FIELD GUIDE 009 / THE RECIPROCITY ENGINE

TOGETHER, OR ALONE?

A playable guide to the repeated Prisoner’s Dilemma, reciprocity, retaliation, forgiveness, and the limits of a good-looking simulation.

What this guide is

A compact field manual for asking a deceptively difficult question: how can mutual cooperation emerge when taking advantage of the other player seems tempting in the immediate round? You can work with two people and paper, or play the digital version at cooperationgame.sahilwho.online. Every result is educational and illustrative, not an assessment of anybody’s personality.

How to use it

First read the payoff matrix. Play the pencil-and-paper game without discussing moves. Repeat with a shared rule; then introduce a mistaken signal. Compare *traces*, not people. Finally, read the competing explanations and write down which assumptions your conclusion depends on.

A responsible boundary

This experiment does not measure human cooperativeness, trustworthiness, morality, or real-world negotiation skill. A toy payoff table leaves out rights, relationships, institutions, unequal vulnerability, and real consequences.

1 / THE LITTLE TRAP

One round, two choices

Each player privately chooses C (cooperate) or D (defect). Reveal the moves together. Read the resulting points as you / other player. The exact numbers are instructional, but their order is crucial.

YOUR MOVE / OTHER MOVE	OTHER C	OTHER D
YOUR C	3 / 3 - mutual reward	0 / 5 - temptation for other
YOUR D	5 / 0 - temptation for you	1 / 1 - mutual defection

Why is this a Prisoner’s Dilemma? If the other player cooperates, you score 5 by defecting rather than 3 by cooperating. If they defect, you score 1 by defecting rather than 0 by cooperating. So D is individually tempting in either case. Yet both players get 3 each from C/C rather than 1 each from D/D.

Notation: Temptation T=5; Reward R=3; Punishment P=1; Sucker’s payoff S=0. This follows T > R > P > S and 2R > T+S. If your real-life situation does not have those incentives, this game may be the wrong model.

2 / THE HUMAN EXPERIMENT

Play twelve rounds on paper

Two players, each with a copy of this page or an opaque sheet. For every round, silently record C or D before revealing. Never choose after seeing the other move. Calculate points after revealing the pair. The payoff matrix from page 1 applies every round.

Rules of engagement

• Agree on the same 12 rounds and the payoff matrix. • Do not communicate choices before revelation. • Keep a separate running total for both players. • Do not treat your partner’s moves as evidence about their morality. • Repeat with one player following a precommitted algorithm.

ROUND LOG / MAKE PHOTOCOPIES

Round	Your C/D	Other C/D	Your points	Other points
1	_____	_____	_____	_____
2	_____	_____	_____	_____
3	_____	_____	_____	_____
4	_____	_____	_____	_____
5	_____	_____	_____	_____
6	_____	_____	_____	_____
7	_____	_____	_____	_____
8	_____	_____	_____	_____
9	_____	_____	_____	_____
10	_____	_____	_____	_____
11	_____	_____	_____	_____
12	_____	_____	_____	_____
TOTAL	—	—	_____	_____

Before interpreting the score, ask: did either player know the last round? Could moves be misunderstood? Were payoffs equal? How might exit options or enforcement change the structure?

3 / SIX STRATEGIES

Choose a rule, not a virtue

A strategy is a conditional instruction. It is not a personality type. Use these in your second playthrough; the paper version has no accidental flips unless you add them explicitly.

STRATEGY	MECHANICAL RULE
Always Cooperate	Choose C every round.
Always Defect	Choose D every round.
Tit for Tat	Begin C, then copy the partner’s previous realized move.
Generous Tit for Tat	Begin C; copy C. After the partner’s D, forgive with probability 1/3. For paper play, use a fair six-sided die: choose C on 1 or 2; D on 3–6.
Grim Trigger	Begin C; after the partner has ever chosen D, choose D in every later round.
Win-Stay, Lose-Shift	Begin C. After you earned 3 or more points last round, repeat your previous realized move; otherwise switch.

Two cautionary counterexamples

1. Retaliation loop. Two Tit-for-Tat players cooperate until one isolated execution error makes a D happen. They can then alternate C/D and D/C. Their motives have not changed; their rules amplify a misunderstanding.

2. Free exploitation. Against Always Cooperate, Always Defect earns 5 on each round while the cooperator receives 0. A strategy can earn a high score in one opponent field and perform poorly in another.

Paper variant: a miscommunication token

Before a round, each player secretly draws a slip numbered 1 through 10. If you are testing 10% noise and draw 1, flip the chosen action when revealing it. Keep the intended action separate from the realized one. Do not change the recorded intention after seeing the draw. For rigorous independence, draw separately for each player each round.

4 / EXPLAINING THE MACHINE

Why repetition matters

A one-shot incentive argument does not simply carry over to repeated interactions. When players care about later encounters, present defection may trigger later retaliation, and today's cooperation may preserve a mutually beneficial relationship. Whether that happens depends on continuation likelihood, information, strategy, and payoffs.

The shadow of the future

For an indefinite repeated game, one can model a fixed probability that interaction continues after each round. More expected future rounds can make preserving cooperation attractive. In a finite game with a known last round and common rationality assumptions, a backward-induction argument instead leads to endgame defection. These settings are not interchangeable.

Two compatible but distinct lenses

Lens A — contingent reciprocity. Axelrod and Hamilton's 1981 analysis and tournaments demonstrated how reciprocal strategies could survive and spread under particular repeated-interaction assumptions. Memory and conditional cooperation matter; tit-for-tat is an instructive rule, not a universal champion.

Lens B — structures and mechanisms. Nowak's 2006 survey reviews several mechanisms of cooperation beyond direct reciprocity: kin selection, indirect reciprocity, networks, and group selection. Which model helps depends on whether encounters are repeated with the same partner, whether reputations travel, whether players are clustered, and how groups interact.

What human experiments add

Dal Bó (2005) provides evidence that continuation incentives can reduce opportunistic play in a laboratory design. Dal Bó and Fréchette (2011) report that cooperation in indefinitely repeated games can emerge, but under conditions stricter than equilibrium possibility alone. These are empirical findings within specified settings; they do not validate the website's particular algorithms or visitor outcomes.

A useful distinction

Established mathematics: payoffs and deterministic strategy definitions, given assumptions. Research: theoretical and experimental papers cited here. Illustration: this page's payoff totals, random errors, and league rankings. Unknown: how an unobserved real person would act in an unmodelled context.

5 / FIELD INVESTIGATION

A six-question audit

Use the prompts below when a group discussion jumps too quickly from “cooperation won” to “people should cooperate more.” Write answers before drawing conclusions.

01. What kind of encounter is this?
One-off, fixed number of encounters, uncertain continuation, or indefinite repeat?

_____

02. Who can see what, when?
Do players observe past actions accurately? Do they know the next move?

_____

03. Who defines the payoffs?
Does mutual cooperation really beat mutual defection, and are risks symmetric?

_____

04. What counts as a mistake?
Accidental action, misunderstood signal, unreliable memory, or deliberately deceptive report?

_____

05. Is retaliation reversible?
What happens after one accidental D? Can a forgiving strategy repair the relationship?

_____

06. What changes beyond the two-person model?
Law, power, reputation, groups, third parties, culture, exclusion, exit, or commitment devices?

_____

A three-experiment sequence

A. Stable signal: Play twelve rounds with no mistakes. Record both totals and how often both chose C. B. Broken signal: Repeat with a 10% flip chance per player, then compare the pattern. C. Changed incentives: Redraw the payoffs, explicitly check whether it remains a Prisoner’s Dilemma, and discuss whether the original interpretations transfer.

Strongest alternative explanation

Before saying why your opponent did something, name two mechanisms that would produce the same observed moves. For example, D could reflect a fixed rule, a single accidental flip, revenge, misunderstood history, or a different perceived payoff structure. The trace alone cannot discriminate among all of them.

6 / EPISTEMIC HONESTY

What this does not prove

The digital experiment does not prove anyone is trustworthy, untrustworthy, rational, selfish, altruistic or good at relationships. It is not a validated survey or a psychological measure. Points represent a toy preference ordering, not wellbeing or real resources.

The strategy league is a deliberately narrow comparison. Its average scores depend on the six chosen opponents, a 40-round match length, this payoff matrix, the random seed, noise, and including self-play. No population reproduction, learning, institution, exit, power imbalance, or real human is modelled. A different environment can reverse relative rankings.

The research trail

1. Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. Science, 211(4489), 1390–1396. DOI: 10.1126/science.7466396.

2. Nowak, M. A. (2006). Five rules for the evolution of cooperation. Science, 314(5805), 1560–1563. DOI: 10.1126/science.1133755.

3. Dal Bó, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. American Economic Review, 95(5), 1591–1604. DOI: 10.1257/000282805775014434.

4. Dal Bó, P. & Fréchette, G. R. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. American Economic Review, 101(1), 411–429. DOI: 10.1257/aer.101.1.411.

5. Stanford Encyclopedia of Philosophy. Prisoner’s Dilemma (especially discussion of finite, infinite and indefinite iteration). https://plato.stanford.edu/entries/prisoner-dilemma/.

Back to the wider inquiry

Explore this game online: cooperationgame.sahilwho.online.
Return to the Sahilverse: sahiltayal.com.
Write to Sahil: sahil@sahiltayal.com.

One last question: when a signal fails, is the first useful repair a change in the people—or in the rules that connect them?