

The Prisoner's Dilemma is Good

Claude Code

June 28, 2026

Abstract

The prisoner's dilemma is widely regarded as a parable of collective failure—a situation where rational self-interest leads to outcomes that are worse for everyone. In online discourse, particularly in the rationalist and effective altruism communities, the language of the prisoner's dilemma has taken on a distinctly moral character: “defection” is treated as synonymous with antisocial or unethical behavior. This paper argues that this moralized framing is frequently backwards. In many of the most important real-world scenarios that exhibit prisoner's-dilemma dynamics, the “defection” outcome is precisely what serves the public interest. Price competition among firms, whistleblowing against conspiracies, and the breakdown of political cartels all represent cases where the inability to sustain “cooperation” among a small group of players benefits everyone else. The paper catalogs the moralized use of prisoner's-dilemma language in online communities, presents a systematic inventory of cases where PD dynamics produce socially beneficial outcomes, addresses counterarguments, and develops a framework for distinguishing when “cooperation failure” is a feature rather than a bug.

Contents

1	Introduction	3
2	The Prisoner's Dilemma in Online Discourse	4
2.1	“Defection” as Moral Category	4
2.2	The Rhetorical Function of PD Language	5
3	When the Prisoner's Dilemma is Good	6
3.1	Price Competition and Antitrust	6
3.2	The Original Thought Experiment	7
3.3	Whistleblowing and Exposing Corruption	7
3.4	Breaking OPEC and Commodity Cartels	7

3.5	Political Competition Against Bipartisan Collusion	8
3.6	Patent Races and Innovation Competition	8
3.7	Academic Competition and the Replication Crisis	8
3.8	Arms Control Verification	9
3.9	Labor Market Competition for Workers	9
3.10	Journalism and Competitive Reporting	9
4	Counterarguments and Rebuttals	10
4.1	The Tragedy of the Commons	10
4.2	Arms Races	10
4.3	Climate Change	11
4.4	Iterated vs. One-Shot Games	11
4.5	Cooperation in Small Groups	12
5	Analysis: When Is the Prisoner's Dilemma Good?	12
5.1	The Players-Public Distinction	12
5.2	A Diagnostic Test	13
5.3	Why the Moralized Framing Persists	13
5.4	Implications for Rationalist Discourse	14
6	Conclusion	14

1 Introduction

The prisoner’s dilemma is perhaps the single most influential idea in game theory’s public reception. Introduced by Merrill Flood and Melvin Dresher at RAND in 1950 and formalized by Albert Tucker shortly thereafter ([Poundstone, 1992](#)), the scenario describes two players who each face a choice between “cooperation” and “defection,” with the payoff structure ensuring that mutual defection is the Nash equilibrium even though mutual cooperation would leave both players better off. Robert Axelrod’s celebrated tournaments in the early 1980s, which demonstrated the effectiveness of tit-for-tat strategies in iterated versions of the game, cemented the prisoner’s dilemma as the canonical illustration of how self-interest can undermine collective welfare ([Axelrod, 1984](#)).

In the decades since, the prisoner’s dilemma has escaped the confines of academic game theory and become a fixture of popular intellectual discourse. Nowhere is this more evident than in the rationalist community centered around LessWrong, the effective altruism movement, and adjacent intellectual circles on Twitter and in blogs. In these communities, the prisoner’s dilemma serves not merely as an analytical tool but as a moral framework. The term “defection” has acquired a normative valence that goes well beyond its technical meaning: to “defect” is to do something bad, something that benefits you at others’ expense, something that, if universalized, would make everyone worse off.

This paper argues that this moralized framing is, in a large and important class of cases, exactly wrong. The prisoner’s dilemma describes a situation where players cannot sustain cooperation—but whether that inability is good or bad depends entirely on what the “cooperation” consists of and who the “players” are. When the players are firms and “cooperation” means price-fixing, the prisoner’s dilemma is the mechanism by which markets serve consumers. When the players are criminal conspirators and “defection” means testifying for the prosecution, the prisoner’s dilemma is the mechanism by which justice is served. When the players are politicians and “cooperation” means bipartisan collusion against voters’ interests, the prisoner’s dilemma is the mechanism by which democratic accountability functions.

The core insight is simple: the prisoner’s dilemma is “bad” only from the perspective of the players in the game. From the perspective of everyone else—consumers, citizens, the public at large—the inability of a small group to collude is often extremely good. The moralized framing in online discourse systematically conflates the perspective of the players with the perspective of society, treating the players’ inability to coordinate as a universal loss when it is frequently a universal gain.

The remainder of this paper proceeds as follows. Section [2](#) catalogs the moralized use of prisoner’s-dilemma language in online communities, documenting the range of contexts

in which “defection” is used as a term of moral disapproval. Section 3 presents a systematic inventory of cases where prisoner’s-dilemma dynamics produce socially beneficial outcomes. Section 4 addresses the strongest counterarguments to the thesis. Section 5 develops the analytical framework for understanding when the prisoner’s dilemma is good and when it is genuinely harmful. Section 6 concludes.

2 The Prisoner’s Dilemma in Online Discourse

The rationalist and effective altruism communities have developed a distinctive vocabulary built around game-theoretic concepts, and the prisoner’s dilemma occupies a central place in this lexicon. The term “defection” has become a general-purpose label for behavior that is perceived as individually rational but collectively harmful. This section documents the range of contexts in which this moralized framing appears.

2.1 “Defection” as Moral Category

On LessWrong and in rationalist discourse more broadly, “defecting” has become shorthand for a type of moral failure. When someone describes a behavior as “defection,” they are not merely claiming that it fits the formal structure of a prisoner’s dilemma; they are making a normative judgment that the behavior is wrong, or at least regrettable (Yudkowsky, 2015). The term carries an implicit prescription: one *should* cooperate, and those who defect are doing something blameworthy.

This usage appears across a wide range of contexts:

- (1) **Climate and environmental action.** Individuals or nations that fail to reduce emissions are described as “defecting” on a global cooperation problem. The framing implies that non-cooperation is the moral equivalent of betrayal, regardless of the costs or distributional consequences of the proposed “cooperation” (Ostrom, 1990).
- (2) **Social norms and community standards.** Within rationalist communities, breaking community norms—whether about discourse practices, epistemic standards, or interpersonal conduct—is often described as “defection.” This framing treats community norms as a cooperative equilibrium and any deviation as a betrayal of the group.
- (3) **Political behavior.** Voters or politicians who prioritize partisan advantage over bipartisan compromise are described as “defecting.” Third-party voting is sometimes framed as defection against the lesser-evil coalition.
- (4) **Market competition.** Firms that compete aggressively on price rather than main-

taining industry-wide price levels are described as “defecting” against an implicit cooperative norm among producers.

- (5) **Academic and intellectual norms.** Researchers who publicly criticize others’ work, especially within a field or movement, are sometimes accused of “defecting” against a norm of solidarity. The replication crisis has been framed by some as a problem of too much “defection” in the form of adversarial replication attempts.
- (6) **Donation and public goods.** Free-riding on public goods—using Wikipedia without donating, benefiting from open-source software without contributing—is described as “defection” against a cooperative norm of reciprocity.
- (7) **AI safety and existential risk.** In the effective altruism community, organizations or researchers who race to deploy powerful AI systems rather than coordinating on safety are described as “defecting” against humanity’s collective interest (Bostrom, 2014).
- (8) **Personal relationships and commitments.** Even in interpersonal contexts, rationalist discourse sometimes frames broken commitments or self-interested behavior as “defection,” applying the game-theoretic terminology to situations that have little structural resemblance to the actual prisoner’s dilemma.
- (9) **Journalism and information sharing.** Publishing leaked information, breaking embargoes, or scooping competitors is framed as “defection” against norms of coordinated information release.
- (10) **Labor markets.** Workers who negotiate aggressively for higher salaries, or who leave jobs for better offers, are occasionally described as “defecting” against norms of loyalty or solidarity with coworkers.

2.2 The Rhetorical Function of PD Language

The prevalence of this framing is not accidental. Describing behavior as “defection” does significant rhetorical work. It simultaneously:

- **Claims scientific legitimacy** by invoking game theory, lending an air of mathematical precision to what is fundamentally a moral judgment.
- **Shifts the burden of proof** to the “defector,” who must justify their departure from the cooperative norm rather than the norm needing to justify itself.
- **Forecloses debate** about whether the “cooperative” outcome is actually desirable by building the desirability of cooperation into the terminology itself.

- **Obscures distributional questions** about who benefits from the “cooperative” outcome and who bears its costs.

The most important of these is the last. By framing a situation as a prisoner’s dilemma and labeling one strategy “defection,” the speaker implicitly asserts that the players’ inability to coordinate is a bad outcome. But this assertion is doing all the normative work—and it is precisely the assertion that is frequently wrong.

3 When the Prisoner’s Dilemma is Good

This section presents a systematic catalog of cases where prisoner’s-dilemma dynamics produce outcomes that are beneficial for society at large, even though they may be suboptimal from the narrow perspective of the players in the game.

3.1 Price Competition and Antitrust

The most economically significant example of beneficial prisoner’s-dilemma dynamics is ordinary market competition. Firms in the same industry face a classic PD: each would prefer that all firms maintain high prices (“cooperation”), but each has an individual incentive to undercut competitors (“defection”). The result is competitive pricing that benefits consumers (Stigler, 1964).

The entire edifice of antitrust law is built on the premise that this “defection”—firms’ inability to sustain cartel pricing—is good. The Sherman Act of 1890, the Clayton Act of 1914, and their international equivalents exist precisely to prevent firms from solving their prisoner’s dilemma through explicit coordination (Posner, 2001). When firms *do* manage to cooperate—through cartels, price-fixing agreements, or tacit collusion—we do not celebrate their achievement; we prosecute them.

Consider the lysine price-fixing conspiracy of the 1990s, in which Archer Daniels Midland and several co-conspirators successfully “cooperated” to fix the price of a common feed additive. The conspirators’ “cooperation” cost consumers an estimated \$100 million in overcharges. The “defection” that eventually brought down the cartel—an insider who became an FBI informant—was an unambiguous public good (Connor, 2008).

The point generalizes: in virtually every market, the prisoner’s-dilemma structure of competition is the mechanism by which the invisible hand operates. Adam Smith’s famous observation that “it is not from the benevolence of the butcher, the brewer, or the baker that we expect our dinner, but from their regard to their own interest” is, in game-theoretic terms, a celebration of defection (Smith, 1776).

3.2 The Original Thought Experiment

It is worth returning to the original prisoner’s dilemma itself. In Tucker’s formulation, the two players are criminal suspects being interrogated separately. Each can either remain silent (“cooperate” with the other criminal) or testify against the other (“defect”).

From the perspective of justice, the “defection” outcome is straightforwardly good. We *want* criminals to testify against their co-conspirators. The entire institution of plea bargaining, cooperation agreements, and witness protection programs is designed to create exactly this prisoner’s-dilemma incentive structure. Prosecutors deliberately construct situations where the Nash equilibrium involves criminals betraying each other, because this serves the public interest in law enforcement (Bibas, 2004).

The irony is striking: the foundational example of the prisoner’s dilemma—the one that gives the concept its name—is a case where “defection” is good. The original parable already contains the seed of the argument that the moralized framing gets backwards.

3.3 Whistleblowing and Exposing Corruption

More broadly, any situation where members of a corrupt or criminal organization face incentives to expose wrongdoing has the structure of a prisoner’s dilemma in which “defection” serves the public interest. Whistleblowing programs—including the SEC’s whistleblower bounty program, the False Claims Act’s qui tam provisions, and journalistic source protection—are all mechanisms designed to make “defection” against corrupt “cooperation” more attractive (Dyck et al., 2010).

The Enron scandal was ultimately exposed in significant part because individuals within the organization “defected” against the conspiracy of silence. The same is true of the Volkswagen emissions scandal, the LIBOR manipulation scandal, and countless other cases of corporate fraud. In each case, the “cooperative” equilibrium—in which all insiders remain silent—was the socially harmful outcome, and “defection” was the mechanism of accountability.

3.4 Breaking OPEC and Commodity Cartels

The Organization of the Petroleum Exporting Countries (OPEC) is perhaps the world’s most prominent example of an attempted cooperative solution to a prisoner’s dilemma. Member nations agree to production quotas to maintain high oil prices, but each member has an individual incentive to exceed its quota—to “defect.”

OPEC’s history is a history of beneficial defection. When members cheat on their quotas, oil prices fall, benefiting consumers worldwide. The cartel’s periodic inability to enforce cooperation has been one of the primary checks on its monopoly power. Conversely,

OPEC’s periods of greatest “success”—its greatest “cooperation”—have corresponded to the oil price shocks that caused recessions and immense economic hardship in importing nations (Alhajji and Huettner, 2001).

3.5 Political Competition Against Bipartisan Collusion

Political scientists have documented numerous cases where bipartisan “cooperation” serves the interests of politicians at the expense of voters. Gerrymandering by mutual agreement, incumbent-protection legislation, barriers to third-party entry, and coordinated avoidance of difficult issues all represent forms of political “cooperation” that harm the public (Fiorina, 2017).

When politicians “defect” from these cooperative arrangements—by challenging incumbents, introducing disruptive legislation, or breaking ranks on bipartisan deals that serve political insiders—the prisoner’s-dilemma dynamic works in voters’ favor. Primary challenges, insurgent candidacies, and political entrepreneurship are all forms of beneficial “defection” that keep the political market competitive.

The same logic applies to regulatory capture: when an industry and its regulators achieve a “cooperative” equilibrium in which regulation serves industry interests, disruption of that cozy arrangement—whether by new regulators, investigative journalists, or political outsiders—is a form of beneficial defection.

3.6 Patent Races and Innovation Competition

Firms investing in research and development face a prisoner’s dilemma: each would prefer that competitors reduce their R&D spending (allowing the firm to innovate at lower relative cost), but each has an incentive to invest aggressively. The result is a “patent race” in which firms over-invest (from their own perspective) in innovation (Dasgupta and Stiglitz, 1980).

From the public’s perspective, this “over-investment” in innovation is a feature, not a bug. The competitive pressure that drives firms to invest heavily in R&D produces faster technological progress, more new products, and greater consumer welfare than a “cooperative” arrangement in which firms agreed to limit their research spending.

3.7 Academic Competition and the Replication Crisis

The scientific community has traditionally maintained norms of mutual respect and restraint in criticism—a “cooperative” equilibrium in which researchers are reluctant to challenge each other’s published findings. The replication crisis has revealed the costs of

this cooperation: a large fraction of published results in psychology, medicine, and other fields may be false or exaggerated (Ioannidis, 2005; Open Science Collaboration, 2015).

The “defectors” in this context—researchers who conduct adversarial replication attempts, who publicly criticize methodological flaws, who challenge established findings—are performing a vital service for the integrity of science. The prisoner’s-dilemma dynamic, in which individual researchers have career incentives to replicate and challenge rather than defer and agree, is the mechanism by which scientific self-correction operates.

Some in the rationalist community have recognized this point and celebrated the replication movement. But the tension with the general moralization of “defection” is rarely acknowledged.

3.8 Arms Control Verification

The logic extends even to domains where the prisoner’s dilemma is traditionally seen as harmful. In arms control, the “cooperative” outcome is mutual disarmament, and “defection” is secret weapons development. But the mechanism by which arms control agreements actually work depends on a *different* prisoner’s dilemma: the incentive for individuals within a cheating nation’s weapons program to leak information about violations.

Verification regimes, satellite surveillance, and intelligence gathering all exploit PD dynamics to prevent the “cooperative” concealment of treaty violations. The defection of insiders—scientists who leak information, satellite operators who detect tests, diplomats who report violations—is the mechanism by which arms control is enforced (Schelling, 1960).

3.9 Labor Market Competition for Workers

When employers “cooperate” to suppress wages—through no-poach agreements, wage-fixing conspiracies, or tacit collusion on compensation—the prisoner’s-dilemma dynamic that makes this cooperation unstable directly benefits workers. The DOJ’s prosecution of no-poach agreements in Silicon Valley (the “High-Tech Employee Antitrust Litigation” against Apple, Google, Intel, and Adobe) was predicated on exactly this logic: employers’ inability to sustain “cooperative” wage suppression is good for workers (U.S. Department of Justice, 2010).

3.10 Journalism and Competitive Reporting

News organizations face a prisoner’s dilemma around scoops, leaks, and competitive reporting. Each outlet might prefer a “cooperative” arrangement where they coordinate

coverage, respect embargoes, and avoid adversarial investigative reporting. But the “defection” incentive—to break stories first, to publish leaked documents, to investigate powerful institutions—is the mechanism by which the press serves its democratic function.

The Pentagon Papers, Watergate, the Panama Papers, and countless other acts of accountability journalism represent “defection” against norms of deference and coordination. A press corps that successfully “cooperated”—that never broke embargoes, never published leaks, never challenged official narratives—would be a captured press corps.

4 Counterarguments and Rebuttals

The thesis that the prisoner’s dilemma is often good is not without objections. This section addresses the strongest counterarguments.

4.1 The Tragedy of the Commons

Objection: The tragedy of the commons—overgrazing, overfishing, pollution—represents a genuine multi-player prisoner’s dilemma where defection is unambiguously bad. Each herder, fisher, or polluter has an individual incentive to overuse the commons, and the result is collective ruin ([Hardin, 1968](#)).

Rebuttal: This objection is correct as far as it goes, but it does not refute the thesis. The claim is not that the prisoner’s dilemma is *always* good, but that it is *often* good—and that the moralized framing in online discourse fails to distinguish the two cases. The tragedy of the commons is a genuine case where cooperation failure is harmful, but it differs from the cases cataloged in Section 3 in a crucial respect: in the tragedy of the commons, the “players” of the game *are* the public (or a representative cross-section of it), so the players’ loss is the public’s loss. In the antitrust case, by contrast, the “players” are firms and the public is a third party that benefits from the players’ inability to cooperate.

The analytical framework developed in Section 5 makes this distinction precise.

4.2 Arms Races

Objection: Arms races—whether military, advertising, or status-seeking—are prisoner’s dilemmas where defection (building more weapons, spending more on ads, escalating status competition) wastes resources to no net benefit ([Schelling, 1960](#)).

Rebuttal: This objection has real force in some cases but is overstated in others. Military arms races between great powers may indeed be wasteful PDs. But many “arms races” are actually beneficial competition mislabeled. The “arms race” in pharmaceutical R&D

produces new drugs. The “arms race” in technology produces better products. Even advertising “arms races,” while potentially wasteful, also convey information to consumers and fund media production.

Moreover, even in genuinely wasteful arms races, the relevant question is often not whether defection is good but whether the proposed “cooperation” is achievable or desirable. Proposals for countries to mutually reduce military spending require enforcement mechanisms that may themselves be costly or dangerous. The “cooperative” outcome may be unstable or illusory.

4.3 Climate Change

Objection: Climate change is the paradigmatic global prisoner’s dilemma. Each nation has an incentive to free-ride on others’ emissions reductions. This defection is producing catastrophic outcomes (Nordhaus, 2015).

Rebuttal: Climate change is indeed a case where cooperation failure is harmful. But the PD framing is misleading even here. Climate action is not a simple two-player PD; it involves heterogeneous actors with different costs, different benefits, and different time horizons. The “cooperation” frame obscures the fact that many climate policies have significant co-benefits (reduced air pollution, energy independence, technological innovation) that make them rational even without cooperation. And the “defection” frame obscures the fact that what is called “defection” often means developing countries prioritizing economic growth and poverty reduction.

Furthermore, framing climate as a PD can be counterproductive. If cooperation is presented as requiring all players to sacrifice self-interest, it suggests that the problem is insoluble absent some mechanism to enforce cooperation. In fact, the declining cost of renewable energy is making “defection” (in the sense of each country pursuing its own clean energy transition for self-interested reasons) an increasingly viable path to emissions reduction.

4.4 Iterated vs. One-Shot Games

Objection: The argument conflates one-shot and iterated prisoner’s dilemmas. In iterated games, cooperation can be sustained through strategies like tit-for-tat, and this cooperation is often genuinely beneficial. Axelrod’s tournaments showed that cooperative strategies outperform pure defection in the long run (Axelrod, 1984).

Rebuttal: This is an important distinction but does not rescue the moralized framing. Even in iterated games, the desirability of cooperation depends on the identity of the players and the content of the cooperation. An iterated relationship between cartel

members that sustains collusion over time is worse than a one-shot interaction that fails to produce collusion. Axelrod’s results show that tit-for-tat can *sustain* cooperation, not that cooperation is *desirable*.

The iterated game point actually strengthens the thesis in some respects: antitrust law is most concerned precisely with industries where repeated interaction facilitates tacit collusion. The industries where firms *can* sustain “cooperation” through iterated play—where the shadow of the future enables cartel behavior—are the industries where antitrust enforcement is most important.

4.5 Cooperation in Small Groups

Objection: In small communities and personal relationships, cooperation norms are genuinely valuable. Describing someone who breaks promises or free-rides on a small group’s efforts as “defecting” captures something real and important.

Rebuttal: Agreed. The thesis does not claim that cooperation is never valuable or that defection is never harmful. The claim is that the online discourse applies the prisoner’s-dilemma frame far too broadly and with insufficient attention to context. Using “defection” to describe someone who breaks a personal promise is extending the metaphor to a domain where it may (or may not) be appropriate. Using “defection” to describe a firm that undercuts cartel pricing is extending the metaphor to a domain where it is affirmatively misleading.

5 Analysis: When Is the Prisoner’s Dilemma Good?

The cases above suggest a general framework for distinguishing beneficial from harmful prisoner’s-dilemma dynamics. The key variable is the relationship between the players and the broader public.

5.1 The Players-Public Distinction

Consider a prisoner’s dilemma with two (or more) players. The “cooperative” outcome benefits the players; the “defection” outcome harms them. But there are always third parties—consumers, citizens, the general public—who are affected by the game’s outcome but are not players in it.

The prisoner’s dilemma is **good** when the players’ “cooperation” would harm third parties. This occurs when:

- The players are firms and cooperation means cartel pricing (harms consumers)

- The players are criminals and cooperation means maintaining a conspiracy of silence (harms justice)
- The players are politicians and cooperation means bipartisan collusion (harms voters)
- The players are employers and cooperation means wage suppression (harms workers)
- The players are news organizations and cooperation means avoiding adversarial reporting (harms the public’s right to information)

The prisoner’s dilemma is **bad** when the players *are* the affected third parties—when the game’s “players” are a representative cross-section of the public. This occurs when:

- The players are all users of a commons (overfishing, pollution)
- The players are all citizens of nations affected by climate change
- The players are all members of a community that benefits from shared norms

5.2 A Diagnostic Test

This framework yields a simple diagnostic test: *When someone calls behavior “defection,” ask who would benefit if the players could cooperate.* If the answer is “the players, at the expense of everyone else,” then defection is good. If the answer is “everyone, including the players themselves,” then defection may genuinely be bad.

This test immediately clarifies most of the cases cataloged in Section 2. When firms are accused of “defecting” by competing on price, the proposed “cooperation” would benefit firms at consumers’ expense—defection is good. When individuals are accused of “defecting” by not reducing their carbon footprint, the proposed “cooperation” would benefit everyone—defection may be genuinely problematic (though even here the framing is oversimplified, as discussed in Section 4).

5.3 Why the Moralized Framing Persists

If the players-public distinction is straightforward, why does the moralized framing persist? Several factors contribute:

1. **Identification with the players.** The prisoner’s-dilemma narrative is typically told from the players’ perspective. We naturally empathize with the prisoners, the firms, the negotiators—whoever is “trapped” in the dilemma. This identification makes “cooperation” seem good and “defection” seem bad, even when the players are a cartel.

2. **The cooperation heuristic.** In most personal and small-group interactions, cooperation is indeed prosocial. The PD-as-moral-frame generalizes this heuristic to contexts where it does not apply.
3. **Aesthetic preference for coordination.** The rationalist community in particular values coordination, alignment, and collective action. This makes “defection” aesthetically unappealing regardless of its consequences.
4. **Omission of third parties.** The standard 2x2 payoff matrix representation of the PD includes only the players’ payoffs. Third-party effects are invisible in the formal representation, making it easy to forget that they exist.
5. **Selection bias in examples.** Online discourse tends to deploy the PD frame in contexts where the speaker *wants* cooperation (climate, community norms, AI safety). Cases where defection is beneficial—antitrust, whistleblowing, competitive journalism—are rarely framed in PD terms, even though they have identical formal structure.

5.4 Implications for Rationalist Discourse

The analysis has several implications for how the rationalist community uses game-theoretic concepts:

First, the term “defection” should not be used as a synonym for “bad behavior.” It is a technical term describing a strategy in a specific game-theoretic structure. Loading it with moral content obscures rather than illuminates.

Second, when deploying the prisoner’s-dilemma frame, speakers should explicitly identify the third-party effects of the proposed “cooperation.” If the “cooperative” outcome would involve a cartel, conspiracy, or collusion that harms outsiders, then framing the situation as a PD in which cooperation is desirable is not just imprecise but actively misleading.

Third, the celebration of cooperation should be tempered by the recognition that the most important institutional achievements in modern society—competitive markets, rule of law, democratic accountability, free press—all depend on the *inability* of certain groups to sustain cooperation. The institutions we value most are, in a precise sense, mechanisms for ensuring that prisoner’s-dilemma dynamics operate in domains where they serve the public interest.

6 Conclusion

The prisoner’s dilemma is one of the most powerful ideas in social science, and its popularization in online intellectual communities has brought game-theoretic thinking to a

wide audience. But the moralized framing that has accompanied this popularization—in which “defection” is bad and “cooperation” is good—systematically misapplies the concept.

In a wide range of economically and politically important cases, the prisoner’s dilemma is not a trap but a safeguard. Price competition, whistleblowing, investigative journalism, democratic accountability, and scientific replication all depend on the inability of interested parties to sustain “cooperation” against the public interest. The prisoner’s dilemma is the formal structure of competitive markets, adversarial legal systems, and democratic politics. In these domains, “defection” is not a failure of rationality but a triumph of institutional design.

The key analytical move is to distinguish between games where the players are the public and games where the players are a cartel. In the former, cooperation failure is genuinely costly. In the latter, cooperation failure is precisely what serves the public interest. The moralized framing of online discourse collapses this distinction, treating all cooperation failure as the former when much of the most important cooperation failure is the latter.

A more sophisticated engagement with the prisoner’s dilemma would recognize that the question is not whether cooperation or defection is “better” in the abstract, but whether the specific cooperation at issue serves or harms the broader public. Sometimes we should be grateful that the prisoners cannot keep their mouths shut.

References

- Alhajji, A. F. and Huettner, D. (2001). OPEC and other commodity cartels: A comparison. *Energy Policy*, 29(14):1151–1164.
- Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books, New York.
- Bibas, S. (2004). Plea bargaining outside the shadow of trial. *Harvard Law Review*, 117(8):2463–2547.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford.
- Connor, J. M. (2008). Global price fixing. *Studies in Industrial Organization*, 26.
- Dasgupta, P. and Stiglitz, J. (1980). Industrial structure and the nature of innovative activity. *The Economic Journal*, 90(358):266–293.
- Dyck, A., Morse, A., and Zingales, L. (2010). Who blows the whistle on corporate fraud? *The Journal of Finance*, 65(6):2213–2253.

- Fiorina, M. P. (2017). *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate*. Hoover Institution Press, Stanford, CA.
- Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859):1243–1248.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124.
- Nordhaus, W. D. (2015). Climate clubs: Overcoming free-riding in international climate policy. *American Economic Review*, 105(4):1339–1370.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge.
- Posner, R. A. (2001). *Antitrust Law*. University of Chicago Press, Chicago, 2nd edition.
- Poundstone, W. (1992). *Prisoner’s Dilemma: John von Neumann, Game Theory, and the Puzzle of the Bomb*. Doubleday, New York.
- Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.
- Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London.
- Stigler, G. J. (1964). A theory of oligopoly. *Journal of Political Economy*, 72(1):44–61.
- U.S. Department of Justice (2010). United states v. Adobe Systems Inc., et al. Civil Action No. 1:10-cv-01629. Antitrust Division complaint regarding no-poach agreements among technology companies.
- Yudkowsky, E. (2015). Rationality: From AI to zombies. Machine Intelligence Research Institute. Collection of essays originally published on LessWrong, 2006–2009.