A Case for Cooperation: Dependence in the Prisoner's Dilemma

Grant Stenger

September 3, 2024

> *"The man who is cheerful and merry has always a good reason for being so,—the fact, namely, that he is so."*
>
> — Schopenhauer, *The Wisdom of Life* (1851)

Abstract

Descriptions of the Prisoner's Dilemma usually suggest that the optimal policy for each prisoner is to selfishly defect instead of to cooperate. I still think the traditional analysis is correct on its own terms. If we model the game in the standard way—simultaneous play, no communication, each player evaluating only the direct causal effects of their individual act—then defection is the dominant strategy and (D, D) is the Nash equilibrium.

Where I part ways with the standard story is one layer up. Real agents are not sampled from nowhere. They come with similar priors, similar information, similar dispositions, and sometimes even similar reasoning procedures. In those settings, my contemplated action can be evidence about yours even when it does not causally affect yours. In the extreme clone case, the relevant question is not "what happens if I alone switch from C to D while everything else stays fixed?" but something more like "what happens if this kind of mind outputs C instead of D?"

This essay separates three notions that are easy to blur together: ordinary causal dependence, evidential dependence, and dependence at the level of a shared policy or decision procedure. Once those are distinguished, the case for cooperation becomes much cleaner.

Contents

TL;DR

Descriptions of the Prisoner's Dilemma usually suggest that the optimal policy for each prisoner is to selfishly defect instead of to cooperate. I still think the traditional analysis is *correct on its own terms.* If we model the game in the standard way—simultaneous play, no communication, each player evaluating only the direct causal effects of their individual act—then defection is the dominant strategy and (D, D) is the Nash equilibrium.

Where I part ways with the standard story is one layer up. Real agents are not sampled from nowhere. They come with similar priors, similar information, similar dispositions, sometimes even similar reasoning procedures. In those settings, my contemplated action can be *evidence* about yours even when it does not causally affect yours. In the extreme clone case, the relevant question is not "what happens if I alone switch from C to D while everything else stays fixed?" but something more like "what happens if this kind of mind outputs C instead of D?"

There is also an important correction to make right away. A bare Pearl-style common-cause model does *not* by itself imply that causal decision theory should cooperate. On a simple common-cause DAG, ordinary interventions on Alice's action leave Bob's action distribution unchanged, so standard causal decision theory still defects. So the real issue is not merely whether the players are correlated. It is what *kind* of dependence a rational agent should treat as decision-relevant.

1 The Traditional Analysis

In Game Theory 101, here is how the Prisoner's Dilemma is usually presented. Alice and Bob are conspirators in a crime. They are caught and brought to separate interrogation rooms. They are presented with a Faustian bargain: to snitch or not to snitch. If neither snitches on the other, they both get a one-year sentence. If one snitches and the other does not, then the snitch goes home free while the non-snitch serves three years. If they both snitch, they each serve two years.

The payoff diagram corresponding to this setup is

	B_C	B_D
A_C	$(-1, -1)$	$(-3, 0)$
A_D	$(0, -3)$	$(-2, -2)$

If Alice cooperates, Bob is better off defecting to get 0 years instead of 1 year. If Alice defects, Bob is better off defecting to get 2 years instead of 3 years. So in either case Bob is better off defecting. A strategy which is optimal regardless of the choices of an opponent is called a *dominant strategy.* Symmetrically, Alice is better off defecting no matter what Bob does. This means that even though they are both happier in the case where they both cooperate, serving just one year each, the Nash equilibrium is the case where they both defect, serving two years each.

We can generalize the payoff matrix a bit to represent all situations that capture the structure of a Prisoner's-Dilemma-like scenario:

	B_C	B_D
A_C	(R, R)	(S, T)
A_D	(T, S)	(Q, Q)

I use the variables Q, R, S, T to keep organized. Q is the Quarrel payoff when they rat each other out. R is the Reward for mutual cooperation. S is the Sucker's payoff if the other player snitches and they do not. T is the Temptation payoff for snitching while the other does not. The standard Prisoner's Dilemma inequalities are

$$S < Q < R < T.$$

1.1 Probabilistic Play

Now let us make the model a bit more formal and extend the binary action space to a probabilistic strategy model.

Instead of discrete actions, let Alice and Bob each choose mixed strategies

$$[p(A_C), p(A_D)] \qquad \text{and} \qquad [p(B_C), p(B_D)],$$

with

$$p(A_C) + p(A_D) = 1, \qquad p(B_C) + p(B_D) = 1.$$

Alice wants to maximize her expected utility:

$$\mathbb{E}[U_A(G)] = \mathbb{E}[U_A(G \mid A_C)]p^*(A_C) + \mathbb{E}[U_A(G \mid A_D)]p^*(A_D).$$

Splitting by Bob's action gives

$$\begin{aligned}\mathbb{E}[U_A(G)] = {} & R\,p^*(A_C)p^*(B_C) + S\,p^*(A_C)p^*(B_D) \\ & + T\,p^*(A_D)p^*(B_C) + Q\,p^*(A_D)p^*(B_D).\end{aligned}$$

Substituting $p(A_D) = 1 - p(A_C)$ and $p(B_D) = 1 - p(B_C)$ yields

$$\begin{aligned}\mathbb{E}[U_A(G)] = {} & R\,p^*(A_C)p^*(B_C) + S\,p^*(A_C)(1 - p^*(B_C)) \\ & + T(1 - p^*(A_C))p^*(B_C) + Q(1 - p^*(A_C))(1 - p^*(B_C)).\end{aligned}$$

This is linear in $p^*(A_C)$. After expanding and regrouping,

$$\mathbb{E}[U_A(G)] = p^*(A_C)\big[p^*(B_C)(Q + R - S - T) + (S - Q)\big] + \big[p^*(B_C)(T - Q) + Q\big].$$

So Alice cooperates only if the coefficient on $p^*(A_C)$ is positive. But

$$p^*(B_C)(Q + R - S - T) + (S - Q) < 0 \iff 0 < p^*(B_C)(T - R) + (1 - p^*(B_C))(Q - S),$$

and the right-hand side is positive because $T > R$ and $Q > S$. So the coefficient is negative, which means Alice should defect. The same analysis applies symmetrically to Bob.

That is the standard result. On the usual unilateral-action reading of the game, defection wins.

I do not think that result is false. I think it is incomplete.

2 The Clone Case for Cooperation

Here is the case that motivates the rest of the essay.

Imagine that Alice and Bob are clones: same dispositions, same information, same reasoning, same environment, same everything relevant. Let us suppose, in the strongest possible version of the thought experiment, that they will deterministically make the same decision.

In that case Alice's estimate of Bob's behavior satisfies

$$\hat{p}_A(B_C \mid A_C) = 1, \qquad \hat{p}_A(B_C \mid A_D) = 0.$$

Equivalently,

$$\hat{p}_A(B_D \mid A_C) = 0, \qquad \hat{p}_A(B_D \mid A_D) = 1.$$

Now recompute Alice's expected utility:

$$\mathbb{E}[U_A(G)] = \mathbb{E}[U_A(G \mid A_C)]p^*(A_C) + \mathbb{E}[U_A(G \mid A_D)]p^*(A_D).$$

Conditioning on Bob's action under Alice's contemplated action gives

$$\begin{aligned}\mathbb{E}[U_A(G)] = {} & \hat{p}_A(B_C \mid A_C)p^*(A_C)R + \big(1 - \hat{p}_A(B_C \mid A_C)\big)p^*(A_C)S \\ & + \hat{p}_A(B_C \mid A_D)(1 - p^*(A_C))T \\ & + \big(1 - \hat{p}_A(B_C \mid A_D)\big)(1 - p^*(A_C))Q.\end{aligned}$$

Plugging in the clone assumptions,

$$\mathbb{E}[U_A(G)] = p^*(A_C)R + (1 - p^*(A_C))Q.$$

So

$$\mathbb{E}[U_A(G)] = p^*(A_C)(R - Q) + Q.$$

Because $R > Q$, Alice maximizes expected utility by choosing

$$p^*(A_C) = 1.$$

She cooperates. So does Bob. The dilemma disappears.

The clone argument is *not* claiming that Alice's act token causally pushes Bob's act token around from another room. The interesting claim is subtler: when the two agents are sufficiently alike, the counterfactual "if I cooperate" may carry information about what Bob does as well. In the strongest version, Alice and Bob are not two independent roulette wheels. They are two outputs of effectively the same underlying procedure.

3 Everything Is Correlated

All right. If Alice knows Bob will make exactly the same decision, she can cooperate fearlessly. So now let us put uncertainty back into the model.

Suppose Alice and Bob are not literal clones but are instead siblings, or agents with similar values, similar information, similar dispositions, similar training, or similar decision procedures. We know the world is full of correlation.[1] It would be surprising if two agents with similar priors and identical payoff matrices were perfectly independent in every decision-relevant respect.

That said, this is exactly where I want to be more careful than the earlier draft. "Everything is correlated" is directionally true, but it is not enough all by itself. Some kinds of correlation matter for cooperation here; some do not.

[1] Gwern Branwen, "Everything Is Correlated."

3.1 Causal Modeling 101

Let V denote a shared upstream factor influencing both Alice and Bob. The causal graph looks like this:

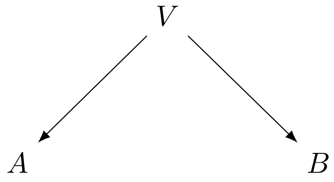

Figure 1: A simple common-cause DAG for the prisoners' choices.

Here V is a common cause of both actions. If we condition on V, Alice and Bob may be independent. If we do not observe V, then their actions become statistically dependent:

$$\mathbb{P}(A, B) = \sum_v \mathbb{P}(A \mid V = v)\mathbb{P}(B \mid V = v)\mathbb{P}(V = v).$$

So far so good. This is exactly the sort of thing Pearl-style causal graphs are good at making explicit.

3.2 What This Does and Does Not Buy Us

This common-cause structure gives us *observational dependence*. For example,

$$\mathbb{P}(B_C \mid A_C) = \sum_v \mathbb{P}(B_C \mid V = v)\mathbb{P}(V = v \mid A_C),$$

which can differ from $\mathbb{P}(B_C)$.

But here is the crucial correction. On the bare common-cause DAG above, intervening on Alice's action does *not* change Bob's distribution. Under Pearl's do-operator,

$$\mathbb{P}(B_C \mid \mathrm{do}(A_C)) = \sum_v \mathbb{P}(B_C \mid V = v)\mathbb{P}(V = v) = \mathbb{P}(B_C),$$

and likewise

$$\mathbb{P}(B_C \mid \mathrm{do}(A_D)) = \mathbb{P}(B_C).$$

So on a pure common-cause model, ordinary causal decision theory still says

$$\mathbb{E}[U_A \mid \mathrm{do}(A_C)] = R\,\mathbb{P}(B_C) + S(1 - \mathbb{P}(B_C)),$$

$$\mathbb{E}[U_A \mid \mathrm{do}(A_D)] = T\,\mathbb{P}(B_C) + Q(1 - \mathbb{P}(B_C)),$$

and because $T > R$ and $Q > S$, Alice still defects.

This is not a bug in Pearl. It is the point.

A bare common-cause graph does not rescue cooperation under standard causal decision theory. If I want the cooperation intuition, I need something stronger than ordinary downstream causation. I need a notion of dependence at the level of evidence, type, reasoning, or shared policy.

4 A Cleaner General Model of Dependence

The cleanest way to write the decision rule is with two parameters:

$$\alpha := \hat{p}_A(B_C \mid A_C), \qquad \beta := \hat{p}_A(B_C \mid A_D).$$

These are Alice's own estimates of Bob's cooperation probability under her two contemplated actions.

Then Alice's expected utility from cooperating is

$$\mathbb{E}[U_A \mid A_C] = \alpha R + (1-\alpha)S,$$

while her expected utility from defecting is

$$\mathbb{E}[U_A \mid A_D] = \beta T + (1-\beta)Q.$$

So cooperation is optimal exactly when

$$\alpha R + (1-\alpha)S > \beta T + (1-\beta)Q.$$

Rearranging gives the master inequality

$$\boxed{\alpha(R-S) - \beta(T-Q) > Q - S.}$$

4.1 Recovering the Standard Independent Case

If Alice treats Bob's action as independent of her own contemplated action, then $\alpha = \beta = \mathbb{P}(B_C)$. In that case

$$\mathbb{E}[U_A \mid A_C] - \mathbb{E}[U_A \mid A_D] = \mathbb{P}(B_C)(R-T) + (1-\mathbb{P}(B_C))(S-Q) < 0,$$

so she defects. That is exactly the classical result again.

4.2 Recovering the Clone Case

If Alice and Bob are perfect clones, then $\alpha = 1$ and $\beta = 0$. The condition becomes

$$R > Q,$$

which is true in any Prisoner's Dilemma. So Alice cooperates.

4.3 Interpretation

This is, to my mind, the cleanest statement of the whole essay.

The case for cooperation is not "correlation, therefore cooperate." It is more specific: if contemplating cooperation makes Bob's cooperation much more likely in your model of the world, and contemplating defection does not similarly swing Bob toward cooperating, then cooperation can beat defection.

Or, more compactly: if your act is strong enough evidence about theirs, cooperate.

5 A Simple Symmetric Noisy Model

If you want a one-parameter version of the dependence story, a convenient symmetric specialization is

$$\alpha = \frac{1+\rho}{2}, \qquad \beta = \frac{1-\rho}{2}, \qquad \rho \in [-1,1].$$

Here $\rho = 1$ is the clone limit, $\rho = 0$ is the independent case, and $\rho = -1$ is perfect anti-correlation.

I do not mean that ρ is the unique or canonical notion of statistical correlation here. It is just a simple symmetric parameterization of the two conditional probabilities.

Substituting into the master inequality gives

$$\mathbb{E}[U_A \mid A_C] - \mathbb{E}[U_A \mid A_D] = \frac{R+S-T-Q+\rho(R-S-Q+T)}{2}.$$

So cooperation is optimal when

$$\boxed{\rho > \frac{Q-R-S+T}{-Q+R-S+T}.}$$

Above a critical level of positive dependence, cooperation becomes optimal.

6 A Shared-Policy Model

The earlier draft had a section called "identical mixed strategies." The basic intuition there was good, but the interpretation needs to be sharpened.

If Bob independently chooses the same mixing rate p as Alice, but Alice still evaluates only her own unilateral deviation while holding Bob's policy fixed, then we are back in the standard linear mixed-strategy setting and defection still wins.

So to get something genuinely different, we need to say something stronger: Alice and Bob instantiate the same underlying mixed *policy*, and the question is which common policy parameter p should that shared procedure output.

That is a different optimization problem. We are no longer asking for Alice's best response while holding Bob fixed. We are optimizing over a *shared policy parameter*.

If the common policy cooperates with probability p, then Alice's expected utility is

$$U_{\text{sym}}(p) = Rp^2 + (S+T)p(1-p) + Q(1-p)^2.$$

Expanding,

$$U_{\text{sym}}(p) = (R-S-T+Q)p^2 + (S+T-2Q)p + Q.$$

6.1 Optimizing the Shared Policy

Differentiate:

$$\frac{dU_{\text{sym}}}{dp} = 2(R-S-T+Q)p + (S+T-2Q).$$

The vertex is at

$$p^* = \frac{2Q-S-T}{2(R-S-T+Q)} = \frac{S+T-2Q}{2(S+T-R-Q)}.$$

Now for the important simplification. Because

$$U_{\mathrm{sym}}(1) = R \qquad \text{and} \qquad U_{\mathrm{sym}}(0) = Q,$$

and $R > Q$, the fully defecting policy $p = 0$ is never optimal in a genuine Prisoner's Dilemma.

The optimum is therefore

$$p^* = \begin{cases} 1, & \text{if } S + T \le 2R, \\ \dfrac{S + T - 2Q}{2(S + T - R - Q)}, & \text{if } S + T > 2R. \end{cases}$$

So in the usual toy cases, the shared policy simply outputs full cooperation.

6.2 Toy Example

For the standard sentence example,

$$R = -1, \qquad S = -3, \qquad T = 0, \qquad Q = -2.$$

Then

$$S + T = -3 \qquad \text{and} \qquad 2R = -2,$$

so $S + T \le 2R$, which implies

$$p^* = 1.$$

The shared policy fully cooperates.

7 Three Different Objects

At this point I think the heart of the disagreement can be stated very cleanly. There are three different objects that are easy to blur together:

$$\mathbb{P}(B \mid A), \qquad \mathbb{P}(B \mid \mathrm{do}(A)), \qquad \mathbb{P}(B \mid \text{shared policy outputs } A).$$

These are not the same.

- $\mathbb{P}(B \mid A)$ is *evidential* dependence.
- $\mathbb{P}(B \mid \mathrm{do}(A))$ is *causal* dependence in Pearl's sense.
- "shared policy outputs A" is dependence at the level of a common decision procedure, which is the right lens for clones, twins, source-code-like agents, or sufficiently similar reasoners.

The standard one-shot Prisoner's Dilemma analysis uses the second object. My case for cooperation leans on the first, and in the strongest clone/twin cases on the third.

That is why the clone argument can be compelling even though Pearl's bare common-cause DAG still tells ordinary CDT to defect. The disagreement is not about algebra. It is about which counterfactuals we think matter.

8 So What Is the Actual Claim?

I do not want to oversell this.

I am not claiming that the standard Prisoner's Dilemma result is simply wrong. If you insist on the standard normal-form game, standard best-response analysis, and standard act-by-act causal counterfactuals, then yes: defect.

What I *am* claiming is this:

1. The clone/twin/sufficiently-similar-agent case is a very real and very important nearby case.
2. In that case, cooperation can be rational.
3. Much of the intuitive force of the standard Prisoner's Dilemma comes from slipping too quickly from "no direct causal influence" to "no decision-relevant dependence at all."
4. Real-world one-shot dilemmas are often closer to the evidential or shared-policy cases than the classroom presentation lets on.

I still think "Everything is correlated" is part of the right intuition. But the strong version of the claim is not that any tiny common-cause correlation dissolves the dilemma. The strong version is that when agents are sufficiently similar that their contemplated actions are informative about one another, or when they instantiate the same effective policy, cooperation becomes intelligible as the rational move rather than as a sentimental mistake.

9 Conclusion

Despite typical Game Theory 101 lectures suggesting 100% selfish play in one-shot games like this, I think there is a real and underappreciated case for cooperation once we stop pretending that "no downstream causal influence" settles the matter.

On the pure causal-decision-theory reading of the problem, defect. Fine. But in clone cases, twin cases, evidentially coupled cases, and shared-policy cases, cooperation is not naive. It is exactly what falls out of the math.

A broader moral of the analysis is that we should be careful about throwing away dependence just because it is not a simple arrow from my act token to yours. There are many ways for decisions to hang together. Some are causal. Some are evidential. Some live at the level of type, reasoning, or policy. If we collapse all of those into "independent," we miss the interesting part.

In this sense, the cooperation argument is a kind of opposite of adverse selection. In many markets, conditional on taking an action, you become less happy because the very fact that you can take it is evidence that something has gone wrong: you win the auction, hire the lemon, buy what no one else wanted. But here there is a kind of *advantageous selection.* If I cooperate, and if I have good reason to think my decision is informative about yours, then I acquire precisely the sort of information I like: someone sufficiently like me likely cooperated too.

For those who remain compelled by the original argument that "still, if you know your opponent is going to cooperate, you are better off defecting to serve no time instead of one year," I think the right reply is now clearer than it was before. That objection is framed in the unilateral-deviation language of the standard model. The whole point of the clone/twin/shared-policy picture is that this is not the only counterfactual on the table.

So yes: on one reading of the problem, the causal decision theorists are stuck in their (D, D) "rational" Nash equilibrium. On another, more interesting reading—the one where sufficiently similar minds should expect their choices to hang together—my co-conspirators and I will be faithfully cooperating and walking free into the warm sun of Pareto optimality.

Acknowledgements. Thank you Alok Singh, Baran Cimen, Keegan McNamara, Jackson Stenger, and Paul Schmidt-Engelbertz for reading through the draft and providing helpful guidance.

References

[1] Gwern Branwen. *Everything Is Correlated.* gwern.net/everything.

[2] Judea Pearl. *Causality.* Cambridge University Press, 2nd edition, 2009.