have no reason to choose Take initially. Second, selfish subjects expect that such players exist, and are willing to gamble on Pass in the hopes that their opponent is an altruist who passes back. To our knowledge, this exact story has not been proposed in the literature; the gang-of-four story comes close but additionally assumes that selfish players strategically build a false reputation for being altruistic, something for which we see no evidence.⁶

Another application of the epistemic experiment methodology is the one-shot Prisoners' Dilemma game. Roughly one-third of subjects choose Cooperate in this game, which is strictly dominated if subjects are payoff-maximizing. Does a simple social preference story explain this behavior as rational? Or is this a good example of the disjunction effect, wherein subjects would prefer to Defect if they knew which strategy their opponent chose, but "as long as the other has not made his decision, mutual cooperation looms as an attractive solution for both players" (Shafir and Tversky, 1992)? Our epistemic experiment shows support for the disjunction effect story: around two thirds of subjects report utilities such that Defect is in fact a dominant strategy, yet about 20% of these players still choose to Cooperate. Overall, we find that about half of all cooperation we observe is irrational—meaning it is not a best response given the stated beliefs and utilities—even among those with other-regarding preferences. It appears that people prefer the *act* of cooperating even though they're quite certain it will lead to a less-preferred outcome.⁷ As philosopher Daniel Dennett put it, "I'd feel better spending \$3 gained by cooperating than \$10 gained by defecting" (Hofstadter, 1983).

⁴We discuss this theory in more detail in Section V; see Dekel and Siniscalchi (2015) for an excellent exposition of Reny's argument.

⁵One traditional way to tackle this question using only strategy choice data is to see which model can fit the data best using something like a cross-validation horse race; see García-Pola et al. (2020), for example. This method cannot help identify new models, however, and may give misleading results if some models are misspecified or if the data are small (Healy and Park, 2023).

⁶The gang-of-four story was also pretty clearly refuted in a similar setting by Kagel and McGee (2016) using team chats and by Cox et al. (2015) using surprise replays of the repeated game.

⁷This is reminiscent of the warm-glow motivation for contributing to public goods suggested by Andreoni (1989).

We also run epistemic experiments on a dominance solvable game and an asymmetric coordination game. In both games almost all players have purely selfish preferences.⁸ In the dominance solvable game we see many subjects whose beliefs 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 outcome, and a certain form of loss aversion may cause them to deviate from that best response. Similarly, in the 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 the payments that they and their opponent receive, and not about the strategies that led to those payments. We refer to this as *consequentialism*, and discuss in Section VIII why we believe our methodology forces us to make this assumption. Second, our notion of rationality 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 2×2 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: almost all second movers are C-EU-rational. Since second movers face no strategic uncertainty, this result suggests that most irrationality is born out of strategic uncertainty.

Relying so heavily on elicitation data necessitates that we provide a lengthy discussion of our elicitation methodology, including how the elicitations were presented, how subjects

⁸We say that someone has “selfish preferences” in a game if their ordinal ranking of outcomes coincides with the selfish ranking. This doesn't mean they don't care about others' payoffs; it just means that in these particular games their other-regarding preferences are not strong enough to alter their ranking of outcomes. This is common, for example, in zero-sum (or nearly zero-sum) games.

were trained to understand the methods and report truthfully, and the possible limitations of using these methods. But this discussion can detract from the results themselves, so we relegate it to a later section. It appears in Section VIII, after the conceptual framework (Section III), experimental design (Section IV), and results (Section V for the centipede games and Sections VI and VII for the 2×2 games). Section IX concludes.

II. RELATED LITERATURE

Weibull (2004) argues 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).⁹ We take Weibull’s criticism as a launching point for our investigation: By eliciting cardinal utilities we can test directly game theoretic concepts without making strong assumptions on subjects’ preferences.¹⁰

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 in the setting where players see each others’ elicited preferences.¹¹ Although focusing on ordinal preferences simplifies the elicitation procedure, it necessarily limits most of their analyses to dominant strategies. For example, in a coordination game the players’ best responses depend on both their beliefs and their cardinal utility. With only ordinal information it is impossible to say whether an observed action is a best response or not. For this reason, we insist on eliciting cardinal utilities despite the added complexity.

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), Bornstein et al. (2004), and

⁹What we call a game form he calls a “game protocol”. The terminology “game form” comes from the mechanism design literature.

¹⁰Weibull (2004) also discusses how players’ preferences might depend on their opponents’ preferences (building on Levine, 1998), and that 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.

¹¹Revealing 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.

Wang (2023) 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.¹²

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.

Our elicitation data in centipede games points to an explanation similar to the “gang of four” story (Kreps et al., 1982; McKelvey and Palfrey, 1992), but without the reputation-building component. One model that roughly captures what we observe is provided by Wang (2023). She first runs an experiment comparing the game with linearly increasing total payoffs to a constant-sum version. She elicits first- and second-order beliefs, but not preferences. As in other studies, she finds less rationality (more passing) in the increasing-payoffs version, though without data on preferences this could also be driven by perfectly rational players who have other-regarding preferences. So, to rationalize the data, she models her experiment as a Bayesian game with two types—a selfish type and a non-selfish type that maximizes the sum of players’ payoffs—and shows that the Bayes-Nash equilibrium predicts passing behavior only in the linearly-increasing version. Our results suggest that the presence of non-selfish players is indeed important, though we rely on neither a specific model of preferences nor the assumption of Bayes-Nash equilibrium.

Beyond the centipede game, we also study several classic 2×2 game forms. An extensive literature review of these games is beyond the scope of this paper; see Gächter et al. (2024) and Mengel (2018) for two recent meta-studies of the Prisoners’ Dilemma, Cooper and Weber (2020) for a survey of coordination game experiments, and Camerer (2003) for an excellent coverage of behavioral game theory more generally.

As discussed above, there are complementary approaches for studying beliefs and rationality using clever experimental designs rather than direct elicitation. For example, Cason

¹²Cox 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.

and Sharma (2007) (among others) use robot opponents to induce mutual knowledge of beliefs, and they can measure the impact of social preferences by varying whether or not a human player receives the robot’s payoffs. As another example, Brocas et al. (2018) and Kneeland (2015) develop a novel “ring game” in which one player has a dominant strategy, the next player has a unique best response, the next has a unique best response to that, and so on. Kneeland (2015) 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. Friedenber 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. Brocas et al. (2018) also study a sequential-move version of the game and find greater rates of rationality, which consistent with our sequential-move results here. Calford and Chakraborty (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.

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 $\Gamma = (I, (S_i)_{i \in I}, \pi)$, where each S_i (for $i \in I$) is the set of strategies available to player i and $\pi : S_1 \times S_2 \rightarrow X$ is the outcome function that specifies a physical outcome for each strategy profile $s \in S = S_1 \times S_2$. Let π_i denote the projection of π 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’ beliefs are therefore probability distributions over the possible states of their opponent. Although states are private, 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 $u_i : X \rightarrow \mathbb{R}$, defined only over physical outcomes. This is elicited via probability equivalents. Specifically, for any x , $u_i(x)$ can be elicited by selecting “good” and

“bad” outcomes \bar{x} and \underline{x} such that $u_i(\bar{x}) \geq u_i(x) \geq u_i(\underline{x})$ and then finding the probability q^* such that player i is indifferent between x and the lottery $(q^*, \bar{x}; 1 - q^*, \underline{x})$, which pays \bar{x} with probability q^* and \underline{x} with probability $1 - q^*$. Assuming expected utility and normalizing $u_i(\bar{x}) = 1$ 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 π (meaning, for each possible outcome of the game form).¹³

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 Γ* (or, simply, *selfish in Γ*) if $x'_i > x_i$ implies $u_i(x') > u_i(x)$ for all x', x in the range of π . It is possible for someone to be consistent with selfishness in some game forms, 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. If a player's cardinal utility is simply an affine transformation of their payoffs (specifically, if $u_i(x_i) = (x_i - \underline{x}_i)/(\bar{x}_i - \underline{x}_i)$) then we say the player is *risk-neutral selfish*.

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.¹⁴

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

¹³To ensure $u_i(\bar{x}) > u_i(x) > u_i(\underline{x})$, we choose \bar{x} and \underline{x} such that $\bar{x}_i > x_i > \underline{x}_i$ for each i for every x in the range of π . If in fact $u_i(\bar{x}) < u_i(x)$ 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.

¹⁴There 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. If opponents know the mixing device is a fair coin then their beliefs should assign equal probability to these two states.

second-order belief.¹⁵ An entire infinite hierarchy of beliefs $\vec{p}_i = (p_i^1, p_i^2, p_i^3, \dots)$ 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 are not certain about their opponent's 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 .

To elicit $p_i^{1s}(s_{-i})$ we simply find the probability q^* such that i is indifferent between an act which pays outcome \bar{x} if $-i$ plays s_{-i} (and \underline{x} otherwise) and a lottery that pays \bar{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.¹⁶

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 maximizes the expected value of their elicited cardinal utility, given the first-order beliefs they report. We combine these to give our formal definition of rationality used in this paper.

Definition 1 (C-EU-rationality). A player i at state $\omega_i = (u_i, s_i, \vec{p}_i)$ (meaning they have cardinal utilities u_i , will play strategy $s_i \in S_i$, and have beliefs $\vec{p}_i = (p_i^1, p_i^2, \dots)$) is *consequentialist-expected-utility rational*, or *C-EU-rational*, if

- (1) the domain of u_i is X , the set of physical outcomes (consequentialism), and
- (2) $s_i \in \arg \max_{\hat{s}_i \in S_i} \sum_{s_{-i}} p_i^{1s}(s_{-i}) u_i(\pi(\hat{s}_i, s_{-i}))$ (EU-maximization).

Let $BR_i(p_i^{1s}|u_i) = \arg \max_{\hat{s}_i \in S_i} \sum_{s_{-i}} p_i^{1s}(s_{-i}) u_i(\pi(\hat{s}_i, s_{-i}))$ be the set of C-EU best replies of player i with marginal first-order belief p_i^{1s} and consequentialist cardinal utility u_i . C-EU-rationality thus requires that $s_i \in BR_i(p_i^{1s}|u_i)$. If player i is not C-EU-rational at ω_i (meaning $s_i \notin BR_i(p_i^{1s}|u_i)$) then we say they are *C-EU-irrational* at ω_i .

To understand the necessity of studying C-EU-rationality, consider instead if we applied the weaker notion of *rationality*, which is utility maximization 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

¹⁵It 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.

¹⁶We also elicit their probability that their best guess is correct. For more details on how this is incentivized—and why it is incentive compatible—see Section VIII.

unknown state $\omega_{-i} = (u_{-i}, s_{-i}, \vec{p}_{-i})$ (specifically, the s_{-i} component of the state) 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 $U_i : S_i \rightarrow \mathbb{R}$. Player i is *rational at ω_i* if $U_i(s_i) \geq U_i(s'_i)$ for all $s'_i \in S_i$.¹⁷ C-EU-rationality simply adds the requirement that $U_i(s_i) = \sum_{s_{-i}} p_i^{1s}(s_{-i}) u_i(\pi(s_i, s_{-i}))$.

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 form Γ 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.¹⁸ Preferences over outcomes, however, can be elicited since the payments do not depend on any decisions made by the other player. Thus, we believe that 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 ω_{-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 \bar{x} if $\omega_{-i} \in R_{-i}$ (and \underline{x} otherwise) and a lottery that pays \bar{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.¹⁹

At state $\omega = (\omega_i, \omega_{-i})$, the (Bayesian) game induced by Γ 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 $\omega_i = (u_i, s_i, \vec{p}_i)$.²⁰ The experimenter selects the game form Γ , and Γ 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 $h^t = (a^1, a^2, \dots, a^{t-1})$, 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 *Take* or *Pass*) chosen by the active player at decision node t' . Strategies map

¹⁷In this more general framework U_i would be included as a component of ω_i .

¹⁸Since U_i is ordinal one might consider testing rationality by testing Sen's conditions α and β , for example. But adding or deleting strategies from a game necessarily changes the game, so such consistency should not be expected.

¹⁹This requires that we assume that i believes the elicitation procedures are incentive compatible for $-i$. The experimental instructions therefore emphasize incentive compatibility multiple times.

²⁰Specifically, 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.

histories into actions (T or P). Let $\{z_1, \dots, z_7\}$ 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 $u_i(\pi(z_t))$.²¹

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 .²² 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 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 2×2 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 plays sequential-move versions of the 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) $\operatorname{argmax}_{u_{-i}} p_i^{1u}(u_{-i}|h^1)$ (the mode of i 's initial belief about $-i$'s utility), and

²¹This is slightly overloaded notation since π 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.

²²This 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.

(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.²³

We describe in Section VIII how we elicit each of these objects in an incentive compatible way. Arguably, the most novel of these elicitation techniques is the elicitation of cardinal utilities over outcomes, denoted by $u_i(\pi(s_i, s_{-i}))$. For example, suppose in a 2×2 game one cell has payoffs $x = (\$5, \$10)$, meaning \$5 for the row player and \$10 for the column player. How can we elicit the row player's cardinal utility for this outcome? As described above, we do this by asking them to consider a lottery that pays the outcome $\bar{x} = (\$20, \$20)$ with probability q and the outcome $\underline{x} = (\$0, \$0)$ with probability $1 - q$, and then finding the q^* that makes them exactly indifferent between this lottery and getting $x = (\$5, \$10)$ with certainty. Again, this indifference point represents their cardinal utility (assuming C-EU rationality) since we can normalize $u_i(\underline{x}) = 0$ and $u_i(\bar{x}) = 1$, giving $u_i(x) = q^*$.²⁴

The indifference point q^* is identified using a choice list procedure, described in detail in Section VIII. This elicited cardinal utility captures risk aversion, since risk averse individuals will have a lower indifference point q^* . It also captures social preferences since payments are made to both players. If $u_i(x) \notin [u_i(\underline{x}), u_i(\bar{x})]$ (perhaps due to inequality preferences) then this would show up in the data as boundary reports of $q^* = 0$ or $q^* = 1$, respectively, though in practice these values are rarely observed in our data. Finally, if the subject is not an expected utility maximizer then q^* should not be interpreted as a cardinal utility, but instead simply as an ordinal measure of their preference for $x = (\$5, \$10)$. Such a subject may then show up in the data as violating C-EU-rationality in this game form, though it's possible a C-EU-irrational subject may falsely appear to be C-EU-rational. Thus, our estimates of the frequency of C-EU-irrationality necessarily represent lower bounds for the actual incidence

²³Recall 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 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.

²⁴The centipede experiments involve higher payoffs at later terminal nodes, so for those experiments we set $\bar{x} = (\$30, \$30)$.

of C-EU-irrationality. All other elicitation procedures are based on choice list procedures, as explained in Section VIII.²⁵

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).

Actions in CENT were labeled Down (instead of Take) and Pass. Strategies in SIM and SEQ were labeled Up, Down, Left, and Right in all five game forms.

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 the opponent choosing Take at the next node leads to small losses for the player but large gains for the opponent then a modestly-altruistic subject might even prefer that outcome over choosing Take themselves. CENT-LO represents a low-cost treatment that has this feature. CENT-HI and CENT-ALL drastically increase the potential 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.²⁶

The SIM and SEQ experiments (featuring the 2×2 game forms) were not computerized; instead, subjects were given a printed booklet consisting of seven pages. Each of the first five pages showed a game form at the top, followed by the eight elicitation questions for that game form 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.

²⁵Following the advice of Danz et al. (2022) and Healy and Kagel (2023), the choice lists are explained in the opening instructions but not shown on the actual decision screens. Instead, subjects are simply asked for their indifference point and a choice list is implemented "behind the scenes." Subjects are told multiple times that it is in their best interest to report truthfully in each elicitation question.

²⁶Two other treatments were run with moderate costs. These are described in the online appendix. As expected, results lie "between" CENT-LO and CENT-HI.

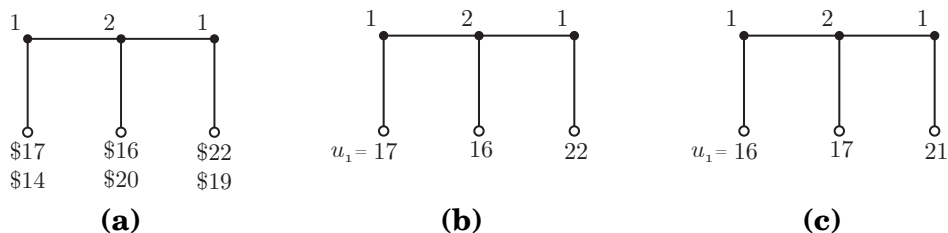


FIGURE II. (a) A three-node segment of a centipede game form. (b) The game with ‘selfish’ utilities. (c) The game with other-regarding preferences.

The sixth and seventh pages contain the 16 individual binary decisions intended to measure the subject’s risk aversion and ambiguity aversion. Each subject filled in their answers to all questions on all pages and turned in the booklet to the experimenter. The experimenter then matched each booklet with that of the corresponding opponent, randomly drew a question to be paid, 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. This method, suggested by Allais (1953), is incentive compatible if we assume subjects’ preferences over gambles satisfy monotonicity with respect to statewise dominance. This is strictly weaker than expected utility unless we further assume subjects reduce compound lotteries; 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 XV and Table XIV 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 II. 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 II shows decision nodes three, four, and five from CENT-LO. Panel (a) shows the segment from 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 risk-neutral selfishness. Panel (c) shows example utilities of a (hypothetical) non-selfish player who exhibits some degree of altruism.²⁷

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 choosing Pass is a best response—is the interval $[1/6, 1]$. We refer to the size of this basin as *SizeBAP*, which here equals $5/6$.²⁸ Roughly speaking, *SizeBAP* provides a measure of how tempted a player may be to Pass.²⁹ If choosing Pass 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.

²⁷Actual cardinal utility values were all required to be in $[0, 1]$; to simplify the illustration we rescale these examples to $[\underline{x}_i, \bar{x}_i] = [0, 30]$.

²⁸This is inspired by a similar measure used by Dal Bó and Fréchette (2011) to analyze repeated prisoners’ dilemmas.

²⁹In the larger game with more than three nodes, *SizeBAP* is only an approximate measure of the true temptation to Pass.

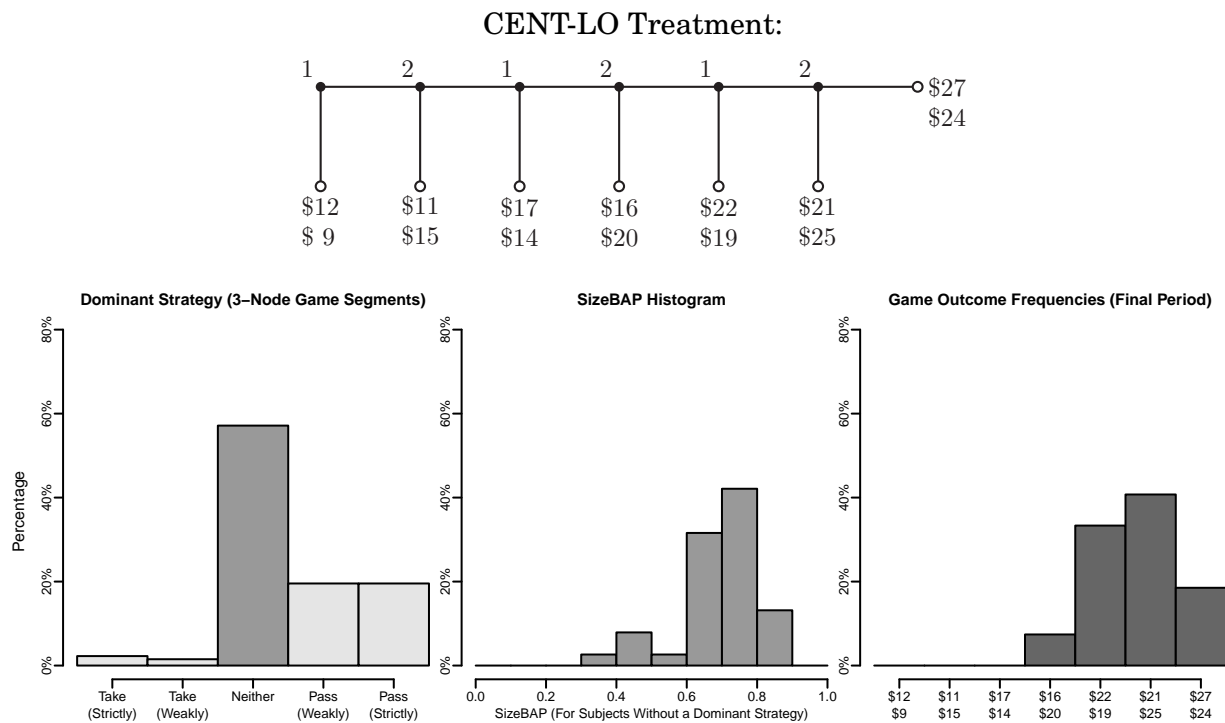


FIGURE III. The CENT-LO treatment. Top: The game form. Bottom left: Across all 3-node game segments, the percentage of subjects who have a dominant strategy to Take or Pass (or neither). Bottom middle: The temptation to Pass for subjects without a dominant strategy, as measured by *SizeBAP*. Bottom right: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

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 III shows the game form and results for the CENT-LO treatment. The top panel displays the game form. The risk to choosing Pass 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 risk-neutral selfish utilities 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 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.³⁰

In the middle histogram we take those subjects who have no dominant strategy and calculate the *SizeBAP* measure for each 3-node segment. These are the subjects who are consistent with selfishness in this game form. Again, the larger the *SizeBAP*, the more likely the subject's best response is to Pass. The *SizeBAP* here is typically large, often nearly as high as the risk neutral selfish value of 0.83. If these subjects are C-EU-rational then they should Pass as long as they believe there exists a reasonable fraction of altruists. Given that 40% of subjects are in fact altruistic (having a dominant strategy to Pass), we should expect that many of these selfish subjects will choose 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 choosing Pass is much higher. In the CENT-HI treatment (shown in Figure IV) choosing Pass 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*. In particular, 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

³⁰This 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.

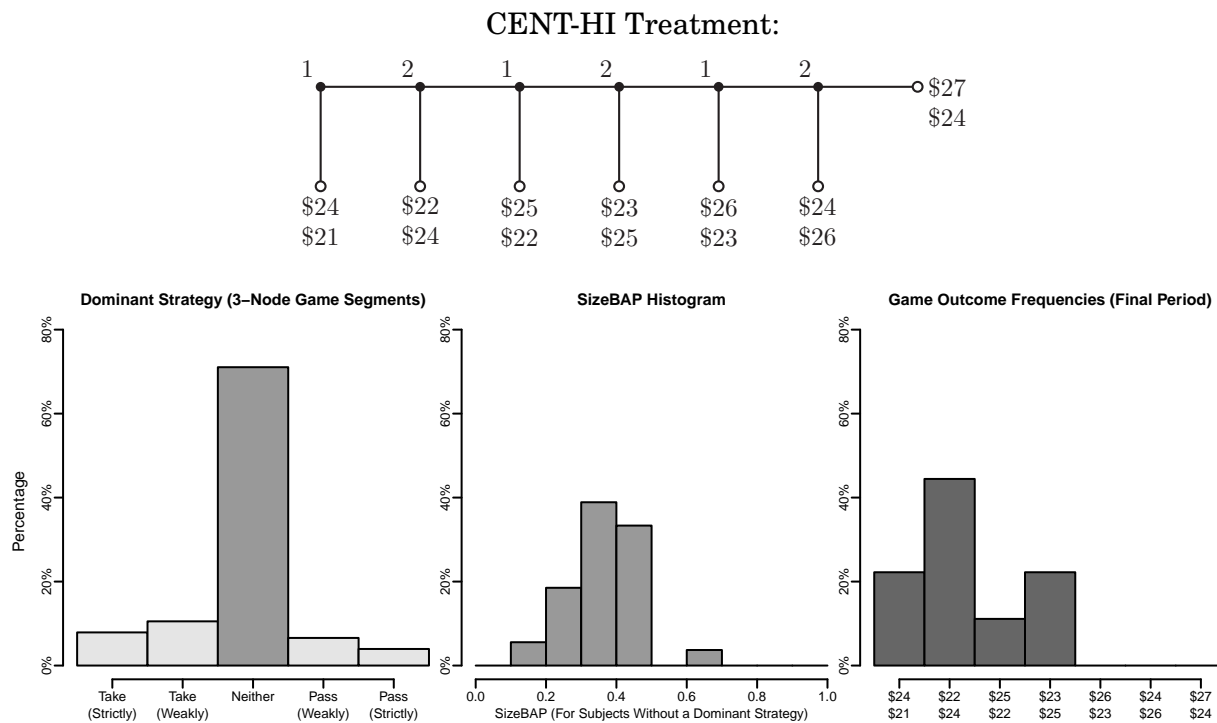


FIGURE IV. The CENT-HI treatment.

the distribution of actual types simply doesn't justify that belief. Indeed, the game typically ends much earlier as a result, with over 2/3 of games ending in the first two nodes.

Finally, we test a winner-take-all version of the centipede game form in our CENT-ALL treatment (Figure V).³¹ Here, Passing risks almost the entire payoff for a gain of \$4. For a risk-neutral selfish subject the *SizeBAP* is 0.22 at their first 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.³² Even the *SizeBAP* measures are close to the risk-neutral selfish 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.³³

³¹This 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 similar results to ours. There are multiple design differences between these studies, making it difficult to pinpoint which features cause lower Pass rates.

³²Again, “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.

³³Recall 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.

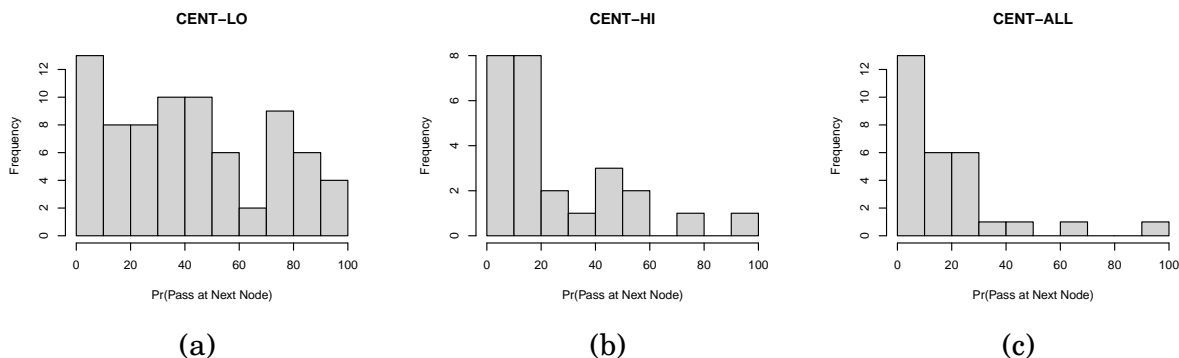


FIGURE VI. Histograms of non-altruists’ belief that the opponent will Pass at the next node.

fraction of subjects who would prefer to pass altruistically when the cost of doing so is low, but not when that cost is increased.³⁴

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 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.³⁵ 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.³⁶

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

³⁴Interestingly, 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.

³⁵Despite the small sample size, Passing in CENT-ALL is significantly lower than in CENT-LO, with a χ^2 p -value of 0.038.

³⁶If 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.

a belief greater than 39.2%. In CENT-HI and CENT-ALL, however, beliefs are significantly lower (see Figure VI), consistent with the lower incidence of altruistic types.³⁷

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 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.³⁸ 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

³⁷Figure VI looks very similar if we restrict attention to those with selfish preferences. This is because selfish types constitute 88% of all non-altruists.

³⁸Here 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.

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 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.³⁹

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 (χ^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).⁴⁰ We do find that the altruists are generally more C-EU-rational than non-altruists in CENT-LO (64.1% vs. 46.3%; χ^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 χ^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.⁴¹

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

³⁹Beliefs 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.

⁴⁰The 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.

⁴¹Similarly, 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.

riskiness of Passing is increased then more players become consistent with selfishness, and so the selfish types no longer have an incentive to Pass.⁴² Consequently, players Take earlier.

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.⁴³

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.

⁴²Recall 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.

⁴³In the online 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.

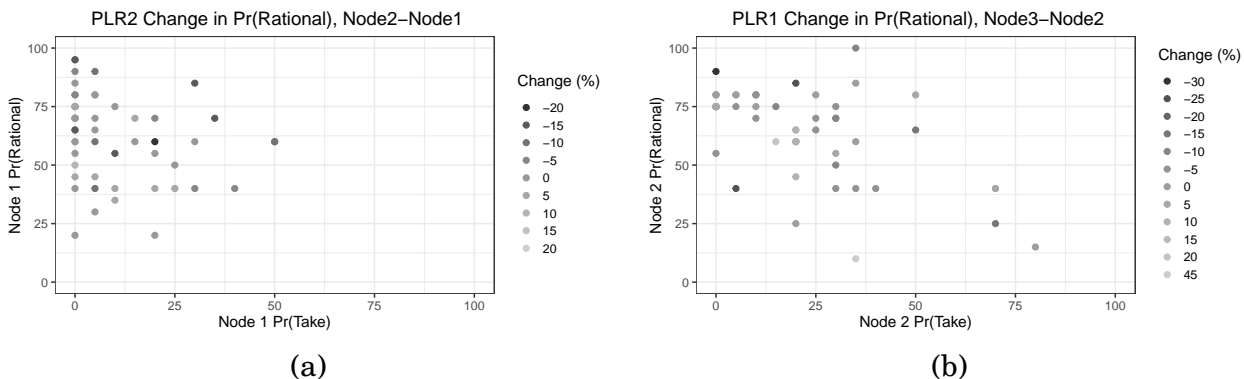


FIGURE VII. Change in belief in rationality in CENT-LO after observing the opponent choose Pass.

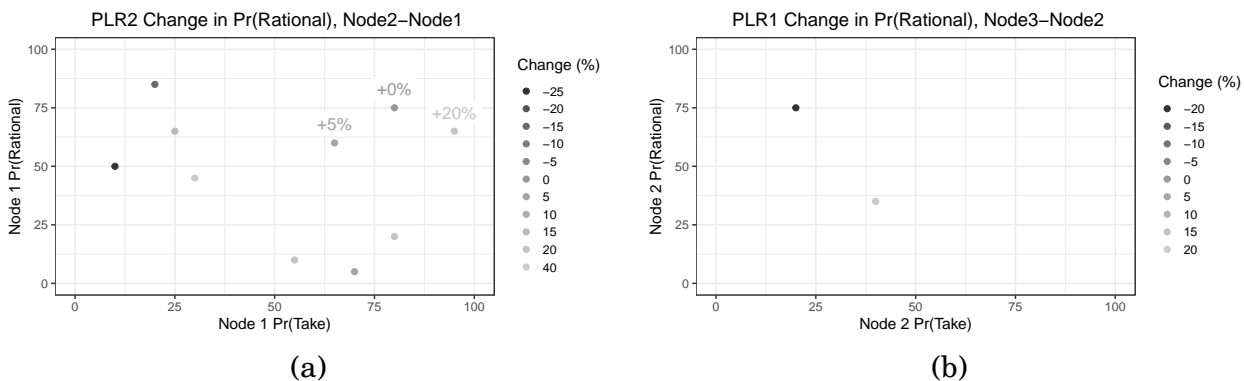


FIGURE VIII. Change in belief in rationality in CENT-ALL after observing the opponent choose Pass.

How does player 2’s belief in rationality change in response? First consider CENT-LO. Panel (a) of Figure VII 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 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.

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

FIGURE IX. The Prisoners' Dilemma Game Form

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 VIII. 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 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 $\Pr(\text{Take})$ and $\Pr(\text{Rational})$ from panel (a) show that the belief in rationality actually increases, not decreases, when the opponent passes.⁴⁴ 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

Next we turn to the SIM experiment, in which subjects played five 2×2 simultaneous-move game forms without feedback. Of those game forms, the only one in which we observe a substantial fraction of non-selfish preferences is the Prisoners' Dilemma, shown in Figure IX.⁴⁵ 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.⁴⁶ 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

⁴⁴In CENT-HI there are also three observations of player 2 having $\Pr(\text{Take})$ and $\Pr(\text{Rational})$ 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 $+20\%$.

⁴⁵In the actual experiment the strategies were labeled "U" and "D".

⁴⁶For 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 $u_i(\pi(D,C)) < u_i(\pi(C,C))$.

Pref. Type	$BR_i(C)$	$BR_i(D)$	% Subj.	$BR_i(p_i^{1s} u_i) = C$		$BR_i(p_i^{1s} u_i) = D$	
				$s_i = C$	$s_i = D$	$s_i = C$	$s_i = D$
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 .

whom we have valid preference and belief data.⁴⁷ 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).

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.”⁴⁸

⁴⁷Two 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 $s_i = D$.

⁴⁸In 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 X. The Dominance Solvable Game Form

Column Players' Types Pref. Type	Column Players' Types		% Subj.	$BR_i(p_i^{1s} u_i) = L$		$BR_i(p_i^{1s} u_i) = R$	
	$BR_i(U)$	$BR_i(D)$		$s_i = L$	$s_i = R$	$s_i = L$	$s_i = R$
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 .

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 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.⁴⁹

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 X. 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

⁴⁹For brevity, analyses of the remaining two game forms are relegated to the online appendix.

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.⁵⁰ Yet 25% of column players violate iterated dominance (in terms of payoffs) by choosing L . The 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.⁵¹ 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.⁵²

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 XI.⁵³ As in the SIM treatment, all row players choose Up, and all do 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

⁵⁰One 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.

⁵¹These 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.

⁵²Ambiguity 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 .

⁵³Two column players are omitted because action choices of the row players were not recorded.

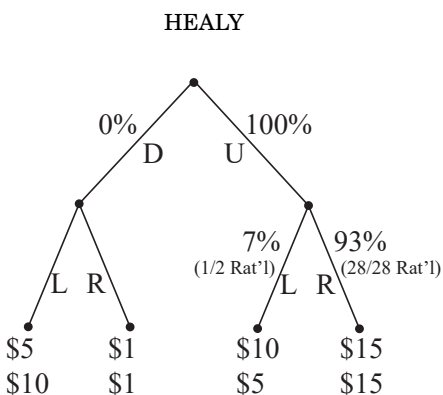


FIGURE XI. Frequencies of choices and fractions of subjects who are C-EU-Rational in the sequential-move Dominance Solvable game form.

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

FIGURE XII. The Asymmetric Coordination Game Form

Row's Type	$BR_1(L)$	$BR_1(R)$	% Subj.	$BR_1(p_1^{1s} u_1) = U$		$BR_1(p_1^{1s} u_1) = D$	
				$s_1 = U$	$s_1 = D$	$s_1 = U$	$s_1 = D$
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.	$BR_2(p_2^{1s} u_2) = L$		$BR_2(p_2^{1s} u_2) = R$	
				$s_2 = L$	$s_2 = R$	$s_2 = L$	$s_2 = R$
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.

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 XII. 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.⁵⁴ 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 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 .⁵⁵ 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.⁵⁶ 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 XV 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 XIII. 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 U , after which 26 out of 28 second movers acquiesce and choose L . Thus, negative reciprocity (or,

⁵⁴Four 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 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.

⁵⁵The median belief for this group is $p_1^{1s}(R) = 0.80$.

⁵⁶Following footnote 52, 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 .

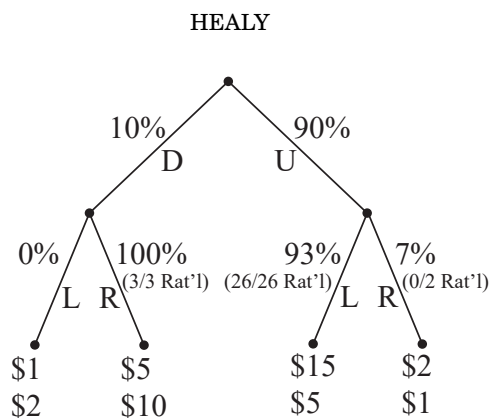


FIGURE XIII. Frequencies of choices and fractions of subjects who are C-EU-Rational in the sequential-move Asymmetric Coordination game form.

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.⁵⁷

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 τ 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

Having presented all of the results, we now return to the discussion of our elicitation methodology, including issues of incentive compatibility, multiple switching, order effects, contamination, and consequentialism. The instructions and experimental interfaces all appear in an online supplemental appendix.

Row #	Option A	or	Option B
0	You get \$5 and they get \$10	or	0% chance you both get \$20
1	You get \$5 and they get \$10	or	1% chance you both get \$20
⋮	⋮	⋮	
99	You get \$5 and they get \$10	or	99% chance you both get \$20
100	You get \$5 and they get \$10	or	100% chance you both get \$20

FIGURE XIV. The choice list used to elicit $u_i(\$5, \$10)$.

Elicitation Via Choice Lists

Every quantity we elicit can be identified as the subject's indifference point in some list of binary choices. For example, as discussed in Sections III and IV, if a subject satisfies C-EU rationality and has cardinal utility $u_i(\$5, \$10) = q^*$, then that subject is indifferent between outcome $(\$5, \$10)$ and a lottery that gives $\bar{x} = (\$20, \$20)$ with probability q^* and $\underline{x} = (\$0, \$0)$ otherwise. To find that indifference, we present the subject with the choice list shown in Figure XIV and ask them at which row they would switch from Option A to Option B. The various elicitation differ only in what appears as Option A and Option B. At the end of the experiment one choice is chosen at random for payment; if that choice is an elicitation then one row from the choice list is randomly chosen and the subject is paid their choice on that row.

Before the experiment begins all subjects are trained on how choice lists work and why they are incentive compatible. The training begins with a simple example of how choice lists could be used to elicit their dollar value for an item such as a cheeseburger. Each row offers a choice between a cheeseburger (Option A) and a certain dollar amount that increases row-by-row (Option B). Subjects learn that they should have a single switch point, that each row is equally likely to be chosen for payment, and that their chosen switch point does not affect which row is paid. Then they are shown via an example how misreporting their value leads to less-preferred payments on some rows without affecting their payment on the remaining rows. The instructions then say that “by lying, you are never made better off on any question, and for some questions you are made worse off” and that “you have no incentive to lie about your value.” Formally, choice lists are incentive compatible as long as subjects respect statewise dominance Azrieli et al. (2018).

Next, they're shown how a similar list can be used to elicit their “probability value” (*i.e.*, their cardinal utility) for a cheeseburger instead of a dollar value. Here, Option B becomes “A $p\%$ chance of \$20,” with p increasing down the list. Then, instead of a cheeseburger, we could use such a list to elicit their probability value for two-person payments such as “you

⁵⁷The 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.

get \$5 and someone else in the room gets \$10.” In this case Option B now pays \$20 to both people, giving exactly the list shown in Figure XIV.

Then we show them how a list can be used to elicit their personal probability that an event will occur. For example, Option A would be “You get \$20 if a coin flip lands heads” and Option B would be “You get \$20 with probability $p\%$.” This method is used to elicit any probabilistic belief. For example, their first order belief $p_i^{1s}(s_{-i})$ is elicited by having Option A be “You get \$20 if your opponent plays s_{-i} .”

Finally, they’re shown by example how we can elicit their most-likely outcome of some random event, and why the list payment method makes it so that reporting their true most-likely outcome maximizes their chances of being paid the given prize. To illustrate, suppose there are two events E^1 and E^2 , they believe E^1 is more likely, and they can pick for which event they’ll report a belief. Then, by comparing the choice list for E^1 to the choice list for E^2 side-by-side, we can see that choice list for E^1 will give a weakly better outcome on every row. Thus, when given a choice, it is incentive compatible to report their belief about the more-likely event. A similar insight appears in Möbius et al. (2022).

The training concludes by reminding them that, in all questions, they always have an incentive to tell the truth. After the elicitation instructions, subjects are then given instructions specific to the games being played.

During the experiment the choice lists are not shown while making actual decisions. Instead, the subjects are told to report their probability value or their belief and the experimenter will fill out the list for them, switching from Option A to Option B at their reported value. An advantage of this method is that it should minimize any possible list ordering effects.⁵⁸ A potential disadvantage is that we enforce a single switch point on all lists; however, in a previous study using the same subject pool Brown and Healy (2018) found that only around 5% of subjects exhibited multiple-switching behavior in similar choice lists.

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.⁵⁹

When eliciting cardinal utilities, we ask the subject for their probability value of each outcome x in the game form’s matrix (or, at each terminal node in an extensive-form tree). Recall that the elicited cardinal utility $u_i(x)$ captures both risk aversion and social preferences, and if $u_i(x) \notin [u_i(\underline{x}), u_i(\bar{x})]$ then we would observe switch points at either 0% or 100%. We use $\bar{x} = (\$20, \$20)$ in the SIM and SEQ treatments and $\bar{x} = (\$30, \$30)$ in all centipede treatments. If a cardinal utility question is chosen for payment then the choice of only one

⁵⁸A similar method is recommended by Danz et al. (2022). See Birnbaum (1992); Harrison et al. (2005); Andersen et al. (2006); Harrison et al. (2007a,b); Freeman et al. (2019); and Beauchamp et al. (2020), for examples of such list ordering effects.

⁵⁹Some authors prefer to perform these draws at the beginning of the experiment to reduce hedging opportunities. In the domain of individual decision-making, Oechssler et al. (2019) and Baillon et al. (2022) find that changing the timing of the draws does not affect choices.

of the two subjects (selected at random) is paid, and then both subjects receive a payment based on that choice. This ensures that i 's choice (if chosen for payment) truly determines the final payment of both i and $-i$.

When asking subjects their best guess of their opponent's utilities or first-order beliefs, we have the subject make a guess of the utility or belief values reported by their opponent along with the subject's probability that their guess is correct. This most-likely guess is properly incentivized as described above. In the SIM and SEQ treatments we also elicited their second-most-likely guess by having them submit a second guess that must differ from the first.⁶⁰

All elicited probabilities (including cardinal utilities and guesses of utilities and beliefs) are restricted to multiples of 5% for simplicity, and to increase the probability of correctly guessing the opponent's reported utilities and beliefs.

In the SEQ and SIM treatments we also elicit the subject's best guess of the ordinal ranking of their opponent's reported utility values. Specifically, they guess to which outcome their opponent gave the highest probability value, the second highest probability value, and so on. And they report their belief that this ordinal ranking is correct. We pay them for this most-likely belief as described above. And we also collect their second-most-likely ordinal ranking. In the centipede treatments we do not elicit these ordinal rankings; instead, when eliciting their own cardinal utility or their best guesses of their opponent's cardinal utility, we show them the implied ordinal ranking and ask them to confirm that this ranking is consistent with their intended report.

Strategy choices in SEQ and SIM are incentivized in the usual way, by paying both subjects for the actual outcome of the game. In the centipede treatments we also elicit at each node their planned strategy for the continuation game, which is simply the node at which they currently plan on playing Take. (If the active player chooses Take at the current node then no future plan is elicited from that player.) If these elicited plans are chosen for payment then one node that was reached in the game is randomly selected, the strategies reported by the players at that node are played out, and both players receive the payoffs from the terminal node that would be reached under those two reported strategies.

First order beliefs in the centipede games are elicited at every node by asking the player's belief that their opponent would choose Take at each upcoming node if it were reached. If paid, these beliefs are compared to the planned strategy submitted by the opponent at that same node. For second-order beliefs we ask each player to report their best guess of their opponent's belief that the player would choose Take at each upcoming node. We also ask their probability that this guess is correct and pay for this most-likely guess as described above. This is elicited from the active player even if they choose Take at the current node.

⁶⁰If a subject reports a higher probability for the second-most-likely event then we switch which is labeled as their highest-probability guess when analyzing the data.

Eliciting Beliefs About 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}_{-i}^{1s} and \tilde{p}_{-i}^{1s} . Utility is known to be \hat{u}_{-i} . Suppose player $-i$ is rational at $(\hat{s}_{-i}, \hat{p}_{-i}^{1s})$ and $(\tilde{s}_{-i}, \tilde{p}_{-i}^{1s})$, but not at $(\hat{s}_{-i}, \tilde{p}_{-i}^{1s})$ or $(\tilde{s}_{-i}, \hat{p}_{-i}^{1s})$. If i reports that both strategies are equally likely and both beliefs are equally likely then their belief in rationality cannot be determined: On one extreme, 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. On the other extreme, it could come from a belief that they are perfectly negatively correlated, putting zero probability on C-EU-rationality. Thus we cannot infer anything about i 's belief in C-EU-rationality from their marginal beliefs alone. Rather than trying to elicit the joint distribution, we instead elicit beliefs about C-EU-rationality directly in our experiment.⁶¹

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 a choice list as described above—because we can observe whether the opponent's choices actually are “consistent” or not.⁶²

Possible 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.⁶³ 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.

⁶¹Wang (2023) elicits this joint distribution, though the second-order beliefs are necessarily quite coarse.

⁶²In the CENT experiment we changed the wording to “consistent with weighted value theory,” where “weighted value theory” refers to expected utility maximization.

⁶³Recall 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 followed elicitation questions from the third.

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	<i>U</i> or <i>L</i>	75	19	72	74	19	26	64	32	66	36
	<i>D</i> or <i>R</i>	0	56	3	1	55	48	9	41	5	38
SIM-NoE	<i>U</i> or <i>L</i>	30	2	28	30	11	9	24	7	19	23
	<i>D</i> or <i>R</i>	0	28	2	0	19	21	6	23	11	7
χ^2 test <i>p</i> -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 online appendix. *Rejection of equality between treatments at the 5% level with either a Bonferroni or Holm-Bonferroni correction.

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.

However, we can also test for contamination by re-running all of these experiments 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.⁶⁴ 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 60.⁶⁵

Figure XV 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 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.

⁶⁴NoE 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 Take or Pass at each node reached.

⁶⁵It 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.

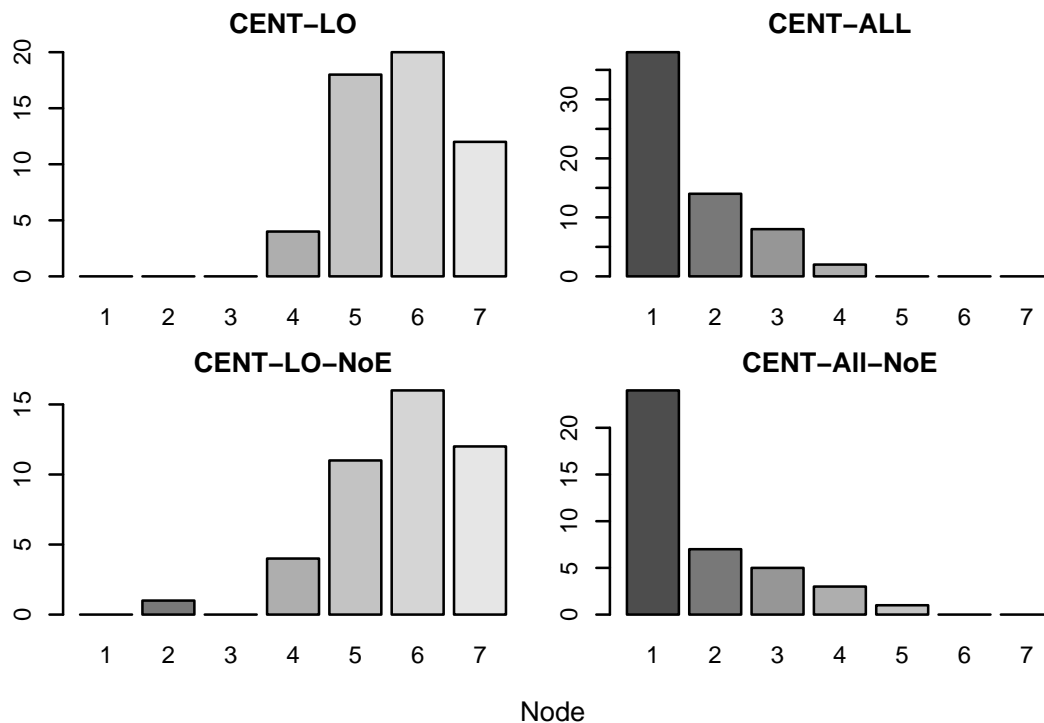


FIGURE XV. Histograms of how frequently each terminal node was reached in each treatment. NoE refers to a “No Elicitation” treatment. Periods 3 and 4 are pooled.

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 χ^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 XVI. A game form in which consequentialism is likely to fail.

Possible Issue #2: Consequentialism

Traditional game theory takes as primitive cardinal payoffs of the form $u_i(s_i, s_{-i})$. This 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 XVI. For the row player it is plausible that $u_1(U, L) \neq u_1(D, R)$. 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, $u_1(U, L) \neq u_1(D, R)$ 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 XVI 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

The intent of this paper is to show how elicitation data—specifically the elicitation of utilities and beliefs—can be used to identify the varied phenomena across game forms, leading to deeper insights than can be gleaned from strategy choice alone. So, 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 examples of irrationality 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.⁶⁶

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 modeled 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. While it is tempting to try to simulate noisy responses to see which biases might emerge in our conclusions, the lessons learned from that exercise are likely to be sensitive

⁶⁶Details of this, and of the tests of level- k and quantal response equilibrium beliefs, are available upon request.

to assumptions about the structure of the noise. And such assumptions are not something we're well equipped to test with our existing data.

Obviously, our elicitation methodology can be applied wholesale to any games of interest. For example, given the findings of Calford and Chakraborty (2022), it might be interesting to run epistemic experiments on games with more than two players to explore whether players believe others have the 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., T. DOHMEN, AND M. WIBRAL (2008): "Do the reciprocal trust less?" *Economics Letters*, 99, 454–457.
- ANDERSEN, S., G. W. HARRISON, M. I. LAU, AND E. E. RUTSTRÖM (2006): "Elicitation Using Multiple Price List Formats," *Experimental Economics*, 9, 383–405.
- ANDREONI, J. (1989): "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, 97, 1447–1458.
- AUMANN, R. (1987): "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.
- AUMANN, R. AND A. BRANDENBURGER (1995): "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1161–1180.
- AZRIELI, Y., C. P. CHAMBERS, AND P. J. HEALY (2018): "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, 126, 1472–1503.
- BAILLON, A., Y. HALEVY, AND C. LI (2022): "Randomize at Your Own Risk: On the Observability of Ambiguity Aversion," *Econometrica*, 90, 1085–1107.
- BEAUCHAMP, J. P., D. J. BENJAMIN, D. I. LAIBSON, AND C. F. CHABRIS (2020): "Measuring and controlling for the compromise effect when estimating risk preference parameters," *Experimental Economics*, 23, 1069–1099.
- BECKER, G. M., M. H. DEGROOT, AND J. MARSCHAK (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.

- BIRNBAUM, M. H. (1992): "Violations of Monotonicity and Contextual Effects in Choice-Based Certainty Equivalents," *Psychological Science*, 3, 310–315.
- BLANCO, M., D. ENGELMANN, A. K. KOCH, AND H.-T. NORMANN (2014): "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis," *Games and Economic Behavior*, 87, 122–135.
- BLANCO, M., D. ENGELMANN, AND H. T. NORMANN (2011): "A within-subject analysis of other-regarding preferences," *Games and Economic Behavior*, 72, 321–338.
- BOLTON, G. E. AND A. OCKENFELS (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.
- BORNSTEIN, G., T. KUGLER, AND A. ZIEGELMEYER (2004): "Individual and group decisions in the centipede game: Are groups more "rational" players?" *Journal of Experimental Social Psychology*, 40, 599–605.
- BROCAS, I., J. D. CARRILLO, AND A. SACHDEVA (2018): "The path to equilibrium in sequential and simultaneous games: A mousetracking study," *Journal of Economic Theory*, 178, 246–274.
- BROWN, A. L. AND P. J. HEALY (2018): "Separated decisions," *European Economic Review*, 101, 20–34.
- BRUNNER, C., T. F. KAUFFELDT, AND H. RAU (2016): "Mutual knowledge of preferences and equilibrium play: experimental evidence," .
- CALFORD, E. M. AND A. CHAKRABORTY (2022): "Higher-Order Beliefs in a Sequential Social Dilemma," .
- CAMERER, C. F. (2003): *Behavioral Game Theory*, Princeton, NJ: Princeton University Press.
- CAMERER, C. F., T.-H. HO, AND J.-K. CHONG (2004): "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119, 861–898.
- CASON, T. N. AND T. SHARMA (2007): "Recommended Play and Correlated Equilibria: An Experimental Study," *Economic Theory*, 33, 11–27.
- CLARK, K. AND M. SEFTON (2001): "The Sequential Prisoner's Dilemma: Evidence on Reciprocation," *Economic Journal*, 111, 51–68.
- COLLINS, S. M. AND D. JAMES (2015): "Response mode and stochastic choice together explain preference reversals," *Quantitative Economics*, 6, 825–856.
- COOPER, D. J. AND R. A. WEBER (2020): "Recent advances in experimental coordination games," in *Handbook of Experimental Game Theory*, ed. by C. M. Capra, R. T. A. Croson, M. L. Rigdon, and T. S. Rosenblat, Cheltenham, UK: Edward Elgar Publishing, 149–183, section: Handbook of Experimental Game Theory.
- COX, C. A., M. T. JONES, K. E. PFLUM, AND P. J. HEALY (2015): "Revealed reputations in the finitely repeated prisoners' dilemma," *Economic Theory*, 58, 441–484.
- COX, J. C. AND D. JAMES (2012): "Clocks and Trees: Isomorphic Dutch Auctions and Centipede Games," *Econometrica*, 80, 883–903.
- (2015): "On Replication and Perturbation of the McKelvey and Palfrey Centipede Game Experiment," in *Replication in Experimental Economics*, Emerald Group Publishing Limited, vol. 18 of *Research in Experimental Economics*, 53–94.
- DAL BÓ, P. AND G. R. FRÉCHETTE (2011): "The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence," *American Economic Review*, 101, 411–429.
- DANZ, D., S. HUCK, AND P. JEHIEL (2016): "Public Statistics and Private Experience: Varying Feedback Information in a Take-or-Pass Game," *German Economic Review*, 17,

359–377.

- DANZ, D., L. VESTERLUND, AND A. J. WILSON (2022): “Belief Elicitation and Behavioral Incentive Compatibility,” *American Economic Review*, 112, 2851–2883.
- DEKEL, E. AND M. SINISCALCHI (2015): “Epistemic Game Theory,” in *Handbook of Game Theory*, ed. by H. P. Young and S. Zamir, Oxford: North Holland, vol. 4, 619–702.
- EYSTER, E. AND M. RABIN (2005): “Cursed Equilibrium,” *Econometrica*, 73, 1623–1672.
- FEHR, E. AND K. M. SCHMIDT (1999): “A Theory of Fairness, Competition, and Cooperation,” *Quarterly Journal of Economics*, 114, 817–868.
- FEY, M., R. D. MCKELVEY, AND T. PALFREY (1996): “An experimental study of constant-sum centipede games,” *International Journal of Game Theory*, 25, 269–287.
- FREEMAN, D. J., Y. HALEVY, AND T. KNEELAND (2019): “Eliciting risk preferences using choice lists,” *Quantitative Economics*, 10, 217–237.
- FRIEDENBERG, A. AND T. KNEELAND (2023): “Is Bounded Reasoning about Rationality Driven by Limited Ability?” .
- FUDENBERG, D. AND D. K. LEVINE (1997): “Measuring Players’ Losses in Experimental Games,” *The Quarterly Journal of Economics*, 112, 507–536.
- GARCÍA-POLA, B., N. IRIBERRI, AND J. KOVÁŘÍK (2020): “Non-equilibrium play in centipede games,” *Games and Economic Behavior*, 120, 391–433.
- GILBOA, I. AND D. SCHMEIDLER (1989): “Maxmin Expected Utility with Non-Unique Prior,” *Journal of Mathematical Economics*, vol. 18, issue 2, pp. 141–153.
- GREINER, B. (2015): “Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE,” *Journal of the Economic Science Association*, 1, 114–125.
- GRETHER, D. M. (1981): “Financial Incentive Effects and Individual Decisionmaking,” Tex.note: Caltech working paper.
- GÄCHTER, S., K. LEE, M. SEFTON, AND T. O. WEBER (2024): “The role of payoff parameters for cooperation in the one-shot Prisoner’s Dilemma,” *European Economic Review*, 166, 104753.
- GÄCHTER, S., D. NOSENZO, E. RENNER, AND M. SEFTON (2012): “Who Makes a Good Leader? Cooperativeness, Optimism, and Leading-by-Example,” *Economic Inquiry*, 50, 953–967.
- HARRISON, G. W., M. LAU, E. E. RUTSTRÖM, AND M. B. SULLIVAN (2005): “Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues,” in *Field experiments in economics*, ed. by J. Carpenter, G. Harrison, and J. List, Emerald Group Publishing Limited, vol. 10 of *Research in Experimental Economics*.
- HARRISON, G. W., M. I. LAU, AND E. E. RUTSTRÖM (2007a): “Estimating Risk Attitudes in Denmark: a Field Experiment,” *Scandinavian Journal of Economics*, 109, 341–368.
- HARRISON, G. W., J. A. LIST, AND C. TOWE (2007b): “Naturally Occurring Preferences and Exogenous Laboratory Experiments: A Case Study of Risk Aversion,” *Econometrica*, 75, 433–458.
- HEALY, P. J. AND J. KAGEL (2023): “Testing Elicitation Mechanisms Via Team Chat,” .
- HEALY, P. J. AND H. PARK (2023): “Model selection accuracy in behavioral game theory: A simulation,” *European Economic Review*, 152, 104362.
- HOFSTADTER, D. (1983): “Metamagical Themas: The calculus of cooperation is tested through a lottery,” *Scientific American*, 248, 14–28.
- HOLT, C. A. AND S. K. LAURY (2002): “Risk Aversion and Incentive Effects,” *American Economic Review*, 92, 1644–1655.

- HOLT, C. A. AND A. M. SMITH (2016): “Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes,” *American Economic Journal: Microeconomics*, 8, 110–139.
- KAGEL, J. H. AND P. MCGEE (2016): “Team versus Individual Play in Finitely Repeated Prisoner Dilemma Games,” *American Economic Journal: Microeconomics*, 8, 253–276.
- KAWAGOE, T. AND H. TAKIZAWA (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.
- KREPS, D. M. (1990): *Game Theory and Economic Modelling*, Oxford University Press, google-Books-ID: qMoTDAAAQBAJ.
- KREPS, D. M., P. MILGROM, J. ROBERTS, AND R. WILSON (1982): “Rational Cooperation in the Finitely Repeated Prisoners’ Dilemma,” *Journal of Economic Theory*, 27, 245–252.
- KROCKOW, E. M., B. D. PULFORD, AND A. M. COLMAN (2015): “Competitive Centipede Games: Zero-End Payoffs and Payoff Inequality Deter Reciprocal Cooperation,” *Games*, 6, 262–272.
- LEVINE, D. (1998): “Modelling Altruism and Spitefulness in Experiments,” *Review of Economic Dynamics*, 1, 593–622.
- MCGRANAGHAN, C., K. NIELSEN, T. O’DONOGHUE, J. SOMERVILLE, AND C. D. SPRENGER (2024): “Distinguishing Common Ratio Preferences from Common Ratio Effects Using Paired Valuation Tasks,” *American Economic Review*, 114, 307–347.
- MCINTOSH, C. R., J. F. SHOGREN, AND A. J. MORAVEC (2009): “Can tournaments induce rational play in the Centipede game? Exploring dominance vs. strategic uncertainty,” *Economics Bulletin*, 29, 2018–2024.
- MCKELVEY, R. D. AND T. R. PALFREY (1992): “An Experimental Study of the Centipede Game,” *Econometrica*, 60, 803–836.
- (1995): “Quantal Response Equilibria for Normal Form Games,” *Games and Economic Behavior*, 10, 6–38.
- MCKELVEY, R. D. AND T. R. PALFREY (1998): “Quantal Response Equilibria for Extensive Form Games,” *Experimental Economics*, 1, 9–41.
- MENGEL, F. (2018): “Risk and Temptation: A Meta-study on Prisoner’s Dilemma Games,” *The Economic Journal*, 128, 3182–3209.
- MIETTINEN, T., M. KOSFELD, E. FEHR, AND J. WEIBULL (2020): “Revealed preferences in a sequential prisoners’ dilemma: A horse-race between six utility functions,” *Journal of Economic Behavior & Organization*, 173, 1–25.
- MYERSON, R. B. (1991): *Game Theory: Analysis of Conflict*, Cambridge, MA: Harvard University Press.
- MÖBIUS, M. M., M. NIEDERLE, P. NIEHAUS, AND T. S. ROSENBLAT (2022): “Managing self-confidence: Theory and experimental evidence,” *Management Science*.
- NAGEL, R. C. (1995): “Unraveling in Guessing Games: An Experimental Study,” *American Economic Review*, 85, 1313–1326.
- OECHSSLER, J., H. RAU, AND A. ROOMETS (2019): “Hedging, ambiguity, and the reversal of order axiom,” *Games and Economic Behavior*, 117, 380–387.
- PULFORD, B. D., A. M. COLMAN, C. L. LAWRENCE, AND E. M. KROCKOW (2017): “Reasons for cooperating in repeated interactions: Social value orientations, fuzzy traces, reciprocity, and activity bias,” *Decision*, 4, 102–122.

- 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., D. GREENE, AND P. HOUSE (1977): "The "false consensus effect": An egocentric bias in social perception and attribution processes," *Journal of Experimental Social Psychology*, 13, 279–301.
- RUBINSTEIN, A. AND Y. SALANT (2016): "'Isn't everyone like me?": On the presence of self-similarity in strategic interactions," *Judgement and Decision Making*, 11, 168–173.
- SHAFIR, E. AND A. TVERSKY (1992): "Thinking through uncertainty: Nonconsequential reasoning and choice," *Cognitive psychology*, 24, 449–474.
- STAHL, D. O. AND P. O. WILSON (1994): "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior and Organization*, 25, 309–327.
- STAHL, D. O. AND P. W. WILSON (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218–254.
- WANG, Y. (2023): "Belief and higher-order belief in the centipede games: An experimental investigation," *Pacific Economic Review*, 28, 27–73.
- WEIBULL, J. W. (2004): "Testing Game Theory," in *Advances in Understanding Strategic Behavior; Game Theory, Experiments and Bounded Rationality. Essays in Honour of Werner Guth*, ed. by S. Huck, Palgrave Macmillan, 85–104.