Achieving Full Cooperation in Finitely Repeated Prisoners' Dilemma via a Deposit Mechanism: Implementability and Robustness

Naoki Takata

September 18, 2025

Abstract

In a finitely repeated Prisoners' Dilemma (PD), backward induction implies defection in all periods. We introduce a simple *deposit mechanism* under which each player posts a refundable deposit that is forfeited if and only if they ever choose D. We show that if the deposit satisfies the threshold $D \geq T - R$ (where T is the temptation payoff and R is the mutual cooperation payoff), then full cooperation $(\mathrm{C}, \mathrm{C}, \ldots, \mathrm{C})$ is a subgame perfect equilibrium (SPE) for any finite horizon, irrespective of discounting, provided forfeiture is immediate. We further derive robustness conditions under observational errors (false forfeitures) and partial slashing, discuss implementation details, and provide a simple simulation procedure.

Keywords: repeated games; mechanism design; commitment; collateral; subgame perfect equilibrium; deposit mechanism; prisoners' dilemma; cooperation; implementation; robustness; simulation; comparative statics; implications

1 Introduction

In finitely repeated PD, the standard prediction is universal defection due to backward induction. In practice, institutions like deposits, sureties, and penalties support cooperation. We formalize a minimal change to the stage environment—a *deposit mechanism*—and study its equilibrium effects. Our main result is a simple threshold: if the private cost of forfeiture exceeds the one-shot deviation gain $T - R$, then full cooperation is sustained as an SPE for *any* finite horizon.

2 Model

Consider a two-player symmetric PD with stage payoffs

$$T > R > P > S, \qquad 2R > T + S.$$

Actions are C (cooperate) and D (defect). The horizon is finite $N \in \mathbb{N}$. Stage actions are perfectly observed[1].

Deposit mechanism Before period 1, each player $i \in \{1, 2\}$ escrows a deposit $D > 0$ with a trusted third party (or a smart contract). Rules:

- If player i ever plays D in any period, their deposit D is immediately forfeited.

[1] We introduce observation errors in a robustness section.

- If i plays C in all periods, their deposit is fully refunded at the end.

Period-t payoffs equal stage payoffs plus any immediate deposit change.

Strategies and equilibrium We allow history-dependent strategies and use subgame perfection. Consider the *grim trigger* profile σ^*:

> Play C as long as no D has occurred in the history; play D forever after the first D.

3 Main result

Theorem 1 (Deposit threshold for full cooperation)**.** *Under the deposit mechanism, if*

$$D \ \geq \ T - R, \tag{1}$$

then the grim-trigger profile σ^* is an SPE for any finite horizon N. On the equilibrium path, players cooperate in all periods.

Proof. By the one-shot deviation principle, consider a history with no prior D. In period t, deviating from C to D yields an immediate gain $T - R > 0$ but triggers immediate forfeiture of D. Future continuation values under grim trigger cancel in the comparison because after any deviation both players play D henceforth, while under compliance the continuation is identical up to the deviation decision at t. Therefore, the deviation's net benefit is $(T - R) - D \leq 0$ whenever (1) holds. In any subgame after some D has occurred, playing D is a stage Nash best reply. Hence σ^* is subgame perfect. Immediate forfeiture makes the result independent of N and discounting. □

Remark (Discounting and payment timing). With immediate forfeiture, any discount factor $\delta \in (0, 1]$ leaves (1) unchanged. If forfeiture were delayed until the end, the current-value cost would be $\delta^{N-t}D$, yielding a history-dependent threshold $D \geq (T - R)\,\delta^{-(N-t)}$. Immediate slashing is preferable in implementation.

4 Robustness

4.1 Partial slashing

If only a fraction $\alpha \in (0, 1]$ of the deposit is forfeited upon any D, the threshold becomes

$$\alpha D \ \geq \ T - R \qquad \Longleftrightarrow \qquad D \ \geq \ \frac{T - R}{\alpha}.$$

4.2 Observation errors (false positives)

Suppose a compliant C player is mistakenly forfeited with probability $\phi \in [0, 1)$ in a period (false positive), while a D player is always forfeited (no false negatives). Then the expected cost of deviating once at t is $(1 - \phi)D$, so the incentive constraint is

$$D \ \geq \ \frac{T - R}{1 - \phi}. \tag{2}$$

Higher false-positive rates require proportionally larger deposits.

5 Comparative statics and implications

- **Simplicity:** The threshold depends only on the stage deviation gain $T - R$, not on horizon length or detailed histories (with immediate slashing).

- **Decentralized implementation:** Smart-contract escrow with immediate slashing preserves the clean threshold in theorem 1.

- **Practical design:** In noisy environments, calibrate D using (2) and include an appeals/review process.

6 A simple simulation procedure

Given (T, R, P, S), sweep D and measure the cooperation rate under grim trigger with error rate ϕ.

Algorithm 1 Monte Carlo evaluation of cooperation vs. deposit level

```
Input: $T, R, P, S$, horizon $N$, error rate $\phi$, deposit grid $\{D_k\}$, trials $M$
for each $D_k$ do
    for $m = 1$ to $M$ do
        history $\leftarrow$ empty; forfeited$\leftarrow$False
        for $t = 1$ to $N$ do
            if forfeited=False then
                both play C
            else
                both play D
            end if
            with prob. $\phi$, randomly forfeit one compliant player (if any); if so, set forfeited=True
        end for
        record cooperation share in this run
    end for
    output average cooperation share for $D_k$
end for
```

7 Extension: false negatives

If a deviator avoids forfeiture with probability $\beta \in (0, 1)$ (false negative), the expected forfeiture cost after deviating is $(1 - \beta)D$. With both false positives ϕ and false negatives β, a sufficient condition is

$$(T - R) \;\leq\; (1 - \beta - \phi + \phi\beta)\, D.$$

Large ϕ and β sharply increase the required D, highlighting the role of *monitoring and evidence design.*

8 Related literature

Collateral and hostage mechanisms in repeated games are classical. Our contribution is to (i) isolate the *minimal* threshold $D \geq T - R$, (ii) emphasize independence from horizon and discounting under immediate slashing, and (iii) provide closed-form robustness conditions under observational imperfections aligned with modern implementations. See, e.g., [**?**, **?**, **?**, **?**].

9 Conclusion

A deposit mechanism overturns the non-cooperation prediction in finitely repeated PD with a minimal institutional change. A simple condition—forfeiture at least as large as the one-shot temptation gain—implements full cooperation as an SPE. In realistic noisy environments, appropriate deposit calibration and appeals procedures maintain robustness.

Acknowledgments

A Stage payoffs and notation

The symmetric PD payoff matrix (row player's payoffs first):

	C	D
C	(R, R)	(S, T)
D	(T, S)	(P, P)

B Rewriting the threshold

Theorem 1's condition can be written as $D \geq \Delta$, where $\Delta := T - R$ is the one-shot deviation gain. With partial slashing α and false positives ϕ simultaneously, a sufficient condition is $\alpha(1 - \phi)D \geq \Delta$.