Judgment and Decision Making, Vol. 9, No. 6, November 2014, pp. 523–547

Ostracism and fines in a public goods game with accidental contributions: The importance of punishment type

Torrin M. Liddell[*] John K. Kruschke[†]

Abstract

Punishment is an important method for discouraging uncooperative behavior. We use a novel design for a public goods game in which players have explicit intended contributions with accidentally changed actual contributions, and in which players can apply costly fines or ostracism. Moreover, all players except the subject are automated, whereby we control the intended contributions, actual contributions, costly fines, and ostracisms experienced by the subject. We assess subject's utilization of other players' intended and actual contributions when making decisions to fine or ostracize. Hierarchical Bayesian logistic regression provides robust estimates. We find that subjects emphasize actual contribution more than intended contribution when deciding to fine, but emphasize intended contribution more than actual contribution when deciding to ostracize. We also find that the efficacy of past punishment, in terms of changing the contributions of the punished player, influences the type of punishment selected. Finally, we find that the punishment norms of the automated players affect the punishments performed by the subject. These novel paradigms and analyses indicate that punishment is flexible and adaptive, contrary to some evolutionary theories that predict inflexible punishments that emphasize outcomes.

Keywords: punishment, public goods game, ostracism, trembling-hand, intention, outcome bias.

1 Introduction

Cooperation is essential to the functioning of human society. Often, cooperation requires subverting the interests of an individual in favor of the interests of her group. The conflict between the individual's and group's interests is often studied via a *public goods game* (PGG), which is an economic game that is similar to the prisoner's dilemma. In a PGG, each round of play begins with an equal endowment allocated to all players. Players then individually decide how much they will invest in the common pool (i.e., the public good). After the investments, the common pool is multiplied by a constant (e.g., the pool is doubled) and then divided equally among all players regardless of their contribution. This structure implies that the best outcome for an individual is to contribute nothing (i.e., defect) while others do contribute to the common pool (i.e., cooperate). Yet the best outcome overall is for all players to contribute their entire initial endowment. Thus the interests of the individual conflict with the interests of the group. Previous research has shown that, in the absence of punishment, cooperation in these games usually starts relatively low and declines over repeated rounds. But when players have the option to punish each other, then contributions start high and remain relatively stable over repeated rounds (e.g., Fehr & Gächter, 2000; Ostrom, Walker & Gardner, 1992). Thus, punishment is a mechanism for maintaining cooperation, at least under some conditions.

Much of the research on punishment has focused on costly fines, whereby a player can impose a fine on another player but only at a cost to herself. In our new experiments, we simultaneously give players the option of costly fine or ostracism. In our contexts, we assume that ostracism has no immediate cost to the player who causes it. Our design also adds a noise component to the contribution procedure, so that the intended contribution is not identical to the actual contribution. Our results show that ostracism and costly fines are used differently in several ways: While costly fines tend to be directed at players whose *actual* contribution is low, ostracism tends to be directed at players whose *intended* contribution is low. Moreover, propensity to ostracize more readily adapts to group norms than propensity to apply costly fines, and ostracism is applied more than costly fine to players who do not increase their contribution after suffering a fine on a previous round. Our results indicate that different types of punishment are flexible and responsive to various aspects of social structure.

The remainder of the introduction provides background about outcome bias in punishment, the influence of group

The authors gratefully acknowledge Lauren Bojarski, Kara Buckingham, Angela Receuver, and Casey Yoder for their assistance administering the experiments. The authors also thank Sam Zorowitz for his helpful discussion regarding experiment design.

[*] Department of Psychological and Brain Sciences, Indiana University, Bloomington 1101 E 10th St, Bloomington, IN 47405 Email: torrin.liddell@gmail.com.

[†] Department of Psychological and Brain Sciences, Indiana University, Bloomington Email:johnkruschke@gmail.com.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

norms on punishment, and the modulation of punishment by its previous success or failure. We then present results from a series of three experiments involving a novel variation of the PGG, in which intended and actual contributions are explicitly dissociated, and players have the option to ostracize or fine. The results are analyzed using hierarchical Bayesian methods, which provide rich information about the relative influences of intended and actual contributions on different types of punishment.

1.1 Punishment type

Punishments in PGGs and other economic games are usually costly fines (e.g., Fehr & Gächter, 2000; Ostrom et al., 1992; Fehr & Schmidt, 1999; Cushman, Dreber, Wang & Costa, 2009). This type of punishment allows players to deduct resources from another player at a cost to themselves. However, another real-world punishment is ostracism. Ostracism entails a refusal of repeat business with the punished party, and by definition has no immediate cost. Ostracism prevents any future transgressions from the punished party. This type of punishment can also motivate cooperation in public goods games (Cinyabuguma, Page & Putterman, 2005; Maier-Rigaud, Martinsson & Staffiero, 2010; Masclet, 2003), but it has been studied relatively rarely.

Baumard has suggested that ostracism is much more representative of everyday punishment than costly fine (Baumard, 2010, 2011; Baumard, André & Sperber, 2013). He cited anthropological literature to argue that in the hunter-gather societies representative of the environments under which humans evolved, costly punishment is exceedingly rare. Furthermore, he argued that human cooperation can be explained by partner choice alone, which is simultaneously inexpensive compared to a costly punishment and prevents any future transgressions. This account is consistent with research on non-human animals that indicates that costly punishment is quite rare and that ostracism is much more frequently observed (Raihani, Thornton & Bshary, 2012; Stevens, Cushman & Hauser, 2005).

There is also evidence from game-theoretic computer simulations that exclusion may be more conducive to the evolution of cooperation than other forms of punishment. The simulations of Sasaki & Uchida (2013) assumed that ostracism of a free rider resulted in immediate benefits for the remaining cooperative group members because of less dilution of group output in subsequent rounds.[1] On the other hand, costly fines produce no direct benefit if the punished individual does not increase his contribution in subsequent rounds.

Given that ostracism has important structural differences from costly fine, and that ostracism may be an especially important form of punishment in the real world, we incorporated both forms of punishment in our novel PGG.

1.2 Outcome bias and outcome emphasis

Outcome bias occurs when the outcome of an event contributes to the evaluation of an action even when all other aspects of the action (e.g., intention of the actor, reasoning of the actor) are held constant and fully known to the evaluator (Baron & Hershey, 1988). When people make a punishment decision they often exhibit outcome bias (though not necessarily in all conditions, see, e.g., Rand Fudenberg & Dreber, 2013). Moreover, they can exhibit an even more extreme behavior pattern we refer to as outcome emphasis, which means that they weigh the actual outcome of the transgression *more strongly* than the transgressor's intended outcome. When there is outcome emphasis, accidental transgressions tend to be punished, but attempted transgressions that fail to occur tend to be excused.

Consider, for example, an economic game developed by Cushman et al. (2009) that involved a so-called "trembling hand," named after gunmen with trembling hands who might intend to hit their target but accidentally miss, or who intend merely to scare their target with a close miss but accidentally hit. One player was given an amount of money to allocate between herself and the second player. After allocation, the second player was allowed in response to apply a monetary penalty or bonus to the allocating player. The unique feature of the game was that the allocating player chose one of three dice instead of an exact allocation. The three dice had different probabilities of (a) selfishly keeping the entire allocation, (b) fairly splitting it, or (c) generously giving it all away to the second player. One die had a 2/3 chance of being selfish, a second die had a 2/3 chance of being fair, and a third die had a 2/3 chance of being generous, with the other two allocations having 1/6 probability in all cases. Thus, the allocator could intend to be selfish, fair, or generous, but accidentally roll an unintended outcome. Crucially, the choice of die was explicitly revealed to the receiving player. Results showed that when making punishment decisions, subjects put much greater weight on the actual outcome as compared to the intended outcome. In other words, people were willing to punish choosers who were accidentally selfish, despite knowing that the chooser intended to be fair or generous.

A phenomenon related to outcome bias in the present context is inequity aversion, wherein people use punishment to enforce equality of outcome (Fehr & Schmidt,

[1] Contrary to the work of Sasaki & Uchida (2013), excluded players were replaced by randomly selected new players in our experiment. This means that it was not guaranteed that ostracism would result in less dilution of group output.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

1999; Bolton & Ockenfels, 2000; Raihani & McAuliffe, 2012) or more generally as a response to inequity (Cook & Hegtvedt, 1983; Yamagishi & Horita, 2009). If individuals are not attempting to punish accidental transgressors and are instead attempting to enforce fairness, outcome bias would be reducible to inequity aversion. To test this possibility, the die experiment summarized above (Cushman et al., 2009) included a condition in which choosers had *no* control over their allocation to the other player, and allocations were explicitly random. The receiving players still punished "selfish" allocations, but the effect of outcome was less than in the trembling-hand condition. Therefore, inequity aversion alone is unlikely to be a complete explanation of outcome bias.

Despite the evidence supporting the existence of outcome emphasis in punishment, we have reason to think that outcome emphasis may be reduced or even eliminated in the case of ostracism. Punishers may be less willing to lose a person who, based on his intentions, is likely to be cooperative in the future, even if his intentions did not yield cooperative behavior in the present encounter.

1.3 Group norms

Social norms and punishment are strongly intertwined. Norms set the bar for what is worthy of punishment and what is not (e.g., Fehr & Fischbacher, 2004; Carpenter, Verhoogen & Burks, 2005). We are concerned with norms that establish which trangressions are punishable, and norms for what type of punishment to apply. Some previous research explored preexisting norms spontaneously used by individuals in experimental situations. For example, Carpenter and Matthews (2009) were able to estimate the punishment norm subjects used in an economic game regarding decisions to punish or not, finding that within one's own group players compare contributions to a high absolute threshold (insensitive to group average), and players that fail to meet this threshold are punished. There are many examples of variations in punishment behavior in laboratory games across cultures (Henrich et al., 2005) including the especially peculiar case of antisocial punishment (Herrmann, Thöni & Gächter, 2008).[2]

Clearly, social norms play an important role in punishment. We report a new experiment that investigates the interplay of punishment norms and punishment type. We are interested in the degree to which punishment is influenced by the social norm, and if this influence depends upon the type of punishment under consideration.

1.4 Efficacy of punishment

Copious evidence indicates that punishment is useful for maintaining cooperative behavior (e.g., Yamagishi, 1986; Fehr & Gächter, 2000; Ostrom et al., 1992; Cinyabuguma et al., 2005; Maier-Rigaud et al., 2010; Masclet, 2003). However, it is less clear whether a specific individual will adjust her punishments in response to the efficacy of the punishment, or if the urge to punish is a fixed response to perceived transgression.

Cushman (2013) argued that punishment should be a fixed response, in that the likelihood that the punishment will change future behavior should be disregarded. If punishers reduced the magnitude or probability of their punishments when the punishment did not affect the behavior of the punished individual, then persistent transgressors would defeat punishment. Consequently, punishment could not evolve as a mechanism for encouraging cooperation. This argument was borne out by evolutionary simulations (Cushman & Macindoe, 2009). Because cooperation has in fact flourished in real populations, it must be (the argument goes) that punishment evolved to be a fixed response.

We ask whether subjects differentially utilize costly fine and ostracism based on the transgressors behavioral change (or lack thereof) in response to punishment. How do punishers respond to players who persistently free ride?

1.5 Automation

To assess many of the research questions of interest it was efficacious to have complete experimental control of the game environment, and therefore we automated all the players other than the single human subject. Automated players have been used in previous research in the context of PGGs (Suri & Watts, 2011; Barclay, 2006), other economic games (e.g., the prisoner's dilemma in Kiesler, Sproull & Waters, 1996), and other forms of experimental games (e.g., a blame attribution game in Gerstenberg & Lagnado, 2010). Moreover, there is evidence that computerized players are treated as human in multiple contexts (e.g., Nass, Fogg & Moon, 1996; Fogg & Nass, 1997; Nass & Moon, 2000).

In all of our experiments, players were told that they may be playing against networked or automated players (see Appendix 6 for verbatim instructions). We did not collect any measures that assessed whether or not subjects believed they were playing against automated or human players. In informal discussions after experiment sessions,

[2] More generally, norms of punishment in the real world vary in other ways. Regions and cultures differ in endorsement of corporal punishment of children (Lansford & Dodge, 2008; Flynn, 1994). Attitudes towards the death penalty have fluctuated greatly in the United States, ranging from 42% supporting capital punishment in 1966 to 80% in 1996 (Jones, 2013; Jacobs & Carmichael, 2002; Zeisel & Gallup, 1989). More recently, punishments intended to humiliate or shame the offender, such as spending time publicly wearing a sign detailing one's crime, have been controversially reintroduced in some American courts. Public humiliation is a form of punishment that some legal scholars have argued is acceptable under our punishment norms, whereas others argued the opposite (Book, 1999; Kahan, 1996, 2006; Whitman, 1998).

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

where all subjects were given the opportunity to present any comments or questions they had regarding the experiment, some subjects expressed uncertainty regarding whether they were playing against human or automated players, but the vast majority did not bring up the topic.

1.6 The present studies

We investigate each of these issues using our novel PGG in which intended and actual contributions are explicitly dissociated. In the first experiment, we verified that outcome emphasis occurs in the new PGG for costly fines. In the second experiment, we introduced ostracism and we assessed how much punishers weigh actual contributions versus intended contributions for each type of punishment. Finally, in the third experiment, we investigated how the two types of punishments are influenced by group punishment norms, the degree of control in accidental contributions, and of the efficacy of punishment.

2 Experiment 1: Emphasis on actual contribution for costly fines

A key innovation for our PGG is that contributions to the public good are affected by a trembling hand. All players see the intended and actual contributions to the public good. We set out to ask whether outcome emphasis in punishment would occur in a trembling-hand PGG, that is, where the bias would manifest as punishments that emphasize the actual contribution more than the intended contribution.

2.1 Methods

160 Indiana University (IU) undergraduates participated in the experiment for course credit. Subjects were recruited from the human subjects pool of the Department of Psychological and Brain Sciences. We assume the subjects were representative of the pool, which is approximately 65% female with ages ranging approximately from 18 to 45 years with a modal age of 19.

Subjects were told they would be playing a game while seated at a computer with other players who might be networked people or automated. Each player was referred to by a static single letter label. At the beginning of each round players were given 10 points and allowed to contribute as many points as they wished to a common pool. This contribution was described as an investment in a group venture. Following this choice, noise was applied to the intended contribution to produce the actual contribution. The noise was a random integer chosen uniformly from the set 2, 3, and 4, and then assigned a positive or negative sign with equal chance. This noise pattern (particularly, the lack of 0 noise) was chosen in order for the influence of the intended and actual contribution to be more easily distinguished. Subjects were told that this random noise reflected real world contingencies such as miscommunications or mistakes. This value was added to the intended contribution to produce the actual contribution. Actual contributions could not be below 0 or above 10. Every game had five players, four of which were automated. The actions of the automated players were randomly selected from a pregenerated list of contribution combinations. A complete list is available in Appendix 6. Each combination consisted of a player who was intentionally low, a player who was accidentally low, a player who was intentionally high, and a player who was accidentally high. The “low” and “high” designations are relative to a baseline of 5. After the subject made her contribution, all of the other players’ contributions were displayed in a table on the computer screen. This table also contained the amount paid out to each player from the pool and each player’s total gain for the round. Payout was equal to the total amount contributed multiplied by 1.6 and then divided equally among all players. Total gain for a player was equal to the payout from the pool plus any amount that the player kept from his initial allocation.

After the contributions and payoffs were displayed, the subject had the opportunity to punish the other players. In phase 1 of the experiment, consisting of 22 rounds, only the subject was given the opportunity to punish other players. Subjects were not made aware of the length of this phase, or even that there would be a second phase. Punishment consisted of deducting points from the player, at a cost of a quarter point per point deducted. Subjects could punish any number of players as long as they did not attempt to spend more points than they gained in the round.

In phase 2, again consisting of 22 rounds, the automated players also applied penalties. Subjects were told that they were starting a new game with new players, and that the other players could apply penalties. Every automated player applied a penalty to every other player in the amount that the other player’s actual contribution was less than the mean original contribution. The subject applied her penalties without seeing the other players’ penalties. After the subject applied her penalties, a table was displayed that showed the punishments applied by all players to all players, along with the net gain after penalties. The purpose of this two-phase design was to be able to observe the behavior of subjects unbiased by the punishment behavior of the automated players (phase 1) and also in the presence of other punishing players (phase 2).

The trembling-hand PGG has several other novelties relative to previous research. In our trembling-hand PGG, punishers are also contributors, unlike in previous work with a different paradigm in which punishers were only responding to the actions of others (Cushman et al.,

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

2009). In the trembling-hand PGG, many rounds are actually played consecutively instead of using the "strategy method" in which hypothetical judgments are solicited from each subject. Finally, in our trembling-hand PGG, the other players are automated to give us complete control of the game environment.

2.2 Results

When deciding to punish, the subject sees three sources of information about herself and the other players, namely their intended contribution, their actual contribution, and their net gain. We are interested in how much each source of information is weighted in the decision to apply a fine. (Note that we are analyzing the probability of applying any fine, not the magnitude of fine applied. We do so to allow comparison with the exclusionary punishment introduced in Experiment 2, which has no magnitude. However, we also performed a linear regression with similar results; this approach is detailed in Appendix 4.) Therefore, to model the probability of applying a fine, we used logistic regression on three predictors: the intended contribution of the targeted player, the actual contribution of the targeted player, and the extent to which the targeted player got more net points than the punisher, which we call "indignation". Colloquially, indignation is a sense of anger or annoyance at perceived injustice. This label is a convenient mnemonic for the numerical predictor but it does not imply that we measured a subjective attitude.

Indignation as we define it here is closely related to the concept of inequity aversion as described by Fehr and Schmidt (1999). In the model utilized by Fehr and Schmidt (1999), inequity is the average difference in payout between a given player and all other players. Inequity so defined is essentially average indignation as defined in our analysis. Thus, our including indignation as a predictor allows the regression model to distinguish the influence of personal inequity (indignation) from the influence of actual contribution.

2.2.1 Bayesian hierarchical logistic regression

The hierarchical model applies logistic regression to each individual and had higher-level distributions across the individual regression parameters to describe group-level tendencies. For a full description of the model, see Appendix 1. The important parameters for our purposes are the normalized group-level regression weights, which indicate the relative influence of the three predictors. The regression weights are denoted β_{act} for the actual-contribution predictor, β_{int} for the intended-contribution predictor, and β_{indig} for the indignation predictor. These beta weights represent the relative importance of the given predictor in determining the probability of applying a fine, at the level of the group tendency. A large magnitude beta weight represents that the predictor is relatively important, and a beta weight near zero indicates that the associated predictor is relatively unimportant for predicting the application of a fine. Furthermore, a positive beta weight indicates that a higher value on that predictor produces a higher probability of fining (as would be expected for indignation) whereas a negative beta weight indicates that a higher value of the predictor produces a lower probability of fining (as would be expected for actual contribution and intended contribution).

We estimate the parameters using Bayesian methods (Gelman et al., 2013; Kruschke, 2015, 2013; Kruschke, Aguinis & Joo, 2012; Ntzoufras, 2009). Bayesian estimation is especially seamless for complex hierarchical models such as the one used here, because it yields a complete posterior distribution of jointly credible parameter values, given the data. There is no need to compute p values from auxiliary sampling assumptions and null hypotheses. We use Markov chain Monte Carlo (MCMC) techniques programmed in R (R Core Team, 2014), JAGS (Plummer, 2003) and runjags (Denwood, 2013) to generate 20,000 representative credible values from the joint posterior distribution on the 649 parameters. The chains were burned in and checked for convergence, and run long enough to produce an effective sample size (ESS) of at least 10,000 for all of the reported results. This yields a stable and accurate representation of the posterior distribution on the parameters.

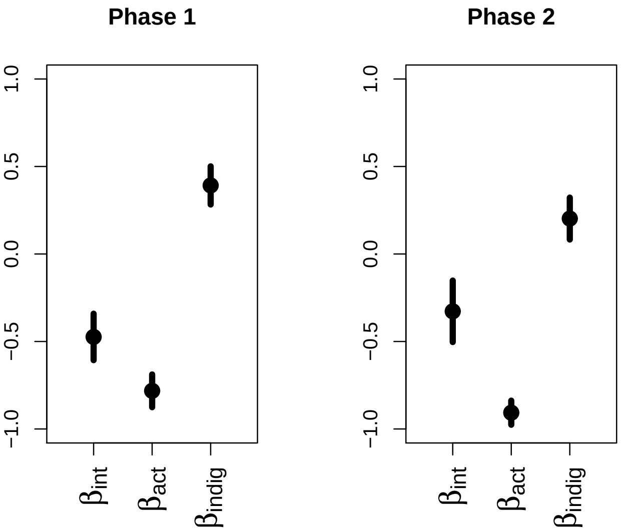
2.2.2 Parameter Estimates

We analyze the data of phase 1 (in which only the subject could apply a fine) separately from the data of phase 2 (in which all players could apply fines). Figure 1 shows the *95% highest density intervals* (HDIs) on the beta weights for each predictor in phases 1 and 2. The 95% HDI contains the 95% most probable parameter values, and is useful as a summary of the posterior distribution, along with the distribution's central tendency. The 95% HDI can also be used as part of a decision rule for rejecting or accepting a null value (Kruschke, 2015, 2013). The decision rule uses a *region of practical equivalence* (ROPE) around the null value, which indicates a band of values that are equivalent to the null for practical purposes. If the HDI falls completely outside the ROPE, the null value is rejected. We will say in this case that the parameter is "credibly" greater than or less than the null value. If the HDI falls completely inside the ROPE, the null value is accepted for practical purposes. In this article, we leave the ROPE tacit, recognizing that the bounds of practical equivalence are not crucial for our claims.

As expected, the regression weights on the intended contribution and actual contribution are negative, meaning

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 1: Parameter estimates from Experiment 1, showing marginal posterior distributions of the normalized group-level regression coefficients. The vertical black bars indicate the 95% highest density interval (HDI) which contains the most credible 95% of the values, with the point indicating the mean. In both phases, the regression weight on actual contribution is of greater magnitude (more negative) than the regression weight on intended contribution.



that the probability of punishing decreases as intended and actual contributions increase. The weight on indignation is positive, meaning that the probability of punishing increases as indignation increases. This positive weighting suggests that inequity aversion influences punishment in our novel PGG, analogous to previous results in different procedures (Cushman et al., 2009).

We are most interested in the relative weights of intended contribution and actual contribution. It is evident from Figure 1 that the regression weight on actual contribution is of greater magnitude (i.e., more negative) than the regression weight on intended contribution. To quantitatively assess the relative weights of these two predictors, we computed the difference of the regression weights at each step of the MCMC chain. In phase 1, the weight on actual contribution is larger than the weight on intended contribution (mean difference = 0.313, 95% HDI from 0.098 to 0.529), and in phase 2 this difference is even stronger (mean difference = 0.831, 95% HDI from 0.656 to 0.992). Thus, we have shown, for the first time in a trembling-hand PGG, that people deciding to fine weigh actual contributions more heavily than intended contributions.

The relative emphasis on actual contribution increases from phase 1 to phase 2. One possible reason is that subjects became more familiar with the task and increased their consistency of responding, allowing the trend to be more clearly expressed. A second possible reason is that behavior in phase 2 reflects mimicking of the automated players, who applied fines based on actual contribution. (See Appendix 3 for an analysis of behavior across rounds in Experiment 2 that is relevant to these possibilities.) Experiment 3 explores this latter possibility, and shows that, although mimicking may play a role in subjects' punishments, heavier weighting of actual contribution is maintained by people even when the automated players punish only on the basis of intended contributions.

3 Experiment 2: Emphasis on intended contribution in ostracism

Having established in Experiment 1 that there is emphasis on actual contribution when deciding to fine in a PGG, we investigated in Experiment 2 if that emphasis is also true when deciding to ostracize. As described previously, some researchers have argued that ostracism is more representative of the punishments employed in early human evolution (e.g., Baumard, 2010, 2011; Baumard et al., 2013). The theory posits that people choose their partners for cooperative ventures based on a potential partner's previous instances of cooperation or defection. This inexpensive strategy is sufficient to facilitate cooperation, and the alternative strategy involving costly direct punishment would not need to have evolved.

We hypothesized that decisions to ostracize would place more emphasis on intended contribution than decisions to fine, because we expected that subjects would be less willing to lose a well-intentioned partner from future rounds because of an accidental outcome. To test this hypothesis, we conducted an experiment very similar to Experiment 1 that included ostracism as a punishment option.

3.1 Methods

351 IU undergraduates participated in the experiment for course credit. Subjects were recruited via the IU Psychological and Brain Sciences human subject pool, with demographics as reported for Experiment 1.

As in Experiment 1, subjects played a two-phase public goods game with four automated opponents. The automated behavior was randomly selected from the same preexisting distribution used for Experiment 1 (Appendix 6). The contribution procedure was identical to Experiment 1. However, the punishment process had an important elaboration, such that punishment could consist of imposing a costly fine, as in Experiment 1, or ostracizing the player from the game at no cost to the punisher. Only one of these punishments could be imposed on any one player. If

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

an automated player was excluded, the player would be replaced by a new automated player on the next round. If the subject was excluded by the automated players, the subject would experience a 15-second time out while a message was displayed that described that the system was searching for a new game. (The instructions, however, did not mention the time out.) The subject would then be put into a new round with all new automated players. The exclusion did not change the total number of rounds played.

In the context of the trembling hand PGG, we refer to ostracism as "exclusion". We expected this term to convey more clearly the nature of this punishment to subjects, as ostracism might connote reputation effects that were not explicit in the game, and we expected that "exclusion" would be more familiar and easier to understand. Thus when referring to the ostracism punishment we refer to "exclusion" and when referring to theoretical results about punishment we refer to "ostracism".

As before, in phase 1 of the experiment (the first 22 rounds) only the subject could apply penalties, whereas in phase 2 of the experiment (the last 22 rounds) the automated players could also apply punishments. The automated players excluded a player if her intended contribution was at least 2 points lower than the mean intended contribution. If a player did not meet the exclusion criterion, the automated players applied costly fines using the same punishment rule as in Experiment 1.

3.2 Results

To analyze the punishment behavior in Experiment 2 we again use a Bayesian hierarchical model that predicts the probability of each punishment choice given the value of the three predictors: actual contribution, intended contribution, and indignation. However, now the analysis concerns a trinary choice, not a binary one. To handle the trinary choices, we use a conditional logistic regression that predicts two choice probabilities. The first is the probability of applying exclusion versus not applying exclusion. The second is the probability of applying a fine, given that no exclusion was applied. The analysis is a *conditional* logistic regression because this second probability is conditional on the first choice (exclusion) not occurring.[3]

3.2.1 Bayesian hierarchical conditional logistic regression

A detailed description of the model is available in Appendix 2. Again, the primary parameters of interest are the normalized group-level beta weights just as in Experiment 1 (see Appendix 5 for an analysis of individual behavior), but each of the three predictors now has two sets of beta weights. For the probability of exclusion, the three weights are denoted $\beta_{exc,act}$, $\beta_{exc,int}$, and $\beta_{exc,indig}$. The second group of beta weights predicts the probability of applying a fine, given that no exclusion occurred. These are denoted $\beta_{fine,act}$, $\beta_{fine,int}$, and $\beta_{fine,indig}$. These beta weights are interpreted just as before, but now each beta weight concerns both a specific predictor and a specific punishment. We again use MCMC techniques to generate 20,000 representative credible values from the joint posterior distribution on the 2,825 parameters in each phase (see Appendix 3 for an analysis of specific rounds). The effective sample size for all results reported below was at least 10,000.

3.2.2 Parameter Estimates

Figure 2 shows the 95% HDIs of the beta weights. As expected and as found in Experiment 1, the weights on the intended contribution and actual contribution are negative, and the weights on indignation are positive.

The two sides of each panel of Figure 2 show the weights for excluding and fining. Importanly, notice that for excluding the weight on the intended contribution is of greater magnitude (more negative) than the weight on the actual contribution. However, for fining, the opposite is true, as it was in Experiment 1. Thus, people emphasize intended contributions more than actual contributions when deciding to ostracize, but emphasize actual contributions more than intended contributions when deciding to fine. Quantitative analysis verifies the apparent differences in Figure 2, as detailed in the following paragraphs.

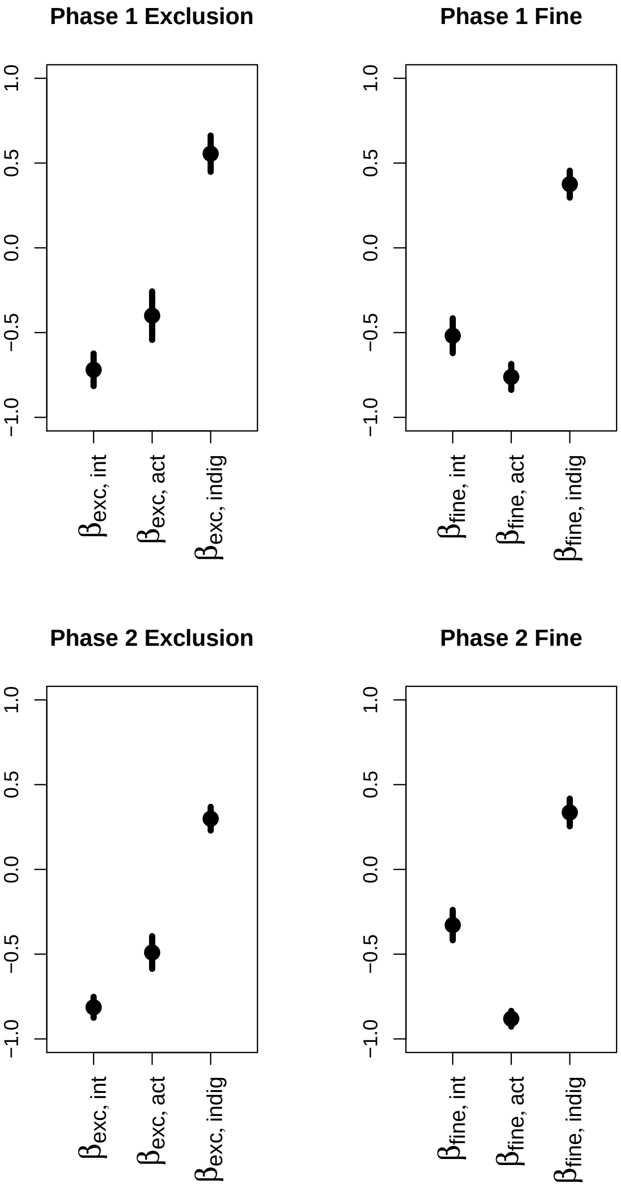
In phase 1, consider the weight on intended contribution, comparing across exclusion and fine (i.e., $\beta_{fine,int}$ versus $\beta_{exc,int}$): the mean difference is 0.202, with 95% HDI from 0.058 to 0.345. Consider the weight on actual contribution, comparing across exclusion and fine (i.e., $\beta_{fine,act}$ versus $\beta_{exc,act}$): the mean difference *in the opposite direction* is 0.365, with 95% HDI from 0.198 to 0.524. Focus now on the weights for exclusion (i.e., $\beta_{exc,int}$ versus $\beta_{exc,act}$): the magnitude of the weight on actual contribution is *less extreme* than the weight on intended contribution, with a mean difference of -0.320, 95% HDI from -0.539 to -0.113. Focusing on fines (i.e., $\beta_{fine,int}$ versus $\beta_{fine,act}$), the weight on actual contribution is *more extreme* than the weight on intended contribution, with a mean difference of 0.247, 95% HDI from 0.072 to 0.416. The same differences are even more pronounced in phase 2.

In phase 2, consider the weight on intended contribution, comparing across exclusion and fine: the mean difference is 0.491, with 95% HDI from 0.382 to 0.599. Con-

[3] A traditional analysis for n-ary choice data is multinomial logistic regression, which models the probabilities of all choices without conditionalizing on any one of them. We instead use conditional logistic regression because the multinomial model assumes the independence of irrelevant alternatives (Luce, 1959, 2008), which we do not have reason to believe applies to our data.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 2: Results from Experiment 2, showing 95% HDIs for the posterior distributions of beta weights for exclusion (left side of each panel) and fining (right side of each panel). Notice that for exclusion, the magnitude of the beta weight on intended contribution is larger (i.e., more negative) than on actual contribution, but for fining the opposite is true.



sider the weight on actual contribution, comparing across exclusion and fine: the mean difference, again in the opposite direction, is 0.392, with 95% HDI from 0.284 to 0.498. Focus now on the weights for exclusion: the magnitude of the weight on actual contribution is again less than the weight on intended contribution, with a mean difference of −0.326, 95% HDI from −0.480 to −0.171. Focusing on fines, the weight on actual contribution is again more than the weight on intended contribution, with a mean difference of 0.557, 95% HDI from 0.428 to 0.682.

These results verify again that actual contributions are weighed heavily when considering to punish by fining, but intended contributions are weighed heavily when considering to ostracize. However, as in Experiment 1, the trends appear to be stronger in phase 2 than in phase 1. As previously discussed, subjects may mimic the punishment behavior of the automated players in phase 2, as the automated players did focus on intention information for exclusion and actual contribution information for fining. However, it is important to note that, even in phase 1, when no automated-player penalties were occurring, the pattern that favored actual contribution for fines and intended contribution for exclusion was clearly present. In the next experiment we directly investigate the possible effect of mimicking.

4 Experiment 3: The effects of contribution control, punishment norms, and punishment efficacy

In our third experiment we addressed three issues. The first regards the effect of group norms on punishment behavior. Recall that in the second phases of the Experiments 1 and 2, trends became more pronounced. This change could have been due to familiarity with the paradigm and stabilization of response tendencies, or it could have been caused by subjects mimicking the punishment tendencies of the automated players. To test this second possibility, we introduced two new punishment rules for the automated players, and we randomly assigned subjects to experience one of the two rules. One rule based punishments only on actual contribution, ignoring intended contribution. Under this punishment rule, the automated players would exclude another player if her actual contribution was 2 or less, otherwise the player would be fined in an amount of how much her actual contribution was less than 8 (with a small amount of random noise applied). If the player contributed at least 8 points, no punishment was applied. The other rule based punishments only on the intended contribution, ignoring the actual contribution, using the same numerical criteria. If subjects mimic the behavior of other players, then subjects in the two groups should differently weigh actual and intended contributions.

The second issue concerned the effect of the player's control of her actual contribution on how she is punished. In Experiments 1 and 2, a player's actual contribution was created by randomly adding or subtracting between 2 and 4 points from the intended contribution. Thus, players

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

retained some control over their actual contributions, although it was imperfect control. One possible explanation for emphasis on actual outcome in fining is that punishers are not responding to the bad outcome *per se*, but to the player's failure to insure against the bad outcome by intentionally contributing high (as suggested by Young, Nichols & Saxe, 2010, regarding blame judgments). Because players had partial control over the outcome, punishers blame the player for not anticipating the possible bad outcome and insuring against it. A bad outcome directs attention towards this risky intentional behavior and also provides evidence that the behavior was indeed risky, and punishment follows. As this explanation posits a potential rational explanation for outcome emphasis in punishment (at least in the context of our experimental design) we will refer to it as the "rationalistic explanation" for outcome emphasis.

To test this possibility, we ran conditions with different types of control over the actual contribution. The first type of control we call "additive" and is very similar to the randomness used in the previous experiments: a random value between -4 and $+4$ was summed to the intended contribution to yield the actual contribution. The additive rule gives the player some control over the actual contribution. The other type of control is called "random". Under the random procedure, fifty percent of the time the actual contribution was exactly the intended contribution, and the other fifty percent of the time the actual contribution was merely a random draw from the range of possible contributions between 0 and 10. Thus, if the actual contribution did not match the intended contribution, it was clear that the player had no control whatsoever over the actual contribution. Subjects were instructed about the nature of their control over the actual contribution. The important feature of the random condition is that it is not possible to explain punishers' emphasis on actual contribution as a response to a lack of care, because, when an accidental low contribution occurs, the contributor could do nothing to prevent it. Therefore, if the rationalistic hypothesis explains outcome emphasis from the previous experiments, we would expect attenuated or non-existent outcome emphasis in the random condition.

The third issue concerned the effect of the punishment's efficacy in changing contribution behavior. Some researchers have argued that punishment behavior must necessarily be unresponsive to the efficacy of punishment in order to evolve as a stable strategy (Cushman & Macindoe, 2009). However, to our knowledge, responsiveness to efficacy of punishment has not been directly tested. In order to do so, the automated players in Experiment 3 were given two types of contribution patterns. The first type we call *punishment-responsive contributors*, who started with relatively high contributions (near 8 points), and reduced their intended contribution by 2 points per round unless a fine was applied, in which case they increased their intended contribution by 2 points. The second contributor type we call *unresponsive contributors*, who started with relatively low contributions (near 3 points) and maintained this low intended contribution consistently, regardless of fines. If efficacy matters to choice of punishment, then unresponsive contributors will be excluded more often and fined less often than responsive contributors, all else being equal.

4.1 Methods

258 IU undergraduates participated in the experiment for course credit. Subjects were recruited via the IU Psychological and Brain Sciences human subject pool, which has demographics as described previously. Subjects were randomly assigned to one of the two types of control over actual contribution (additive or random) and to one of the automated-player punishment rules (actual-contribution focused or intended-contribution focused), resulting in approximately 64 subjects per combination.

Except where noted here, all aspects of the design of Experiment 3 were identical to those of Experiment 2.

All subjects in all conditions started each game with two responsive contributors and two unresponsive contributors.

Because this experiment is directly interested in the influence of punishments norms, it did not include an initial phase during which only the subject punished. Instead, all players, including automated ones, were given the full range of punishment options throughout the entirety of the game, which lasted 30 rounds. In addition, Experiment 3 also involved several minor changes to increase the feeling of playing with real people. All players were labeled on screen with a random name (instead of a single letter). The names were drawn from the 500 most popular baby names, for males and females, at the United States Social Security baby name data base (http://www.ssa.gov/oact/babynames/). The automated-player contribution and punishment choices had realistic timers before they were displayed on screen, such that they appeared in an asynchronous cascade after the subject entered her intended contribution or punishment.

4.2 Results

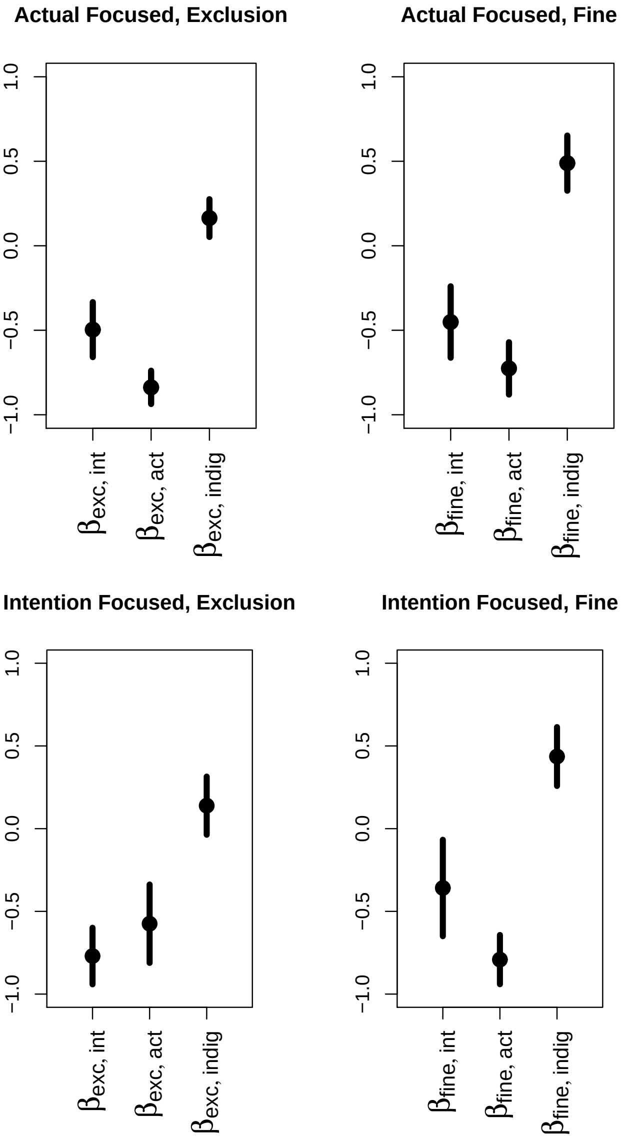
We use Bayesian conditional logistic regression as in Experiment 2. We devote a separate analysis and discussion to each of the three main research questions.

4.2.1 Punishment Norms

To assess the effect of the automated-player punishment norms, we performed a separate conditional logistic re-

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 3: Results of Experiment 3 for automated players punishing on the basis of actual contribution (top panel) or on the basis of intended contribution (bottom panel). Notice that the beta weights for fining (right side of each panel) are similar across the conditions, but the beta weights for excluding (left side of each panel) are different across the two conditions. The exclusion decisions in the actual-focused condition shows more emphasis on actual contribution than in the intention-focused condition.



gression analysis on each of the two punishment-norm groups. Because it takes exposure to several examples to experience and learn the punishment norms of the other players, we exclude the initial 10 rounds from the analysis, using the remaining 20 rounds. (See Appendix 3 for an analysis of behavioral change across rounds.)

Figure 3 plots the beta weights for the groups who experienced automated players punishing on the basis of actual or intended contribution. Notice that the beta weights for fining are similar across the conditions, but the beta weights for excluding are different across the two conditions. The exclusion decisions in the actual-focused condition shows more emphasis on actual outcome than in the intention-focused condition.

To quantitatively assess differences in the beta weights on intended contribution and actual contribution, we subtracted each weight in the intention-focused condition from its corresponding weight in the actual-focused condition. For fining, there was no major difference in the weight on intended contribution (mean difference = −0.085, 95% HDI from −0.457 to 0.271) or on actual contribution (mean difference = 0.071, 95% HDI from −0.156 to 0.287). In contrast, for excluding there was a difference in weights across the two conditions for both intended contribution (mean difference = 0.282, 95% HDI from 0.028 to 0.518) and actual contribution (mean difference = −0.263, 95% HDI from −0.528 to −0.011).

When these results are compared to the first phase of Experiment 2 (see Figure 2), where there was no automated player to mimic, a clear pattern emerges. First, fining consistently emphasizes actual contributions, regardless of the punishment norms of the other players. Second, excluding seems to emphasize intended contribution by default, but can be changed to mimic the punishment norms of the group.

4.2.2 Effect of Noise Rule

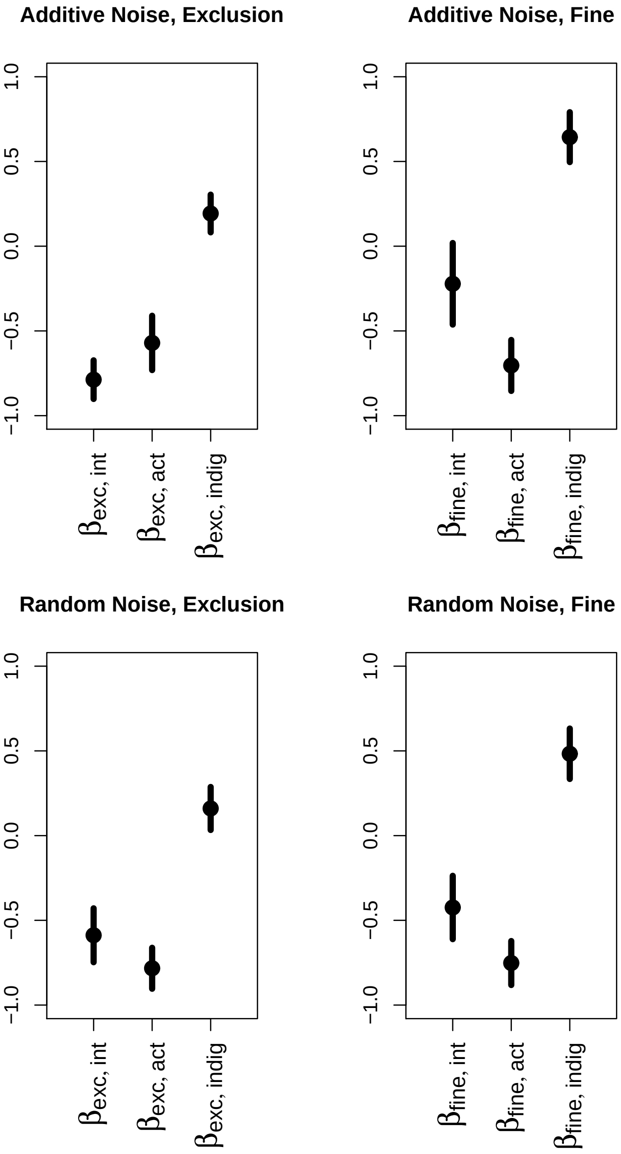
We conducted separate Bayesian conditional logistic regressions for the two noise-rule conditions. The noise-rule was explicitly described in the instructions to subjects, and thus did not have to be learned. On this rationale we included all 30 rounds in the analysis.

Figure 4 shows the group-level beta weights for the additive- and random-noise groups. In comparing weights for fining across the two conditions, it is clear that there is no major difference. In particular, there is no notable trend toward a reduced weighing of actual contribution in the random-noise condition. This is not consistent with the hypothesis that punishers are blaming failure to insure against accidental loss. In comparing weights for excluding across the two conditions, again there is no trend toward reduced weighting of actual contribution in the random-noise condition; indeed there is a weak trend in the opposite direction.

Quantitative comparisons of the regression weights show the following. For exclusion, the magnitude of the weight on intended contribution was greater for additive noise than random noise (mean difference = 0.206, 95% HDI from 0.009 to 0.407) and the magnitude of the weight on actual contribution was less for additive

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 4: Experiment 3 results for subjects with additive noise on their contributions (top panel) or for subjects with random noise (bottom panel). Notice that the beta weights for fining (right side of each panel) are similar across the conditions, but the beta weights for excluding (left side of each panel) are different across the two conditions. The exclusion decisions in the additive noise condition demonstrate the usual pattern favoring intention, but in the random noise condition demonstrate a comparatively greater weight on actual contribution and a comparatively lesser weight on intended contribution.



noise than random noise (mean difference = −0.216, 95% HDI from −0.415 to −0.010). In contrast, for fining we found no credibly non-zero difference for intended contribution (mean difference = −0.192, 95% HDI from −0.501 to 0.113) and no credibly non-zero difference for actual contribution (mean difference = −0.046, 95% HDI from −0.251 to 0.154).

Thus we find no evidence of a reduction in outcome emphasis under the random noise rule. On the contrary, for exclusion we find greater weight on actual contribution (and less on intended contribution) in the random noise condition. This result goes against the rationalistic explanation for outcome emphasis in the trembling-hand PGG. But what might produce the trend opposing the predictions of the rationalistic explanation? One possibility is that inconsequential intentions in the random noise condition tended to be ignored in favor of other sources of information such as the actual contribution. For random noise, when the intended and actual contributions differ from one another, subjects know that the intended contribution had no effect on the final outcome of the round, causing subjects to discount its importance and exhibit more outcome emphasis. This effect is more pronounced in exclusion because, under usual conditions of at least partial control, exclusion is strongly influenced by the intended contribution. Another possibility is that, despite instructions and experiences to the contrary, subjects assume that something could have been done to prevent a low actual contribution. This hypothesis is difficult to rule out, but if it is impossible to prevent individuals from punishing a lack of care even in cases where it is completely clear that no lack of care occurred, the rationalistic explanation seems to become less of an explanation for outcome emphasis but more a re-description of it. In any case, both of these hypotheses require further research in order to determine which, if either, is occurring.

4.2.3 Efficacy

We again performed two separate conditional logistic regressions; one on all the punishment choices in which the target was a responsive contributor, and one on all the punishment choices in which the target was an unresponsive contributor. Because the presence of two contributor types had to be learned by subjects we included only the final 20 rounds in the analysis.

We predict that responsive contributors will be more likely to be fined but less likely to be excluded than unresponsive contributors, *given equal values of the predictors: intended contribution, actual contribution, and indignation*. This hypothesis is agnostic about the relative weights of the predictors, as it concerns only the relative propensities to apply a fine and to apply exclusion across the two types of contributor. To assess this prediction, we took all of the actually occurring combinations of predictor values (intended contribution, actual contribution, and indignation) and computed the propensities to apply exclusion and fine predicted by the two regressions. There

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Table 1: Selected predictor values and the corresponding 95% HDIs on the difference between the responsive contributors and unresponsive contributors in probability of exclusion and probability of fine. A positive difference means the probability is higher for responsive contributors than for unresponsive contributors.

Predictor Values			Δ Excluding Prob. $P(Exclude)_{\text{responsive}}$ $-P(Exclude)_{\text{unresponsive}}$	Δ Fining Prob. $P(Fine\|\neg Exclude)_{\text{responsive}}$ $-P(Fine\|\neg Exclude)_{\text{unresponsive}}$
Intended	Actual	Indignation	95% HDI	95% HDI
1	2	5	−0.34 to −0.06	0.04 to 0.36
4	1	9	−0.31 to −0.04	0.02 to 0.32
2	2	3	−0.25 to −0.03	0.05 to 0.37
1	1	7	−0.37 to −0.04	0.04 to 0.32
1	2	8	−0.40 to −0.09	0.02 to 0.31

were 20,640 such predictor combinations. At each step of the MCMC chain, we used the parameter estimates at that step along with the values of the predictors to compute posterior predicted probability of exclusion, $P(Exclude)$, and probability of fine given there was not exclusion, $P(Fine|\neg Exclude)$. We then computed the difference of the predicted $P(Exclude)$ and $P(Fine|\neg Exclude)$ values across the two contributor types.[4]

Table 1 displays some predictor sets selected to illustrate the differences between responsive contributors and unresponsive contributors. Consider, for example, the bottom row of Table 1, which indicates a player for whom the intended contribution was 1, the actual contribution was 2, and the indignation was 8. The regression analyses reveal that the probability of excluding that player was about 25 percentage points less if that player was a responsive contributor than if that player was an unresponsive contributor. The 95% HDI on the difference extends from -0.40 to -0.09 (as shown in the table). The probability of fining that player was about about 17 percentage points more if that player was a responsive contributor than if that player was a unresponsive contributor. The 95% HDI on the difference extended from $+0.02$ to $+0.31$ (as shown in the table).

In this set of analyses, we utilized a ROPE of $\pm .02$. This means that in order for us to consider two probabilities to be credibly different, the 95% HDI on their difference must be entirely less than $-.02$ or greater than .02. Using this criterion, we found that, for 26% of all the predictor sets, the responsive contributors were credibly less likely to be excluded than unresponsive contributors. No predictor sets showed a credible difference in the opposite direction. Furthermore, 99% of the predictor sets had a mean difference favoring exclusion of unresponsive contributors. In 15% of the predictor sets, responsive contributors were credibly more likely to be fined than unresponsive contributors, and no predictor sets showed a credible opposite trend. 51% of the predictor sets had a mean difference favoring fining of responsive contributors.

These results suggest that subjects are sensitive to the efficacy of their punishments, because they punish responsive contributors differently from unresponsive contributors. Players who improved their contributions in response to being fined were not excluded from the game. By contrast, players who maintained their low contribution even after fines were excluded from the game. Thus, punishment was not just an automatic response to freeloading but took into account the potential benefit that could be expected from applying different types of punishment.

5 General Discussion

In a series of experiments we found that outcome emphasis, while robust in costly fines, is not present in ostracism. Furthermore, we found that the norms used by the group affect ostracism but not fines, that outcome emphasis occurs for fines even for uncontrollable outcomes, and that subjects choose punishments that are most useful given the responsiveness of the individual being punished.

5.1 Ostracism, outcome bias, and outcome emphasis

We have found that outcome emphasis, that is, weighing actual contribution more than intended contribution, is not present in ostracism. However, throughout all of our experiments, actual contribution has been a non-zero predic-

[4] The average effective sample size (ESS) of the MCMC chain was 8,710 for the estimate of the difference in $P(Exclude)$ across conditions and 5,614 for the estimate of the difference in $P(Fine|\neg Exclude)$ across conditions.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

tor of punishment, even after taking intended contribution and indignation into account in the regression analyses; that is, outcome *bias* is present in all forms of punishment. And for costly fines outcome emphasis is present under all of the manipulations presented here.

This raises two issues for further investigation. First, why does ostracism demonstrate intention emphasis whereas fining does not? As discussed previously, it could be due to an unwillingness to lose access to a well-intentioned cooperation partner. However, this explanation does not provide a reason for why the relative weight on intention is susceptible to changes in the controllability and the group norms of punishment. It may be that ostracism is a more flexible mode of punishment for adapting to the needs of the social environment, whereas applying costly fine is comparatively a more stable vengeance for personal injury.

5.2 Automation and our results

In the introduction we summarized literature suggesting that the use of automated players is unlikely to strongly alter subjects' behavior. Our results are consistent with the hypothesis that subjects were reacting to the automated players as if they were human. If the subjects were treating the automated players merely as unfeeling computer-generated numbers that should be handled in whatever way maximizes personal points, then it is difficult to explain why players should mimick the punishment behavior of the automated players in Experiment 3, or why players should administer any costly fines at all in Experiment 1.

Nevertheless, it could be valuable to pursue trembling-hand PGGs with groups of human players. One follow-up could be a simple replication with human players, in which all but one are confederates of the experimenter who are trained to contribute and punish according the automated rules. We would expect results like those we reported here. Another follow-up could involve all naive subjects, with the goal being to investigate the contribution and punishment norms that spontaneously arise.

5.3 Efficacy

We have presented evidence that individuals are sensitive to the efficacy of their punishments when deciding what punishments to apply. Recall that Cushman (2013) argued that punishment behavior that is sensitive to efficacy is easily exploitable by cheaters who ignore punishment, and these arguments were supported by results from evolutionary simulations (Cushman & Macindoe, 2009). While this claim seems at odds with the results presented here, it might be attributable to the authors' assumptions that the only choice available to punishers is to apply a punishment or not. However, in the present experiment and many real world interactions, there is a range of potential punishment responses. Moreover, ostracism is a free or inexpensive method by which to avoid being cheated in the future, thus preventing exploitation by unresponsive cheaters. Thus the arguments and simulations presented by Cushman and colleagues do not apply to this situation. It is likely that there are some real world scenarios where ostracism is an option and other scenarios that more closely match the conditions set out by Cushman and colleagues. Future research is necessary to assess these conditions, and the degree to which individuals adjust their punishment behavior based on these differing environmental factors. Moreover, evolutionary models of punishment behavior must be expanded to take into account the possibility of alternative punishments like ostracism.

5.4 Indignation and inequity aversion

In all of our analyses the regression weight on indignation has been credibly non-zero for both fining and ostracism. This suggests the some form of inequity aversion plays a role in all punishment decisions in a PGG, and this role is relatively insensitive to the factors manipulated in the experiments presented here. The importance of inequity aversion as a motivation is consistent with results from previous work (e.g., Cushman et al., 2009; Fehr & Schmidt, 1999). This inequity motivation demands explanation, especially in the case of ostracism, which does not restore equity in a direct sense. One potential explanation for this persistent role of indignation is that inequity captures attention and causes individuals to immediately consider whether punishment is necessary, after which other information (such as the intended and actual contribution) attenuates or exacerbates the initial impulse to punish.

5.5 Other forms of punishment

Everyday punishment take many forms other than those discussed here, including reputation damage, inflicting emotional pain, inflicting physical pain, and probably many others. Future research is needed to determine the relations of these punishment modalities and what environmental factors elicit them.

References

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of personality and social psychology, 54*(4), 569.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior, 27*(5), 325–344.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Baumard, N. (2010). Has punishment played a role in the evolution of cooperation? A critical review. *Mind & Society, 9*(2), 171–192. http://dx.doi.org/10.1007/s11299-010-0079-9.

Baumard, N. (2011). Punishment is not a group adaptation. *Mind & Society, 10*(1), 1–26. http://dx.doi.org/10.1007/s11299-010-0080-3.

Baumard, N., André, J.-B., & Sperber, D. (2013). A mutualistic approach to morality: the evolution of fairness by partner choice.*The Behavioral and Brain Sciences, 36*(1), 59–78. http://dx.doi.org/10.1017/S0140525X11002202.

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review, 90*(1), 166–193.

Book, A. S. (1999). Shame on you: An analysis of modern shame punishment as an alternative to incarceration. *William & Mary Law Review, 40*(2), 653–686.

Carpenter, J., & Matthews, P. H. (2009). What norms trigger punishment? *Experimental Economics, 12*(3), 272–288. http://dx.doi.org/10.1007/s10683-009-9214-z.

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters, 86*(3), 393–398. http://dx.doi.org/10.1016/j.econlet.2004.08.007.

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics, 89*(8), 1421–1435. http://dx.doi.org/10.1016/j.jpubeco.2004.05.011.

Cook, K., & Hegtvedt, K. (1983). Distributive justice, equity, and equality. *Annual Review of Sociology, 9*(1983), 217–241.

Cushman, F. (2013). The role of learning in punishment, prosociality, and human uniqueness. In K. Sterelny, R. Joyce, B. Calcott, & B. Fraser (Eds.), *Cooperation and its evolution (Life and mind: Philosophical issues in biology and psychology)* 333–372. Cambridge, MA: MIT Press.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS one, 4*(8), e6699. http://dx.doi.org/10.1371/journal.pone.0006699.

Cushman, F., & Macindoe, O. (2009). The Coevolution of Punishment and Prosociality Among Learning Agents. In *Proceedings of the 31st annual conference of the cognitive science society.*

Denwood, M. J. (2013). runjags: An R Package Providing Interface Utilities, Distributed Computing Methods and Additional Distributions For MCMC Models in JAGS. *Journal of Statistical Software.*

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87. http://dx.doi.org/10.1016/S1090-5138(04)00005-4.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review, 90*(4), 980–994.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.

Flynn, C. (1994). Regional differences in attitudes toward corporal punishment. *Journal of Marriage and the Family, 56*(2), 314–324.

Fogg, B. J., & Nass, C. (1997). How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI'97 extended abstracts on Human factors in computing systems*, 331–332.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). Chapman and Hall/CRC Texts in Statistical Science.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. Cognition, 115(1), 166–171.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362–1367.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences, 28*(06), 795–815.

Jacobs, D., & Carmichael, J. (2002). The political sociology of the death penalty: A pooled time-series analysis. *American Sociological Review, 67,* 109–131.

Jones, J. M. (2013). U.S. death penalty support lowest in more than 40 years. *Gallup Politics.*

Kahan, D. M. (1996). What do alternative sanctions mean? *University of Chicago Law Review, 63*(2), 591–653.

Kahan, D. M. (2006). What's really wrong with shaming sanctions. *Yale Law School Legal Scholarship Repository Faculty Scholarship Series, 102.*

Kiesler, S., Sproull, L., & Waters, K. (1996). A prisoner's dilemma experiment on cooperation with people and human-like computers. Journal of Personality and Social Psychology, 70(1), 47.

Kruschke, J. K. (2015). *Doing Bayesian data analysis, Second edition: A tutorial with R, JAGS, and Stan.* Waltham, MA: Academic Press / Elsevier.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology. General, 142*(2), 573–603. http://dx.doi.org/10.1037/a0029146.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods, 15*(4), 722–752. http://dx.doi.org/10.1177/1094428112457829.

Lansford, J. E., & Dodge, K. A. (2008). Cultural norms for adult corporal punishment of children and societal rates of endorsement and use of violence. *Parenting: Science and Practice, 8*(3), 1–11. http://dx.doi.org/10.1080/15295190802204843.Cultural.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Luce, R. D. (2008). Luce's choice axiom. *Scholarpedia, 3*(12), 8077. (Revision # 121550)

Maier-Rigaud, F. P., Martinsson, P., & Staffero, G. (2010). Ostracism and the provision of a public good: experimental evidence. *Journal of Economic Behavior & Organization, 73*(3), 387–395. http://dx.doi.org/10.1016/j.jebo.2009.11.001.

Masclet, D. (2003). Ostracism in work teams: a public good experiment. *International Journal of Manpower, 24*(7), 867–887. http://dx.doi.org/10.1108/01437720310502177.

Ntzoufras, I. (2009). Bayesian modeling using WinBUGS. Wiley.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates?. *International Journal of Human-Computer Studies, 45*(6), 669–678.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.

Ostrom, E., Walker, J. M., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *The American Political Science Review, 86*(2), 404–417.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing.*

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Raihani, N. J., & McAuliffe, K. (2012). Human punishment is motivated by inequity aversion, not a desire for reciprocity. *Biology Letters(July)*, 18–21. http://dx.doi.org/10.1098/rsbl.2012.0470.

Raihani, N. J., Thornton, A., & Bshary, R. (2012). Punishment and cooperation in nature. *Trends in Ecology & Evolution, 27*(5), 288–295. http://dx.doi.org/10.1016/j.tree.2011.12.004.

Rand, D. G., Fudenberg, D., & Dreber, A. (2013). It's the thought that counts: The role of intentions in reciprocal altruism. *Available at SSRN 2259407.*

Sasaki, T., & Uchida, S. (2013). The evolution of cooperation by social exclusion. *Proceedings. Biological sciences / The Royal Society, 280*(1752), 20122498. http://dx.doi.org/10.1098/rspb.2012.2498.

Stevens, J. R., Cushman, F., & Hauser, M. D. (2005). Evolving the Psychological Mechanisms for Cooperation. *Annual Review of Ecology, Evolution, and Systematics, 36*(1), 499–518. http://dx.doi.org/10.1146/annurev.ecolsys.36.113004.083814.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS One, 6*(3), e16836.

Whitman, J. Q. (1998). What is wrong with inflicting shame sanctions? *Yale Law School Legal Scholarship Repository Faculty Scholarship Series, 655.*

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology, 51*(1), 110–116.

Yamagishi, T., & Horita, Y. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Science, 106*(28), 11520–11523.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Review of Philosophy and Psychology, 1*(3), 333–349. http://dx.doi.org/10.1007/s13164-010-0027-y.

Zeisel, H., & Gallup, A. (1989). Death penalty sentiment in the United States. *Journal of Quantitative Criminology, 5*(3), 285–296.

Appendix 1: Details of the model for Experiment 1

The application of a fine by player i to player j is denoted $y^{[i,j]}$ and is an integer from the set $\{1, 2\}$ where 1 denotes no penalty was applied, and 2 denotes that a fine was applied. It should be noted that this analysis models the propensity to apply a fine, not the *amount* of the fine. (See Appendix 4 for an analysis that incorporates the amount of fine with analogous results.) The model describes the action, $y^{[i,j]}$, as a random draw from a categorical distribution (i.e., multinomial distribution with $N = 1$) with category probabilities $\pi^{[i,j]}_{none}$ and $\pi^{[i,j]}_{fine}$ (with the constraint that $\pi^{[i,j]}_{none} = 1 - \pi^{[i,j]}_{fine}$), which is denoted

$$y^{[i,j]} \sim \text{cat}\left(\pi^{[i,j]}_{none}, \pi^{[i,j]}_{fine}\right) \tag{1}$$

where the symbol "∼" is read "is distributed as," and where cat indicates a categorical distribution.

The model uses three predictors. One predictor is the intended contribution by player j, denoted $x^{[j]}_{int}$. Intuitively, as the intended contribution increases, the probability of punishing the player should decrease. The second predictor is the player's actual contribution, denoted $x_{act}^{[j]}$. Intuitively, as the actual contribution increases, the probability of punishing the player should decrease.

The third predictor is what we call the "indignation" of subject i toward player j, which is the net gain of player j minus the net gain of subject i, denoted $x_{indig}^{[i,j]}$. Intuitively, as indignation increases, the probability of punishment might increase. Indignation is included as a predictor to reflect inequity aversion, and because the net gain was explicitly displayed on the screen along with intended and actual contributions. As described in the introduction, previous experiments have demonstrated inequity aversion in punishment. By including a separate predictor for inequity aversion, the independent influences of actual and intended contributions can be better assayed.

We used a standard logistic regression model for describing each individual player. For each subject i, we compute a weighted combination of the predictors for the underlying tendency to apply a fine to a player j, denoted $\lambda_{fine}^{[i,j]}$:

$$\begin{aligned}\lambda_{fine}^{[i,j]} = \beta_0^{[i]} & \\ &+ \beta_{int}^{[i]}\left(x_{int}^{[j]} - \overline{x}_{int}\right)\\ &+ \beta_{act}^{[i]}\left(x_{act}^{[j]} - \overline{x}_{act}\right)\\ &+ \beta_{indig}^{[i]}\left(x_{indig}^{[i,j]} - \overline{x}_{indig}\right)\end{aligned} \tag{2}$$

Equation 2 shows that the predictors were mean-centered by subtracting their overall means across all trials and players. This mean centering makes the intercept, $\beta_0^{[i]}$, better interpretable as baseline behavior at the mean of the predictors, and makes shrinkage from the hierarchical model (to be described below) apply at the mean instead of at zero.

The underlying tendency to apply a fine is converted to a probability via the conventional logistic function:

$$\phi_{fine}^{[i,j]} = 1 \Big/ \left(1 + \exp(-\lambda_{fine}^{[i,j]})\right) \tag{3}$$

To produce the final probability of applying a fine, the model accounts for "oops" errors by mixing the probability of Equation 3 with a random-choice probability of 1/2, using a mixing coefficient α:

$$\pi_{fine}^{[i,j]} = \alpha\,(1/2) + (1-\alpha)\left(\phi_{fine}^{[i,j]}\right) \tag{4}$$

Effectively, Equation 4 makes the logistic function have asymptotes at $\alpha/2$ and $1-\alpha/2$ instead of at 0 and 1. It is worth noting that estimates of the guessing rate α were quite small, with typical values not exceeding 0.009 in any analysis. Nevertheless, including a non-zero α is important to account for rare outlying responses that could otherwise force the regression coefficients to be artificially small in magnitude.

We use a hierarchical model in which individual $\beta^{[i]}$ coefficients are assumed to come from higher-level distributions that describe group-level tendencies. Each individual's coefficients are assumed to be t-distributed across the group:

$$\begin{aligned}\beta_0^{[i]} &\sim \mathrm{t}(\mu_0, \tau_0, \nu{=}5)\\ \beta_{int}^{[i]} &\sim \mathrm{t}(\mu_{int}, \tau_{int}, \nu{=}5)\\ \beta_{act}^{[i]} &\sim \mathrm{t}(\mu_{act}, \tau_{act}, \nu{=}5)\\ \beta_{indig}^{[i]} &\sim \mathrm{t}(\mu_{indig}, \tau_{indig}, \nu{=}5)\end{aligned} \tag{5}$$

where τ is the precision (reciprocal of squared scale) of the t-distribution, and where ν is the normality of the distribution, often referred to as the degrees-of-freedom parameter. Preliminary analyses indicated considerable unsystematic outliers in the individual-level predictor coefficients. Therefore we choose the relatively low value of 5 for ν to allow the group-level coefficients to be robust against individual outliers. The use of t distributions to accommodate outliers is routine in statistical modeling (e.g., Kruschke, 2013).

The primary focus of the analysis is on the group-level means of the regression coefficients in Equation 5. The estimate of μ_{int}, for example, is the group-level mean value for $\beta_{int}^{[i]}$, which is the weight placed on the intended contribution for applying a fine. It should be noted that in all the figures we plot a normalized reparameterization of the regression weights according to the following formula, where $pred$ is a placeholder for int, act, or $indig$:

$$\mathrm{Normalized}(\mu_{pred}) = \frac{\mu_{pred}}{\sqrt{\mu_{int}^2 + \mu_{act}^2 + \mu_{indig}^2}} \tag{6}$$

The normalization across predictors is reasonable because the scales of the three predictors are the same: monetary points. The normalized regression weights represent the values of the raw regression weights relative to one another. This allows easier comparison across regression weights, and in later experiments across conditions. We refer to the normalized group-level μ_{pred} parameters as "beta weights" because they denote the typical values of the coefficients in Equation 2.

The hierarchical structure of the model rationally imposes shrinkage on the individual estimates. The estimate of each $\beta_{pred}^{[i]}$ is influenced by subject i's responses and by the estimates of the higher-level μ_{pred} and τ_{pred} parameters. The higher-level parameters are influenced by data from all subjects, hence each individual's estimate is a compromise between the individual's data and the typical

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

group data. Hierarchical models are an especially useful way to estimate group-level tendencies, without assuming that all individuals have identical behavior, and without assuming that all individuals are mutually uninformative (e.g., Gelman & Hill, 2007; Kruschke, 2015). It is important to note that due to the mean centering of the predictors (see Eqn. 2) the intercept expresses baseline behavior at the mean values of the predictors and thus shrinkage applies to the mean-centered intercepts and slopes. This makes the shrinkage more meaningful than applying it to intercepts located arbitrarily at zero monetary points, which for the actual and intended contribution predictors essentially never occurred in the experiment.

We establish vague, noncommittal prior distributions for the means and precisions of the group distributions:

$$\begin{aligned}\mu_{pred} &\sim \text{normal}(0, 1e-10)\\ \tau_{pred} &\sim \text{gamma}(1.10512, 0.010512)\end{aligned}$$

where the shape and rate constants in the gamma distribution give it a mode of 10 and a standard deviation of 100. These broad prior distributions imply that the prior has minimal influence on the posterior distribution. The α parameter also had a noncommittal prior, $\alpha \sim$ uniform$(0, .1)$.

Appendix 2: Details of the model for Experiments 2 and 3

As before, we denote the punishment applied by subject i to player j as $y^{[i,j]}$, which is now an integer from the set $\{1, 2, 3\}$ where 1 indicates no punishment, 2 indicates a fine, and 3 indicates exclusion. The model assumes that $y^{[i,j]}$ can be described as a random draw from a categorical distribution:

$$y^{[i,j]} \sim \text{cat}\left(\pi^{[i,j]}_{none}, \pi^{[i,j]}_{fine}, \pi^{[i,j]}_{exc}\right) \tag{7}$$

We use the same predictors and logistic function as in the analysis of Experiment 1. Thus, we treat the underlying tendency for subject i to apply exclusion to player j, $\lambda^{[i,j]}_{exc}$ as a weighted combination of the predictors:

$$\begin{aligned}\lambda^{[i,j]}_{exc} = \beta^{[i]}_{exc,0} &+ \beta^{[i]}_{exc,int}\left(x^{[j]}_{int} - \overline{x}_{int}\right)\\ &+ \beta^{[i]}_{exc,act}\left(x^{[j]}_{act} - \overline{x}_{act}\right)\\ &+ \beta^{[i]}_{exc,indig}\left(x^{[i,j]}_{indig} - \overline{x}_{indig}\right)\end{aligned} \tag{8}$$

Furthermore, $\lambda^{[i,j]}_{fine}$ is the underlying tendency for subject i to apply a fine to player j, given that subject i did not exclude player j:

$$\begin{aligned}\lambda^{[i,j]}_{fine} = \beta^{[i]}_{fine,0} &+ \beta^{[i]}_{fine,int}\left(x^{[j]}_{int} - \overline{x}_{int}\right)\\ &+ \beta^{[i]}_{fine,act}\left(x^{[j]}_{act} - \overline{x}_{act}\right)\\ &+ \beta^{[i]}_{fine,indig}\left(x^{[i,j]}_{indig} - \overline{x}_{indig}\right)\end{aligned} \tag{9}$$

These underlying tendencies are converted to choice probabilities as follows:

$$\begin{aligned}\phi^{[i,j]}_{exc} &= 1 \Big/ \left(1 + \exp(-\lambda^{[i,j]}_{exc})\right)\\ \phi^{[i,j]}_{fine} &= \left[1 \Big/ \left(1 + \exp(-\lambda^{[i,j]}_{fine})\right)\right]\left(1 - \phi^{[i,j]}_{exc}\right)\\ \phi^{[i,j]}_{none} &= 1 - (\phi^{[i,j]}_{fine} + \phi^{[i,j]}_{exc})\end{aligned} \tag{10}$$

The conversion to choice probabilities in Equation 10 is what makes the model *conditional* logistic regression, because the probability of fining is the logistic of the fining tendency multiplied by the probability of not excluding. It should also be noted that due to the way $\phi^{[i,j]}_{exc}$ and $\phi^{[i,j]}_{fine}$ are defined, $\phi^{[i,j]}_{none}$ cannot be less than zero. The regression coefficients of Equations 8 and 9 are estimated using the conditional probabilities of Equation 10.

As in the analysis in Experiment 1, the logistic probabilities of Equation 10 are mixed with random choices (1/3) to accommodate occasional off-task responses or "oops" errors:

$$\pi^{[i,j]}_{action} = \alpha\,(1/3) + (1-\alpha)\left(\phi^{[i,j]}_{action}\right) \tag{11}$$

The probabilities of Equation 11 are used to model the trinary choices in Equation 7.

In summary, for each individual we have two sets of beta weights, one set describing the propensity to apply exclusion, and the other set describing the propensity to apply a fine given that exclusion was not applied.

We again use a hierarchical model in which individual beta coefficients are assumed to come from higher-level distributions that describe group-level tendencies. Each individual's coefficient is assumed to be t distributed across the group:

$$\beta^{[i]}_{pen,pred} \sim \text{t}(\mu_{pen,pred}, \tau_{pen,pred}, \nu = 5) \tag{12}$$

where the subscript pen stands in for either of the two possible penalities (exclude or fine) and the subscript $pred$ stands in for any of the three predictors (intended contribution, actual contribution, or indignation). As in Experiment 1, the primary focus of the analysis is on the group-level means $\mu_{pen,pred}$ in Equation 12.

For the Bayesian estimation, we use the noncommittal prior distributions that were used for the analysis of Experiment 1. And, like the analysis of Experiment 1, we

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

use MCMC techniques to generate 20,000 representative credible values from the joint posterior distribution on the 2,825 parameters in each phase. Unless otherwise noted, the effective sample size for all results reported in the article was at least 10,000.

Appendix 3: Behavior across rounds

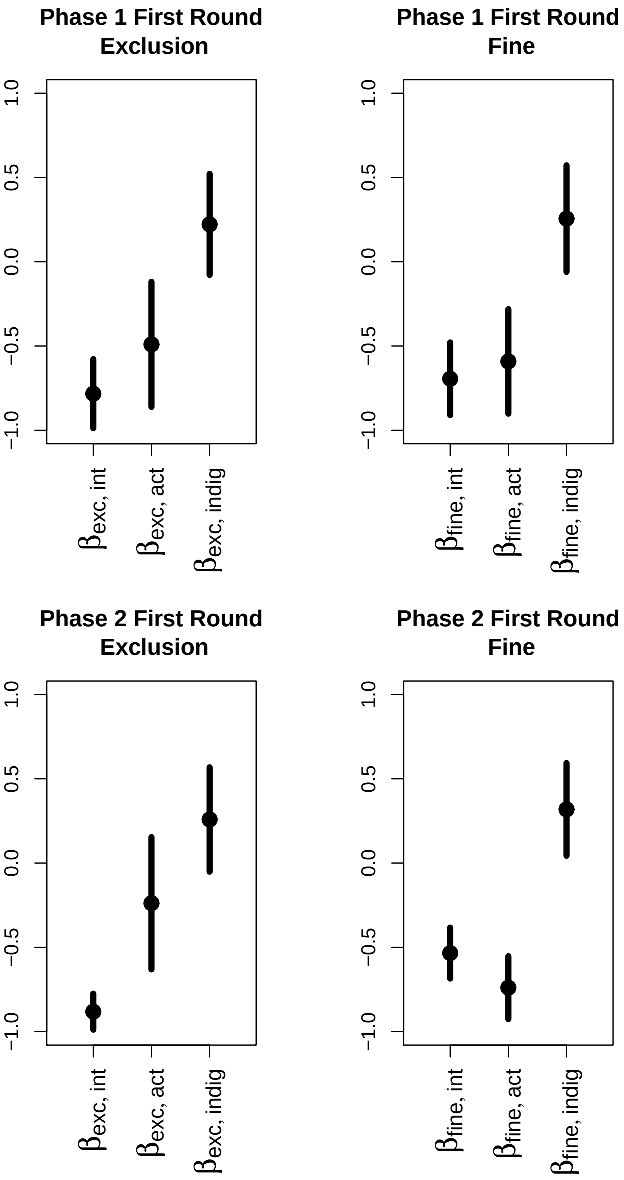
The analyses reported in the main body used the data from all of the rounds in a given block or condition (or the final 20 rounds in the case of our analyses of punishment norms and efficacy). In this section, we analyze behavior in smaller subsets of blocks to assess how behavior changes throughout the course of the experiment. Specifically, we investigate changes in the pattern of intended versus actual weights in Experiment 2 over time (with special attention to the role played by the automated player punishment scheme) and how the effect of group norm on player behavior changes early in Experiment 3 versus late in Experiment 3.

Behavior across rounds in Experiment 2

Here, we apply an identical Bayesian analysis to singular rounds in Experiment 2. Recall that in Experiment 2, the pattern observed (intention emphasis in exclusion and actual emphasis in fines) was stronger in phase 2 (with automated player punishment) than in phase 1 (without automated player punishment). This could be due to an increased familiarity with the task and thus more consistent behavior, or it may be that subjects were mimicking the automated players (which were also displayed this pattern). To rule out the hypothesis that the difference was entirely due to mimicking we investigate behavior in two singular rounds: the first round of phase 1, and the first round of phase 2. In particular, the behavior in the first round of phase 2 is of special interest as it represents the only round of the second phase where the subject has not yet seen the punishment behavior of the other players.

Figure 5 shows the 95% HDIs of the beta weights in the first round of phase 1 and the first round of phase 2.[5] Notice that in the first round, the weights on intended and actual contribution are similar to each other in magnitude for both exclusion and fine. A quantitative comparison of these parameters indicates no credible differences within any punishment type or across the two punishment types. However, in the first round of phase 2, the pattern of intended contribution emphasis for exclusion and actual contribution emphasis for fine more readily emerges.

Figure 5: Beta weights for the first round of phase 1 and the first round of phase 2. Notice that in phase 1 there is a large amount of overlap across the beta weights on intended and actual contribution. Compare to the first round of phase 2, where this overlap is lessened.



The weight on intention is smaller in fine than in exclusion (mean difference = −0.35, 95% HDI from −0.56 to −0.15) and the weight on actual contribution is larger in fine than in exclusion (mean difference = 0.55, 95% HDI from 0.07 to 1.04). Moreover, in exclusion the weight on actual contribution is smaller than the weight on intention (mean difference = −0.66, 95% HDI from −1.13 to −0.15) and in fine the opposite pattern is marginally present (mean difference = 0.24, 95% HDI from −0.10 to 0.55).

[5] The effective sample size of the chains for the comparisons reported here was low because of autocorrelation resulting from the small data set. The four chains were well overlapping, however, so we trust that the chains converged and report the parameter values to fewer decimal places.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Norm acquisition across rounds in Experiment 3

Experiment 3 provided some evidence that in the case of exclusion, subjects were mimicking the punishment rule used by the automated players. This hypothesis also entails that there would be little difference across conditions in earlier rounds, but that in later rounds the observed differences would emerge. Unfortunately, there are not sufficient data in a singular round for such comparisons to have a reasonable precision; in other words, the data in a single round lead to estimates that are very uncertain. Instead, we conducted two analyses, one using the first ten rounds, and one using the last ten rounds.

Figure 6 shows the 95% HDIs of the beta weights for the actual-focused and intended-focused conditions in the first ten and last ten rounds. To quantitatively assess any differences in the beta weights on intended contribution and actual contribution, we subtracted each weight in the intention-focused condition from its corresponding weight in the actual-focused condition.

In the first ten rounds, there were no credible differences in any of the parameters, matching the qualitative assessment. For fining, there was no major difference in the weight on intended contribution (mean difference = -0.181, 95% HDI from -0.564 to 0.192) or on actual contribution (mean difference = 0.141, 95% HDI from -0.198 to 0.464). For exclusion there was also no difference in weights across the two conditions for intended contribution (mean difference = -0.011, 95% HDI from -0.267 to 0.231) or actual contribution (mean difference = -0.027, 95% HDI from -0.359 to 0.291).

In contrast, the pattern of differences in the analysis of the last ten rounds more closely matched the analyses presented in the main body. For fining, there was again no difference in the weight on intended contribution (mean difference = -0.025, 95% HDI from -0.419 to 0.372) or on actual contribution (mean difference = 0.034, 95% HDI from -0.212 to 0.292). In exclusion, there was a difference in the weight on intended contribution (mean difference = 0.329, 95% HDI from 0.049 to 0.599) and a marginal but non-credible difference in the weight on actual contribution (mean difference = -0.25, 95% HDI from -0.569 to 0.043).

In summary, we found no credible differences between the two conditions in the first ten round, suggesting that the differences observed were not present in the groups before the exposure to the group norm, and the changes due to group norm took a somewhat significant exposure time for any behavior changes to be detectable. Analysis of the final ten rounds demonstrated a pattern more similar to the results presented in the main body, although the difference in the weight on actual contribution had reduced certainty because of the smaller dataset.

Appendix 4: Analysis of fine amount

The models presented in the main text use a logistic analysis of punishment choice. This is to allow a direct comparison across the analysis of fine (which can be treated as a metric value or nominal outcome) and the analysis of exclusion (which necessarily is treated as a nominal outcome). In this appendix we analyze the data from Experiments 1 and 2 using a linear regression on the magnitude of fine, and (in Experiment 2) a separate logistic regression predicting exclusion. We find that actual contribution is weighed more heavily than intention in the linear regression predicting fine, while the opposite is true for the logistic regression predicting exclusion. These results are consistent with the results presented in the main body of the article.

Analysis of fine amount in Experiment 1

The model used here is identical in all ways to the model described in Appendix 1 except for the following. First, each fine amount from player i to player j, denoted $y^{[i,j]}$, is treated as a random draw from a normal distribution with mean $\mu^{[i,j]}$ and precision τ:

$$y^{[i,j]} \sim \text{normal}\left(\mu^{[i,j]}, \tau\right) \tag{13}$$

The central tendency of fine amount $\mu^{[i,j]}$ is a linear function of the familiar predictors:

$$\begin{aligned}\mu^{[i,j]} = {} & \beta_0^{[i]} \\ & + \beta_{int}^{[i]}\left(x_{int}^{[j]} - \overline{x}_{int}\right) \\ & + \beta_{act}^{[i]}\left(x_{act}^{[j]} - \overline{x}_{act}\right) \\ & + \beta_{indig}^{[i]}\left(x_{indig}^{[i,j]} - \overline{x}_{indig}\right)\end{aligned} \tag{14}$$

These predictors have hierarchical priors identical to those specified in Appendix 1, with each subject i having an individual weight that is distributed in a higher level distribution defined by group level parameters. The final new component of this model is the prior on τ, representing the precision of the normal distribution around the predicted tendency $\mu^{[i,j]}$. We use the vague prior we have used previously when estimating the τ values:

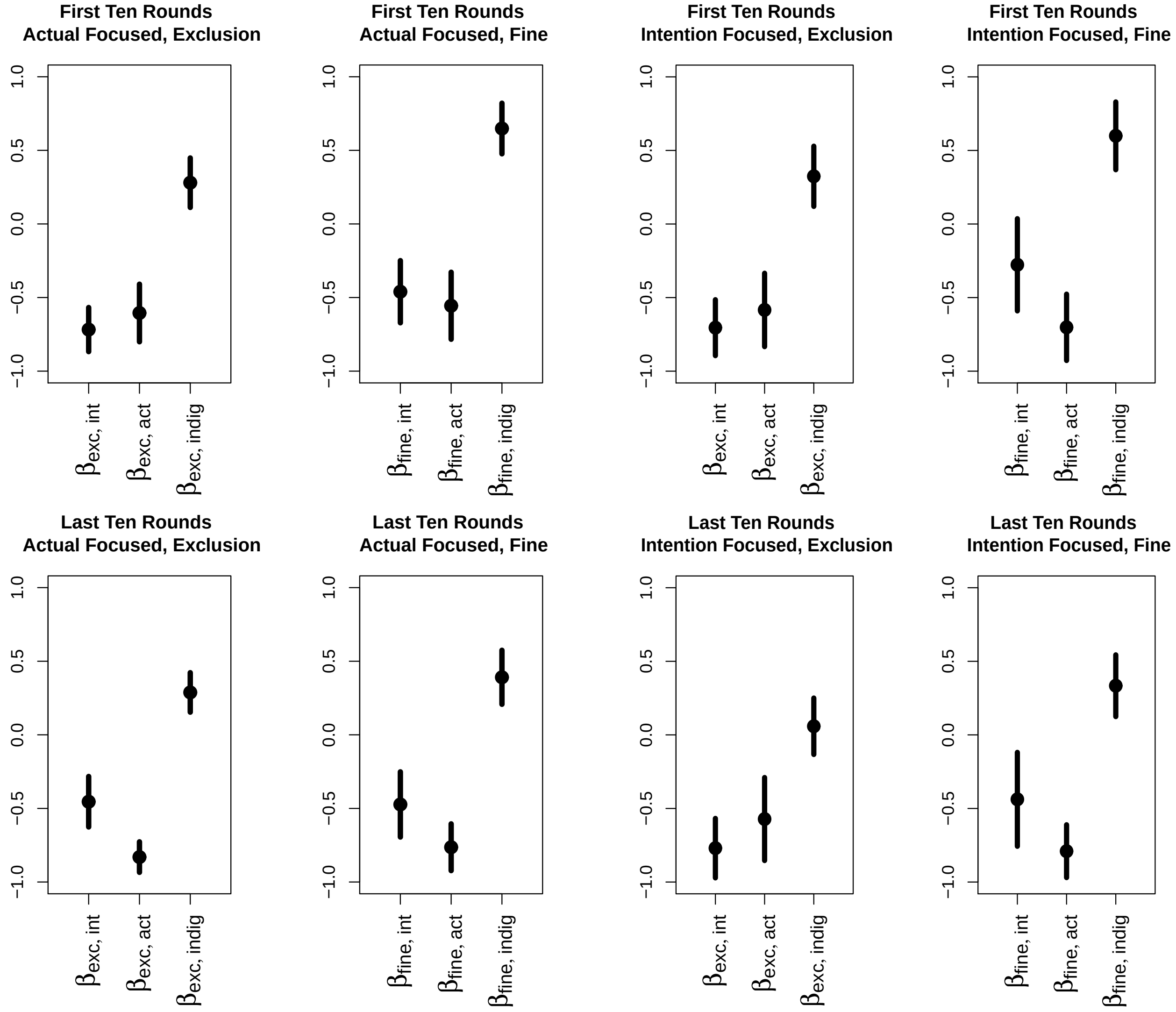
$$\tau \sim \text{gamma}(1.10512, 0.010512)$$

Results

The results of this analysis were analogous to those presented in the main text. Figure 7 indicates the HDIs of

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 6: Top two panels: The beta weight estimates in the actual-focused and intention-focused conditions during the first ten rounds. Notice that all the weights on all predictors are qualitatively similar across the two conditions. Bottom two panels: beta weight estimates in the actual-focused and intention-focused conditions during the last ten rounds. Notice that the weight on intended for exclusion is qualitatively smaller in the actual-focused condition than in the intention-focused condition, and the weight on actual for exclusion is qualitative larger in the actual-focused condition than in the intention-focused condition.



the three group level weights in phases 1 and 2. As usual, we confirmed the qualitative difference with a quantitative comparison: in phase 1 the weight on actual contribution is greater than the weight on intended contribution (mean difference = 0.483, 95% HDI from 0.227 to 0.726) and the pattern is the same in phase 2 (mean difference = 0.607, 95% HDI from 0.383 to 0.825).

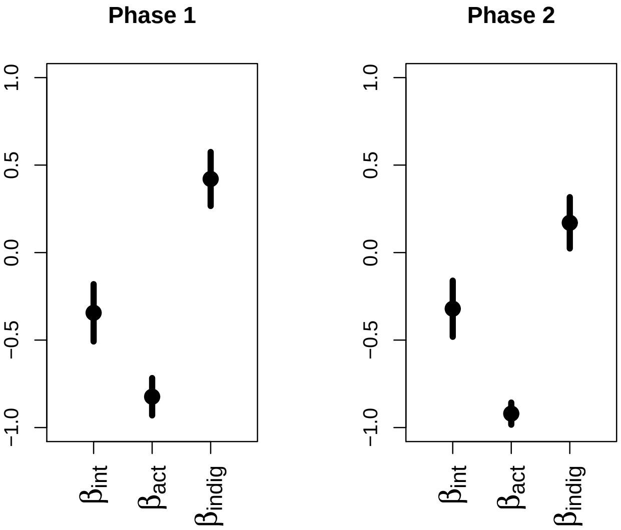
Fine amount analysis in Experiment 2

The analysis of the Experiment 2 data utilizes the analysis described in the previous section, with the addition of an independent logistic regression applied to exclusion behavior. This logistic regression is identical to the model described in Appendix 2, except that the categorical behavior denoted $y^{[i,j]}_{cat}$ can now only take on two values, with 1 representing no punishment or a fine, and 2 representing exclusion:

$$y^{[i,j]}_{cat} \sim \text{cat}\left(\pi^{[i,j]}_{none,fine}, \pi^{[i,j]}_{exc}\right) \tag{15}$$

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Figure 7: Posterior estimates of the group level beta weights using linear regression of the amount of fine. Notice that in both phase 1 (left panel) and phase 2 (right panel) the weight on the actual contribution is greater in magnitude than the weight on intended contribution.



Thus in the logistic portion of this model only a single set of beta weights need to be estimated:

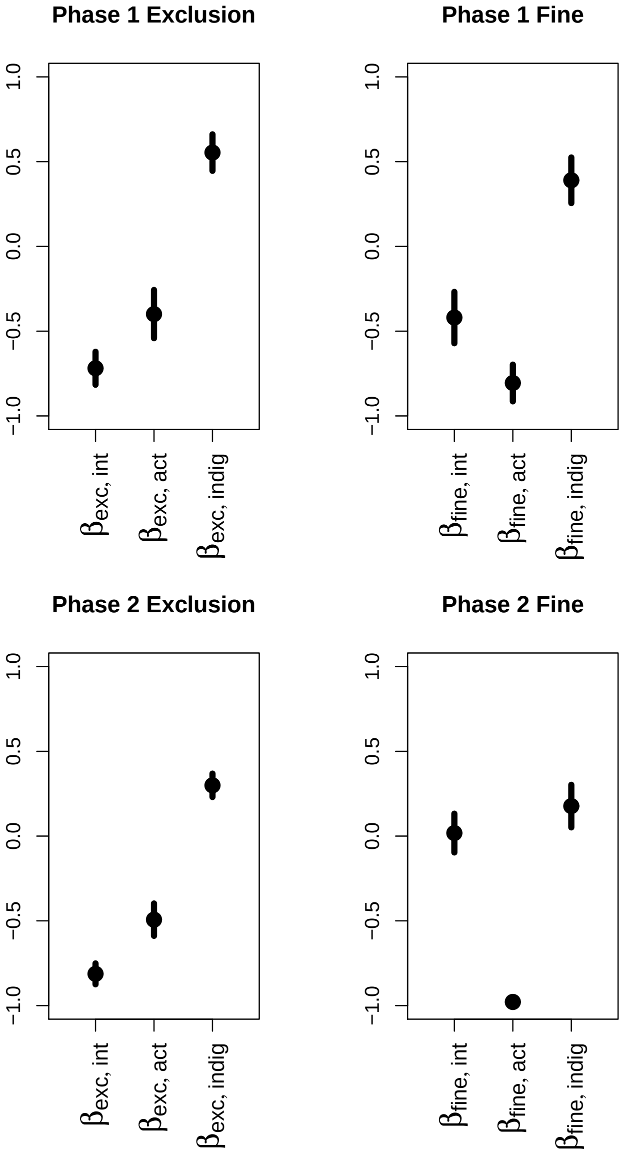
$$\begin{aligned}\lambda^{[i,j]}_{exc} = {} & \beta^{[i]}_{exc,0} \\ & + \beta^{[i]}_{exc,int}\left(x^{[j]}_{int} - \overline{x}_{int}\right) \\ & + \beta^{[i]}_{exc,act}\left(x^{[j]}_{act} - \overline{x}_{act}\right) \\ & + \beta^{[i]}_{exc,indig}\left(x^{[i,j]}_{indig} - \overline{x}_{indig}\right) \qquad (16)\end{aligned}$$

The conversion from this tendency $\lambda^{[i,j]}_{exc}$ to a choice probability, as well as the hierarchical priors for each of the beta weights, are identical to the model described in Appendix 2.

Results

Figure 8 indicates the HDIs of the three group level weights in phases 1 and 2. In phase 1, quantitative comparison of parameters indicate that the weight on actual contribution is less than the weight on intention for exclusion (mean difference = −0.321, 95% HDI from −0.531 to −0.102) and the weight on actual contribution is more than the weight on intention for fine (mean difference = 0.395, 95% HDI from 0.143 to 0.639). In phase 2 these patterns are the same or more pronounced (exclusion: mean difference = −0.325, 95% HDI from −0.472 to −0.163 ; fine: mean difference = 1.0, 95% HDI from 0.885 to 1.12).

Figure 8: Posterior estimates of the group level beta weights using a linear regression on fine amount and a logistic regression on exclusion. Notice that in both phase 1 (top panel) and phase 2 (bottom panel) the weight on the actual contribution is greater in magnitude than the weight on intention for fines, but the opposite is true for exclusion.



Summary

In the main text we utilize logistic models of fine behavior for ease of comparison with categorical exclusion behavior. This appendix instead uses linear regression to analyze fine, and a separate logistic analysis to analyze exclusion. The results of these analyses are consistent with the pure logistic model.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Appendix 5: Individual Analysis and Correlations

This appendix investigates individual behavior in Experiment 2. The purpose of this analysis is to assess whether the differences in predictor emphasis observed in the group level behavior across the two punishment types is potentially an artifact of distinct patterns of individual behavior. In particular, we investigate the hypothesis that some subjects attend to intention and others to actual contribution (overall), and that the observed differences across the two punishment types are due to the differential usage of the two punishment types among these two types of subjects. To investigate this hypothesis, we investigate the correlation between measures of intended contribution emphasis and general propensity to punish.

Results

Measure of intended contribution emphasis

The analysis presented in the main body estimates beta weight parameters for each individual. However, in the model those parameters were distributed in a single t distribution described by group level parameters. This hierarchical structure means that the group level parameters are influenced by the individual level parameters, and in turn the individual level parameters are influenced by the group level parameters; any given individual's parameters are partially influenced by the behavior of the group. Because the focus of the present analysis does not assume a single, heavy-tailed cluster of subjects, we removed this hierarchical structure. Instead, we estimate each individual's parameters separately, with the prior on each individual parameter being a fixed but vague normal distribution identical to the prior on the group level parameters of the hierarchical model.

We calculate the degree of intended contribution emphasis (or actual contribution emphasis) for each individual for both exclusion and fine. For each subject i, we calculate the posterior distributions on $\beta^{[i]}_{exc,int} - \beta^{[i]}_{exc,act}$ (subject i's intended contribution emphasis for exclusion) and $\beta^{[i]}_{fine,int} - \beta^{[i]}_{fine,act}$ (subject i's intended contribution emphasis for fine). We then take the mean of these posterior distributions of the differences to represent subject i's intended contribution emphasis for exclusion and fine, respectively. Note that despite the label this value can indicate intended contribution emphasis or actual contribution emphasis: a positive value represents intended contribution emphasis, and a negative value indicates actual contribution emphasis.

Measure of propensity to punish

We use three measures of subject i's propensity to punish: proportion of interactions where i applied a fine, proportion of interactions where i applied exclusion, and proportion of interactions where i applied any punishment at all.

Table 2 shows the correlations of these measures. The correlation of intended contribution emphasis across the two punishment types is moderately positive. Also note that the tendency to fine is slightly negatively related to the tendency to exclude, meaning that subjects may favor one punishment type over the other. These both could be consistent with the hypothesis that the observed group level differences could be due to different individual punishment types. However, the correlations between intended contribution emphasis and the various propensity to punish measures casts doubt on this hypothesis. Intended contribution emphasis for exclusion has essentially no correlation with any of the punishment propensity measures, meaning that individuals with higher intended contribution emphasis in exclusion are neither more nor less likely to use either punishment type. Moreover, intended contribution emphasis for fine has small correlations with the punishment measures, but in the opposite trend predicted by the hypothesis: those with more emphasis on intended contribution (that is, less emphasis on actual contribution) are actually less likely to exclude and more likely to fine. This would mean that intention emphasizing fines are associated with more fining and less excluding, which at the group level should lead to more intention emphasis for fines, the opposite of the observed pattern.

Appendix 6: Detailed experiment procedure

This appendix includes the verbatim instructions for each experiment and the exact set of contribution behavior sets used by the automated players in Experiments 1 and 2.

Verbatim instructions

In all three experiments, subjects read an identical general description of the task they would be completing during the consent process:

"You will sit comfortably in front of a standard desktop computer, located in a booth in the adjacent room. Complete instructions will be displayed on the computer screen. You will play a game with four other players, who may be other subjects on networked computers or automated players. The object is to gain as many points as you can. You will periodically be given points which you can choose to either keep or invest in a common pool. You

Table 2: Correlations between intended contribution emphasis and propensity to punish. Note: ICE Exc = intended contribution emphasis for exclusion, ICE Fine = intended contribution emphasis for fine, Prop Exc = propensity to exclude, Prop Fine = propensity to fine, and Prop Punish = propensity to punish (including both exclusion and fine).

	ICE Exc	ICE Fine	Prop Exc	Prop Fine	Prop Punish
ICE Exc	1.0				
ICE Fine	.320	1.0			
Prop Exc	−.023	−.092	1.0		
Prop Fine	.033	.237	−.126	1.0	
Prop Punish	.010	.125	.606	.713	1.0

will also be given the opportunity to respond to the contributions of other players. The experiment will take less than an hour. At the end of the experiment, a more detailed explanation of the experiment will be provided. We cannot reveal more before the experiment, because doing so could affect the manner in which you play the game."

After the consent process was completed, subjects sat down one of the five computers in a partially enclosed booth. The remaining directions were administered via the experiment program one paragraph at a time. These directions varied based on the experiment, and based on condition.

Experiments 1 and 2 verbatim instructions

The instructions for Experiments 1 and 2 were identical except for the bold and italicized sections below, which describe the punishments available to the subject. Subjects in Experiment 1 saw the italicized section, whereas subjects in Experiment 2 saw the bold version.

"In this experiment, you will be playing a game with four other players. In each round of the game, you and the other players are given 10 new points. You and the other players then have the option to contribute some of your points to a mutual investment pool. After all the players have made their contributions, the pool is grown by a certain amount and then the total is divided equally among all players. The amount you get back from the pool depends on how much every player invests. The more everyone invests, the more everyone gets. But even if you invest nothing, you will still receive an equal share of the common investment.

At the beginning of each round, you select the amount of points you would like to contribute to the pool. Unfortunately the amount that actually gets contributed could be somewhat higher or lower than what you intended. This accidental variation reflects real-world contingencies such as miscommunications. All players are subject to the same accidental discrepancy between the intended contribution and the actual amount.

After you see what every player has contributed and how much they have gotten you have the opportunity to penalize other players by deducting points from them, but for every point you deduct you have to pay 1/4 of a point.

After you see what every player has contributed and how much they have gotten you have the opportunity to penalize other players. There are two ways to penalize. First, you can deduct points from another player, but for every point you deduct you have to pay 1/4 of a point. Second, you can vote to exclude a player from the game, without cost to you. If a player is excluded from your game, he or she is replaced with a new player.

If you have any questions, please ask the experimenter now. Otherwise, PLEASE WAIT FOR THE EXPERIMENTER TO CLOSE THE CURTAIN BEHIND YOU. After the curtain is closed, press the space bar when you are ready to begin."

Subjects also were given the following instructions when they completed phase 1 of the experiment:

"From this point on in the game the other players will be able to penalize you and each other. Their penalty options are identical in all ways to the penalty options you have had access to throughout the game. Your ability to penalize remains unchanged. After you assign all of your penalties, you will see the penalties that the other players assigned."

Experiment 3 verbatim instructions

The instructions for Experiment 3 were identical to the instructions for Experiment 2, except that for half the subjects (those in the random noise rule condition) the paragraph describing the accidental variation in the contribution process was replaced with the following:

"At the beginning of each round, you select the amount of points you would like to contribute to the pool. Unfortunately sometimes the contributed amount will not be what you intended, and will instead be some other random value

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

unrelated to the amount you meant to contribute. Whenever this random event occurs, your intended amount has no effect on the random actual contribution. This accidental change reflects real-world contingencies such as miscommunications. All players have the same chance of their intended amount not matching their actual amount."

Pregenerated contribution patterns

Experiments 1 and 2 utilized pregenerated contribution patterns, with each round containing one intentionally high contributor, one intentionally low contributor, one accidentally high contributor, and one accidentally low contributor. Both the order of these behavior sets, and the placement of each contributor type was randomly determined for each subject. The automated players were not consistent; for example, an automated player could be the intentionally high contributor one round, and the accidentally low contributor the next. Table 3 is a complete list of the behavior patterns utilized in experiments 1 and 2.

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press

Table 3: A complete list of all the behavior patterns presented to subjects in Experiments 1 and 2.

Intentionally High		Accidentally High		Intentionally Low		Accidentally Low	
Intended	Actual	Intended	Actual	Intended	Actual	Intended	Actual
6	8	4	6	4	2	6	4
8	6	4	6	2	4	6	4
6	8	3	5	3	1	6	4
8	6	4	6	1	3	5	3
7	10	3	6	3	0	7	4
10	7	4	7	0	3	6	3
8	10	4	8	4	0	8	4
10	8	4	8	0	4	8	4
8	10	3	7	3	0	8	4
10	8	4	8	0	3	7	3
7	10	4	7	4	1	7	4
10	7	4	7	1	4	7	4
6	9	3	6	3	0	6	3
9	6	3	6	0	3	6	3
6	10	2	6	2	0	6	2
10	6	2	6	0	2	6	2
7	9	4	6	4	2	7	5
9	7	5	7	2	4	6	4
6	9	4	7	4	1	6	3
9	6	3	6	1	4	7	4
8	10	3	6	3	0	8	5
10	8	5	8	0	3	6	3
6	10	4	8	4	0	6	2
10	6	2	6	0	4	8	4
6	10	3	7	3	0	6	2
10	6	2	6	0	3	7	3
7	10	3	7	3	0	7	3
10	7	3	7	0	3	7	3
6	8	2	4	2	0	6	4
8	6	4	6	0	2	4	2
8	10	4	7	4	1	8	5
10	8	5	8	1	4	7	4
7	10	4	8	4	0	7	3
10	7	3	7	0	4	8	4
6	9	2	5	2	0	6	3
9	6	3	6	0	2	5	2
7	10	2	6	2	0	7	3
10	7	3	7	0	2	6	2
8	10	4	6	4	2	8	6
10	8	6	8	2	4	6	4
7	10	2	5	2	0	7	4
10	7	4	7	0	2	5	2
8	10	2	6	2	0	8	4
10	8	4	8	0	2	6	2

https://doi.org/10.1017/S1930297500006409 Published online by Cambridge University Press