The prisoner's dilemma

You and your partner (the person sitting next to you) have been in business running drugs for the last few months. You've just been arrested by the police, who are interrogating you in separate rooms.

Here's what you know about your situation: you know that if **both** you and your partner confess, given the evidence that the police will then possess, you'll each get about 5 years for your crimes.

On the other hand, if you confess and your partner doesn't, you'll as a reward get off scot free, and your partner will be stuck serving 10 years in jail. (And the reverse if your partner confesses, and you do not.)

If you both keep quiet - and **neither** of you confesses - then the police will have only have the evidence to convict the two of you for a lesser crime - for which, you estimate, you'll have to serve 2 years in jail.

Take a second and think about your decision. Then write down on a piece of paper either "confess" or "stay silent."

Now show the paper to your partner, and take note of how many years of each of you will have to serve in jail.

Some years later, you and your partner are both free, and back to your old tricks. Unsurprisingly, you again get arrested, and the police again offer you the same deal.

Again take a second and think about your decision. Then again write down on a piece of paper either "confess" or "stay silent."

Now show the paper to your partner, and take note of how many years of each of you will have to serve in jail.

Some years later, you and your partner are, once again, both free, and, once again, back to your old tricks. Unsurprisingly, you again get arrested, and the police again offer you the same deal.

Again take a second and think about your decision. Then again write down on a piece of paper either "confess" or "stay silent."

Now show the paper to your partner, and take note of how many years of each of you will have to serve in jail.

Finally, you are both out of jail, and by this point are quite old. However, you decide to get together for one last big deal. But you're caught, and the police again offer you the same deal. You're getting very tired of this, but, given your age, at least **you know that this is the last time that you are your partner will ever be arrested together.**

Again write down on a piece of paper either "confess" or "stay silent" and, when you're both done, show the paper to your partner, and calculate the results.

You and your partner (the person sitting next to you) have been in business running drugs for the last few months. You've just been arrested by the police, who are interrogating you in separate rooms.

Here's what you know about your situation: you know that if **both** you and your partner confess, given the evidence that the police will then possess, you'll each get about 5 years for your crimes.

On the other hand, if you confess and your partner doesn't, you'll as a reward get off scot free, and your partner will be stuck serving 10 years in jail. (And the reverse if your partner confesses, and you do not.)

If you both keep quiet - and **neither** of you confesses - then the police will have only have the evidence to convict the two of you for a lesser crime - for which, you estimate, you'll have to serve 2 years in jail.

The situation in which you and your partner were placed is a **prisoner's dilemma**. Simple prisoner's dilemmas are games in which two agents face a decision between two courses of action, A and B, with the following properties: for each player, no matter what the other player does, B will provide a better outcome than A; but a situation in which both players do A is mutually preferable to a situation in which both do B.

The version of the prisoner's dilemma just described can be modeled by the following chart:

Courses of action	Possibility 1: Your partner confesses	Possibility 2: Your partner stays silent
Confess	5 years in jail	go free
Stay silent	10 years in jail	2 years in jail

Once you represent the choice in this way, one important fact about cases of this sort becomes clear: confessing **dominates** silence. No matter what your opponent does, you are better off confessing.

This means that there is a very strong argument, using dominance reasoning, for the conclusion that the rational thing to do in a prisoner's dilemma is to confess (or, more generally, to perform the action such that if you both do it the mutually less preferable outcome results).

If this seems plausible, then why think that there is anything paradoxical about the prisoner's dilemma?

Courses of action	Possibility 1: Your partner confesses	Possibility 2: Your partner stays silent
Confess	5 years in jail	go free
Stay silent	10 years in jail	2 years in jail

Once you represent the choice in this way, one important fact about cases of this sort becomes clear: confessing **dominates** silence. No matter what your opponent does, you are better off confessing.

This means that there is a very strong argument, using dominance reasoning, for the conclusion that the rational thing to do in a prisoner's dilemma is to confess (or, more generally, to perform the action such that if you both do it the mutually less preferable outcome results).

If this seems plausible, then why think that there is anything paradoxical about the prisoner's dilemma?

The problem arises from the fact that there also seems to be a plausible argument available in favor of silence. For consider: you and your partner are very similar, and are being presented with exactly the same choice. Therefore it seems quite plausible that **whatever choice you make, your partner will make the same choice**. So you should approach your choice knowing this, and you should choose whatever action will lead to the best outcome, given the assumption that your partner will pursue the same course of action. Since you are better off if both of you remain silent rather than both confess, you should remain silent.

This is, in a way, analogous to the reasoning used to support 1 boxing in response to Newcomb's problem. There the idea was that, although your decision does not **cause** the Predictor to put money in the box (or not), we do have good reason to believe that there is a correlation between the Predictor's decision and yours. Therefore, the 1 boxer argues, we should pursue that course of action which will maximize the money obtained on the supposition that if we 1 box, the Predictor puts money in the box, and that if we 2 box, he does not.

And so we can give a reply to this argument for silence which is the analogue of the 2 boxers reply to the 1 boxer. Remember that the 2 boxer was inclined to say something like this:

> We do have good reason to believe that your choice is well-correlated with the Predictor's prediction. But your choice does not cause the Predictor to do anything - whether the money is in the box or not does not causally depend on how many boxes you choose. So now, after the money is in the box, it is rational to take both boxes. (This is so even if it is rational, **before the Predictor decides what to put in the box**, to get yourself into a 1 boxing frame of mind, or do whatever you think might increase the odds of the Predictor taking you to be a 1 boxer - even if this involves trying to convince yourself that 1 boxing really is the best course of action.)

Courses of action	Possibility 1: Your partner confesses	Possibility 2: Your partner stays silent
Confess	5 years in jail	go free
Stay silent	10 years in jail	2 years in jail

The problem arises from the fact that there also seems to be a plausible argument available in favor of silence. For consider: you and your partner are very similar, and are being presented with exactly the same choice. Therefore it seems quite plausible that **whatever choice you make, your partner will make the same choice**. So you should approach your choice knowing this, and you should choose whatever action will lead to the best outcome, given the assumption that your partner will pursue the same course of action. Since you are better off if both of you remain silent rather than both confess, you should remain silent.

This is, in a way, analogous to the reasoning used to support 1 boxing in response to Newcomb's problem. There the idea was that, although your decision does not **cause** the Predictor to put money in the box (or not), we do have good reason to believe that there is a correlation between the Predictor's decision and yours. Therefore, the 1 boxer argues, we should pursue that course of action which will maximize the money obtained on the supposition that if we 1 box, the Predictor puts money in the box, and that if we 2 box, he does not.

And so we can give a reply to this argument for silence which is the analogue of the 2 boxers reply to the 1 boxer. Remember that the 2 boxer was inclined to say something like this:

> We do have good reason to believe that your choice is well-correlated with the Predictor's prediction. But your choice does not cause the Predictor to do anything - whether the money is in the box or not does not causally depend on how many boxes you choose. So now, after the money is in the box, it is rational to take both boxes. (This is so even if it is rational, **before the Predictor decides what to put in the box**, to get yourself into a 1 boxing frame of mind, or do whatever you think might increase the odds of the Predictor taking you to be a 1 boxer - even if this involves trying to convince yourself that 1 boxing really is the best course of action.)

And we can say a similar thing about the argument for silence in the prisoner's dilemma:

> We do have good reason to believe that your choice is well-correlated with that of the other prisoner. But your choice does not cause the other prisoner to do anything - whether or not you confess does not cause him to confess, or stay silent. So it is rational to stay confess since, whatever he does, you will be better off by confessing.

Is this convincing? Are there any important disanalogies between the prisoner's dilemma and Newcomb's problem, or for consistency must 2 boxers be in favor of confessing, and 1 boxers of silence?

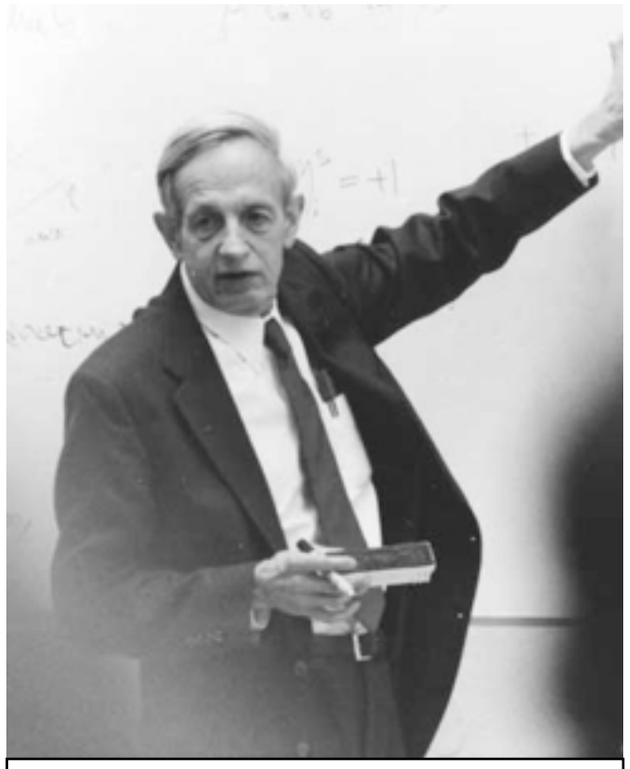
Courses of action	Possibility 1: Your partner confesses	Possibility 2: Your partner stays silent
Confess	5 years in jail	go free
Stay silent	10 years in jail	2 years in jail

To understand why the prisoner's dilemma is interesting and also problematic, it is useful to introduce the notion of a Nash equilibrium (named after John Nash).

Roughly, a Nash equilibrium is a set of strategies for a given task which is such that, given what every other player is doing, no player would be better off by changing his strategy. Intuitively, it seems that a Nash equilibrium will be a rational approach to the task, a kind of stable system of cooperation - after all, if none of the participants can improve their lot by improving their strategy, then mustn't each be happy with the way things are going?

Suppose that we have the above 2 player prisoner's dilemma: what combination of "silence" and "confess" is a Nash equilibrium?

The only Nash equilibrium is the state of both players confessing. But this is not, intuitively, an outcome with which either player should be happy; but it is also in a way the only stable strategy, since given any other combination of strategies, at least one of the players will be better off changing their strategy.



Would you have cast Russell Crowe to play this man in a movie?

Many have thought that this sort of problem is mirrored in many "real life" problems. Consider, for example, the so-called "tragedy of the commons", which might be illustrated by the following story:

> A number of dairy farmers live in a town. All have insufficient land for their purposes, so each would be better off if they could let their cows graze on the town common. But if each of them does this with as many cows as they can, the commons will be ruined for everyone.

It is easy to think of other "free rider" problems - for example, having to do with the environment, or trying to get out of jury duty, or taking the time to vote - which can be thought of as instances of the prisoner's dilemma as well. The worry that the prisoner's dilemma illustrates is that if each person is rational, everyone will do the things that, collectively, lead to the worst consequences.

Courses of action	Possibility 1: Your partner confesses	Possibility 2: Your partner stays silent
Confess	5 years in jail	go free
Stay silent	10 years in jail	2 years in jail

It is easy to think of other “free rider” problems - for example, having to do with the environment, or trying to get out of jury duty, or taking the time to vote - which can be thought of as instances of the prisoner’s dilemma as well. The worry that the prisoner’s dilemma illustrates is that if each person is rational, everyone will do the things that, collectively, lead to the worst consequences.

But matters are not quite so straightforward. So far we have just considered a “single play” prisoner’s dilemma - but, as you may have noticed in the game with which we started, the case seems importantly different when we imagine repeated prisoner’s dilemmas involving the same people, since there is the possibility that what those partner’s do in future versions of the prisoner’s dilemma will depend on what you do in this one. And it is plausible that many real-world prisoner’s dilemmas are best thought of as multiple-play versions of the dilemma rather than single-play ones.

So even if confession is the rational course of action in the one-play dilemma, it remains an open question what the rational strategy is for multiple-play prisoner’s dilemmas.

Answering this question is not nearly so simple as addressing the one-play version of the dilemma. But some interesting data can be taken from computer simulations designed to answer this question. One can think of a strategy for a multi-turn prisoner’s dilemma as a set of rules for whether to confess or stay silent based on various features of the particular “turn” of the game being played. For example, simple rules might include: “always confess”, “always stay silent”, and “alternate confessing and staying silent.”

One can then run a computer program which simulates repeated interactions between agents following rules of this sort, and see which rules are most successful in the long term. (Using our example, this would be accumulating the fewest total years served in jail, but one could obviously use other examples as well.) Interestingly, in one pioneering simulation of this sort (performed by Robert Axelrod in the 1980’s) the most successful strategies proved to be the ones which are **nice** strategies, in the sense that they are never the first to confess in a several-turn prisoner’s dilemma.

The most successful strategy overall was the one that Axelrod called “tit for tat”: the strategy which begins with silence on the first turn and then does on every subsequent turn what its opponent did on the preceding turn.

One might regard results of this sort as suggestive for thinking about the evolution of social cooperation: if the most advantageous strategies for dealing with multiple-play prisoner’s dilemmas involve (to continue with our example) sometimes staying silent even though confessing would be the best short-term strategy, perhaps this can help explain how creatures with a disposition to cooperate in this way could have arisen through natural selection.

So even if confession is the rational course of action in the one-play dilemma, it remains an open question what the rational strategy is for multiple-play prisoner's dilemmas.

Answering this question is not nearly so simple as addressing the one-play version of the dilemma. But some interesting data can be taken from computer simulations designed to answer this question. One can think of a strategy for a multi-turn prisoner's dilemma as a set of rules for whether to confess or stay silent based on various features of the particular "turn" of the game being played. For example, simple rules might include: "always confess", "always stay silent", and "alternate confessing and staying silent."

One can then run a computer program which simulates repeated interactions between agents following rules of this sort, and see which rules are most successful in the long term. (Using our example, this would be accumulating the fewest total years served in jail, but one could obviously use other examples as well.) Interestingly, in one pioneering simulation of this sort (performed by Robert Axelrod in the 1980's) the most successful strategies proved to be the ones which are **nice** strategies, in the sense that they are never the first to confess in a several-turn prisoner's dilemma.

The most successful strategy overall was the one that Axelrod called "tit for tat": the strategy which begins with silence on the first turn and then does on every subsequent turn what its opponent did on the preceding turn.

One might regard results of this sort as suggestive for thinking about the evolution of social cooperation: if the most advantageous strategies for dealing with multiple-play prisoner's dilemmas involve (to continue with our example) sometimes staying silent even though confessing would be the best short-term strategy, perhaps this can help explain how creatures with a disposition to cooperate in this way could have arisen through natural selection.

One might also think that results of this sort can help to explain the evolution of morality. Whether or not this idea is ultimately correct, it at least faces some initial stumbling blocks, given that moral codes at least sometimes mandate the performance of some actions which would not be successful in multi-play prisoner's dilemmas (such as sacrificing your life for someone else).

So far we have discussed single-play and multiple-play versions of two person prisoner's dilemmas. But as the tragedy of the commons illustrates, there can be prisoner's dilemmas which involve indefinitely many players. Somewhat surprisingly, versions of the prisoner's dilemma can also arise involving just one player.

One such example is, arguably, Warren Quinn's example of the self-torturer.

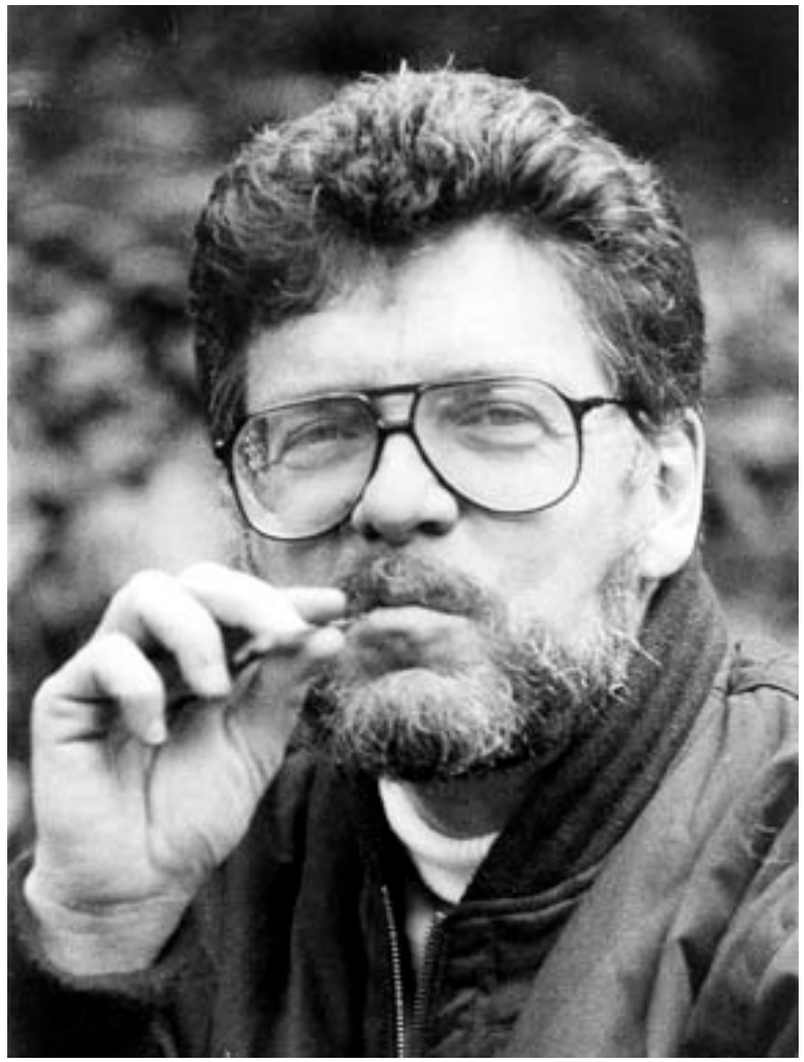
So far we have discussed single-play and multiple-play versions of two person prisoner's dilemmas. But as the tragedy of the commons illustrates, there can be prisoner's dilemmas which involve indefinitely many players. Somewhat surprisingly, versions of the prisoner's dilemma can also arise involving just one player.

One such example is, arguably, Warren Quinn's example of the self-torturer.

Suppose there is a medical device that enables doctors to apply electric current to the body in increments so tiny that the patient cannot feel them. The device has 1001 settings: 0 (off) and 1 . . . 1000.[1] Suppose someone (call him the self-torturer) agrees to have the device, in some conveniently portable form, attached to him in return for the following conditions: The device is initially set at 0. At the start of each week he is allowed a period of free experimentation in which he may try out and compare different settings, after which the dial is returned to its previous position. At any other time, he has only two options — to stay put or to advance the dial one setting. But he may advance only one step each week, and he may *never* retreat. *At each advance he gets $10,000.*

Since the self-torturer cannot feel any difference in comfort between adjacent settings, he appears to have a clear and repeatable reason to increase the voltage each week. The trouble is that there *are* noticeable differences in comfort between settings that are sufficiently far apart. Indeed, if he keeps advancing, he can see that he will eventually reach settings that will be so painful that he would then gladly relinquish his fortune and return to 0.[2]

The self-torturer is not alone in his predicament. Most of us are like him in one way or another. We like to eat but also care about our appearance. Just one more bite will give us pleasure and won't make us look fatter; but very many bites will. And there may be similar connections between puffs of pleasant smoking and lung cancer, or between pleasurable moments of idleness and wasted lives.



Suppose that we separate the self-torturer's life with the device into 1000 decisions about whether to stay put or advance. Consider each such person-at-a-moment as a separate "person-stage". Now ask: is each person-stage better off, or worse off by advancing? Clearly, for the reasons Quinn gives, better. But the whole collection of person-stages is clearly worse off if everyone advances than if none of the person-stages do.

Let's think a bit more about the relationship between this and the original presentation of the prisoner's dilemma.

Suppose that we separate the self-torturer's life with the device into 1000 decisions about whether to stay put or advance. Consider each such person-at-a-moment as a separate "person-stage". Now ask: is each person-stage better off, or worse off by advancing? Clearly, for the reasons Quinn gives, better. But the whole collection of person-stages is clearly worse off if everyone advances than if none of the person-stages do.

Let's think a bit more about the relationship between this and the original presentation of the prisoner's dilemma.

Let's begin by thinking about a simpler version of Quinn's example. Suppose that the machine has just three settings, instead of 1001: 0, 1, and 2. Suppose further that setting 0 is indistinguishable from setting 1, and setting 1 is indistinguishable from setting 2, whereas setting 0 is distinguishable from setting 2. Let's call this the **simple machine**. Suppose, further, that the gap between 0 and 2 is big enough that one would not be willing to advance from 0 to 2 for $10.

Now let time 1 be a time at which you are considering advancing from 0 to 1. Should you do it for $10,000? It seems that you should, given the above remarks. We can think of the choice like this:

Courses of action	Possibility 1: At time 2, you will advance the device	Possibility 2: At time 2, you will not advance the device
Advance the simple machine	Be at 2 with $10	Be at 1 with $5
Don't advance	Be at 1 with $5	Be at 0 with $0

Given that 2 is indistinguishable from 1, it looks like advancing dominates not advancing.

Now let time 2 be a subsequent time at which you are considering advancing from 1 to 2. Should you do it for $10,000? Just the same reasoning shows that you should.

Does this show that there is something wrong with dominance reasoning? After all, this seems to lead to setting 1000 on Quinn's device.

Remember what we said about the original prisoner's dilemma: whereas there is a pretty straightforward dominance argument in favor of confessing in the single-play dilemma, matters are not so straightforward in the multiple-play versions of the dilemma, since there one's choice can affect the probabilities of the various possibilities in later versions of the dilemma.

Does this show that there is something wrong with dominance reasoning? After all, this seems to lead to setting 1000 on Quinn's device.

Remember what we said about the original prisoner's dilemma: whereas there is a pretty straightforward dominance argument in favor of confessing in the single-play dilemma, matters are not so straightforward in the multiple-play versions of the dilemma, since there one's choice can affect the probabilities of the various possibilities in later versions of the dilemma.

Can we say something parallel here? Perhaps we can say that the reason why you should **not** advance at time 1 is that your doing so will give you a later choice which you know will lead to a worse outcome.

Then perhaps we could treat time 2 as the last iteration of a multiple-play prisoner's dilemma, in which it does seem rational to confess (or here, advance the simple machine). But this is not paradoxical, since it really does seem that you are better off spending the rest of your life at setting 1 with $5 than at setting 0 with $0 (so long as you never get another chance to advance the machine). (This is one respect in which Quinn's dilemma is more like real-life versions of the prisoner's dilemma: if one knows that one has the chance to advance the machine every week, and one is not sure when one is going to die, then one can never be quite sure when the last play of the game is.)

But there is still something a little bit weird going on here. We are saying that we can reject the dominance argument in favor of advancing the machine from 0 to 1 at time 1 on the basis of the fact that, if we do so, a future application of dominance reasoning will lead us to a bad result. What we're doing, it sounds like, is saying that we can ignore what a rule R tells us is rational to do if following R in this case would put us in a position to apply rule R again --- and, if we did **that**, we'd be worse off. If this is true, doesn't it just follow that R is not in general a reliable rule governing rational decision making? (Here, of course, R=the rule of dominance.)

This is something to think about. But one might defend the dominance principle against examples like Quinn's in a different way, and that is just to say that the idea of a machine like the one he describes is incoherent: maybe it is impossible to progress from a pain-free state to a state of intolerable pain via a series of steps each of which is indistinguishable from the last. What do you think?