Mapping the Effect of Altruistic Punishment on Cooperation

Keywords: cooperation, public goods games, punishment, experiments, behavioral economics

Extended Abstract

Introduction

Since the experimental work of Fehr and Gächter on the role of punishment in sustaining costly cooperation [1], many studies have shown that the efficiency of punishment (the net social benefit after subtracting punishment costs) can vary depending on contextual factors such as length of interaction [2], information on peer outcomes [3], and the efficacy of punishment [4].

In many cases, studies that explore how the effect of punishment on efficiency interacts with public goods game (PGG) design factors do so by manipulating a single factor (e.g. total number of rounds) in isolation. Relatedly, factors assumed to have no effect are typically not varied (e.g. many experimental studies have 4 players). Consequently, the current literature on the role of altruistic punishment in sustaining cooperative behavior may be blind to higher-order, untheorized interactions, and may not be robust to variations along assumed “nuisance” dimensions.

In this work, we adopt an integrative experimentation approach [5], in which we identify 20 parameters that define the space of PGG experimental designs and execute experiments across this space systematically. By doing so, we aim to confirm known interactions and describe novel, higher-order dependencies by modeling the heterogeneity of the effect of altruistic punishment across the 20-dimensional design space of public goods games.

Experimental Design

To probe the main effects and interactions of interest, a highly-parameterized public goods game was implemented in Empirica [6], as shown in Figure 1. The parameter space includes dimensions of game structure (e.g. number of players, length of interaction), informational structure (e.g. whether peer information is displayed, and whether groups can engage in “cheap talk” via a chat window), and incentive structure (e.g. the marginal per capita return, and the existence/cost/effectiveness of peer punishment/reward).

An initial batch of experimental conditions is chosen by a space-filling sample of the experimental design space, then executed with participants recruited from Prolific and Amazon Mechanical Turk. From this initial batch, a model of the effectiveness of punishment across experimental conditions is learned and used to select subsequent experiments that are predicted to maximize/minimize the effectiveness of punishment, to validate the model’s ability to predict the effect of the intervention on yet unseen contexts.

Separately, a panel of experts will be surveyed about their predictions of the main effects of PGG design factors on net social benefit, the interactions between these factors and the effect of altruistic punishment, and the regions of design space where the benefit of altruistic punishment is maximized. This survey will allow us to compare our findings to commonly held beliefs and provide a basis to compare the predictive power of human- and machine-generated theories of cooperation in the PGG context.

Preliminary Results

At the time of submission, we have completed a pilot experiment with 5,500 contribution decisions made during a total of 1,000 rounds of public goods games, across 32 different game configurations. From this data, we have found patterns of heterogeneity validating our premise, as illustrated in Figure 2, showing that the effect of punishment on the size of contributions can vary from negative to positive based on combinations of factors such as anonymity, group size, and punishment effectiveness.

By the time this work is presented, we plan to have executed experiments with a total of ~400 different PGG configurations, allowing us to analyze the effect of punishment on various outcomes and at different scales of analysis. Through this study, we aim to demonstrate the epistemological value and practical application of an integrative experimental design, validated by generating and testing novel predictions about cooperative behavior. Also, given that the experimental conditions in our study are sampled independent of the current set of theories in the field, the resulting data are potentially informative not just about existing theories, but about theories that are yet to be proposed. Consequently, we believe this data will have greater longevity than data generated by traditional experiments.

References

1. Fehr E, Gächter S. Cooperation and Punishment in Public Goods Experiments. Am Econ Rev. 2000;90: 980–994.

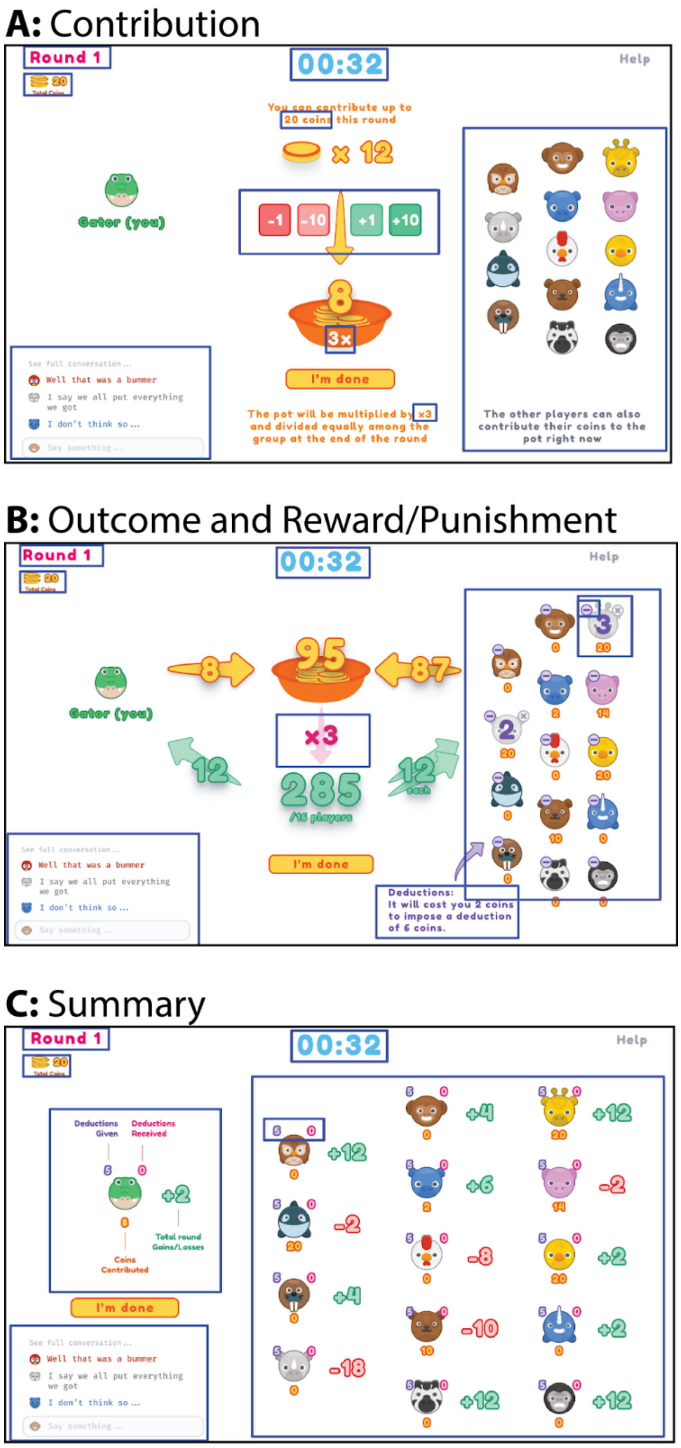
2. Gächter S, Renner E, Sefton M. The long-run benefits of punishment. Science. 2008;322: 1510.

3. Nikiforakis N. Feedback, punishment and cooperation in public good experiments. Games Econ Behav. 2010;68: 689–702.

4. Nikiforakis N, Normann H-T. A comparative statics analysis of punishment in public-good experiments. Exp Econ. 2008;11: 358–369.

5. Almaatouq A, Griffiths TL, Suchow JW, Whiting ME, Evans J, Watts DJ. Beyond Playing 20 Questions with Nature: Integrative Experiment Design in the Social and Behavioral Sciences. Behav Brain Sci. 2022; 1–55.

6. Almaatouq A, Becker J, Houghton JP, Paton N, Watts DJ, Whiting ME. Empirica: a virtual lab for high-throughput macro-level experiments. Behav Res Methods. 2021;53: 2158–2171.



You can contribute up to 20 coins this round

The pot will be multiplied by x3 and divided equally among the group at the end of the round

The other players can also contribute their coins to the pot right now

See full conversation ...

Well that was a bummer

I say we all put everything we got

I don't think so ...

Say something ...

Deductions: It will cost you 2 coins to impose a deduction of 6 coins.

Figure 1: A highly-parameterized implementation of a public goods game (PGG) in the virtual lab. Each round of a PGG has three phases: (A) Contribution, in which players choose how much to add to the public fund; (B) Outcome & Punishment/Reward, in which players observe their peers' contributions, and can decide to punish/reward them; (C) Summary, in which all actions taken in the previous stages are presented. Each element of the interface enclosed in a blue box is parameterized by one or more factors governing functionality, information display, and game structure.

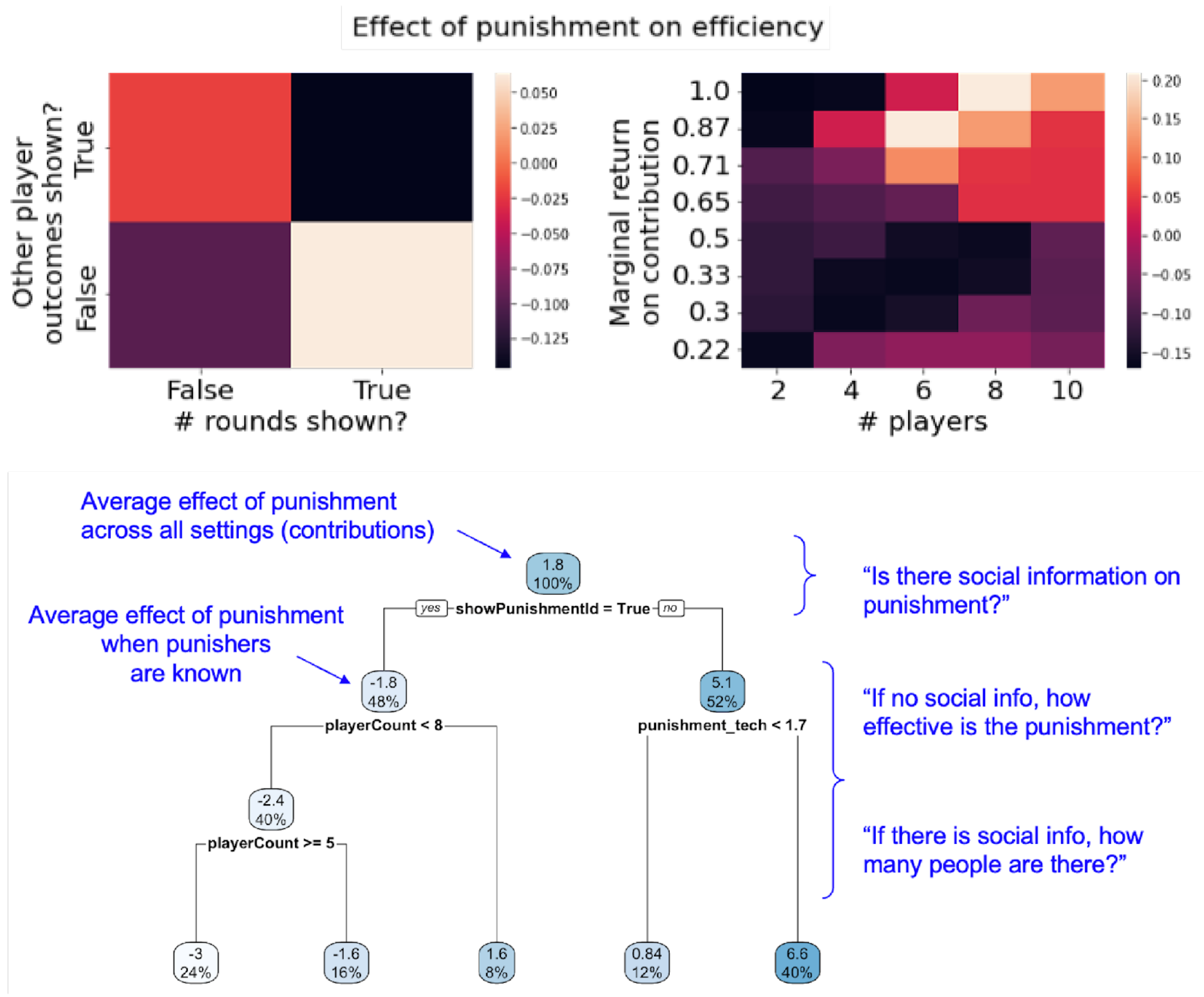


Figure 2: The effect of punishment on cooperation is heterogeneous. The pairwise heatmaps and causal tree in this figure illustrate the highly contingent nature of punishment's effectiveness across different parameter values. These early results are a clear indicator that the impact of punishment varies significantly depending on the specific conditions of the public goods game.