The Price of Justice: Altruistic Punishment and Human Cooperation

Based on Fehr & Gächter (Nature, 2002)

Timothy Nguyen

EC323: Behavioral Economics
Professor Jawwad Noor

August 6, 2025

Agenda

1. Introduction: The Puzzle of Cooperation

Agenda

1. Introduction: The Puzzle of Cooperation
2. Theoretical Framework & Predictions

Agenda

1. Introduction: The Puzzle of Cooperation
2. Theoretical Framework & Predictions
3. Experimental Design: The Public Goods Game

Agenda

1. Introduction: The Puzzle of Cooperation
2. Theoretical Framework & Predictions
3. Experimental Design: The Public Goods Game
4. Analysis of Experimental Results

Agenda

1. Introduction: The Puzzle of Cooperation
2. Theoretical Framework & Predictions
3. Experimental Design: The Public Goods Game
4. Analysis of Experimental Results
5. Discussion & Final Conclusions

Agenda

1. Introduction: The Puzzle of Cooperation
2. Theoretical Framework & Predictions
3. Experimental Design: The Public Goods Game
4. Analysis of Experimental Results
5. Discussion & Final Conclusions
6. Q&A

The Puzzle of Cooperation

- Humans often help strangers, even at a personal cost.

The Puzzle of Cooperation

- Humans often help strangers, even at a personal cost.
- This is a puzzle because standard economic models assume people are purely self-interested.

The Puzzle of Cooperation

- Humans often help strangers, even at a personal cost.
- This is a puzzle because standard economic models assume people are purely self-interested.
- So, why do we punish "free-riders" when we get nothing out of it?

The Puzzle of Cooperation

- Humans often help strangers, even at a personal cost.
- This is a puzzle because standard economic models assume people are purely self-interested.
- So, why do we punish "free-riders" when we get nothing out of it?
- This is called **altruistic punishment**.

Connecting to Our Class: Social Preferences

This experiment challenges the basic assumption of pure self-interest.

- In our lectures, we've discussed how people are not always selfish. They have **social preferences**.

Connecting to Our Class: Social Preferences

This experiment challenges the basic assumption of pure self-interest.

- In our lectures, we've discussed how people are not always selfish. They have **social preferences**.
- People care about fairness and what others get. This is sometimes called **inequity aversion**.

Connecting to Our Class: Social Preferences

This experiment challenges the basic assumption of pure self-interest.

- In our lectures, we've discussed how people are not always selfish. They have **social preferences**.
- People care about fairness and what others get. This is sometimes called **inequity aversion**.
- Altruistic punishment is a powerful example of this: people are willing to *pay money* just to reduce the payoff of someone they think is unfair.

Connecting to Our Class: Social Preferences

This experiment challenges the basic assumption of pure self-interest.

- In our lectures, we've discussed how people are not always selfish. They have **social preferences**.
- People care about fairness and what others get. This is sometimes called **inequity aversion**.
- Altruistic punishment is a powerful example of this: people are willing to *pay money* just to reduce the payoff of someone they think is unfair.
- This behavior shows that a sense of justice can be a stronger motivator than money.

The Experiment: A "Public Goods" Game

The Setup:

- Anonymous groups of 4.
- Each person gets 20 Money Units (MUs).
- You can invest in a group project.

The Experiment: A "Public Goods" Game

The Setup:

- Anonymous groups of 4.
- Each person gets 20 Money Units (MUs).
- You can invest in a group project.

The Dilemma:

- For every 1 MU you invest, everyone in the group gets 0.4 MUs back.
- Your best strategy is to be a **free-rider**: invest nothing and just collect the rewards from others' investments.
- But if everyone free-rides, everyone is worse off!

The Two Conditions

To see if punishment works, subjects were split into two scenarios.

- **Condition 1: No Punishment**
  - You play the game, see what others invested, and that's it. You can't do anything about it.

The Two Conditions

To see if punishment works, subjects were split into two scenarios.

- **Condition 1: No Punishment**
  - You play the game, see what others invested, and that's it. You can't do anything about it.
- **Condition 2: Punishment Allowed**
  - After investing, you have the option to punish others.
  - **The Cost of Justice:** You can spend 1 MU to make someone else lose 3 MUs.
  - This is "altruistic" because it costs you money and gives you no direct financial reward.

The Standard Model: A Selfish Player's Choice

Standard economic theory assumes utility (U_i) equals monetary payoff (π_i).

- **Payoff Function:** $\pi_i = (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j$

The Standard Model: A Selfish Player's Choice

Standard economic theory assumes utility (U_i) equals monetary payoff (π_i).

- **Payoff Function:** $\pi_i = (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j$
- **The Rational Choice (Marginal Utility):**
  - The change in utility for contributing one more MU is:

$$\frac{\partial U_i}{\partial c_i} = -1(\text{cost}) + 0.4(\text{your share}) = -0.6$$

  - Your utility goes down for every unit you contribute.

The Standard Model: A Selfish Player's Choice

Standard economic theory assumes utility (U_i) equals monetary payoff (π_i).

- **Payoff Function:** $\pi_i = (20 - c_i) + 0.4 \sum_{j=1}^{4} c_j$
- **The Rational Choice (Marginal Utility):**
  - The change in utility for contributing one more MU is:

$$\frac{\partial U_i}{\partial c_i} = -1(\text{cost}) + 0.4(\text{your share}) = -0.6$$

  - Your utility goes down for every unit you contribute.
- **Prediction:** A purely self-interested player will always choose $c_i^* = 0$. This model predicts cooperation will collapse.

A Social Preferences Model (Utility View)

To explain the results, we need utility functions that include social factors.

The Cooperator's Choice to Punish

- Their utility is reduced by observing unfairness (D_j).

A Social Preferences Model (Utility View)

To explain the results, we need utility functions that include social factors.

The Cooperator's Choice to Punish

- Their utility is reduced by observing unfairness (D_j).
- Utility: $U_i = \pi_i - \alpha(D_j)$
  (π_i: payoff, $\alpha(D_j)$: anger from unfairness)

The Free-Rider's Choice to Cooperate

- Their utility is reduced by the expected loss (L_j^e) from being punished.
- Utility: $U_j = \pi_j - L_j^e(c_j)$
  (π_j: payoff, L_j^e: expected loss)

A Social Preferences Model (Utility View)

To explain the results, we need utility functions that include social factors.

The Cooperator's Choice to Punish

- Their utility is reduced by observing unfairness (D_j).
- Utility: $U_i = \pi_i - \alpha(D_j)$
  (π_i: payoff, $\alpha(D_j)$: anger from unfairness)
- They will pay a cost (P_i) to punish if the psychological relief is greater than the monetary cost.

The Free-Rider's Choice to Cooperate

- Their utility is reduced by the expected loss (L_j^e) from being punished.
- Utility: $U_j = \pi_j - L_j^e(c_j)$
  (π_j: payoff, L_j^e: expected loss)
- They will cooperate if the anticipated pain of punishment (L_j^e) is greater than the gain from free-riding.

Result 1: Punishment Boosts Cooperation

The experimental data rejects the standard model and supports the social preferences model.

- **Without Punishment:** Cooperation collapsed, as predicted by the self-interest model.

Result 1: Punishment Boosts Cooperation

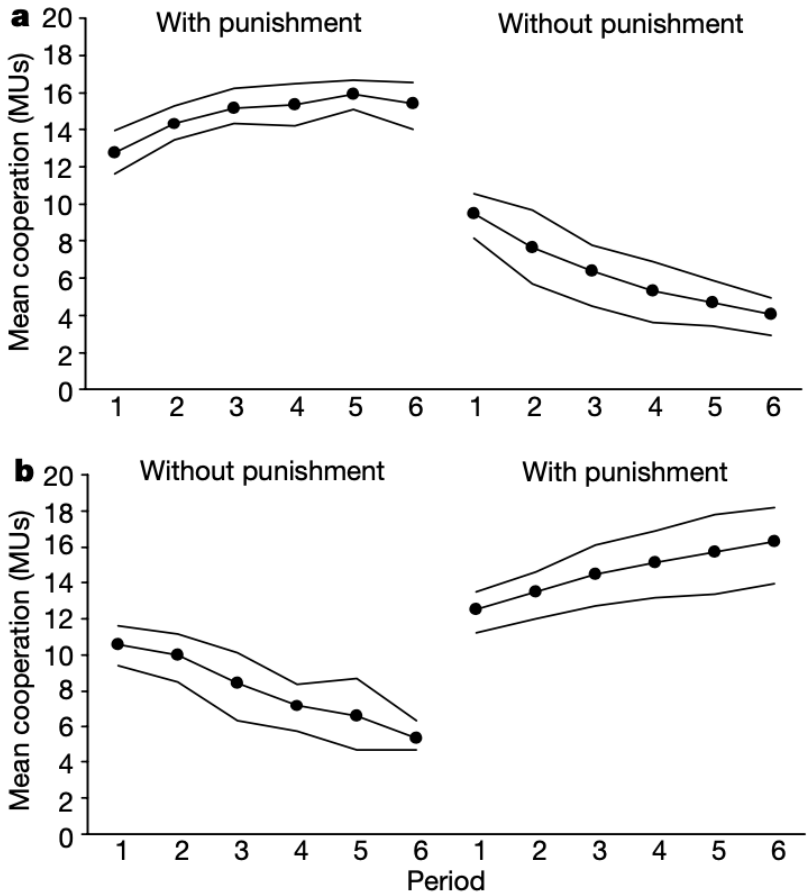
The experimental data rejects the standard model and supports the social preferences model.

- **Without Punishment:** Cooperation collapsed, as predicted by the self-interest model.
- **With Punishment:** Cooperation was high and even *increased* over time.

Result 1: Punishment Boosts Cooperation

The experimental data rejects the standard model and supports the social preferences model.

- **Without Punishment:** Cooperation collapsed, as predicted by the self-interest model.
- **With Punishment:** Cooperation was high and even *increased* over time.
- The graph shows this clearly. Cooperation thrives only when punishment is possible.



Result 2: Punishment Is Targeted

Punishment wasn't random; it was a direct response to unfairness.

- People consistently punished free-riders.

Result 2: Punishment Is Targeted

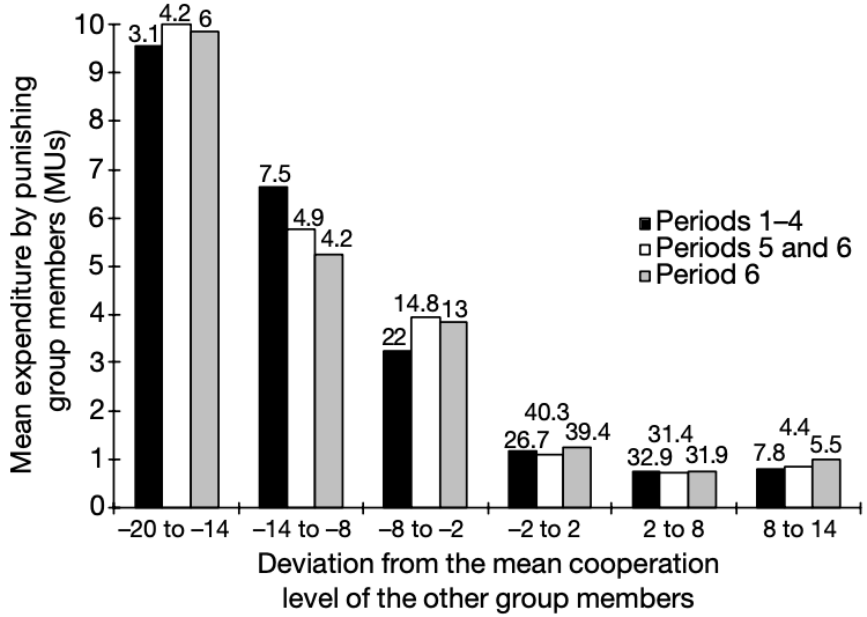
Punishment wasn't random; it was a direct response to unfairness.

- People consistently punished free-riders.
- The more someone free-rode, the more heavily they were punished.

Result 2: Punishment Is Targeted

Punishment wasn't random; it was a direct response to unfairness.

- People consistently punished free-riders.
- The more someone free-rode, the more heavily they were punished.
- This graph shows that the biggest free-riders (far left) received the most punishment, matching the predictions of the social preferences model.



Personal Comments & Discussion

My Reactions:

- It's fascinating that a "negative" emotion like anger can actually lead to better outcomes for the group.
- The simple design of the experiment makes the results very powerful and easy to understand.
- This research has huge real-world implications for everything from managing climate change to building online communities.

Questions for Discussion:

- Can you think of real-world examples of altruistic punishment?

Connection 1: The 'Standard Model' of Selfishness

Our lecture notes first introduce the standard economic model, which assumes people are purely selfish and only care about their own payoff.

- This "selfishness model" makes a clear and testable prediction for the Public Goods Game: **a rational player will contribute nothing ($c_i = 0$)**.

Connection 1: The 'Standard Model' of Selfishness

Our lecture notes first introduce the standard economic model, which assumes people are purely selfish and only care about their own payoff.

- This "selfishness model" makes a clear and testable prediction for the Public Goods Game: **a rational player will contribute nothing ($c_i = 0$)**.
- As we saw in the utility function, this is because every dollar contributed leads to a personal loss of 60 cents.

Connection 1: The 'Standard Model' of Selfishness

Our lecture notes first introduce the standard economic model, which assumes people are purely selfish and only care about their own payoff.

- This "selfishness model" makes a clear and testable prediction for the Public Goods Game: **a rational player will contribute nothing ($c_i = 0$)**.
- As we saw in the utility function, this is because every dollar contributed leads to a personal loss of 60 cents.
- Therefore, this experiment can be seen as a direct test of the standard model. If people cooperate, the model is incomplete.

Connection 2: Social Preferences - Fairness & Reciprocity

The behavior in the experiment is better explained by models of **social preferences**, as discussed in the lecture notes.

- **Inequity Aversion:** The notes mention that people may dislike unfair outcomes. The punishment of free-riders is a perfect example. Players pay a personal cost to reduce the payoff of someone who earned "unfair" money, which is consistent with models like Fehr-Schmidt.

Connection 2: Social Preferences - Fairness & Reciprocity

The behavior in the experiment is better explained by models of **social preferences**, as discussed in the lecture notes.

- **Inequity Aversion:** The notes mention that people may dislike unfair outcomes. The punishment of free-riders is a perfect example. Players pay a personal cost to reduce the payoff of someone who earned "unfair" money, which is consistent with models like Fehr-Schmidt.
- **Negative Reciprocity:** Reciprocity is defined as responding to kind actions with kindness and hostile actions with hostility. Free-riding can be seen as a hostile act against the group. Punishment is therefore a classic example of negative reciprocity.

Connection 2: Social Preferences - Fairness & Reciprocity

The behavior in the experiment is better explained by models of **social preferences**, as discussed in the lecture notes.

- **Inequity Aversion:** The notes mention that people may dislike unfair outcomes. The punishment of free-riders is a perfect example. Players pay a personal cost to reduce the payoff of someone who earned "unfair" money, which is consistent with models like Fehr-Schmidt.
- **Negative Reciprocity:** Reciprocity is defined as responding to kind actions with kindness and hostile actions with hostility. Free-riding can be seen as a hostile act against the group. Punishment is therefore a classic example of negative reciprocity.
- **Isolating Emotional Reciprocity:** The experiment used strangers who only interacted once, so people couldn't be acting selfishly to gain future benefits. This shows that when people punish unfairness, they're doing it because they care about fairness, not because they expect something in return.

Thank You Questions?