Behavioral Mechanism Design as a Benchmark for Experimental Studies

David K. Levine[1]

Abstract

I introduce the idea of behavioral mechanism design where in addition to the usual selfish players there are noisy players who play randomly and ethical players who actively seek to maximize social welfare and are willing, up to a point, to “do their bit” to achieve that goal. I calibrate this model using data on risk aversion and giving in dictator games. I then use it to study twelve different (out of sample) experiments involving stag hunt games, ultimatum bargaining games, and public goods games with and without punishment. I show that this simple calibrated model makes sharp predictions and does a good job both qualitatively and quantitatively in explaining the data from those experiments. The theory also identifies quantitative anomalies in the data pointing the way to future improvements. I conclude that this simple calibrated model might be a good benchmark for other experiments.

---

1. Introduction

You and three friends are on your way to the experimental laboratory to meet eight other students to be randomly matched to play an ultimatum bargaining game. You and your friends are public spirited in the sense you would like to maximize the *ex ante* expected utility of the participants - provided it is not too costly for yourselves. You and your friends also know that while the other students are, like you, risk averse, unlike you they are not so public spirited. About half are selfish and will try to get what they can for themselves; the other half will have other agendas, such as worrying

---

[1] Department of Economics, RHUL

Acknowledgements: First version: February 14, 2024. I would like to thank Rohan Dutta, Drew Fudenberg, John Mair, Andrea Mattozzi, Salvatore Modica, Tom Palfrey, Klaus Ritzberger, Catharina Tilmans, and Myrna Wooders. The Leverhulme Trust provided financial support for which I am grateful. A particular debt is owed to John Duffy, Nick Feltovich, Nikos Nikiforakis and Hans-Theo Normann not only for their intellectual contributions to this work, but for providing me with their original experimental data.

about what they will do over the weekend, or trolling the experimenter. Knowing that you will get to play a number of times what should you and your friends agree to do? This is a prototypical example of a behavioral mechanism design problem: behavioral because in addition to selfish types there are two behavioral types: ethical players like the friends in the example, and noise players with other agendas.

In this paper I analyze the behavioral mechanism design problem and provide solutions for a number of games that have been played in real laboratories. In the example, you and your friends should offer an even split as first mover, should accept offers of four dollars or more, and for each dollar less increase the rejection rate by about 30%. The striking fact is that in this and the other games I study the observed play in the laboratory resembles the idealized solution of the behavioral mechanism design problem both qualitatively and quantitatively. To be clear: it is unlikely that if there are ethical players in these experiments they are able to collude or that they know in advance what game they will play. Never-the-less play by experienced participants in the laboratory experiments I study may reasonably be described "as if" it is the solution to a behavioral mechanism design problem.

The setting for the formal model is a finite normal or extensive form game. In that game players are drawn from a population with three types. Selfish types are "standard" players who care only about their own utility. Noisy types are like behavioral or commitment types in the reputation literature or noise traders in the finance literature and play according to a fixed exogenous strategy. Ethical types are like ethical or group rule-utilitarian voters. One the one hand they are willing (to an extent) to sacrifice their individual utility for the common good. On the other hand they act as mechanism designers, picking an equilibrium that maximizes social welfare and optimally deploying their largesse. Below in the literature review I indicate that none if these types are new, and that they are adopted from the existing literature.

The main application of the model is to calibrate it and propose it as a benchmark for analyzing experimental data for standard stakes experiments involving college student participants. A benchmark model in my view is a model that is not estimated from data, but converts experimental instructions into quantitative predictions about play. The point of a benchmark model is to detect anomalies: if the experiment is what is predicted by the benchmark then there is little reason to search for new theories or modify old ones. The standard Nash (or subgame perfect) model with selfish risk neutral agents is an example of a benchmark model, and is widely used as such. It is a low bar because vast numbers of anomalies are known and it is easy to find new ones. The Levine (1986) calibrated model of signalling spite and altruism is a benchmark model albeit it has not proved a very useful one. The Fehr and Schmidt (1999) calibrated model is also a benchmark model and has proven more useful.

To use the behavioral mechanism design model as a benchmark model it must be calibrated. In the calibrated version of the model I make the uniformity assumption that all types are equally likely, the social welfare function puts equal weight on all types, and the noise players maximize a measure of entropy at each information set. In addition all players have the same risk averse utility function for money income. This and the largesse of the ethical types are calibrated to data on individual decisions for games that are non-strategic in the sense that strategies are ordered by strict dominance. I particularly want to empasize the role of risk aversion because efficiency creates a demand for insurance and this in turn means that "fair" allocations are preferred to "unfair" ones.

Having provided a calibrated model I use it to benchmark twelve different experimental treatments. All are classical experiments that have been replicated many times. The first application is to stag hunt for which there are four treatments. In this application social preference in the form of largesse plays no role, but the role of noise players and equilibrium selection is highlighted. In stag hunt theories lacking noise players do poorly and despite the fact they do not make precise predictions are wrong in the few predictions that they do make. The second application is to ultimatum bargaining for which there are two treatments. These experiments highlight the role of risk aversion in generating a demand for fairness. They also provide evidence that players are not merely reacting to unfair or unkind behavior by their opponents but are acting as mechanism designers and actively seek to achieve social goals. The final application is to public goods games with and without punishments of varying costs. This application demonstrates how the constraint on largesse interacts with the possibility of punishment to generate "the law of demand." The ultimatum and public goods contribution games are chosen not only to illustrate specific points about the theory but because they have been widely used to assess models of social preferences. In each case the behavioral mechanism design benchmark is qualitatively and quantitively on the mark, albeit with some quantitative anomalies that I explore.

What, then, is the marginal contribution of this paper to the existing literature? First, with respect to theory, this paper advances a different point of view than most existing models. With rare exceptions, behavioral models take account of psychological factors such as desire for fairness, reciprocity and altruism and build theories of what should be considered kindness and fairness. The theory here approaches these issues from a different angle. Take the willingness to punish those who fail to contribute to the common good. Standard behavioral theories build this into preferences as a kind of desire for revenge against those who fail to do their fair share, who are unkind, or in order to improve equity. In the mechanism design model here punishment is a means to an end - ethical players are willing to punish others to provide them with incentives to contribute to the common good. Fairness is not in conflict with efficiency, but in the presence of risk aversion fairness is demanded by efficiency. As I indicate below in the literature review this is not a new idea, but the model here through its simplicity and starkness provides the basis for a benchmark calibration which earlier models do not.

The second contribution of the paper is to the experimental literature. It provides a simple qualitative and quantitative (and new) explanation of a wide variety of experimental results and can be used as benchmark for detecting anomalies. It enables us to ask and answer questions such as: is risk aversion is sufficient to explain the demand for fairness or is there trade-off between efficiency and fairness?

The model has two ingredients: noise players and the idea that punishments are issued in order to provide incentives. I provide evidence for both of these ideas. In the stag hunt game models that lack noise players predict only that all players should choose the same action. In fact after nine periods of play more than 27% fail to play the modal action. In ultimatum bargaining models of fairness and kindness predict that the frequency with which an offer should be rejected should not depend upon how frequently that offer is made. In fact, in the same population, when the frequency of \$3.00 offers increases from 3% to 31% the frequency of rejections drops from 85% to 14%. Mechanism design, by contrast, says that punishments should not be issued if they do not accomplish the purpose of discouraging ungenerous offers.

2. Literature Review

As I indicated, the viewpoint of this paper has precedent and I would be remiss not to acknowledge the extent to which it builds on my earlier work with Rohan and Salvatore in Dutta, Levine and Modica (2021). That paper had ethical players (there called acolytes) and selfish players, but no noise players. Instead it had a noisy signalling technology that (as acknowledged in the paper) makes sense outside the laboratory but not inside the laboratory. Although we did try to calibrate that model, the calibration was clumsy due to the mismatch between the model and the laboratory and there was very little out of sample testing of the calibration. Here I have dropped the signalling technology as it is irrelevant to the laboratory and replaced it with noise players who are. This leads to a cleaner model and one that can be calibrated using only data from non-strategic settings and used as a benchmark (out of sample) in strategic settings.

The work here is also in the spirit of recent work, for example Fudenberg and Karreskog Rehbinder (2024), exploring how experimental data can be explained by models that are both simple and sensible. The idea of using a numerical target (here welfare) to measure consistency of the theory with data is reminiscent of the idea of measuring losses in Fudenberg and Levine (1997).

Ingredients of the Model: Ethical Players

As I indicated, the features of the model are not new and the types of players have ample precedent in the literature. In the empirical literature Coase (1960), Ostrom (1990) and Townsend (1994) argue that small groups are good at self-organizing to find solutions to mechanism design problems. The formal model of an ethical player is taken directly from the literature on ethical voters, including the theoretical model of Feddersen and Sandroni (2006) and the voting study of Coate and Conlin (2004). Other theoretical and applied uses of these models can be found in Herrera, Morelli and Nunnari (2016) and Levine and Mattozzi (2020) among others.

The idea that ethical players are willing to "do their bit" but only up to a limit is also closely related to the experimental literature on "warm glow" giving. Examples are Andreoni (1990) and Palfrey and Prisbrey (1997). It is similar also to the idea of "revoking costs" used in the bargaining literature such as Dutta (2012).

Ingredients of the Model: Noise Players

As indicated, noise players are not new either. They have been extensively used in the reputational literature, including but not limited to, Kreps and Wilson (1982), Milgrom and Roberts (1982), Fudenberg and Levine (1989), and Mailath and Samuelson (2001). Noise traders are widely used in the finance literature: a quick overview can be found in the Palgrave article by Down and Gorton (2008). Noise players are also related to the quantal response players of McKelvey and Palfrey

(1995) and especially the extensive form version in McKelvey and Palfrey (1998). Their behavior is similar to the limiting case as sensitivity to incentives grows small.

Finally, the role of noise players has parallel in the notion of risk dominant equilibrium, especially in the evolutionary literature: see for example Kandori, Mailath and Rob (1993), Young (1993), and more recently Peski (2010) among many others. Although the time frame for evolution and mutations is quite different than considered here, the role of noise is the same.

Psychological Models

As I have indicated the main alternative to the theory here are the many psychological theories of fairness and or reciprocity. In general these are qualitative analyses of experimental data and are not suitable as benchmark models. To mention a few of the more popular theories: Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) develop models of fairness with which they do qualitative analyses for a variety of experiments, and Fehr and Schmidt (1999) do quantitative analyses as well. Charness and Rabin (2002) introduce a psychological theory of fairness with many factors and do a set of experiments determine which are the most important. Falk and Fischbacher (2006) use higher order beliefs to model intentions and reciprocity. This is primarily a qualitative analysis. Dufwenberg and Kirchsteiger (2004) similarly model intentions, kindness and reciprocity. Along somewhat different lines Levine (1986) models intentions that are inferred from type signalling and uses it to analyze several experiments quantitatively.

Quantitative Calibration: Benchmark Models

As indicated, in addition to the selfish risk-neutral Nash model, there are two models that are potential benchmark models.

Levine (1986) calibrates a type signalling model on ultimatum and a public goods contribution game. There are two parameters describing the three types: altruistic, spiteful, and selfish (constituting 52% of the population). There are two out of sample analyses, that of the centipede game and that of a market game. Both are relatively successful.

Fehr and Schmidt (1999) calibrate a preference for fairness from ultimatum game data. There are two parameters one of which can take on four values and one three. The calibration is not complete as they do not specify the correlation between the two parameters. They conduct three out of sample analyses, that of a market game and that of a public good contribution games with and without punishment. All three are relatively successful. They also give a qualitative analysis of the trust game.

Fehr and Schmidt (1999) also examine dictator, which fails. Levine (1986) did not discuss dictator, but it is easy to apply the model and it fails even worse than Fehr and Schmidt (1999). By contrast, behavioral mechanism design does fine, although this is not an out of sample test as the largesse of ethical players is calibrated from dictator data.

I should add that I have not analyzed a market game as done in those two papers because that fruit is very low-lying: virtually any model predicts that with enough competition we will see competitive equilibrium. In a similar vein I have avoided analyzing games such as best shot where there is no real scope for social preferences, where subgame perfection performs well, and where other theories do equally well. The only exception is the public goods game without punishment: I include that to provide a contrast to the games with punishment.

In both the Levine (1986) and Fehr and Schmidt (1999) models there can be multiple equilibria so that the predictions of those theories are not sharp. I want to emphasize that in contrast to these other models - including selfish risk neutral Nash - the theory here makes sharp predictions. There is a single number - the optimal social welfare - that is spit out be the model from experimental instructions and can easily be compared to the theoretical data. The play leading to that optimum need not be unique but often is, including in settings where the other models make few useful predictions.

3. The Model

The setting is that of a game. Although this may be an extensive form game to limit notation I formally describe only the normal form. There are n player roles and each player role has a finite strategy space $s^i \in S^i$ with payoffs $u^i(s^i, s^{-i})$. Mixed strategies are denoted by σ^i and $u^i(\sigma)$ is the expected utility. Each player role is drawn privately from a single population in which there are three types: (S)elfish, (N)oise and (E)thical, where $\tau \in \{S, N, E\}$ denotes the type and $\phi_\tau > 0$ is the fraction of the population that is type τ with the obvious property that $\phi_S + \phi_N + \phi_E = 1$. Player roles are partitioned into classes of roles that are indistinguishable and mixed strategies for a type are feasible if and only if they are symmetric within each class. For example, in a fourteen player public goods contribution game with identical players there are 14 player roles, but players cannot distinguish what their "player number" is.

As expected the selfish types are standard players who try to maximize their utility $u^i(s^i, s^{-i})$ for the player role i they are assigned. The noise type plays according to a fixed probability distribution σ_N with $\sigma_N^i(s^i) > 0$. The ethical players are public spirited and act as mechanism designers, choosing incentive compatible strategies for themselves and the selfish types to maximize a social welfare function as I now explain.

To be specific, for given mixed strategies for each type σ_τ denote the mixture by $\sigma = \sum_\tau \phi_\tau \sigma_\tau$. The mechanism design problem can be stated as a choice of σ_S, σ_E to maximize the expected per capita social welfare function

$$E \sum_\tau w_\tau \frac{\sum_{i=1}^n u^i(\sigma_\tau^i, \sigma^{-i})}{n}$$

where the welfare weights $w_\tau \geq 0$ and $\sum_\tau w_\tau = 1$. For selfish types there are incentive constraints for $i = 1, \ldots, n$ and $s^i \in S^i$

$$u^i(\sigma_S^i, \sigma^{-i}) \geq u^i(s^i, \sigma^{-i}).$$

In addition the willingness of the ethical players to contribute to the public cause is not unlimited and the ethical players are characterized by a utility limit γ, the *largesse*, on how much they are willing to sacrifice. This gives additional incentive constraints

$$u^i(\sigma_E^i, \sigma^{-i}) + \gamma \geq u^i(\sigma_S^i, \sigma^{-i}).$$

Discussion of the Model

Two aspects of the model deserve mention. First, I have not assumed that the welfare weights are all positive. It might be, for example, that the ethical players do not care about the noise players, viewing them as being deviant. Or they might care only about the welfare of the ethical types.

Second: the behavior of the ethical types (and possibly of the selfish types as well) is not individualistic. Coate and Conlin (2004) refer to ethical voters as "group rule-utilitarian" and this is accurate. That is, ethical players ask: what would we like to happen (given the incentive constraints) and how can we do our share to make it happen. In particular: even if $\gamma = 0$, or, as is the case in stag-hunt if γ is irrelevant, if there are multiple equilibria the ethical players get to select the most favorable equilibrium. It is for this reason that one of my applications is to the stag hunt game. Ordinarily this is viewed as a failure of the hypothesis that most favorable equilibria are selected. As I will show this is not the case for the calibrated behavioral mechanism design model: the presence of the noise player changes the calculus of both equilibrium and welfare and is consistent with what is seen in stag hunt experiments.

Incentive Constraints

The incentive constraints are applied after player roles have been assigned and types determined, but, if the game is sequential, before moves take place. Hence the equilibrium concept is Nash rather than subgame perfect. In this setting with noise players who play everything with positive probabiity this distinction is meaningless: every information set feasible given a player's strategy is reached with positive probability, Bayes law always applies, and every Nash equilibrium is sequential.

The *ex ante* nature of the incentive constraints does have implications for the behavior of the ethical types. That is, their willingness to accept a utility loss of γ is measured *ex ante*. For example, if they move second they may be required in response to an unlikely move of the first player to take a greater loss than γ. The ethical types are committed to do "whatever it takes" when the time comes, provided the *ex ante* expected loss from doing so is not too great. I should note that in experimental treatments where one round is chosen at random to be paid the commitment is automatic: at the time the decision is made the action chosen is purely hypothetical and will involve an actual loss only with some probability - after the fact it is impossible to renege.

Existence of a Solution

The one relevant theoretical fact is that the behavioral mechanism design problem has a solution.

Theorem 3.1. *The problem of maximizing social welfare subject to the incentive constraints has a solution.*

Proof. This follows if the expected utility functions are continuous in the strategies and the constraint set is closed and non-empty. Continuity of expected utility in strategies follows from the fact that the game is finite so they are multi-linear. The constraint set is closed because the utility functions are continuous and the constraints are defined by weak inequalities. The only substantive issue is whether the constraint set is non-empty. Since the noise players act as "nature" there is a Nash equilibrium for the selfish and ethical players in which the ethical players act selfishly: this satisfies all the constraints. □

4. Overview

Before analyzing the experiments in detail I first give an overview of the calibrated model and results. The utility function $u(m)$ for monetary payoffs and γ are calibrated to data. This is done below, but I want to indicate that this calibration is for standard stakes with students as participants: those are the applications I am going to consider. I suspect that for other stakes and with other populations this calibration would not "work." In addition the theory is an equilibrium theory and we only observe something resembling equilibrium in the laboratory when participants have an adequate chance to play and learn. Consequently, in the applications I will only look at data from late stages of repeated matches.

The Calibrated Model

The besides the monetary payoff function, which is given by the experimental instructions, the mechanism design problem depends upon the utility $u(m)$ for monetary payoffs m, the largesse γ of ethical players, the weights w_τ in the social welfare function, the fractions of types ϕ_τ, and the strategy of the noise types σ_N.

Here is the calibrated model. Utility is given by

$$u(m) = 1 - (1 + m/C)^{1-\rho}$$

where $C = 40$ and $\rho = 9$. If there are τ paid rounds then largesse in each round is $u^{-1}(\gamma) = \$1.00/\tau$. For the utility weights and fractions the simplest assumption and the one I will adopt is the *uniformity hypothesis*: this is $w_\tau = \phi_\tau = 1/3$. For the behavior of the noise players σ_N I will adopt the *maximum entropy hypothesis.* I first partition and order the actions at each information set by weak dominance. Within each weak dominance class actions are chosen with equal probability; and each weak dominance class has the same probability as the combination of all lower weak dominance classes. For example, the probability some weakly undominated strategy is chosen is equal to the probability that some weakly dominated strategy is chosen.

Results

The strong prediction made by solving the mechanism design problem concerns welfare. This is reported below in Table 4.1 for the three experiments and twelve treatments analyzed in this paper.[2]

		period(s)	theory	data	actual err	SGP err	FS err
stag	$n = 2p$	7	\$1.18	\$1.18	\$0.00	\$0.22*	
	$n = 2s$	5	\$1.18	\$0.91	\$**0.27**	\$0.39*	
	$n = 14$	10	\$0.64	\$0.60	\$0.04	\$0.70*	
	$n = 15$	10	\$0.60	\$0.66	−\$0.06	\$0.64*	
	$n = 16$	10	\$0.60	\$0.61	−\$0.01	\$0.69*	
ult	no obs	21 − 40	\$3.45	\$3.46	\$ − 0.01	\$0.06	\$1.36
	obs	21 − 40	\$3.45	\$3.43	\$0.02	\$0.09	\$1.39
pub	no pun	10	\$1.51	\$1.51	\$0.00	−\$0.01	
	pun 1	10	\$1.81	\$1.64	\$**0.17**	−\$0.14	
	pun 2	10	\$1.88	\$1.78	\$0.10	−\$0.28	\$0.63*
	pun 3	10	\$1.91	\$1.99	−\$0.08	−\$0.49	\$0.42*
	pun 4	10	\$1.92	\$1.91	\$0.01	−\$0.41	\$0.50*

Table 4.1: Welfare

n in stag hunt is number of players, for $n = 2$ the s denotes strangers and p partners
in ultimatum bargaining no obs is the standard treatment and obs is the treatment where the play of another player is observed
in the public goods game pun represents the punishment factor (or no punishment)

The experiments are stag (hunt), ult(imatum bargaining) and pub(lic good contributions). Welfare is reported in certainty equivalent units by applying u^{-1} to the expected utility of a player in the game generated by the theory. I then computed the actual utility from the data in the same units and the difference between the theory and the data (actual err). This in itself proves little: it is possible to develop theories that generate predictions that do not depend upon the data at all: for example, the maximum possible payoff in the game. It is important to know that there is a wide range of possible predictions for welfare, that is, that the theory can be wrong. To this end, as I explain in Online Appendix 2, I computed welfare for two other benchmark theories, selfish risk neutral subgame perfect equilibrium (subgame perfection or SGP) and the calibrated Fehr and Schmidt (1999) (FS) model. In the final two column I then computed the error for each of these other theories. In cases where there were multiple equilibrium (marked with a *) I followed Fehr and Schmidt (1999) and picked the most efficient one.

---

[2] To avoid informational overload I do not report standard errors here. They are discussed in the context of specific experiments in the text and in Online Appendix 4. They add little to the information presented in the table.

For one game, the public good game with no punishment, all the theories agree that there will be very little contribution. In nine of the other eleven experiments the actual error is no more than \$0.10 in absolute terms. By contrast, the other theories come within \$0.10 of empirical welfare only for subgame perfection in ultimatum bargaining and generally have much larger errors. Overall, I take this to mean that behavioral mechanism design does fairly well in predicting welfare.

In the table I have highlighted the two anomalies identified by the calibrated model. These are the stag hunt game with strangers (players are randomly matched each period) and the punishment factor one public goods game. These I will examine below, but for the moment note that the first and worst anomaly, the stag hunt anomaly, occurs with relatively inexperienced players who got to play only five periods. For the punishment factor one public goods game both subgame perfection and Fehr-Schmidt do better than mechanism design underpredicting welfare by \$0.14 rather than overpredicting it by \$0.17, but none of thsee theories do terribly well.

The one case in which an alternative theory does relatively well - subgame perfection in ultimatum bargaining - is, unfortunately, a case of the broken clock being right twice a day: subgame perfection makes two offsetting errors. On the one hand it underpredicts the generousity of offers, predicting \$1.00 offers as against at least \$3.63 in the data. This lowers welfare. On the other hand it also underpredicts rejections, predicting that no offers will be rejected, while the actual rejection rate in the data is about 20%. This raises welfare and the two errors more or less cancel out. In contrast the Fehr-Schmidt model does poorly with ultimatum welfare, overpredicting by more than \$1.25. Unlike subgame perfection Fehr-Schmidt gets the distribution of offers fairly accurate for one of the two ultimatum games, but overpredicts welfare because it gets the rejection rate too low. This shows that the details are important, and I will go through the details of the mechanism design model shortly.

Before turning to the details of the theory and data, I must explain how the calibration is done.

5. Benchmark Calibration for Long-Term Play

The utility $u(m)$ for monetary payoffs m, the largesse γ of ethical players, the welfare weights w_τ, the fractions of types ϕ_τ and the strategy of the noise types σ_N all must be calibrated. As indicated for the utility weights and fractions are not calibrated to data, rather I adopt the uniformity hypothesis: this is $w_\tau = \phi_\tau = 1/3$. Similarly as I describe below the strategy of the noise players is derived from the maximum entropy hypothesis. Then I calibrate σ_N and $u(m)$, as these are needed for calibrating γ, and conclude by calibrating γ.

I want to emphasize that in this calibration I have taken data from standard experiments using best practices that have been replicated many times. In addition I use only data from non-strategic settings. By this I mean games where strategies are ordered by strict dominance with respect to monetary payoffs: dictator and public goods contribution games.

Entropy Maximation in the Agent Normal Form

Strictly speaking I do not calibrate σ_N at all, rather I assume that it is noisy in the sense of maximizing a measure of entropy. As it is the behavior of noise players that matters, it makes sense to talk of behavior strategies and the most straightforward assumption is that the noise players randomize uniformly over actions at each information set. This leads to absurd play in some settings, so I instead adopt the maximum entropy hypothesis which I now describe.

To motivate the maximum entropy hypothesis, consider the public goods game with punishment studied by Fehr and Gachter (2000). Here in the second stage of a game a player must decide how to allocate 20 “punishment points” among three opponents. These are costly both to the punisher and the punished.

What does this structure means in terms of the information set where punishment is allocated? There is one action in which no punishment points are allocated. There are three actions in which one punishment point is allocated among the three opponents, and in general there are $(k+1)(k+2)/2$ actions which allocate k punishment points among three opponents. The point is that a uniform distribution over actions at this information set implies that large numbers of punishment points are far more likely than small numbers because there are many more ways to allocate them. In particular the probability that six or fewer punishment points are assigned is less than 5% while the probability that 16 or more punishments points are assigned is more than 50%. This is not reasonable and is grossly inconsistent with the play of laboratory participants.

To provide a more "reasonable" description of the play of noise players, I instead categorize actions and assume that entropy is first maximized between categories, then within categories. Specifically, working in the agent normal form so as to deal with behavior strategies and actions at information sets, for each information set I at which player i is playing, strategies can be divided into those that are weakly dominated $W^0(I)$ and those that are not $N^0(I)$. By *zero order reasonableness* I mean that it should not be more likely to play a weakly dominated strategy than a weakly undominated strategy:

$$\Pr(N^0(I)) \geq \Pr(W^0(I)).$$

This criterion should be applied recursively: we can define $W^1(I)$ as the subset of $N^0(I)$ that are weakly dominated by a strategy in $N^0(I)$ and $N^1(I)$ as those which are not, and continuing in this way define $W^k(I), N^i(I)$ until we run out of strategies. The *reasonableness constraints* are

$$\Pr(N^k(I)) \geq \Pr(W^k(I)).$$

The maximum entropy hypothesis then asserts that entropy should be maximized among categories subject to the reasonableness constraints: this says that the constraints should bind and that each category except the last has half the probability of the preceding category with the final two categories having equal weight. It then asserts that entropy within each category should be maximized so that actions within each category are chosen with equal probability.

In the example the maximum entropy hypothesis gives rise to the punishment strategy for the noise players: the probability of issuing k punishment points is $(1/2)^{k+1}$ for $k \leq 19$ and $(1/2)^{20}$ for $k = 20$. For each level of punishment k there are 3^k ways of allocating those punishments among three opponents, and each of these has equal probability.

Risk Aversion

It has long been observed that players are risk averse over the small stakes in laboratory experiments. Risk aversion plays a key role in the theory both because there are risks and because it induces a demand for fairness. That is, if agents are risk averse, maximizing *ex ante* expected utility of a player means that an equal split provides both players with insurance. The social optimality of the equal split plays a key role in the analysis of ultimatum bargaining.

To get a particular utility function I followed Fudenberg and Levine (2011) who derive a "short-run" laboratory utility function in a way that is consistent with risk aversion outside the laboratory. Specifically, this is the CES or constant relative risk aversion function

$$u(m) = 1 - (1 + m/C)^{1-\rho}$$

where $C = \$40.00$ is an estimate of daily "pocket cash" and ρ is a coefficient of relative risk aversion determined from laboratory choices over gambles. The bottom line here is that I take $\rho = 9.0$.

To calibrate ρ I used data from two different experimental approaches: the risky investment approach of Gneezy and Potters (1997) and the multiple price list approach popularized by Holt and Laury (2002). These methods are discussed in the review paper Charness, Gneezy and Imas (2013) and are the two methods used as objective measures of risk aversion in the large scale standardized survey of Snowberg and Yariv (2021). In both cases I used data from the original papers.

Gneezy and Potters (1997) give 84 participants an endowment of \$1.20 and ask them to decide how much to invest in a risk project that pays nothing with probability 2/3 and pays 3.5 times the investment with probability 1/3. They played nine times: the average investment was $x = \$0.30$ and did not vary much from round to round. Differentiating the objective function

$$(2/3)u(1.20 - x) + (1/3)u(1.20 - x + (3.5)x)$$

with respect to x, equating to zero, substituting $x = \$0.30$, and solving for ρ yields the estimate $\rho = 8.7$.

Second, following Fudenberg and Levine (2011), I use data from Holt and Laury (2002)'s normal stakes experiments. They provide 187 participants with a menu of paired lottery choices where the first is a lottery between \$2.00 and \$1.60 the second between \$3.85 and \$0.10. The menu gives different probabilities between the first and second prize. They find when the odds are $50 - 50$ that 70% of participants take the safe choice, while when the odds are $60 - 40$ only 45% of participants take the safe choice. For the first lottery indifference requires $\rho = 4.2$ and for the latter $\rho = 12.5$.

The median individual lies between these two, presumably closer to the top. This is generally consistent with the $\rho = 8.7$ from the Gneezy and Potters (1997), so, to avoid spurious precision, I take $\rho = 9$. Such an individual is indifferent on the Holt and Laury (2002) list at $56 - 44$.

Those familiar with the literature on risk aversion in the laboratory may be puzzled by the fact that these values of ρ are much higher than appear in other studies. This is because I have assumed a "wealth" of $C = \$40.00$ while other studies assume much smaller "wealth." With larger wealth risk aversion must be larger to fit the data. Over the relevant range it makes little difference what utility function is fit to the data. In Appendix 1 I have plotted along with the calibrated utility function a CARA utility function fit to the Gneezy and Potters (1997) data: it looks the same over the relevant range of zero to ten dollars.

One additional remark is important for interpreting the numerical values of welfare and utility: they are reported in certainty equivalent units, that is, by applying u^{-1}.

Largesse

How willing are ethical players to sacrifice for the common good, or to say the same thing, what is γ? To answer this question I use data only from non-strategic settings where actions are completely ordered by strict dominance with respect to monetary payoffs: these are the dictator game, the one shot Prisoner's Dilemma game, and public goods contribution games without punishment. The bottom line is that if there are τ paid rounds then I take $u^{-1}(\gamma) = \$1.00/\tau$.

I am interested in games where experienced players have played many times. A robust finding from many studies is that willingness to give declines substantially over time. Figure 5.1 below plots contributions over time from Fehr and Gachter (2000)'s repeated public goods contribution game with about 66 strangers, and another with about 44 partners. In the final period the two are quite similar and the average of the two $\mu = 0.268$ I take to be the long-term ratio. For comparative purposes I also show the fraction of the population cooperating in Dal Bo (2005)'s one-shot prisoner's dilemma game with 390 strangers. This is quantitively quite similar to the Fehr and Gachter (2000) stranger treatment and stabilizes in about the 7th round with the average over the last four rounds equal to 0.242.

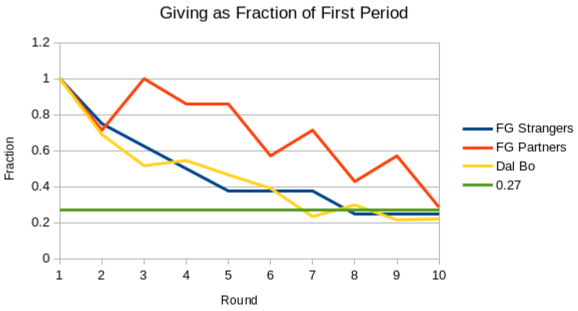

Figure 5.1: Willingness to Give Declines with Experience

I view the long-term ratio μ as discount factor that multiplies giving in a first time game to determine giving for a pool of experienced players. To determine first time giving I use data from the dictator game.

In the dictator game one player allocates a fixed amount between themselves and one other player. The basic source of information about dictator is the Engel (2011) meta-study based on 83 papers with a total of $20,813$ observations. The key relevant finding is that with student participants the donation rate is about 25%. The most common dictator game in the laboratory is for \$10 stakes where giving is in whole dollars.

In the standard \$10 dictator game with students in the laboratory Engel (2011)'s data indicates we can expect an average contribution of \$2.50. Discounting this by $\mu = 0.268$ and taking account of the fact that the experienced noise players each contribute on average \$1.00 yields the formula for the willingness of the ethical players to contribute

$$u^{-1}(\gamma) = 3(2.50)\mu - 1 = 1.00. \tag{5.1}$$

The given value of γ makes sense when one round is chosen randomly to be paid. When all rounds are paid it makes sense that the given value of γ applies to the entire game. That is, if the game is played ten times it makes no sense that each time it is played the ethical players are willing to sacrifice \$1.00, but rather that they are willing to sacrifice that much over the entire course of play, that is, \$0.10 for each round. More generally, if there are τ paid rounds I take $u^{-1}(\gamma) = \$1.00/\tau$.

I note that there is an issue with the γ constraint failing to bind which would invalidate these computations - in Online Appendix 1 I show that the calculations here are robust to this concern.

6. Stag Hunt

The first experiments I analyze are the stag hunt games of Van Huyck, Battalio and Beil (1990) designed to illustrate how coordination on efficient equilibria can fail. This class of games is interesting because the standard benchmark theories Fehr-Schmidt and subgame perfection have little to say about these games, and what little they do say is wrong.

The games studied in Van Huyck, Battalio and Beil (1990) are simultaneous move n player games in which each player chooses effort in dollars from $q^i = \{0.10, 0.20, \ldots, 0.70\}$. The monetary payoff of i is given by

$$m^i(q^i, q^{-i}) = .60 + 2.0 \min\{q^j\} - q^i.$$

Players are paid for every period. There are two treatments: one with a large fixed population that plays for ten periods with $n \in \{14, 15, 16\}$. Three sessions were conducted with $n = 16$ and two each with $n = 14, 15$. The other treatment is for a small population with $n = 2$: this is done both with a fixed population (partners) and randomly matched players (strangers).

Qualitative Analysis. In the stag hunt game no individual player, nor even a third of them, have a substantial chance of raising the minimum, so social preferences including largesse play no role. Rather it is the play of the noise players together with equilibrium selection that is crucial. With a large population (14 or more players) the chances one player messes it up for everyone by choosing a low effort is high and it is impossible to sustain high levels of effort. With a small population the chance of the one other player messing it up is not so great and high effort levels are sustainable. Hence the theory predicts low effort levels in the large population and high effort levels in the small population. This is characteristic of stag hunt experiments. Risk dominance makes similar predictions but involves hypothetical players as opposed to noise players. Characteristic of noise players is that, unlike in other theories, there should be dissidents who fail to play the modal action, and indeed that nearly a third of the population should be dissidents. This is, in fact, true. Note that this analysis provides a strong rationale for explicitly including noise players in substantial numbers: if noise players and equilibrium selection were added to existing models such as Nash equilibrium or models of social preferences the results would be identical to those found here.

Description of the Solution. In all cases the mechanism design problem has a unique solution. All the selfish and noise players choose the same target level of effort. When $n = 2$ noise players are rare and every target effort level is an equilibrium. Welfare is increasing in the target and so the optimum is maximal effort \$0.70. When $n = 14, 15, 16$ the chances of at least one noise player are high and there are only equilibria with low effort levels. Specifically, when $n = 14$ the effort levels \$0.10, \$0.20 are equilibria and the optimum is \$0.20. When $n = 15, 16$ the only equilibrium is \$0.10 so this is the optimum.

Below in Table 6.1 I summarize the theoretical solution and the data from the final period of play. Note that the maximum attainable joint money payoff is \$1.30 per player.

n	strangers	welfare		mean effort		period	participants
		theory	data	theory	data		
2	yes	1.18	1.18	0.60	0.64	7	28
2	no	1.18	0.91	0.60	0.53	5	16
14	no	0.64	0.60	0.27	0.19	10	28
15	no	0.60	0.66	0.20	0.14	10	30
16	no	0.60	0.61	0.20	0.18	10	48

Table 6.1: Summary of Stag Hunt

Qualitatively the theory does extremely well capturing the fact that welfare and effort are higher with fewer players. Quantitatively the theory does reasonably well: however when $n = 2$ with strangers the theoretical welfare is substantially greater than welfare in the data.

I turn now to a more detailed analysis of the mechanism design problem in these stag hunt games.

The Large Population Game

I should start by noting that the large population games are played with a fixed set of players: a partners rather than strangers treatment. However, data is from the final period so that there are no repeated game effects.

To solve the mechanism design problem observe that there are no weakly dominated strategies, so that the noise players randomize uniformly over contributions. Using this, I compute the utility of a selfish player for a particular contribution under the assumption that that no selfish or ethical player is choosing a smaller contribution: I refer to this as the *popular minimum*, as it is the minimum for 2/3rds of the population, although it need not be the minimum at all when noise players are accounted for. As the combinatorics of the noise players is complicated, I computed utility by matching players in a Monte Carlo simulation with $1,000,000$ draws. The results are below in Table 6.2 for the case $n = 16$.

popular minimum	$n = 16$	welfare
0.70	0.27	0.36
0.60	0.37	0.43
0.50	0.46	0.50
0.40	0.56	0.55
0.30	0.64	0.60
0.20	0.695	0.63
0.10	0.700	0.60

Table 6.2: Large Population Game: Selfish Payoff

It follows from these utilities that for selfish players each wants to reduce the popular minimum: the only equilibrium behavior is for all to contribute the minimum \$0.10. In Online Appendix 3 I show that this is also the optimum for the ethical players.

From Table 6.2 it should be clear that the equilibrium at \$0.10 rather than at \$0.20 is delicate: this is why I used the full risk averse utility function even though risk aversion is minor over these stakes. With $n = 14$ there is an equilibrium at \$0.20 as well as \$0.10 and this would be chosen by the ethical players.

Also from Table 6.2 observe that the gain in social welfare of moving from \$0.10 to \$0.20 is small: it is about \$0.03. This highlights a limitation of the mechanism design: while the prediction of welfare is strong, even if equilibrium is unique it may be delicate in the sense that a small perturbation of the parameters may cause it to jump. Moreover, it seems unrealistic that an equilibrium could jump with respect to such a tiny change: in the $n = 14$ game the loss to a selfish player to erroneously choosing effort \$0.10 rather than \$0.20 is less than half a cent, and if even a modest fraction of them wrongly decide \$0.10 is best they all want to switch. This is known problem with mechanism design: it allows the designer to choose equilibria that are not terribly robust.

Dissidents

Consider the Fehr-Schmidt and subgame perfection theories. For these minimum games fairness is not at issue and a player with social preferences behaves no differently than a selfish player: there is no benefit from increasing effort above the minimum or decreasing effort below the minimum. In other words, in the usual way with coordination games, every common effort level is an equilibrium. The only prediction made by these theories is that there should be no dissidents in the sense that every player should play the modal effort level. In Table 6.3 below I provide information about the modal effort levels for the $n = 15, 16$ games and the fraction of dissidents. As can be seen the prediction of no dissidents fails badly as more than a quarter of the population are dissidents. By contrast the behavioral mechanism design benchmark makes precise and correct predictions about the modal effort levels and matches the number of dissendents in the large population games quite well.

n	modal effort		dissidents	
	theory	data	theory	data
$15, 16$	0.10	0.10	29%	27%

Table 6.3: Dissidents are those not choosing the modal effort

It is interesting also to take a look at what the dissidents do. Below in Figure 6.1 I plot the theoretical ($n = 15$ or $n = 16$) and empirical distribution (all large population sessions pooled) of contributions conditional on contributing more than \$0.10. As is assumed, the theory is flat. What is interesting is the data: there is a slight bias towards lower contributions and against intermediate contributions. What is striking though is the high fraction who are contributing the maximum: \$0.70. This is especially the case since in no round of any session was the minimum ever close to \$0.70. Perhaps these noise players are making a statement?

I should indicate that the data here is weak. There are 106 observations and 7 of them have effort \$0.70. In the theory each player has a 1/3 chance of being a noise player, and a noise player has a 1/7 chance of providing effort \$0.70. The binomial probability of getting 7 or more such draws in 106 trials is fairly large by the standards of statistical significance: 13.3%.

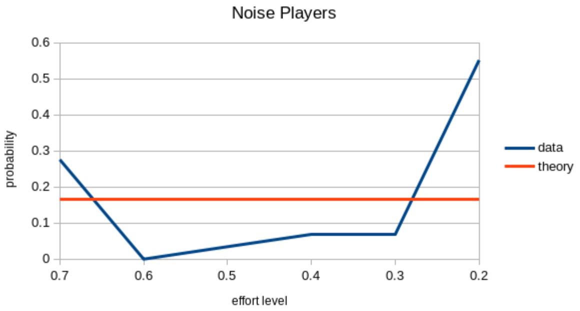

Figure 6.1: Dissidents

Small Population Game

To compute the unique equilibrium in the small population game, the popular minimum payoff from Table 6.2 is recomputed for $n = 2$ in Table 6.4 below using $10,000$ Monte Carlo draws. Here there is a Nash equilibrium at \$0.70 and as it maximizes welfare it is chosen by the ethical players giving the unique solution reported below in Table 6.4.

popular minimum	$n = 2$	welfare
0.70	1.09	1.18
0.60	1.05	1.11
0.50	1.00	1.03
0.40	0.94	0.94
0.30	0.87	0.84
0.20	0.79	0.72
0.10	0.70	0.60

Table 6.4: Small Population

In the small population game and the partners treatment the theory does well. I will therefore focus on the strangers treatment in which players play against randomly matched opponents. This is the worst anomaly in Table 4.1. Since the individual matches were not reported I used a Monte Carlo to randomly match the players $10,000$ times in order to compute welfare from the data.

As observed above when $n = 2$ in the strangers treatment the theory predicts too high a level of welfare and too low a rate of dissidence. However, in addition to relatively little data (16 observations) the game was played only 5 times so the participants cannot be considered experienced. It is important then to ask: how has the game progressed over time? Did they start by trying to cooperate at $s^i = 0.70$ and then this gradually unravelled? Or has cooperation increased over time so that as experience is gained play more closely resembles the prediction of the theory. Below in Table 6.5 I report the distribution of play between the first and fifth round: by every measure play is moving towards that predicted by the theory as player become increasingly successful at coordinating on \$0.70.

	1st round	5th round	theory
minimum $= 7$	10%	25%	51%
minimum ≥ 4	39%	66%	73%
minimum $= 1$	44%	12%	6.5%
dissidents	69%	50%	29%

Table 6.5: Evolution of Play Over Time

Overview

The most important anomalies are

- In the $n = 14$ game the equilibrium has mode 0.20 rather than 0.10 as in the data.
- With $n = 2$ in the strangers treatment the theory indicates far more coordination than in the data.
- The distribution of the dissidents is very different in the data than in the theory.

None of these anomalies are terribly important.

7. Ultimatum Bargaining

Many ultimatum bargaining experiments have been conducted with similar results. In ultimatum bargaining game the first mover proposes the division of a fixed amount of money, usually \$10.00, and the second mover either accepts and both are paid as agreed, or rejects and both get nothing. Here I analyze data from Duffy and Feltovich (1999) for the important reason that players got to play 40 times rather than the usual 10. This is important because play after round 10 is different than earlier, but remains largely constant during the final 30 periods indicating that this is “the long-run” with experienced players. The experimental design is also a clean one with the standard

$10 stakes, offers in whole dollars, no zero offer, and one randomly chosen round paid. The whole dollars greatly eases the analysis of the data: when offers are in $.05 increments we see things like a single offer of $4.60 rejected and one of $4.55 accepted. In other words, to make sense of it the data has to be aggregated into cells and this is always fraught.

Another useful feature of Duffy and Feltovich (1999) is that there are two treatments: one the standard treatment (nobs, 32 participants), and a second in which players get to observe the results of one other match each period (obs, 40 participants) - a treatment that they and I expect to enhance learning. In all cases I use data from the final 30 rounds.

Qualitative Analysis. Ultimatum bargaining highlights the importance of risk aversion in generating a demand for fairness. Without noise players the ethical players would simply insist on the efficient outcome which is an equal split and back this up by rejecting less generous offers. Without noise players this punishment is entirely hypothetical and has no cost. With noise players enforcing more generous offers increases the number of offers that must be rejected, so imposes a social cost offsetting the gain in fairness. Hence the theory predicts that offers should be generous but many should fall short of an equal split. This is characteristic of ultimatum game experiments. The theory also predicts a substantial rejection rate, also characteristic of ultimatum experiments.

Description of the Solution. The behavioral mechanism design problem has a unique solution for this game. There is a target for the selfish first mover. This is supported by the ethical players rejecting offers less generous than the selfish target at an increasing rate. The rejection rate should be as small as possible subject to incentive compatibility. These facts are proven in Appendix 2. I then compute that the welfare maximizing target is $4.00 for the selfish players while the ethical players themselves offer $5.00.

I report the key statistics of the solution below and contrast it with the data for both the obs and nobs treatments. For the rejection rate I also reported for the obs treatment the final 10 periods only. This is to verify that the noise as measured by rejected offers is not declining over time.

mean offer			rejection rate				welfare		
theory	obs	nobs	theory	obs	nobs	obs 10	theory	obs	nobs
4.83	4.46	3.63	0.18	0.20	0.19	0.18	3.45	3.46	3.43

Table 7.1: Ultimatum Bargaining

For welfare and the rejection rate the theory and data match well. The mean offers for the obs treatment is reasonably close to the theory but the mean offer for the nobs treatment is anomalous.

Offer Distribution

Figure 7.1 below provides detail with the theoretical and empirical offer densities.

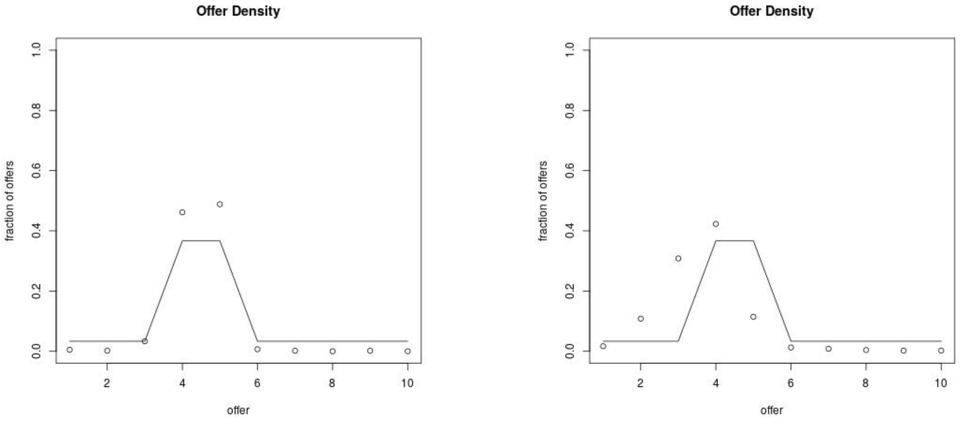

Figure 7.1: Observed Offers (left), Unobserved Offers (right)

The empirical distribution for the obs case looks similar to the theory. The nobs case looks much like the obs case but with offers shifted a dollar to the left except for a modest number that remain at \$5.00. In othe words, the nobs case looks as if the target for the selfish players is 3 rather than 4 and that the majority of the ethical players are offering \$4.00 rather than \$5.00.

Play in the obs case looks quite different than the theory. I want to emphasize, however, that there is very little welfare loss in doing this. Theoretically welfare should be \$3.51 and in the obs case it is \$3.49, not much worse. To understand this better I compute below in Table 7.2 below the welfare corresponding to each target offer for the selfish type. This was needed in any case to find the optimum. Note that the constraint only binds on the ethical players when target is \$1.00 in which

target	5	4	3	2	1
welfare	3.41	3.45	3.42	3.33	3.17

Table 7.2: Welfare

case they must offer $4.00 rather than $5.00 as they do for the other targets. As can be seen setting a target of $3.00 rather than $4.00 results in the same small drop in welfare as in the obs treatment. The same cannot be said for other targets: as the target is lowered below $3.00 welfare drops off fairly rapidly.

The $3.00 target mechanism also does not match the obs data since all the ethical players are offering $5.00. Again, however, the welfare consequences of their the ethical players switching to $4.00 is quite small: I computed this and it lowers welfare from $3.42 to $3.40.

I note that Duffy and Feltovich (1999) argue that the difference between the two treatments is because the learning process is changed by the additional information about other player's play. That makes sense the context of mechanism design as well: I do not imagine that the players make some sort of exact calculation of the solution to the mechanism design problem in their heads, although I imagine they have some general ideas, such as "we must reject bad offers so as to encourage good ones." In particular ethical players may be unsure what "their bit" is supposed to be: some may think $5.00 while others think $4.00 would be enough. Observing the offers of others might well convince those making $4.00 offers that they are not doing their bit, and so switch to $5.00 offers.

Good Offers

There is an anomaly in the offer distribution that is hard to see in the figures. That is that there are far too many good offers. The theory predicts 16.7% of all offers with be for $6.00 or more. In the data this is true for only 20 out of 1080. Moreover, the same as has been found in hundreds of ultimatum experiments: the only apparent exception is the experiment conducted with the whale hunting Lamalera reported in Henrich et al (2004). In that case, however, the noise player was the experimenter - the "low offers plotted for the Lamalera were sham offers created by the investigator."

A good robustness check is to ask what happens if *ad hoc* and by fiat I were to assume that players cannot make offers better than $5.00. This raises the cost to the ethical players of providing incentives to the selfish players because the noise players now make more bad offers. It changes the optimal mechanism: the target for the selfish players drops from \$4.00 to \$3.00 and, accounting for the poorer offers by the noise players, reduces the mean offer from \$4.83 to \$3.67. This mechanism now mirrors the data for the noobs case where the mean offer is \$3.63. The reduction in welfare due to increased rejection of bad offers is offset by the fact that uniform offers between \$1.00 and \$5.00 are more efficient than between \$6.00 and \$10.00 and both welfare and the rejection rate are unchanged. Overall, the theory with this ad hoc modification does about as well as the original.

Associated with the anomaly in the good offers is an associated anomaly with the rejection rates. By pooling all the good offers out of 20 only 1 is rejected, a rejection rate of 5%. According to the theory the rejection rate should be 17% - much higher.

Rejection Rates

I turn finally to the theoretical and empirical rejection rates. At the aggregate level the theory matches the data quite well. Below I report the conditional rejection rates. Qualitatively the theory matches the nobs data well: declining until the target of \$4.00 is reached, then flat. Not unexpectedly for the nobs case the decline ends when the target of \$3.00 is reached. I will focus on the nobs case.

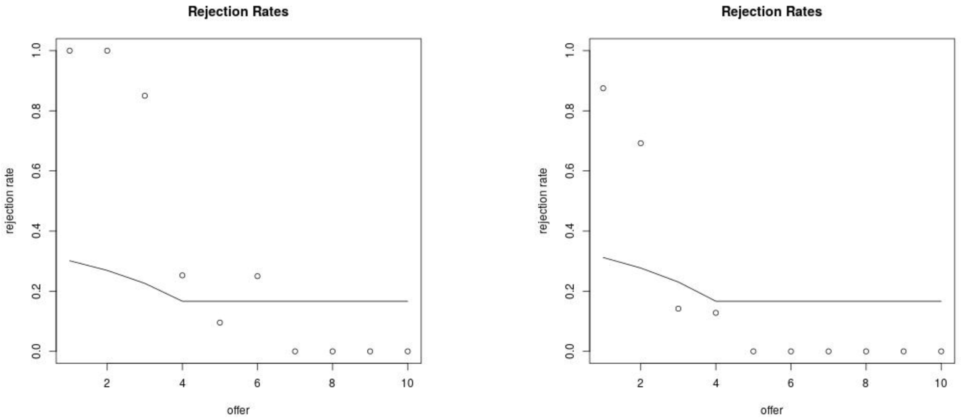

Figure 7.2: Observed Offers (left), Unobserved Offers (right)

Quantitatively the theory does poorly. Good offers are rejected far too often and bad offers not often enough. The low theoretical rate of rejection of bad offers is not in itself surprising. The theoretical rates make the selfish players exactly indifferent. This is a general problem in mechanism design theory. In practice with heterogeneity and noise to get selfish players to behave themselves it is wise to give them stronger incentives than exact indifference.

Assuming the ethical players reject more frequently than the theory says is inadequate to explain the data, however. In the data we may call an offer bad if it is less than \$4.00 in the obs treatment and less than \$3.00 in the nobs treatment. In the data there are 84 bad offers and 64, that is 76% of them, are rejected. By contrast the theory says that the maximum possible rejection rate when ethical players reject, selfish players accept, and noise players reject half the time is only 50%. Even if all the noise players and ethical players rejected the offers the probability of seeing so many rejections 64 is less than 4%. This suggests that the selfish players might be ethical players albeit with a much smaller put still positive value of γ, so willing to punish unlikely bad offers.

Fairness, Kindness and Reciprocity

Models of fairness or kindness and reciprocity such as Levine (1986), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Falk and Fischbacher (2006) and Dufwenberg and Kirchsteiger (2004) predict that an offer should be equally likely to be rejected regardless of how often it is made. Whether an offer is fair or kind does not depend upon how likely it is to be made. The left shift of equilibrium going from the obs to the nobs case provides a test of this hypothesis: the frequency of unfair and unkind offers of \$3.00 jumps up. Table 7.3 below shows that in fact when the frequency of \$3.00 offers jump up from 3% to 31% the rejection rate plunges.

	obs	nobs
frequency of offers	3%	31%
rejection rate	85%	14%

Table 7.3: Offers of 3

As indicated above, while mechanism design prediction is consistent with the obs data it is not with the nobs data. From a broader perspective however, what we see in the nobs data is an incentive compatible mechanism that is not the best, but very close: welfare of \$3.40 against \$3.45. On the other hand emotional players who rejected offers of \$3.00 at the rate seen in the obs case would do terribly against first movers who acted like those in the nobs case: welfare would plunge to \$1.79. Mechanism designers by contrast are pragmatic: even dropping incentives entirely and allowing the selfish types to offer one would result in much higher welfare of \$3.16.

Overview

The most important anomalies are

- In the no-observation case the offers distribution is shifted about \$1.00 to the left from the theory.
- The data has too few good offers and too low a rejection rate for them.
- The data rejects bad offers far too much.

I argued that the no-observation shift is not terribly important. That the theory produces too many good offers and rejects them too frequently is entirely due to the behavior of the noise players so can be isolated. The rejection of bad offers, however, implies that even selfish players must sometimes reject bad offers.

8. Public Goods with Punishment

Public goods experiments with punishment have been much studied and replicated since Fehr and Gachter (2000) and an overview can be found in Chaudhuri (2011). These studies show that without punishment little contribution occurs, but with punishment contribution levels are quite high. I used data from Nikiforakis and Normann (2008) who vary the cost of punishment and use a relatively easy to analyze linear cost structure.

Four partners play ten times but are randomly relabeled each period. The game has two stages. In the first stage money payoffs are given by

$$m^i(1) = 1.50 - s^i + 0.4\sum_{j=1}^{n} q^j$$

where $q^i \in \{0, 0.075, 0.150, \ldots, 1.50\}$ and $n = 4$. There is a punishment factor $\lambda \in \{0, 1, \ldots, 4\}$ and if $\lambda > 0$ there is a second stage

$$m^i(2) = m^i(1) - \sum_{j \neq i} p^{ij} - \lambda \sum_{j \neq i} p^{ji}$$

where $p^{ij} \in \{0, 0.075, 0.150, \ldots\}$ is a punishment assigned by player i to player j. There is also a constraint on individual punishment$\sum_{j \neq i} p^{ij} \leq m^i(1)$. As indicated the punishments have a common cost to the sender but differ in how costly they are for the recipient. Besides the ability to do comparative statics over λ this experiment also has a linear cost structure making it easier to analyze than some earlier experiments.

How to Spend It. Insight into the solution can be gained by considering the problem for an ethical player of optimally deploying their largesse for the lowest punishment factor λ =1. For a private cost of 0.6 an ethical player can increase their own contribution by 1. Suppose instead that the ethical player increases punishment for anyone who chooses the current and lower levels of contributions by 0.6. There are three opponents: roughly, another ethical player, a selfish player, and a noise player. The noise player will contribute the current level or less at some of the time, so the expected cost of the increased punishment is something less than 0.6. On the other hand, both the other players will be induced to increase their output by 1 resulting in a total output increase of 2. In other words, for somewhat lower expenditure of largesse, the ethical player can achieve a greater increase in contribution using it for punishment rather than for their own contribution. Higher punishment factors make largesse even more attractive. Hence punishment is always the right way to expend largesse.

Qualitative Analysis. The importance of these type of experiments is that they show how contributions jump up when punishment is an option. The analysis of the optimal use of largesse provides one explanation. Without punishment there is not enough largesse to lead to substantial contributions. With punishment using largesse for punishment, as shown above, is an effective way to generate high contribution rates. Moreover, unlike in the other experiments, here the largesse constraint binds in an important way. Higher levels of contribution require more punishment and while the gain in contribution outweighs the loss of punishment socially it does not do so individually. The amount of punishment sent by the ethical players does not depend on the contribution rate, rather it is determined by the largesse constraint. With lower punishment factors the punishment received is smaller so less high contribution rates can be sustained: this is the "law of demand" identified by Nikiforakis and Normann (2008).

Description of the Solution. As the stakes are not great, for simplicity of analysis I abstract here from risk aversion (see Appendix 1 for discussion). With this simplification it is easy to compute the solution to the mechanism design problem for each set of parameters. There is a single target for the selfish players. The ethical players provide incentives by punishing contributions below target. This punishment should be as small as possible subject to incentive compatibility. These facts are proven in Appendix 3. Computationally I find that welfare is increasing in the target while the cost of punishment also increases, so the target should be chosen as high as possible subject to the γ constraint for the ethical players. Because of the integer constraint of the selfish players this may not exhaust the largesse off the ethical players: any additional largesse is spent with a probability of the next higher contribution level. The equilibrium described in terms of expected punishment is unique but there may be several mixtures over punishment levels by the ethical players that give the same expected punishment and any of these is an equilibrium. Using these facts I computed the welfare optimal target for each punishment factor.

Below in Table 8.1 I report the optimum and compare it to the data the final (tenth) period. Note that "punishment" is the punishment sent by the punisher not the punishment received by the punished. For comparative purposes, note also that the maximum possible welfare is all contributing the maximum \$1.50 and there is no punishment: it is \$2.40.

punishment factor	contribution		punishment		welfare		participants
	theory	data	theory	data	theory	data	
4	1.16	1.24	0.06	0.07	1.92	1.91	24
3	1.05	1.16	0.06	0.05	1.91	1.99	24
2	0.90	0.68	0.06	0.04	1.88	1.78	24
1	0.70	0.24	0.06	0.01	1.81	1.64	24
none	0.07	0.03	0.00	0.00	1.51	1.51	24

Table 8.1: Public Goods Contribution with Punishment

Before comparing the theory to the data I want to comment on the big picture. An important reason public goods with costly punishment has been frequently studied is because of the stark contrast between the no punishment and punishment case. This can be seen in the data, where contributions jump from practically none with no punishment to 83% of the maximum with punishment factor four. The discussion of how to spend it above provides an explanation. Without punishment the ethical players have no choice but to spend their largesse on increased contributions, but as their largesse is limited this has little impact. Punishment, by contrast, is far more cost effective: the same largesse that has little impact on voluntary contributions has a big impact when used to provide incentives in the form of costly punishments.

Turning back to the comparison between the theory and the data, as expected from Table 4.1 the theory and data match quite well for the higher punishment factors $(3, 4)$ and when there is no punishment. It does, however, underpredict contributions by about \$0.10. Qualitatively the model gets right the declining contributions as the punishment factor declines. However, the theory does poorly from a quantitative point of view for lower punishment factors $(1, 2)$. In both cases actual contributions are substantially lower than predicted by the theory with correspondingly lower welfare.

The contribution schedule in the theory is flatter than in the data. This is partly due to the fact that in all treatments the noise players each make an expected contribution of \$0.75 regardless of the punishment factors. Below in Table 8.2 are the per capita contributions of the ethical and selfish players

punishment		4	3	2	1
contribution	ethical+selfish	1.36	1.20	0.98	0.68
	data	1.24	1.16	0.68	0.24

Table 8.2: If noise players do not contribute

When the punishment factor is 1 the empirical punishment is also much lower than the theory says is needed to sustain high contributions.

Distribution of Contributions

Below in Table 8.3 I provide greater detail for the punishment factor three and four cases. To provide some smoothing with 24 observations spread over so many categories I aggregated the 21 contribution levels into 7 by grouping them in blocks of 3 contribution levels.

effort	frequency		punishment		effort	frequency		punishment	
	theory	data	theory	data		theory	data	theory	data
1.50	0.71	0.71	0.03	0.00	1.50	0.05	0.46	0.03	0.00
1.13	0.05	0.04	0.06	0.00	1.13	0.71	0.21	0.03	0.05
1.03	0.05	0.13	0.09	0.00	1.00	0.05	0.13	0.08	0.08
0.67	0.05	0.00	0.13	?	0.78	0.05	0.13	0.12	0.00
0.53	0.05	0.00	0.16	?	0.53	0.05	0.00	0.16	?
0.30	0.05	0.00	0.19	?	0.38	0.05	0.04	0.21	0.00
0	0.05	0.13	0.23	0.55	0	0.05	0.04	0.26	0.83

Table 8.3: Left: Punishment Factor 4 - Right: Punishment Factor 3

The distribution of effort between the theory and data is reasonably good. The data on punishment is quite noisy since there are only 24 observations so less than 3 individuals in all but the top cell. Never-the-less the broad picture fits the theory: there is increased punishment for contributing less than the target, and probably that increases as distance to the target grows.

Low Punishment Factors

I would like to draw attention instead to what happens with punishment factors two and one. These are reported in Table 8.4 below.

effort	frequency		punishment		effort	frequency		punishment	
	theory	data	theory	data		theory	data	theory	data
1.50	0.05	**0.33**	0.03	0.00	1.50	0.05	0.08	0.03	0.00
1.28	0.05	0.00	0.03	?	1.28	0.05	0.00	0.03	?
1.02	**0.71**	0.08	0.03	0.00	1.05	0.05	0.02	0.03	0.00
0.75	0.05	0.08	0.06	0.08	0.75	**0.71**	0.05	0.04	0.11
0.52	0.05	0.00	0.09	?	0.52	0.05	0.00	0.07	?
0.30	0.05	0.08	0.12	0.00	0.26	0.05	0.08	0.11	0.00
0.02	0.05	**0.42**	0.15	**0.09**	0.03	0.05	**0.75**	0.16	**0.06**

Table 8.4: Left: Punishment Factor 2 - Right: Punishment Factor 1

The theory says ethical and selfish types all share the same target and that 71% of the population should be contributing that target. In the theory column I have highlighted the modal contribution levels implied by the theory: \$0.98 for punishment factor two and \$0.68 for punishment factor one. For punishment factor one the equilibrium seems to have collapsed to a mode of zero rather than \$0.68.

The data for punishment factor two has two peaks at the highest and lowest contribution levels. I have highlighted these as well in Table 8.4. This led me to wonder if there was not a mechanism with two peaks, but there is not: if the ethical players can be induced to contribute at the highest level then the selfish players will contribute not much less. Moreover, the punishment levels in the data are not nearly incentive compatible - the punishment for contributing at the lowest levels - also highlighted - is far too low to make it unprofitable for any type to deviate from the highest to lowest level. As I explain below this anomaly may be due to pooling across sessions.

Partners

As indicated these experiments were conducted using a partners treatment in which the same four players remained together for the entire ten periods. There is no issue with repeated game effects as data is only used from the final period. There is, however, another element of the partners treatment: the possibility of learning about the types of opponents. According to the model types are drawn once and for all at the beginning of the game and noise players randomize independently each period. Hence repeated play provides information about opponents' types. Learning about types provides a possible resolution to the anomaly found in the distribution of contributions for punishment factor two. Note that while the model provides a precise way of analyzing the learning it would not make sense to apply it to this data. While the players may be experienced with the one shot game after nine rounds, they are not experienced in learning about their opponents. To apply Bayesian updating about types I would want data with repeated repeated play as in Dal Bo (2005).

Specifically, with low punishment factors, the optimal mechanism is sensitive to how many ethical and noise players there are. In the extreme, if there are no ethical players, then the modal play is the least contribution while if there are no noise players the modal play is the greatest contribution. Each of these has a 20% chance. More generally with learning and low punishment factors, there will be some groups in which most players make the least contribution and others in which they make greatest contribution but there should rarely be groups in which many players do both. Never-the-less when the sessions are averaged it will give the anomalous pattern seen in Table 8.4 for punishment factor two.

To determine the frequency of anomolous groups in which many players simultaneously make both the lowest and highest contribution, I disaggregated the data for the punishment factor two treatments by group. For each group I calculated the minimum frequency of the highest and lowest contribution cells (\$1.50 and \$0.05 in Table 8.4). If groups tend to cluster at the top or the bottom but not both then these minima should be small. Below in Table 8.5 I report the fraction of punishment factor two groups that correspond to the different minima. There are 4 observations per group in the final period and 20 over the last five periods. For example, if the minimum frequency is 0.05 in the table for the last five periods one of two things is true: only one time out of 20 was the minimum contribution made or only one time out of 20 was the maximum contribution made. In other words the anomaly in Table 8.4 is not present in that session.

minimum frequency	final period	last five periods
$0-5\%$	67%	67%
$20-25\%$	33%	33%

Table 8.5: Top and Bottom Effort Levels

As can be seen, most groups do not simultaneously have high fractions of both the highest and lowest effort, so clustering and aggregation are largely responsible for the anomaly reported in Table 8.4 for punishment factor two. This clustering may be due learning in the partners treatment.

Robustness

As the public goods games with punishment are the only games in which γ plays a role indetermining the solution of the mechanism design problem I want to examine robustness with respect to the calibrated value of γ. Specifically, as observed above, the discount ratio μ for first period largesse is $\mu = 0.242$ in the final periods of the Dal Bo (2005) data. From equation 5.1 this corresponds to a value of $u^{-1}(\gamma) = 0.082$ substantially less than the calibrated 0.10. Below the welfare results are reported both for the calibrated $u^{-1}(\gamma) = 0.10$ and for the alternative $u^{-1}(\gamma) = 0.08$. As expected tightening the constraint reduces welfare - but very little. It results in a slightly worse fit for the high punishment factors and a slightly better fit for the low punishment factors.

welfare/punishment factor	4	3	2	1	none
$u^{-1}(\gamma) = 0.10$	1.92	1.91	1.88	1.81	1.51
$u^{-1}(\gamma) = 0.08$	1.89	1.88	1.84	1.80	1.51
data	1.91	1.99	1.78	1.64	1.51

Table 8.6: Effect of $\mu = 0.242$

Overview

There is one important anomaly: for low punishment factors players are far less successful at achieving high contribution levels than indicated by the theory. This might be explained by learning in the partners treatment.

9. Conclusion

I have presented a simple and stark calibrated benchmark model and documented its successes and failures across a range of different experiments. What do the anomalies tell us about how the model can be improved?

With respect to the selfish and ethical players their sharp optimization could be softened in several ways. The quantal response model of McKelvey and Palfrey (1995) rather than exact best response would eliminate the "good" equilibrium in the large population stag hunt game, for example. However, quantal response models are difficult to use for benchmarking purposes as they require an intensity parameter that is hard to predict without looking at the data. Another softening would be to introduce a trade-off for the ethical players between social welfare and own utility rather than a sharp limit. As a practical matter it seems unlikely that ethical players are willing to sacrifice as much for a small gain in social welfare as for a large gain. This could help in the public goods experiments with low punishment factors: perhaps ethical players with a smoother trade-off would be willing to forgo completely selfish behavior with only modest punishment because the social welfare loss is so great. Similarly, ethical players in public goods experiments where the punishment factor is low might be less willing to forgo selfish behavior when the social welfare gains are modest.

I think that softening the play of the selfish and ethical players while improving fit with the data would not make a good benchmark model: the additional parameters and complexity would not make sense when the model is doing reasonably well with these players already. Where the action is, I would say, is with the noise players.

Modeling the Noise Players

The noise players play a key role in the theory. They serve to keep the ethical players "honest" by forcing them to bear real costs of providing incentives through punishment. They also introduce an element of risk dominance that is important in equilibrium selection. Moreover the number of dissenters in the stag hunt and public goods contribution games are consistent with the calibration and show that indeed, a substantial fraction of the population are in some sense "noise" players.

In many respects the noise players are what differentiates this theory from others. If we just had ethical and selfish types with risk aversion the model would not be so different from using the Fehr and Schmidt (1999) model and choosing the best equilibrium as they suggest. In a similar vein if we introduced noise players into the Fehr and Schmidt (1999) it would likely fix the equilibrium selection problem in the large population stag hunt game and increase the rejections for ultimatum bargaining bringing the model into closer alignment with the data.

Roughly speaking simple entropy maximization delivers the basics at an aggregate level and so is useful as a benchmark. It does less well with the details. In large population stag hunt the dissidents appear to be split between those "making a statement" by providing the highest level of effort and those providing slightly more effort than the minimum. In ultimatum the most important anomaly is caused by the noise players making far too many good offers. They also reject far too many good offers and not nearly enough bad ones. In the public goods contribution games the noise players flatten the contribution schedule below that observed in the data.

Let me highlight some of the theoretical issues with the model of noise players. In the public goods games there is an abrupt change in the behavior of the noise players depending on whether or not there is punishment. Without punishment higher contributions are dominated by lower ones, so the probability of contributions falls off exponentially, and expected contributions by the noise players is quite small. By contrast with punishment no contribution level is weakly dominated so the probability of contributions is uniform and expected contributions jump up from near zero to $0.75. This fits well with the no punishment case and the high punishment factor case, but fares poorly with low punishment factors.

There is also a denomination issue. If there are two actions, one that earns zero and one that loses a dollar both are equally likely. If we add an intermediate denomination so that there is an action that loses fifty cents, then the probability of choosing zero remains at a half, but the probability of losing one falls to a quarter. As we add more intermediate denominations the expected loss falls. This might be true to an extent - but surely not when denominations become hard to distinguish.

In a related vein, when there are just two actions ordered by dominance both have equal probability. As I indicated this is already problematic in ultimatum as it leads to rejections rates too high for good offers and too low for bad offers. It is problematic for other games with two actions, for example, the one shot prisoner dilemma game. In that game suppose that the γ constraint binds on the ethical players so that they cheat. However, the noise player will cooperate half the time leading to a 17% cooperation rate. In fact (see, for example, Dal Bo (2005)) the cooperation rate with experienced players is more like 6%.

To "work" the noise players need to be impulsive and willing to do things "just for the heck of it." This should be clear from the roughly 30% of dissenters in stag hunt and in the public goods contribution games with punishment (see Table 8.3). But the data suggests that noise players are not, as the theory assumes, completely oblivious to what other players are doing. As I have indicated in the literature review, psychological models have become popular: surely an important role for these models is to better understand the behavior of noise players - to model their emotion and impulsivity?

Last Words

In considering the role of psychology I want to indicate that if I were designing a functional and effective human being I would design a person who was emotional and sometimes would be angry. The point is that emotions serve as a commitment device: because we are angry we carry out punishments that *ex post* are not in our interest. But we can control our anger, tailoring it to circumstances: an ethical player would use anger as an ends to a mean, carrying out the commitment to punish, but determining the level of punishment to provide efficient incentives. Perhaps the difference between ethical players and noise players is that the former have greater control over their emotions?

Finally, I want to conclude by emphasizing that the model is calibrated to western college students playing for standard stakes in the laboratory. We know from the work of Snowberg and Yariv (2021) and others that the general population is both more risk averse and more generous than college students. We also know from crosscultural studies such as Henrich et al (2004) that behavior in experiments differs between cultures. However, I have no reason to think that the fraction of ethical players is invariant to culture, and indeed in Dutta, Levine and Modica (2021) we developed a model of how the fraction of ethical players might vary depending on the nature of public goods problems faced by different cultures. One of the advantages of a benchmark model is that by identifying anomalies it provides hints as to what might be different between different populations, stakes, or cultures.

Appendix 1: Risk Aversion for Low Stakes

Figure 9.1 below plots the CES utility function with $\rho = 9$ (the dots) and a CARA utility function fit to the Gneezy and Potters (1997) data (solid line) along with a risk neutral utility function normalized to match utility at \$3.31.

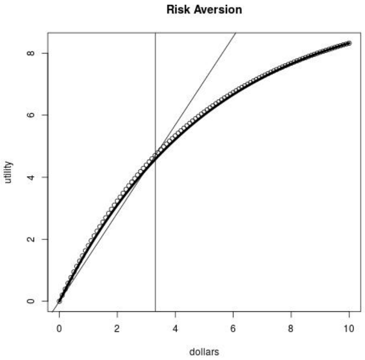

Figure 9.1: CES and CARA Utility

The linear utility function is a good approximation in the public games experiments, including stag hunt, because participants were paid for every period so the stakes were quite low. In Nikiforakis and Normann (2008) the greatest possible monetary payoff in a single period is \$3.31 marked with a vertical line.[3] As can be seen while the utility function has substantial curvature over the entire range, it is minimal over $[0.00, 3.31]$ where a straight line is an accurate approximation.

As it enormously simplifies computations I treated the players in the Nikiforakis and Normann (2008) experiments as risk neutral. I did not do so in stag hunt although it also makes very little difference: however in stag hunt as indicated above the benefit of deviating is so small in the large population games that I wanted to be sure that the calculations were valid when risk aversion was accounted for.

---

[3] Negative monetary payoffs are possible be this never occured in practice.

Appendix 2: Optimal Mechanism for Ultimatum

Proposition. *Selfish first movers use a pure strategy and no selfish or ethical players offers more than* \$5.00. *The rejection rates are as small as possible subject to incentive compatibility.*

Proof. SELFISH PLAYERS NEVER REJECT OFFERS. This is obvious.

THERE IS AN OPTIMAL MECHANISM IN WHICH NO SELFISH OR ETHICAL PLAYER MAKES A GOOD OFFER GREATER THAN \$5.00 WITH POSITIVE PROBABILITY. Suppose not. Set all rejection rates by ethical players for good offers and \$5.00 to zero.

If there is a selfish player making a good offer move all ethical and selfish players to \$5.00. This is the first best so certainly weakly welfare improving and is incentive compatible since the selfish player making the good offer is at least as willing to offer \$5.00 without punishment than a better offer.

If the only good offers are by ethical players, move all ethical players to \$5.00. This is certainly weakly welfare improving. it is incentive compatible for first movers. The utility of selfish first movers does not change. Since originally all ethical players had the same utility and the ethical player making the greater offer has at least as much utility after the move their utility weakly increases, so is incentive compatible. It is incentive compatible for second movers. For offers at or above \$5.00 the utility of selfish second movers does not change and the utility of ethical second movers weakly increases. For offers below \$5.00 the utility of ethical second movers decreases less than that of selfish first movers since the move is to a weakly lower rejection rate for the ethical second movers.

THERE IS AN OPTIMAL MECHANISM WITH MINIMUM INCENTIVE COMPATIBLE REJECTION RATES. Fix the first mover strategies. Lowering rejection rates subject to those strategies remaining incentive compatible improves welfare and improves second mover incentive compatibility for ethical players.

THERE IS AN OPTIMAL MECHANISM IN WHICH SELFISH FIRST MOVERS USE A PURE STRATEGY. If they are indifferent moving them to a higher offer does not change their utility, but increases the utility of the second mover who gets a better offer with lower probability of rejection. □

Appendix 3: Optimal Mechanism for Public Goods with Punishment

Proposition. *There is a single target for the selfish players. The ethical players provide incentives by punishing contributions below target. This punishment should be as small as possible subject to incentive compatibility.*

Proof. THERE IS AN OPTIMAL MECHANISM IN WHICH p^{ji} DEPENDS ONLY ON q^i. Take an optimal mechanism $\hat{q}, \hat{p}$and define $p^{ji}(q^i) = E[\hat{p}^{ji}|q^i]$. Then $\hat{q}, p$ is also incentive compatible and yields the same welfare.

Define $\hat{q}$ to be the maximum in the support of selfish player contributions. THERE IS AN OPTIMAL MECHANISM IN WHICH THERE IS NO PUNISHMENT FOR $q^i > \hat{q}$. This is welfare improving and incentive compatible for selfish types. Because I showed in the text that largesse is best spend on providing incentives it is also incentive compatible for the ethical types.

THERE IS AN OPTIMAL MECHANISM WITH MINIMUM INCENTIVE COMPATIBLE PUNISHMENTS. Fix the first period strategies. Lowering punishment rates subject for those strategies remaining incentive compatible improves welfare.

THERE IS AN OPTIMAL MECHANISM IN WHICH SELFISH FIRST MOVERS USE A PURE STRATEGY. If they are indifferent moving them to a higher contribution level does not change their utility, but increases the utility of everyone else. □

References

Andreoni, J. (1990): "Impure altruism and donations to public goods: A theory of warm-glow giving," *Economic Journal* 100: 464-477.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," *American Economic Review* 91: 166-193.

Gintis, H., Bowles, S., Boyd, R. and Fehr, E. (2003): "Explaining altruistic behavior in humans," *Evolution and Human Behavior*, 24: 153-172.

Charness, G., U. Gneezy, U. and A. Imas (2013): "Experimental Methods: Eliciting risk Preferences," *Journal of Economic Behavior and Organization* 87: 43-51.

Charness, G., and M. Rabin, M. (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics* 117: 817-869.

Chaudhuri, A. (2011): "Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature," *Experimental Economics* 14: 47-83.

Coase, R. H. (1960): "The Problem of Social Cost," *Journal of Law and Economics* 3: 1-44.

Coate, S. and M. Conlin (2004): "A Group Rule-Utilitarian Approach to Voter Turnout: Theory and Evidence," *American Economic Review* 94: 1476-1504.

Dow, J. and G. Gorton (2008): “Noise Traders,” *The New Palgrave: A Dictionary of Economics*, 2nd Edition (Palgrave Macmillan: New York), edited by Steven N. Durlauf and Lawrence E. Blume.

Duffy, J. and Feltovich, N. (1999): “Does observation of others affect learning in strategic environments? An experimental study,” *International Journal of Game Theory* 28: 131-152.

Dufwenberg, M., and G. Kirchsteiger (2004): “A theory of Sequential Reciprocity,” *Games and Economic Behavior* 47: 268-298.

Bó, P. Dal (2005): “Cooperation under the shadow of the future: experimental evidence from infinitely repeated games,” *American Economic Review* 95: 1591-1604.

Dutta, Rohan (2012): “Bargaining with Revoking Costs,” *Games and Economic Behavior*, 74: 144-153.

Dutta, R., D. K. Levine and S. Modica (2021): “The Whip and the Bible: Punishment Versus Internalization,” *Journal of Public Economic Theory* 23: 858-894

Engel, C. (2011): “Dictator games: A meta study,” *Experimental Economics* 14: 583-610.

Falk, A., and U. Fischbacher (2006): “A Theory of Reciprocity,” *Games and Economic Behavior* 54: 293-315.

Feddersen, T., A. Sandroni (2006): “A Theory of Participation in Elections,” *American Economic Review* 96: 1271–1282.

Fehr, E. and S. Gächter (2000): “Cooperation and Punishment in Public Goods Experiments,” *American Economic Review* 90: 980-994.

Fehr, Ernst and Klaus M. Schmidt (1999): “A Theory of Fairness, Competition and Cooperation”, *Quarterly Journal of Economics* 114: 817-868.

Fischbacher, U. and S. Gachter (2010): “Social preferences, beliefs, and the dynamics of free riding in public goods experiments,” *American Economic Review* 100: 541-56.

Fudenberg, D., and G. Karreskog Rehbinder (2024): “Predicting Cooperation with Learning Models,” *American Economic Journal: Microeconomics* 16: 1-32.

Fudenberg, D. and D. K. Levine (1997): “Measuring Players’ Losses in Experimental Games,” *Quarterly Journal of Economics* 112: 507-536.

Fudenberg, D. and D. K. Levine (2011): “Risk, Delay, and Convex Self-Control Costs,” *AEJ Micro* 3: 34–68.

Fudenberg, D. and D. K. Levine (1989): “Reputation and Equilibrium Selection in Games with a Patient Player,” *Econometrica* 57: 759-778

Gneezy, U. and J. Potters (1997): “An Experiment on Risk Taking and Evaluation Periods,” *Quarterly Journal of Economics* 112: 631-645.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2004): “Overview and Synthesis,” *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, ed. Henrich, J. P., Boyd, R., Bowles, S., Fehr, E., Camerer, C., and Gintis, H., Oxford University Press on Demand.

Herrera, H., M. Morelli, M. and S. Nunnari (2016): “Turnout Across Democracies,” *American Journal of Political Science* 60: 607-624.

Holt, C. and S. Laury (2002), “Risk Aversion and Incentive Effects,” *American Economic Review* 92:1644-1655.

Kandori, M., F. Mailath and R. Rob (1993): “Learning, Mutation, and Long Run Equilibria in Games,” *Econometrica:* 29-56.

Kreps, D. and R. Wilson (1982): “Reputation and Imperfect Information,” *Journal of Economic Theory* 50: 253-79.

Levine, D. K. (1986): “Modeling altruism and spitefulness in experiments,” *Review of Economic Dynamics* 1: 593-622.

Levine, D. K. and A. Mattozzi (2020): "Voter Turnout with Peer Punishment," forthcoming, *American Economic Review.*

Mailath, G. J. and L. Samuelson (2001): “Who wants a good reputation?” *Review of Economic Studies* 68: 415-441.

McKelvey, R. D. and T. R. Palfrey(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.

Milgrom, P. and J. Roberts (1982): “Predation, reputation, and entry deterrence,” *Journal of Economic Theory* 27: 280-312

Nikiforakis, N. and H. T. Normann (2008): “A Comparative Statics Analysis of Punishment in Public-good Experiments,” *Experimental Economics* 11: 358-369.

Ostrom, Elinor (1990): *Governing the commons: The evolution of institutions for collective action*, Cambridge university press.

Palfrey, T. R. and Prisbrey, J. E. (1997): “Anomalous behavior in public goods experiments: How much and why?” *American Economic Review*, 829-846.

Peski, M. (2010): “Generalized Risk-dominance and Asymmetric Dynamics,” *Journal of Economic Theory* 145: 216-248.

Snowberg, E., and L. Yariv (2021): “Testing the Waters: Behavior Across Participant Pools,” *American Economic Review* 111: 687-719.

Townsend, R. M. (1994): “Risk and insurance in village India,” *Econometrica*, 539-591.

Van Huyck, J. B., R. C. Battalio and R.O. Beil (1990): “Tacit coordination games, strategic uncertainty, and coordination failure,” *American Economic Review* 80: 234-248.

Young, H. P. (1993): “The Evolution of Conventions,” *Econometrica*: 57-84.

ONLINE Appendix 1: Willingness to Sacrifice

For the calibration of γ in the text to be valid the γ constraint should bind both for the dominant strategy experiments in which willingness to give is measured over time and for the one-off dictator game in which base willingness to give is measured.

Take first the public goods experiments in Fehr and Gachter (2000). From the information given concerning overall earnings, we find in dollars

$$m^i = 1 - (0.6)s^i + (0.4)\sum_{j=1}^{n} s^j.$$

where $s^i \in \{0, .05, 0.10 \ldots, 1.00\}$ and $n = 4$. The derivative of social welfare with respect to s^i is given as

$$\frac{dS}{ds^i} = (1/n)\sum_{j=1}^{n} u'(m^j)\left((0.40) - \mathbf{1}(i = j)(0.6)\right)$$

where

$$u'(m) = -(1/C)(1-\rho)(1+m/C)^{-\rho}.$$

Observe that if an ethical player contributes the maximum of \$1.00 and nobody else contributes they get \$0.80 and everyone else gets \$1.40. Marginal utility for the former is $u'(0.8) = 0.0206$ and the latter $u'(1.4) = 0.0183$. Hence

$$\frac{dS}{ds^i} \geq (0.0183)(0.40) - (0.0206)(0.25)(0.6) = .010 > 0,$$

that is, the fact that a one dollar increased contribution results in a six dollar increase in welfare dominates the (modest) unfairness of ethical players contributing more than others, so an ethical player should contribute the most they can: that is, the constraint should bind.

In the dictator experiments, by contrast, fairness dictates that an ethical player contribute no more than half the maximum, and this constraint might bind. However, in \$10.00 dictator experiments with \$1.00 increments average contributions are \$2.50 while in \$5.00 dictator experiments with \$0.50 increments such as List (2007) they fall to about half that, to \$1.33 in List (2007) (24 participants). If the γ constraint was strictly binding that should not happen as the selfish types continue to contribute zero, the noise types halve their contribution, and the ethical types reduce their contribution by less than a half.

However, List (2007) provides us with extra information: he considers allowing a "take" option of taking up to \$5.00 (48 participants). If the constraint is not binding when the take option is available then contributions by the ethical type should remain fixed at \$2.50, while the contributions of the selfish player should fall by \$5.00 and that of the noise players by approximately that amount. That is, contributions should fall by \$3.33. However, they fall by \$3.82 which indicates that the constraint does bind in the take treatment. The actual contribution in this treatment is −\$2.48, or \$2.52 above the floor. Conveniently this is about the same as the approximately \$2.50 contribution in the standard \$10.00 treatments, so we conclude that \$2.50 is a good estimate of contributions when the constraint does bind.

ONLINE Appendix 2: Welfare Comparison

The Fehr-Schmidt calculations are based on Fehr and Schmidt (1999).

For ultimatum I computed rejection rates, then optimal offers for each type. This led to 70% offering 5 and having it accepted and the remainder offering 4 and having it accepted 90% of the time.

For public goods contributions games I used their Proposition 5. It requires $\beta = 0.6$ for the non-selfish type. The correlation between α and β for their calibrated model is not specified, but even assuming independence there are at least 12% of the population with $\alpha \geq 1$ and $\beta = 0.6$ and this guarantees that the condition in Proposition 5 is satisfied for all the relevant punishment factors.

There is also a constraint that links the best possible equilibrium to the greatest possible punishment. Let Q be the best possible equilibrium, ψ the fraction of types with $\beta = 0.6$ and $\alpha \geq 1$, and punishment factor λ. Recall that the greatest possible punishment if everyone chooses Q in the first state is $1.6Q$. From Fehr and Schmidt (1999) Proposition 5 the constraint is

$$\frac{Q}{4\psi - 1/\lambda} \leq 1.6Q.$$

The fraction of types with $\beta = 0.6$ and $\alpha \geq 1$ could be anywhere from 0.12 to 0.40. At $\lambda = 1$ the necessary ψ is greater than 0.40 so the constraint fails. I took the fraction

ψ to be greater than 0.28 as the constraints for $\lambda > 1$ are satisfied and this best fits the data.

ONLINE Appendix 3: Ethical Players in Stag Hunt

Given that selfish players are contributing \$0.10 should ethical players they raise their contribution to a greater amount, for example \$0.20 instead of \$0.10? This results in a sure loss to themselves (33% of population) of \$0.10. However, if there are no selfish players and all the noise players contribute \$0.20 or more, it does raise the income of all players by \$0.10.

The implication is that if and only if there is a greater than 1/3 chance both that there are no selfish players and that all the noise players contribute \$0.20 or more then it is welfare improving for the ethical players to increase their contribution. However, the chance of this happening is much less than 1/3: the chance that there are no selfish players by itself is less than 0.3%.

The same analysis applies to contributions greater than \$0.20, and indeed above \$0.30 welfare would be reduced even if the selfish players were also willing to contribute that amount. Hence the unique equilibrium as reported above and to contributions above \$0.20 when that is the equilibrium.

ONLINE Appendix 4: Sampling Error

Hypothesis testing is fraught and standard errors are often subject to misinterpretation. The gap between theory and practice is large, and I refer the interested reader to the relevant literature and in particular Leamer (1983) and Imbens (2021).

Before explaining how standard errors they should be interpreted in this setting, let me first describe how to compute them. Standard errors are computed with respect to a model: in this case the "hypothesis" is a point hypothesis - the calibrated model with given and known coefficients that are not estimated and certainly not from the data being analyzed. From the theoretical model samples are drawn in exactly the same way as in the data and this gives the entire distribution of sampling error including the standard error. A convenient way to do this is via Monte Carlo.

There are two possible results of finding confidence intervals based on the standard errors. First, the confidence intervals may be small and the data may lie outside the confidence intervals. Since the theory is certainly wrong and is expected to be wrong with enough data this will always be the case, and so it is meaningless. In smaller samples the confidence intervals might be large relative to the distance from the theory to the data. When the theory is closer to the data than the standard errors suggest the only reasonable conclusion that can be drawn is that I manipulated either the theory or the data: for example, I cherry-picked the experiments to fit the theory, or I chose the calibration after looking at the data. For this reason there is some importance in reporting the standard errors.

In the stag hunt games I analyzed sampling error in the context of dissidents where it was relevant. I have not computed standard errors in the ultimatum games because the sample is large and there is an issue of whether there is serial correlation over the thirty periods of data used. Below in Table 9.1, in the public goods games with punishment, I add to Table 4.1 the standard errors (se) computed using a Monte Carlo with $1,000$ draws.

	theory	data	**se**	actual err	SGP err	FS err
pun 1	1.81	1.64	0.07	**0.17**	−0.14	
pun 2	1.88	1.78	0.10	0.10	−0.28	0.63
pun 3	1.91	1.99	0.13	−0.08	−0.49	0.42
pun 4	1.92	1.91	0.15	0.01	−0.41	0.50

Table 9.1: Welfare and Standard Errors for Public Goods with Punishment

Specifically the procedure is this. For each treatment each Monte Carlo iteration draws six matches. In each match four players are randomly drawn from the player types and the noise players randomly draw contributions and punishments. The punishments of the noise players by the ethical players is then computed, and welfare for the match is determined and averaged over the six matches. This is done $1,000$ times to find the distribution of the draws.

The standard errors are large enough that except in the case of punishment factor one the data is within two standard deviations of the theory. Except in the case of punishment factor four the data is not exceptionally close to the theory compared to the standard error. In addition the standard errors are small enough that with the exception of punishment factor one they exclude SGP and FS which lie considerably more than two standard deviations from the data. This argues that if we were to engage in hypothesis testing the test would have some power. We may also wish to conclude that the anomolous data point - the increase in welfare going from punishment four to three - might be due to sampling error.

Overall the standard errors, while computable, add only modestly to our understanding of the theory and the data.

References

Imbens, G. W. (2021): “Statistical Significance, p-values, and the Reporting of Uncertainty,” *Journal of Economic Perspectives* 35: 157-174.

Leamer, E. E. (1983): “Let’s take the con out of econometrics,” *American Economic Review* 73: 31-43.

List, J. A. (2007): “On the interpretation of giving in dictator games,” *Journal of Political Economy* 115: 482-493.