



dergipark.org.tr/tr/pub/alinterisosbil

Sosyal Bilimler Dergisi

Journal of Social Sciences

| e-ISSN 2602-229X |



<https://prensip.gen.tr>

RESEARCH ARTICLE

ARAŞTIRMA MAKALESİ

An Evolutionary Game-Theoretic Perspective on the Prisoner's Dilemma

Mahkûm İkilemine Evrimsel Oyun Teorik Bir Bakış

Aras YOLUSEVER

İstanbul Kültür Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonomi Bölümü, Ekonomi Pr., İstanbul, Türkiye

Geliş Tarihi/Received Date: 11.05.2026

Kabul Tarihi/Accepted Date: 08.06.2026

ABSTRACT

The Prisoner's Dilemma is the canonical model in microeconomics for the tension between individual rationality and collective welfare. Standard game-theoretic analysis predicts that mutual defection is the unique Nash equilibrium, despite being Pareto-dominated by mutual cooperation. This study revisits the Prisoner's Dilemma through the lens of evolutionary game theory, in which strategy frequencies in a population evolve according to differential payoffs. We derive the replicator equation for the one-shot game and show analytically that the all-defectors state is the unique asymptotically stable equilibrium and the unique evolutionarily stable strategy. We then extend the analysis to the infinitely repeated game with continuation probability $w \in (0,1)$, contrasting the reactive strategy Tit-for-Tat (TFT) with Always Defect (ALLD). We derive the critical continuation probability $w^* = (T - R)/(T - P)$ above which TFT becomes an ESS within the restricted two-strategy space (that is, resists invasion by ALLD), the bistable replicator dynamics that emerge in that regime, and a closed-form expression for the unstable interior equilibrium $y^*(w)$ that separates the two basins of attraction. The bifurcation structure is interpreted economically as a precise formalisation of Axelrod's "shadow of the future." All results are derived from first principles and illustrated with numerical examples.

Keywords: Evolutionary Game Theory, Evolutionarily Stable Strategy, Prisoner's Dilemma, Repeated Games.

ÖZ

Mahkûm İkilemi, mikro ekonomide bireysel rasyonellik ile toplumsal refah arasındaki gerilimi modelleyen kanonik bir yapıdır. Standart oyun-teorik analiz, karşılıklı işbirliği tarafından Pareto baskılanmasına rağmen karşılıklı ihanetin tek Nash dengesi olduğunu öngörmektedir. Bu çalışma, mahkûm ikilemini popülasyondaki strateji frekanslarının farklı kazanımlara bağlı olarak evrildiği evrimsel oyun teorisi perspektifinden yeniden ele almaktadır. Tek seferlik oyun için replikatör denklemi türetilmekte ve "tümü-ihanet" durumunun yegâne asimptotik kararlı denge ve yegâne evrimsel kararlı strateji olduğu analitik olarak gösterilmektedir. Analiz, devam etme olasılığı $w \in (0,1)$ ile sonsuz tekrarlı oyunu genişletilerek, tepkisel strateji kısasa kısas ile her zaman ihanet stratejilerini karşılaştırmaktadır. Kısasa kısasın kısıtlanmış iki stratejili uzayda (yani her zaman ihanet stratejisinin istilasına karşı) evrimsel durağan strateji hâline gelmesi için kritik devam etme olasılığı $w^* = (T - R)/(T - P)$ türetilmekte, bu rejimde ortaya çıkan iki kararlı dengeli (bistable) replikatör dinamikleri incelenmekte ve iki çekim havzasını ayıran kararsız iç denge $y^*(w)$ için kapalı-form bir ifade elde edilmektedir. Çatallaşma yapısı, Axelrod'un "geleceğin gölgesi" kavramının kesin bir biçimselleştirmesi olarak ekonomik açıdan yorumlanmaktadır. Tüm sonuçlar ilk ilkelerden türetilmekte ve sayısal örneklerle desteklenmektedir.

Anahtar Kelimeler: Evrimsel Oyun Teorisi, Evrimsel Durağan Strateji, Mahkûmlar İkilemi, Tekrarlı Oyunlar.

Please cite this paper as follows/Atıf için:

Yolusever, A. (2026). An evolutionary game-theoretic perspective on the prisoner's dilemma. *Alinteri Sosyal Bilimler Dergisi*, 10(3), 316-336. <https://doi.org/10.30913/alinterisosbil.1949249>



INTRODUCTION

The Prisoner's Dilemma (PD) occupies a privileged place in microeconomic theory. Since its formalisation by Tucker (1950) in and its systematic treatment by Rapoport and Chammah (1965), the PD has served as the archetypal model for situations in which the pursuit of individually rational strategies produces a collectively suboptimal outcome. The dilemma is sharp: each player has a strictly dominant action (defection) that yields a higher payoff than cooperation against any opponent strategy, yet the resulting equilibrium is Pareto-dominated by the cooperative outcome. The PD therefore captures, in its simplest possible form, the central problem of social cooperation that animates economic theories of market failure, public goods, oligopolistic collusion, and contract enforcement (von Neumann & Morgenstern, 1944).

The classical resolution offered by non-cooperative game theory—mutual defection as the unique Nash equilibrium (Nash, 1950, 1951)—is descriptively unsatisfying. Empirically, cooperation is pervasive in biological and economic systems where the assumption of common-knowledge rationality is implausible. The evolutionary approach, originating in the seminal work of Maynard Smith and Price (1973), replaces the cognitive postulate of rationality with a population-level dynamic of selection: strategies that yield higher payoffs reproduce faster (whether through biological inheritance, imitation, or learning), and the long-run composition of the population is governed by differential equations of motion rather than by introspective best-response calculations (Hofbauer & Sigmund, 1998; Taylor & Jonker, 1978; Weibull, 1995).

This paper provides a self-contained mathematical analysis of the PD from the evolutionary perspective. The classical PD is first reviewed and its Nash equilibrium characterised via strict dominance. The formal apparatus of evolutionary game theory is then introduced, including the evolutionarily stable strategy (ESS) and the replicator equation. This apparatus is subsequently applied to the one-shot PD, where the evolutionary prediction is shown to coincide with the classical one. The analysis is then extended to the infinitely repeated game, deriving the conditions under which Tit-for-Tat becomes evolutionarily stable, the bistable phase portrait that ensues, and the unstable interior equilibrium. The bifurcation structure in the continuation probability is subsequently examined and given an economic interpretation, before closing remarks bring the discussion to a conclusion. It should be emphasised at the outset that the paper advances no new theoretical result. Each of the results assembled below—the replicator dynamics of the one-shot dilemma, the evolutionary-stability threshold for Tit-for-Tat, and the bistability of the repeated game—is well established in the evolutionary game theory literature. The contribution is accordingly expository and synthetic: to gather these results into a single, self-contained treatment derived from first principles, in a uniform notation and with an explicit economic interpretation.

THE CLASSICAL DILEMMA

Setup and Payoff Structure

Consider a symmetric two-player normal-form game with strategy set $\mathcal{S} = \{C, D\}$, where C denotes *cooperate* and D denotes *defect*. The payoff matrix, written from the row player's perspective, is

$$\mathbf{A} = \begin{pmatrix} R & S \\ T & P \end{pmatrix} \quad (1)$$

so that $a_{CC} = R$ (reward), $a_{CD} = S$ (sucker's payoff), $a_{DC} = T$ (temptation), and $a_{DD} = P$ (punishment). The game is a Prisoner's Dilemma if the payoffs satisfy the two conditions

$$T > R > P > S \quad (2.1)$$

$$2R > T + S \quad (2.2)$$

Condition (2.1) ensures that defection is the dominant action, while (2.2) ensures that mutual cooperation is socially preferred to alternating exploitation, which becomes relevant in the iterated version (Section 5). The canonical numerical example, used throughout this study, is $(T, R, P, S) = (5, 3, 1, 0)$ (Axelrod, 1984).

Strict Dominance and Nash Equilibrium

Let $p \in [0, 1]$ denote the probability with which the opponent plays C . The expected payoff to action C is

$$u(C, p) = pR + (1 - p)S \quad (2.3)$$

and the expected payoff to action D is

$$u(D, p) = pT + (1 - p)P \quad (2.4)$$

The payoff *difference* is

$$u(D, p) - u(C, p) = p(T - R) + (1 - p)(P - S) \quad (2.5)$$

Under (2.1), both $T - R > 0$ and $P - S > 0$. Hence (2.5) is a convex combination of strictly positive quantities, so

$$u(D, p) - u(C, p) > 0 \quad \text{for all } p \in [0, 1] \quad (2.6)$$

Defection therefore strictly dominates cooperation. By iterated elimination of strictly dominated strategies, the unique Nash equilibrium of the one-shot PD is (D, D) , yielding the payoff profile (P, P) .

The Social Dilemma

Although (D, D) is the unique Nash equilibrium of the one-shot Prisoner's Dilemma, the alternative outcome (C, C) yields the strictly higher payoff (R, R) to both players, with $R > P$ by the ordering condition $T > R > P > S$. The equilibrium is therefore Pareto-inefficient: there exists an alternative profile that makes both players strictly better off without making either worse off, and yet this profile is not sustained by individually rational play. This is the formal expression of the dilemma at the heart of the model. Rational individual optimisation, far from delivering the collectively best

outcome, leads in a strategically interdependent environment to a profile that both parties would unanimously prefer to abandon if only they could credibly commit to doing so jointly. The paradox is sharpened by the observation that each player's reasoning is impeccable in isolation: defection strictly dominates cooperation, regardless of the opponent's choice, so neither player can be faulted for selecting it. The pathology lies not in any agent's deliberation but in the structure of the payoff matrix itself, which decouples private incentive from social value.

This tension between private rationality and collective welfare is the conceptual seed from which a vast body of microeconomic and political-economic theory has grown. It motivates the analysis of repeated interaction, in which the prospect of future retaliation can sustain cooperation that is unattainable in a single encounter; the study of formal and informal enforcement mechanisms, by which external sanctions realign individual incentives with the social optimum; the design of contracts, property rights, and regulatory institutions that internalise externalities; and the broader literature on the governance of common-pool resources, public goods, and collective action problems (Axelrod, 1984; Ostrom, 1990). In each of these strands, the Prisoner's Dilemma serves both as a diagnostic—identifying precisely where the market or the unaided strategic interaction will fail—and as a benchmark against which proposed remedies must be measured. The evolutionary analysis that follows takes this diagnostic as its point of departure and asks a different question: whether selection pressures operating on strategy frequencies in a population can, under specifiable conditions, accomplish what individual rationality alone cannot.

FOUNDATIONS of EVOLUTIONARY GAME THEORY

Population Framework

Evolutionary game theory replaces the assumption of individually rational players with that of a large population of agents, each programmed to play a particular (pure or mixed) strategy. Let $\mathbf{x} = (x_1, \dots, x_n) \in \Delta^{n-1}$ denote the population state, where x_i is the frequency of strategy i and $\Delta^{n-1} = \{\mathbf{x} \in \mathbb{R}_+^n : \sum_i x_i = 1\}$ is the $(n-1)$ -dimensional simplex. Agents are matched pairwise at random, and the expected payoff to a player using strategy i in population state $\mathbf{x}$ is

$$f_i(\mathbf{x}) = \sum_{j=1}^n a_{ij} x_j = (\mathbf{Ax})_i \quad (3.1)$$

where a_{ij} is the payoff to strategy i when matched against strategy j . The average payoff in the population is

$$\bar{\phi}(\mathbf{x}) = \sum_{i=1}^n x_i f_i(\mathbf{x}) = \mathbf{x}^\top \mathbf{Ax} \quad (3.2)$$

Evolutionary Stable Strategy

A strategy $\sigma^* \in \Delta^{n-1}$ is an evolutionarily stable strategy (ESS) if for every $\sigma \neq \sigma^*$ there exists $\bar{\varepsilon} \in (0,1)$ such that, for all $\varepsilon \in (0, \bar{\varepsilon})$,

$$\sigma^{*\top} \mathbf{A}[\varepsilon \sigma + (1 - \varepsilon) \sigma^*] > \sigma^\top \mathbf{A}[\varepsilon \sigma + (1 - \varepsilon) \sigma^*] \quad (3.3)$$

The intuition is that of an invasion-proofness condition. Imagine a population whose members all play the incumbent strategy σ^* , and suppose that a small fraction ε of mutants playing some alternative strategy σ appears in the population. The post-invasion

environment is the mixed population $\varepsilon\sigma + (1 - \varepsilon)\sigma^*$. The left-hand side of (3.3) is the expected payoff to an incumbent encountering a random member of this mixed environment; the right-hand side is the expected payoff to a mutant in the same environment. The inequality demands that, for every conceivable mutant strategy and every sufficiently small invasion size, the incumbent earns strictly more than the invader. Under any selection process that favours higher-earning strategies—biological reproduction, behavioural imitation, reinforcement learning—the mutants will therefore decline in frequency and the population will return to σ^* . The ESS concept thus captures the robustness of a strategy against evolutionary perturbation, without invoking any rationality or deliberation on the part of the agents.

The standard equivalent characterisation, due to Maynard Smith (1982), expresses (3.3) as a pair of conditions that are typically easier to verify in concrete games. First, the Nash condition:

$$\sigma^{*\top} A \sigma^* \geq \sigma^\top A \sigma^* \text{ for all } \sigma \in \Delta^{n-1} \quad (3.3a)$$

This says that σ^* is a best response to itself: no alternative strategy earns strictly more than σ^* when matched against a population playing σ^* . Every ESS is therefore a (symmetric) Nash equilibrium of the underlying game. The converse, however, does not hold, which is why a second requirement is needed. The stability condition addresses the borderline case in which an alternative strategy ties the incumbent against the incumbent itself:

$$\text{if } \sigma^{*\top} A \sigma^* = \sigma^\top A \sigma^*, \text{ then } \sigma^{*\top} A \sigma > \sigma^\top A \sigma. \quad (3.3b)$$

In words, if a mutant earns the same as the incumbent against the incumbent population, then the incumbent must earn strictly more than the mutant when matched against the mutant itself. The two conditions together ensure that any neutral mutation that survives the first hurdle is selected against in the second. The Nash condition alone is sometimes called weak evolutionary stability, while the conjunction of (3.3a) and (3.3b) defines strict or standard evolutionary stability. The hierarchy of solution concepts is therefore $\{\text{ESS}\} \subseteq \{\text{symmetric Nash equilibria}\}$, with the inclusion typically strict.

Two further remarks are worth recording. First, a strict Nash equilibrium—one in which $\sigma^{*\top} A \sigma^* > \sigma^\top A \sigma^*$ for all $\sigma \neq \sigma^*$ —is automatically an ESS, because the strict inequality in (3.3a) makes the stability condition (3.3b) vacuous. This will prove useful when we examine the one-shot Prisoner's Dilemma, where the all-defectors state is a strict Nash equilibrium and hence an ESS without any further computation. Second, ESS need not exist for every game, and when they do exist they need not be unique; the iterated Prisoner's Dilemma analysed later in this study furnishes an example with two coexisting ESS, giving rise to the bistable dynamics that constitute one of the central findings of the paper.

The Replicator Equation

The ESS concept is static: it tells us which strategy compositions are immune to invasion but says nothing about how a population away from such a composition evolves over time. The dynamic counterpart of the ESS is the replicator equation of Taylor and Jonker (1978):

$$\dot{x}_i = x_i[f_i(x) - \bar{\phi}(x)], i = 1, \dots, n. \quad (3.4)$$

Equation (3.4) admits a transparent interpretation. The quantity $\dot{x}_i/x_i$ is the per-capita growth rate of strategy i , and the bracketed expression on the right-hand side is the gap between the payoff that strategy i currently earns, $f_i(x)$, and the mean payoff in the population, $\bar{\phi}(x)$. Strategies that out-perform the population average therefore grow exponentially in relative frequency, while those that under-perform shrink, with the rate of change proportional to both the current frequency and the magnitude of the payoff differential. The equation thus formalises a Darwinian principle: differential reproduction follows differential fitness. Although the canonical interpretation is biological—strategies are heritable phenotypes and payoffs translate into reproductive success—the same equation arises in economic and social settings in which agents periodically revise their strategies by imitating more successful peers, with imitation rates proportional to observed payoff differences (Sandholm, 2010; Schlag, 1998).

Several mathematical properties of (3.4) deserve emphasis, because they structure all the analysis that follows. The first is that the unit simplex $\Delta^{n-1} = \{x \in \mathbb{R}_+^n : \sum_i x_i = 1\}$ is *forward-invariant* under the replicator flow. To see this, sum (3.4) over i :

$$\sum_{i=1}^n \dot{x}_i = \sum_{i=1}^n x_i f_i(x) - \bar{\phi}(x) \sum_{i=1}^n x_i = \bar{\phi}(x) - \bar{\phi}(x) = 0 \quad (3.5)$$

so the total mass $\sum_i x_i$ is conserved along trajectories, and any initial condition on the simplex remains on the simplex for all time. The second property is that each boundary face $\{x_i = 0\}$ is itself invariant: if strategy i is initially absent from the population, then $\dot{x}_i = 0$ at that face, so it remains absent forever. The replicator equation cannot, by itself, introduce new strategies—it only redistributes mass among those already present. This is a substantive limitation: cooperation cannot emerge from a population of pure defectors under (3.4) alone, even when cooperation would be evolutionarily stable as an alternative equilibrium. Mechanisms such as mutation, migration, or stochastic finite-population effects (formalised in the Moran process, a stochastic birth–death model of a finite population in which, at each step, one individual is selected to reproduce with probability proportional to its payoff and its single offspring replaces another individual chosen at random) are required to break this invariance, a point that will recur in the discussion of the iterated game.

The third property is the substantive bridge between the dynamic equation and the static ESS concept, sometimes called the folk theorem of evolutionary game theory (Hofbauer & Sigmund, 1998). It states that every ESS is an asymptotically stable rest point of the replicator equation. That is, if σ^* satisfies the invasion-proofness condition (3.3), then any small perturbation of the population away from σ^* decays back to σ^* under the replicator flow. The converse, importantly, is false: there exist asymptotically stable rest points of (3.4) that are not ESS, and even Nash equilibria that fail to be rest points at all when they lie on the interior of a face on which the dynamics

push outward. The relationship between the two solution concepts is therefore one of sufficiency without necessity: ESS is a stronger condition than asymptotic stability, and the replicator dynamics admit a richer phase portrait than the ESS concept alone reveals. For the purposes of this study, however, the implication runs the right way: when we identify an ESS in the one-shot or iterated PD, the folk theorem guarantees that the corresponding population state is dynamically stable, so the evolutionary and dynamical analyses reinforce each other rather than diverging.

Finally, it bears noting that the replicator equation is not the only payoff-monotonic dynamic available. The broader class of *imitation dynamics*, *best-response dynamics*, *logit dynamics*, and *adjustment dynamics* studied in the evolutionary game theory literature (Sandholm, 2010; Weibull, 1995) all share the qualitative feature that better-performing strategies grow in frequency, but they differ in the precise functional form of the growth rate. Many of the qualitative conclusions drawn from (3.4)—in particular, the existence and stability of pure-strategy equilibria in dominance-solvable games like the PD—are robust across this entire class, which lends the analysis a degree of generality beyond the specific functional form chosen. The replicator equation is used here because it is the simplest, most widely studied, and most directly connected to the ESS concept; the more elaborate dynamics serve as a reminder that the qualitative findings do not hinge on the particular dynamical specification.

REPLICATOR DYNAMICS in the ONE-SHOT PRISONER'S DILEMMA

The One-Dimensional Replicator Equation and Fixed Points

Let $x \in [0,1]$ denote the frequency of cooperators in the population, so $1 - x$ is the frequency of defectors. From (3.1),

$$f_C(x) = xR + (1 - x)S \quad (4.1)$$

$$f_D(x) = xT + (1 - x)P \quad (4.2)$$

The average payoff is $\bar{\phi}(x) = xf_C(x) + (1 - x)f_D(x)$. Substituting into (3.4) and using the identity $f_C - \bar{\phi} = (1 - x)(f_C - f_D)$, the replicator equation reduces to the one-dimensional form

$$\dot{x} = x(1 - x)[f_C(x) - f_D(x)] \quad (4.3)$$

Computing the payoff difference,

$$f_C(x) - f_D(x) = x(R - T) + (1 - x)(S - P) \quad (4.4)$$

By (2.1), $R - T < 0$ and $S - P < 0$, so the right-hand side of (4.4) is a convex combination of strictly negative numbers. Therefore

$$f_C(x) - f_D(x) < 0 \quad \text{for all } x \in [0,1] \quad (4.5)$$

A state x^* is a fixed point of (4.3) if $\dot{x}|_{x=x^*} = 0$. From (4.3) this requires $x^*(1 - x^*) = 0$ or $f_C(x^*) = f_D(x^*)$. The latter is impossible by (4.5). Hence the only fixed points are

$$x_1^* = 0 \quad \text{and} \quad x_2^* = 1 \quad (4.6)$$

Stability Analysis

Having established that the one-dimensional replicator equation $\dot{x} = F(x)$ for the one-shot Prisoner's Dilemma admits exactly two fixed points, $x_1^* = 0$ and $x_2^* = 1$, the question of their local stability arises naturally. A fixed point is asymptotically stable if trajectories that begin sufficiently close to it converge back to it as $t \rightarrow \infty$, and unstable if at least some such trajectories diverge from it. For one-dimensional autonomous systems of the form $\dot{x} = F(x)$, the standard tool for resolving stability is *linearisation*: one computes the derivative $F'(x^*)$ at the fixed point and reads off the sign. If $F'(x^*) < 0$, the fixed point is asymptotically stable; if $F'(x^*) > 0$, it is unstable; and if $F'(x^*) = 0$, the linearisation is inconclusive and higher-order terms or global arguments must be invoked.

Let us write the right-hand side of (4.3) as

$$F(x) := x(1-x)[f_C(x) - f_D(x)],$$

and apply the product rule to differentiate:

$$F'(x) = (1-2x)[f_C(x) - f_D(x)] + x(1-x) \frac{d}{dx}[f_C(x) - f_D(x)]. \quad (4.7)$$

The second term vanishes at both $x = 0$ and $x = 1$ because of the factor $x(1-x)$, which simplifies the analysis considerably. Evaluating at the first fixed point,

$$F'(0) = (1)[f_C(0) - f_D(0)] + 0 = S - P \quad (4.8)$$

By the PD ordering condition $T > R > P > S$, we have $S - P < 0$, and therefore $F'(0) < 0$. The fixed point $x_1^* = 0$ is consequently asymptotically stable: small invasions of cooperators into an all-defector population decay back to zero under the replicator flow. Economically, this is the formal expression of the impossibility of cooperative invasion: a rare cooperator earns S when matched against the prevailing defector, while the defector earns P against its own kind, and since $P > S$, the cooperator under-earns and its frequency declines.

Evaluating at the second fixed point,

$$F'(1) = (1-2)[f_C(1) - f_D(1)] + 0 = -[R - T] = T - R \quad (4.9)$$

By the same ordering condition, $T - R > 0$, and therefore $F'(1) > 0$. The fixed point $x_2^* = 1$ is unstable: small invasions of defectors into an all-cooperator population grow under the replicator flow, and the population diverges from the cooperative state. The logic is the mirror image of the previous case. A rare defector matched against the cooperator majority earns T , while a cooperator matched against another cooperator earns R , and since $T > R$, the defector out-earns the incumbent population and its frequency rises.

Although the local linearisation establishes the stability of the two endpoints, a stronger global conclusion is available because the system is one-dimensional and F has no interior zeros. Recall from (4.5) that $f_C(x) - f_D(x) < 0$ for every $x \in [0,1]$, and that $x(1-x) > 0$ for every $x \in (0,1)$. Consequently,

$$F(x) = x(1-x)[f_C(x) - f_D(x)] < 0 \text{ for all } x \in (0,1) \quad (4.10)$$

The flow on the open interval is therefore *strictly* leftward, with no rest points to interrupt the motion. Every trajectory beginning at any interior initial condition $x_0 \in (0,1)$ is monotonically decreasing in t and, being bounded below by zero, must converge to a limit. That limit can only be a fixed point of F , and the only fixed point reachable from the interior is $x = 0$. Hence

$$\lim_{t \rightarrow \infty} x(t) = 0 \text{ for every } x_0 \in [0,1).$$

The all-defectors state is therefore not merely locally but globally asymptotically stable on $[0,1)$: its basin of attraction is the entire phase space minus the unstable endpoint. The unstable fixed point $x_2^* = 1$ is a knife-edge equilibrium with measure-zero basin, reachable only from itself.

This global conclusion strengthens the result of the linearisation in an important sense. Local stability alone would only assure us that a population already near pure defection remains there; the global result rules out the possibility of cooperative attractors—limit cycles, alternative interior equilibria, or any other dynamical structure that might support partial cooperation. The phase line of the one-shot PD is dynamically trivial: a single flow from the unstable repeller at $x = 1$ to the stable attractor at $x = 0$, with no intermediate structure. Whatever the initial mix of cooperators and defectors, evolution erodes cooperation monotonically until it is extinguished entirely.

It is also worth noting that this stability analysis is fully consistent with—and indeed implied by—the ESS analysis sketched in §4.5: defection is a strict Nash equilibrium, hence a strict ESS, and by the folk theorem of evolutionary game theory every ESS is an asymptotically stable rest point of the replicator equation. The linearisation calculation above can therefore be viewed as an explicit verification of the general theorem in this particular case. The convergence is unconditional in the sense that no assumption on the initial conditions, beyond $x_0 < 1$, is required; the dynamical conclusion does not depend on the population being "close enough" to defection at the outset.

Phase Portrait

Figure 1 displays $\dot{x}$ as a function of x for the canonical parameters $(T, R, P, S) = (5, 3, 1, 0)$. The curve lies strictly below the horizontal axis on $(0,1)$, vanishes at the two endpoints, and the flow arrows confirm monotone convergence to the all-defectors state.

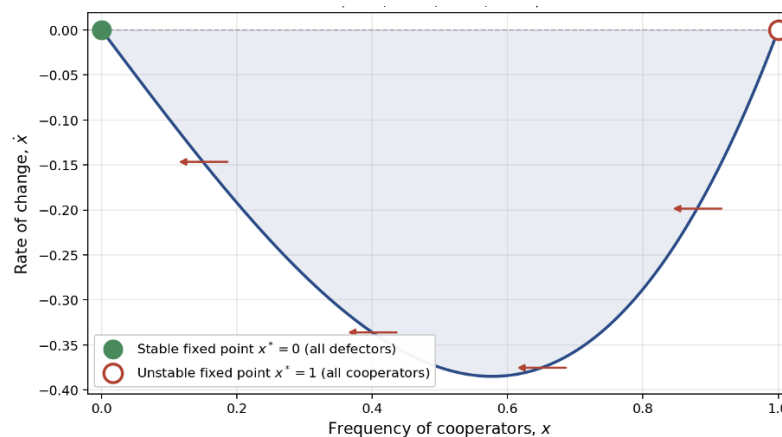


Figure 1. Replicator dynamics of the one-shot Prisoner Dilemma.

The all-defectors state is the unique ESS of the one-shot PD: defection strictly dominates cooperation, so $\sigma^* = D$ satisfies condition (3.3) with $\bar{\epsilon} = 1$. The evolutionary prediction therefore coincides exactly with the classical rational-choice prediction. This is a striking result: an analysis that dispenses with rationality, common knowledge, and best-response computation nevertheless arrives at the same equilibrium. The reason is structural rather than cognitive—defection earns more than cooperation against *any* opponent, so it earns more on average in *any* mixed population, and therefore expands under any payoff-monotonic dynamic. Cooperation, in the one-shot game, has no evolutionary foothold.

COOPERATION THROUGH REPETITION: THE ITERATED PD

Infinitely Repeated Framework

The one-shot analysis just completed shows that cooperation has no evolutionary foothold when each encounter is isolated. Most economically significant interactions, however—trading relationships, oligopolistic competition, the enforcement of informal norms—recur over time, and it is precisely this recurrence that can make the threat of future retaliation credible. The natural next step is therefore to ask whether embedding the stage game in a repeated interaction can restore the possibility of cooperation that the single encounter forecloses. Suppose the stage game of the previous section is played in periods $t = 0, 1, 2, \dots$, with a per-period continuation probability $w \in (0, 1)$. After each round, the interaction continues to the next round with probability w and terminates with probability $1 - w$. The expected number of rounds is $1/(1 - w)$, and the expected discounted payoff of a stream $\{u_t\}_{t \geq 0}$ is $\sum_{t \geq 0} w^t u_t$ (Axelrod & Hamilton, 1981).

Because a strategy in a repeated game may, in principle, condition its current action on the entire prior history of play, the full strategy space is uncountably infinite and admits no closed-form evolutionary analysis. We therefore restrict attention to two focal reactive strategies, the most extensively studied conditional cooperator, Tit-for-Tat, and its unconditional counterpart, always defect. This minimal two-strategy system is chosen because it already exhibits the qualitative phenomenon of interest—the coexistence of a stable cooperative equilibrium alongside the defection equilibrium—while remaining analytically tractable on the one-dimensional simplex; the restriction is expository rather than substantive, since the threshold and bistability results derived below are known to persist, in modified form, under richer strategy sets (see the concluding discussion):

ALLD (Always Defect): play D in every round.

TFT (Tit-for-Tat): play C in the first round; in every subsequent round, play the opponent's previous action (Axelrod, 1984; Rapoport & Chammah, 1965).

Under random matching of TFT against TFT, both players cooperate in every round, yielding the discounted total

$$V(\text{TFT}, \text{TFT}) = \sum_{t=0}^{\infty} w^t R = \frac{R}{1-w} \quad (5.1)$$

TFT against ALLD: TFT cooperates in round 0 and receives S ; from round 1 onward TFT plays D (mimicking the opponent's previous defection) and both defect, yielding

$$V(\text{TFT}, \text{ALLD}) = S + \sum_{t=1}^{\infty} w^t P = S + \frac{wP}{1-w} \quad (5.2)$$

By symmetry,

$$V(\text{ALLD}, \text{TFT}) = T + \frac{wP}{1-w}, \quad V(\text{ALLD}, \text{ALLD}) = \frac{P}{1-w} \quad (5.3)$$

The induced payoff matrix of the repeated game is

$$\mathbf{B}(w) = \begin{pmatrix} \frac{R}{1-w} & S + \frac{wP}{1-w} \\ T + \frac{wP}{1-w} & \frac{P}{1-w} \end{pmatrix} \quad (5.4)$$

Evolutionary Stability of TFT and ALLD

With the induced payoff matrix $\mathbf{B}(w)$ of the indefinitely repeated game in hand, we now ask which of the two pure strategies—TFT or ALLD—qualifies as an evolutionarily stable strategy in the sense of (3.3). Throughout, evolutionary stability is to be understood relative to the restricted strategy set $\{\text{TFT}, \text{ALLD}\}$: the question is whether each strategy resists invasion by the other, not whether it is an ESS against the full space of repeated-game strategies. Because each pure strategy is a strict Nash equilibrium of the repeated game whenever the corresponding diagonal entry of $\mathbf{B}(w)$ strictly exceeds the off-diagonal entry in the same column, the analysis reduces in each case to a single algebraic inequality. We treat the two candidates in turn.

TFT is an ESS against ALLD invasion if and only if a TFT incumbent earns strictly more against a TFT opponent than an ALLD invader would earn against the same TFT opponent. In matrix terms this is the condition $b_{11} > b_{21}$, which from (5.4) reads

$$\frac{R}{1-w} > T + \frac{wP}{1-w} \quad (5.5)$$

The left-hand side is the long-run discounted payoff of mutual cooperation, $R/(1-w)$; the right-hand side decomposes into the one-period temptation gain T from defecting against a cooperator in round 0, plus the discounted stream $wP/(1-w)$ of mutual-punishment payoffs that ensues from round 1 onward, once the TFT opponent has detected the defection and switched to permanent retaliation. The inequality therefore expresses, in an exact form, the trade-off at the heart of repeated-game cooperation: short-term gain from defection versus long-term loss from triggered punishment.

To convert (5.5) into a clean condition on w , multiply both sides by the strictly positive quantity $(1-w) > 0$, preserving the direction of the inequality:

$$R > T(1-w) + wP \quad (5.5a)$$

Expanding the right-hand side gives $R > T - wT + wP$, and rearranging to isolate the terms containing w on one side yields

$$w(T - P) > T - R \quad (5.5b)$$

Finally, since the PD axiom $T > P$ ensures that the coefficient $T - P$ is strictly positive, we may divide both sides by $T - P$ without reversing the inequality, arriving at the celebrated threshold condition

$$w > w^* := \frac{T-R}{T-P} \quad (5.6)$$

The critical continuation probability w^* is one of the foundational quantities of evolutionary cooperation theory, first identified explicitly by Axelrod and Hamilton (1981) and subsequently rediscovered in numerous economic and biological settings. For the canonical Axelrod parameters $(T, R, P, S) = (5, 3, 1, 0)$, one obtains $w^* = (5 - 3)/(5 - 1) = 2/4 = 1/2$, so cooperation supported by TFT is evolutionarily viable as soon as the expected number of remaining rounds, $1/(1 - w)$, exceeds two. The economic interpretation is that the numerator $T - R$ measures the *temptation premium*—the immediate gain from unilateral defection against a cooperator—while the denominator $T - P$ measures the *amplitude* of the conditional payoff structure, that is, the gap between the temptation payoff and the punishment payoff that retaliation imposes. The ratio is the smallest weight on the future that makes the deferred punishment outweigh the immediate temptation. When relationships are too short—when w falls below this critical value—TFT loses its evolutionary footing, and the iterated game collapses dynamically to the same defection trap as the one-shot game.

ALLD is an ESS against TFT invasion if and only if the symmetric condition $b_{22} > b_{12}$ holds, i.e., an ALLD incumbent earns strictly more against an ALLD opponent than a TFT invader would earn against the same ALLD opponent. From (5.4) this becomes

$$\frac{P}{1-w} > S + \frac{wP}{1-w} \quad (5.7)$$

The interpretation here is the converse of the previous case. The left-hand side is the discounted stream of mutual-punishment payoffs that ALLD incumbents earn against one another, $P/(1 - w)$. The right-hand side is what a TFT invader would obtain against an ALLD population: the sucker's payoff S in round 0, when TFT extends an unrequited offer of cooperation, followed by the discounted stream of mutual-punishment payoffs $wP/(1 - w)$ once TFT has detected the defection and switched to retaliation. The inequality asks whether the punishment stream that the incumbent earns from round 0 is strictly larger than the sucker's payoff plus the discounted punishment stream that the invader earns starting one round later.

To simplify, subtract $wP/(1 - w)$ from both sides of (5.7):

$$\frac{P}{1-w} - \frac{wP}{1-w} > S \quad (5.7a)$$

which collapses to

$$\frac{P(1-w)}{1-w} = P > S \quad (5.7b)$$

The condition $P > S$ is one of the defining PD axioms (2.1) and is satisfied by assumption throughout the analysis. We have therefore established the following striking conclusion: ALLD is an ESS for every $w \in (0, 1)$.

In other words, the stability of universal defection against TFT invasion does not depend on the continuation probability at all. No matter how long the expected relationship—indeed, even in the limit $w \rightarrow 1$ where the game becomes effectively infinitely repeated—a population of defectors cannot be invaded by a small fraction of TFT players. The economic reason is intuitive: a lone TFT player in a sea of defectors pays the sucker's penalty S in the very first round, and although it learns from this and switches to defection from round 1 onward, it never recoups the round-0 loss. The incumbent defector, by contrast, never pays this start-up cost. As long as $P > S$ —as long, that is, as mutual defection is strictly better than being cheated—the start-up cost is a permanent disadvantage to the invader, regardless of how many future rounds remain in which to amortise it.

The bistability of the super-critical regime (bistability being the property whereby the dynamics possess two locally stable equilibria, here universal cooperation and universal defection, each attracting an open set of initial conditions and separated by an unstable interior threshold) is the central conceptual achievement of evolutionary repeated-game analysis. It rescues cooperation from the impossibility verdict of the one-shot game without rendering it automatic. Two communities playing the same game, with the same long horizon, and the same payoffs, can therefore inhabit qualitatively different long-run states depending on the historical accident of their initial strategy mix. This co-existence of two stable orders—one cooperative, one antagonistic—prefigures the bifurcation analysis of §6 and provides the formal microfoundation for the empirical observation that cooperative norms, once lost in a population, are difficult to restore even when they would be sustainable were they somehow re-established.

A final remark concerns the role of the stability condition (3.3b) versus the Nash condition (3.3a). In both the TFT and the ALLD analyses above, the relevant inequalities ((5.5) and (5.7)) are strict, which means that TFT and ALLD are *strict* Nash equilibria of the induced repeated game in their respective regimes of stability. By the observation recorded in §3.2, every strict Nash equilibrium is automatically an ESS, and the stability condition (3.3b) is vacuous. The evolutionary stability of pure-strategy equilibria in the iterated PD is therefore robust in the strongest sense: not only is no mutant strictly more profitable, but every mutant is strictly less profitable. This robustness will be reflected in the replicator dynamics analysed next, where both pure strategies are not merely stable but *exponentially* attracting in their respective basins.

Replicator Dynamics

Let $y \in [0,1]$ denote the frequency of TFT. Because only two strategies are present, the general n -strategy replicator system (3.4) again collapses onto the one-dimensional simplex parametrised by the single frequency y , and the dynamics reduce to the same scalar form obtained for the one-shot game in Section 4.1. By analogy with (4.3),

$$\dot{y} = y(1-y)[f_{\text{TFT}}(y) - f_{\text{ALLD}}(y)] \quad (5.8)$$

Define the *invasion differentials* at each pure-strategy edge:

$$\Delta_1 := b_{11} - b_{21} = \frac{R-T+w(T-P)}{1-w} = \frac{(T-P)(w-w^*)}{1-w} \quad (5.9)$$

$$\Delta_0 := b_{12} - b_{22} = S - P < 0 \quad (5.10)$$

Note that $\Delta_1 > 0$ iff $w > w^*$. The payoff difference is linear in y :

$$f_{\text{TFT}}(y) - f_{\text{ALLD}}(y) = y\Delta_1 + (1 - y)\Delta_0 \quad (5.11)$$

The Interior Equilibrium and Bistability

Setting (5.11) equal to zero and solving for y yields a unique interior equilibrium whenever Δ_0 and Δ_1 have opposite signs—precisely the bistable case $w > w^*$:

$$y^*(w) = \frac{-\Delta_0}{\Delta_1 - \Delta_0} = \frac{P - S}{\Delta_1 + (P - S)} \quad (5.12)$$

Substituting (5.9),

$$y^*(w) = \frac{(P - S)(1 - w)}{(P - S)(1 - w) + (T - P)(w - w^*)} \quad (5.13)$$

For the canonical parameters and $w = 0.7$,

$$y^*(0.7) = \frac{(1)(0.3)}{(1)(0.3) + (4)(0.2)} = \frac{0.3}{1.1} \approx 0.2727.$$

Stability of the three fixed points. Linearising the right-hand side of (5.8) at each fixed point gives

$$F'(0) = \Delta_0 < 0 \quad (\text{stable}) \quad (5.14)$$

$$F'(1) = -\Delta_1 < 0 \quad \text{for } w > w^* \quad (\text{stable}) \quad (5.15)$$

$$F'(y^*) = y^*(1 - y^*)(\Delta_1 - \Delta_0) > 0 \quad (\text{unstable}) \quad (5.16)$$

The interior equilibrium y^* is therefore a *tipping point*: trajectories with $y_0 < y^*$ converge to $y = 0$ (universal defection), and trajectories with $y_0 > y^*$ converge to $y = 1$ (universal cooperation). Figure 2 illustrates the phase portrait for $w = 0.7$.

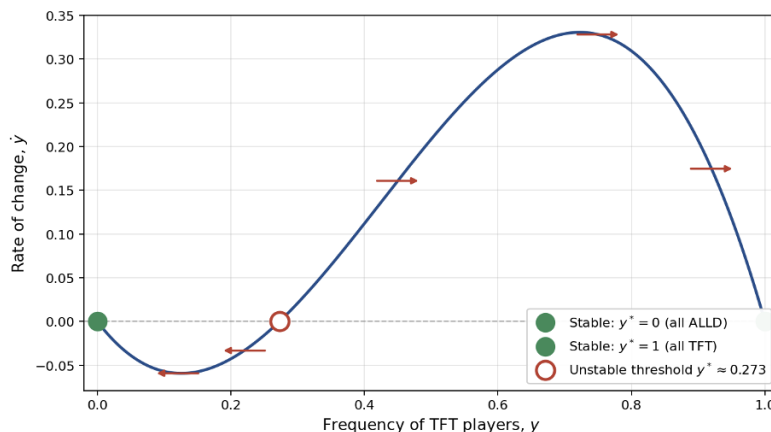


Figure 2. Replicator dynamics in the iterated PD.

The replacement of the one-shot game by an indefinitely repeated game thus alters the strategic landscape qualitatively. The introduction of the future (parametrised by w) creates the possibility of conditional behaviour: TFT punishes defection by withdrawing future cooperation, and this prospect of retaliation can deter opportunism. Mathematically, this is reflected in the sign change of Δ_1 as w crosses w^* . Below the threshold, defection still strictly dominates; above the threshold, the game becomes a

coordination game in the (TFT, ALLD) strategy space, with two basins of attraction whose boundary is given by (5.13).

BIFURCATION STRUCTURE and COMPARATIVE STATICS

The Bifurcation Diagram

A *bifurcation* is a qualitative change in the structure of a dynamical system—in the number, location, or stability of its fixed points—produced by the smooth variation of a parameter, here the continuation probability w . Equation (5.13) defines a continuous function $y^*: (w^*, 1) \rightarrow (0, 1)$ that describes the location of the unstable interior equilibrium as a function of the continuation probability. Two limit results follow immediately from (5.13):

$$\lim_{w \rightarrow w^*+} y^*(w) = 1, \quad \lim_{w \rightarrow 1^-} y^*(w) = 0 \quad (6.1)$$

At the critical threshold $w = w^*$, the interior equilibrium collides with $y = 1$, and the TFT equilibrium loses its stability via a *transcritical bifurcation*. A transcritical bifurcation is a qualitative change in which two fixed points meet at a critical parameter value and exchange stability, both branches persisting on either side of the crossing. The bifurcation here is of precisely this type: as w crosses w^* , the interior equilibrium $y^*(w)$ collides with the boundary fixed point $y = 1$; the linearised eigenvalue $F'(1) = -\Delta_1$ changes sign (since, by (5.9), $\Delta_1 > 0$ if and only if $w > w^*$), while $F'(y^*)$ carries the opposite sign throughout, so the two fixed points pass through one another and exchange stability rather than appearing or annihilating in a pair—the signature that distinguishes a transcritical from a saddle-node bifurcation. As w increases beyond w^* , the threshold y^* retreats toward zero, expanding the basin of attraction of cooperation. In the limit $w \rightarrow 1$ (no termination), an arbitrarily small initial seed of cooperators suffices for the population to converge to universal cooperation.

Differentiating (5.13) confirms that y^* is monotonically decreasing in w :

$$\frac{dy^*}{dw} = -\frac{(P-S)(T-S)}{[(P-S)(1-w) + (T-P)(w-w^*)]^2} < 0 \quad (6.2)$$

Figure 3 displays the bifurcation diagram.

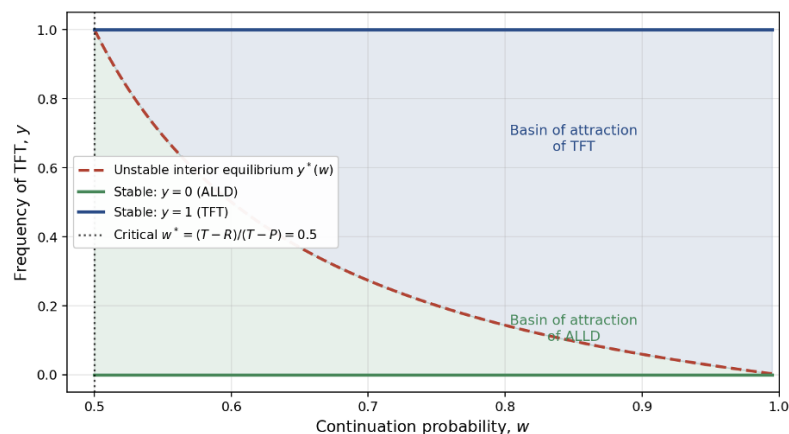


Figure 3. Bifurcation structure of the iterated PD.

The critical condition $w > (T - R)/(T - P)$ has a transparent economic reading. The numerator $T - R$ is the one-period gain from unilateral defection against a cooperator—the *temptation premium*. The denominator $T - P$ is the difference between this temptation and the punishment payoff—the *amplitude* of the conditional payoff structure. The ratio is the smallest continuation probability for which the expected future loss from triggering retaliation (i.e., a sequence of mutual punishment payoffs P instead of mutual rewards R) outweighs the one-shot gain. Cooperation is therefore evolutionarily sustainable precisely when the *shadow of the future* is long enough relative to the immediate incentive to defect (Axelrod, 1984; Fudenberg & Maskin, 1986).

This delivers a quantitative microfoundation for several stylised facts in microeconomic theory: cooperation tends to be more robust in long-lived business relationships than in one-off transactions; collusive arrangements in repeated oligopoly are sustainable only when firms are sufficiently patient; and contractual norms unravel near a known terminal date, when w effectively collapses to zero.

The Role of Initial Conditions

A second, and less frequently emphasised, lesson of the bifurcation analysis concerns the asymmetry between sustainability and emergence. The evolutionary stability of cooperation in the super-critical regime $w > w^*$ is a statement about the resilience of a cooperative population to small perturbations: once the population has arrived at the TFT basin and is composed predominantly of conditional cooperators, no minority of defecting invaders can dislodge it under the replicator flow. This is the content of (5.6) and the asymptotic stability of $y = 1$ established in (5.15). What this statement does not assert, however, is that a population starting from arbitrary initial conditions will find its way to the cooperative basin. Above the threshold w^* the dynamics are bistable, meaning that the phase space $[0, 1]$ is partitioned into two disjoint basins of attraction by the unstable interior equilibrium $y^*(w)$: trajectories starting with $y_0 > y^*$ converge to universal cooperation, while trajectories starting with $y_0 < y^*$ converge to universal defection. A population dominated by defectors—any population with an initial TFT frequency below the threshold y^* —does not spontaneously transition to the cooperative basin under the replicator flow alone, because $y = 0$ is itself locally asymptotically stable. Defection, once entrenched, perpetuates itself even when a strictly preferable cooperative equilibrium exists in the same dynamical system.

This observation has substantive consequences for the theory and the practice of cooperation, and it deserves to be made precise. The replicator equation (3.4) is deterministic and path-dependent: the long-run state is a function of the initial condition, and there is no mechanism within the equation itself that allows a trajectory to escape the basin in which it begins. The boundaries $y = 0$ and $y = 1$ of the simplex are invariant under (3.4), which means that a strategy initially absent from the population remains absent forever; the equation cannot, by its own structure, introduce strategies that are not already present. In a population of pure defectors, therefore, no level of w , however close to unity, can rescue cooperation under the replicator flow alone. The dynamical bistability is real but its biased: the system possesses two attractors but only one of them can be reached from the historically common starting point of a non-cooperative society. The folk theorem of evolutionary game theory guarantees that cooperation is evolutionarily

stable, but it makes no parallel guarantee that cooperation is evolutionarily reachable. Reachability and stability are independent properties, and the bistable structure of the iterated PD separates them sharply.

Some mechanism beyond the bare replicator equation is therefore required to bootstrap cooperation in a previously defecting population—to cross the threshold y^* from below. Several such mechanisms have been studied in the literature, and it is instructive to catalogue them briefly, because each corresponds to a substantive modelling choice with empirical implications. Spatial structure (Nowak & May, 1992) is one of the most powerful: when interactions are local rather than well-mixed, cooperators can form clusters in which they preferentially encounter other cooperators, raising the effective payoff to cooperation within the cluster above what the global population average would suggest, and allowing cooperation to invade from a single seed under conditions where well-mixed dynamics would predict its extinction. Assortative matching, more generally, captures any process by which agents with similar strategies are disproportionately likely to interact, whether through kinship, reputation, voluntary partner choice, or institutional sorting; the formal effect is to raise the average payoff differential in favour of cooperation by a factor related to the degree of assortment (Bergstrom, 2003; Eshel & Cavalli-Sforza, 1982). Finite-population stochastic fluctuations, formalised in the Moran process and related birth-death dynamics (Nowak, 2006), allow the population to drift across the threshold y^* by chance: with positive probability, even an all-defector population will, after a sufficiently long waiting time, accumulate enough cooperators through random sampling to enter the basin of attraction of cooperation.

The expected waiting time grows rapidly with population size and with the distance to be traversed, but it is finite, and over evolutionary timescales it can be the decisive mechanism by which cooperative regimes originate. Mutation, modelled as a constant low rate of strategy switching, plays a similar role: it perpetually injects small amounts of every strategy into the population, ensuring that the boundary $y = 0$ is never strictly invariant in the long run, and allowing the system to explore both basins probabilistically. Finally, institutional intervention—the deliberate engineering of initial conditions by external policy, education, or norm enforcement—operates outside the evolutionary model proper but is arguably the most directly relevant mechanism for human societies. Laws, contracts, religious teachings, professional codes of conduct, and public information campaigns can be understood, in this framework, as devices for raising the initial frequency of cooperative strategies above the threshold y^* , after which the replicator dynamics complete the work.

The conceptual upshot is that the conditions for cooperation, properly stated, are joint rather than separate. The continuation probability w must be high enough that cooperation is sustainable in the evolutionary sense, $w > w^*$, and the initial frequency of cooperative strategies must be high enough that the population lies in the right basin of attraction, $y_0 > y^*(w)$. Either condition alone is insufficient: a long-lived relationship between defectors will not generate cooperation, and a high initial frequency of cooperators in a one-shot or short-lived setting will not preserve it. This double-conditionality is one of the principal lessons of the evolutionary analysis of the Prisoner's

Dilemma, and it explains why empirically observed patterns of cooperation are correlated with both relational variables—the duration, frequency, and reliability of repeated interaction—and historical or institutional variables—pre-existing trust, founding norms, prior cooperative experience, and the legacy effects of past institutions. The microeconomic theory of repeated games, focusing as it traditionally does on the existence of cooperative equilibria, captures only one half of this picture; the evolutionary analysis, by foregrounding the basins of attraction and the role of initial conditions, supplies the other half.

A final observation underscores the practical significance of these mechanisms. The unstable threshold $y^*(w)$ derived in (5.13) is not a fixed quantity but a function of the continuation probability, declining monotonically from unity at $w = w^*$ to zero at $w = 1$, as established in (6.1) and (6.2). The basin of cooperation therefore widens as the shadow of the future lengthens, and the threshold-crossing problem becomes progressively easier as w increases. For continuation probabilities close to unity, even a modest perturbation—a small cluster of cooperators arising by chance, a single generation of assortative matching, a brief institutional intervention—can suffice to place the population in the cooperative basin. For continuation probabilities just barely above w^* , by contrast, the threshold y^* lies close to unity, and the required perturbation is correspondingly large. This implies a complementarity between the relational and institutional determinants of cooperation: investments in longer-term relationships (raising w) and investments in initial cooperative norms (raising y_0) reinforce each other, and a sufficient quantity of either can compensate for a deficit of the other. The bifurcation structure thereby translates an abstract dynamical property into a substantive piece of comparative-statics guidance about how cooperation can be cultivated.

DISCUSSION and CONCLUSION

The evolutionary analysis presented in this study delivers three principal results that complement and refine the standard microeconomic treatment of the Prisoner's Dilemma.

The first concerns the one-shot game. When the strategic interaction is reduced to a single encounter, evolutionary dynamics recover precisely the classical Nash prediction without invoking any postulate of rationality. The all-defectors state emerges as the unique evolutionarily stable strategy and the unique asymptotically stable equilibrium of the replicator equation, and it does so unconditionally on the initial composition of the population. This robustness of the dilemma across two very different solution concepts—one grounded in individual deliberation, the other in population-level selection—underscores that the impossibility of cooperation in one-shot interactions is a structural feature of the payoff geometry rather than an artefact of any particular behavioural assumption. The convergence to defection does not require the agents to be clever; it requires only that defection out-earn cooperation against every possible opponent, which is precisely what the ordering of the payoffs guarantees. Cooperation, in a single encounter, has no evolutionary foothold whatsoever.

The second result transforms the picture once the interaction is allowed to repeat. In the indefinitely repeated game, cooperation becomes evolutionarily stable when the continuation probability exceeds an explicit threshold determined entirely by the payoff

parameters of the stage game. Below this threshold, defection remains the unique evolutionarily stable strategy, and the iterated game is dynamically indistinguishable from its one-shot counterpart. Above the threshold, however, the strategic landscape changes qualitatively. Both pure strategies—conditional cooperation in the form of Tit-for-Tat, and unconditional defection—become evolutionarily stable simultaneously, and the dynamics become bistable. Two distinct long-run states therefore coexist as viable evolutionary outcomes within a single dynamical system, and the population must settle into one or the other. The closed-form expression for the unstable interior equilibrium that separates these two basins of attraction makes precise the trade-off between the duration of the relationship and the initial conditions required for cooperation to take hold. The threshold is not a binary switch but a continuous quantity that depends sensitively on the underlying parameters of the interaction.

The third result emerges from the bifurcation analysis. Cooperation, once feasible, becomes progressively *easier* to sustain as the relationship lengthens. The basin of attraction of conditional cooperation expands monotonically as the continuation probability rises, until in the limit of a vanishingly small probability of termination an arbitrarily small initial seed of cooperators suffices for the population to converge to universal cooperation. The "shadow of the future," in Axelrod's evocative phrase, is therefore not a binary determinant of cooperative outcomes but a continuous one, and its lengthening produces a graded rather than a discrete improvement in the prospects for cooperation. This continuity has substantive empirical implications. It predicts, for instance, that cooperation will be more robust in long-lived business relationships than in one-off transactions, that collusive arrangements in repeated oligopoly will be sustainable only when firms are sufficiently patient, and that contractual norms will tend to unravel as known terminal dates approach—each a stylised regularity of microeconomic and political-economic life that the formal model recovers without ad hoc supplementation.

Taken together, these three findings yield a unified perspective on a question that has occupied microeconomic theory since the formalisation of the Prisoner's Dilemma in the mid-twentieth century. The puzzle is not why cooperation is difficult—the structure of the one-shot game already explains that—but why cooperation is nevertheless observed so widely in biological and economic systems where the assumption of common-knowledge rationality is implausible. The evolutionary framework supplies an answer that does not depend on rationality at all. Selection pressures operating on strategy frequencies in a population can, under specifiable conditions, accomplish what individual rationality alone cannot, but only when two conditions are jointly satisfied: the future must matter sufficiently for retaliation to be a credible deterrent, and the initial distribution of strategies in the population must place it in the right basin of attraction. The conditions for cooperation are therefore double rather than single, and the failure to recognise this dual structure is one source of the persistent puzzle that cooperation, while *sustainable* under broad conditions, is not thereby automatically *attained*.

Several extensions of the model are natural and have been pursued extensively in the subsequent literature. The strategy space can be enlarged to include richer reactive strategies—Generous Tit-for-Tat, Win-Stay-Lose-Shift, Pavlov, and various forms of forgiving or memory-one strategies—that perform differently under noise and that resolve some of the brittleness of pure Tit-for-Tat in the presence of trembling-hand

errors. Finite population sizes introduce stochastic effects that allow escape from local attractors via genetic drift, formalised most fruitfully in the Moran process and the adaptive dynamics framework, and these provide one of the principal mechanisms by which cooperation can bootstrap itself in populations that the deterministic replicator equation would condemn to perpetual defection. Spatial and network structure replaces well-mixed matching with locally biased interactions, which can dramatically alter the conditions for cooperation by allowing cooperators to form clusters whose internal payoff differentials favour their own strategy. Each of these extensions builds on, rather than supplants, the analytical scaffolding developed here; the qualitative features of the bistable iterated game—its threshold, its basins of attraction, and its sensitivity to initial conditions—survive in modified form across the more elaborate models.

The unifying message of the analysis is that evolutionary game theory offers a mathematically disciplined and behaviourally parsimonious lens through which to study one of the foundational problems of microeconomics. The Prisoner's Dilemma is not merely a curiosum of rational-choice theory but a generative model whose dynamical structure—its fixed points, its stability properties, and its bifurcations—encodes the conditions under which the social value of cooperation can or cannot be realised. The classical analysis identifies the dilemma; the evolutionary analysis specifies the conditions for its resolution. Both perspectives are required for a complete picture, and their convergence on the all-defectors outcome in the one-shot game, alongside their joint identification of the threshold and basin-of-attraction structure in the iterated game, is itself a result worth emphasising. Theories that arrive at the same conclusions from very different starting premises gain credibility from that convergence, and the agreement between the rational-choice and evolutionary treatments of the Prisoner's Dilemma is one of the more reassuring instances of such convergence in contemporary microeconomic theory.

REFERENCES

Published Works

- Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.
- Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390-1396. <https://doi.org/10.1126/science.7466396>
- Bergstrom, T. C. (2003). The algebra of assortative encounters and the evolution of cooperation. *International Game Theory Review*, 5(3), 211-228. <https://doi.org/10.1142/S0219198903001021>
- Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, 79(4), 1331-1335. <https://doi.org/10.1073/pnas.79.4.1331>
- Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3), 533-554. <https://doi.org/10.2307/1911307>
- Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139173179>
- Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511806292>

- Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, 246, 15-18. <https://doi.org/10.1038/246015a0>
- Nash, J. F. (1950). Equilibrium points in n -person games. *Proceedings of the National Academy of Sciences*, 36(1), 48-49. <https://doi.org/10.1073/pnas.36.1.48>
- Nash, J. F. (1951). Non-cooperative games. *Annals of Mathematics*, 54(2), 286-295. <https://doi.org/10.2307/1969529>
- Nowak, M. A. (2006). *Evolutionary dynamics: Exploring the equations of life*. Belknap Press of Harvard University Press. <https://doi.org/10.2307/j.ctvjghw98>
- Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, 359, 826-829. <https://doi.org/10.1038/359826a0>
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511807763>
- Rapoport, A., & Chammah, A. M. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan Press. <https://doi.org/10.3998/mpub.20269>
- Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.
- Schlag, K. H. (1998). Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory*, 78(1), 130-156. <https://doi.org/10.1006/jeth.1997.2347>
- Taylor, P. D., & Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2), 145-156. [https://doi.org/10.1016/0025-5564\(78\)90077-9](https://doi.org/10.1016/0025-5564(78)90077-9)
- Tucker, A. W. (1950). *A two-person dilemma*. Unpublished manuscript, Stanford University.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.
- Weibull, J. W. (1995). *Evolutionary game theory*. MIT Press.