MAKING GOOD CHOICES: AN INTRODUCTION TO PRACTICAL REASONING

CHAPTER 12: PRACTICAL REASONING IN POTENTIALLY COOPERATIVE DECISIONS – THE STAG HUNT AND THE PRISONER'S DILEMMA

This chapter continues the topic of Chapter 11: potentially cooperative games. We will examine decision making in two games, the stag hunt and the prisoner's dilemma. Each represents a challenge to practical reasoning to arrive at and justify a rational choice, but one that is different than the challenge contained in the clash of wills and chicken.

12.1 Stag hunt

If you were doing group work that required everyone in the group to pitch-in in order to be successful, and you found out that one (or more) members of the group left to do other things, what would you do? Most people would follow the others and abandon the group project, for unless everyone did his or her share there is no chance of success. But what if no one had left the group (yet); nevertheless, you strongly expected one (or more) would leave before completion, dooming the project. Increasingly you feel that your continued time and effort on this group project will be wasted, if (as you suspect) other members abandon the project. Would you be the first to leave, realizing that your departure would make others likewise "jump ship"? Or, would you continue your efforts until someone else became the first to leave? If there are things that can only be done together, and you can't trust others to do their part, then it seems that you might be better off doing things you can do alone.

But of course, we all realize that we can get a lot more done working together than we can do working alone. This is the case on the smallest level of mutual assistance where, for example, two people might join together to help each other push a stalled car to the side of the road, or three friends might get together to move a large heavy piece of furniture to another room . But as soon as one helper abandons the task, it no longer makes sense for the other(s) to continue. And, if others can't be trusted to pull their weight, it seems to make no sense to join such a collective effort in the first place.

We also realize the benefits of working together on a local or neighborhood level that involves small groups of volunteers contributing to a project. Let's say that the local school play is over and there is a call for volunteers to help clear all 400 folding chairs from the gym floor. The building must close in a half hour, so the job must be done quickly. As soon as the first few volunteers start, others join to help ease their task. As the number of volunteers grows, others are attracted to lend a hand for there is less each has to do. This typically snowballs to the point where many are pitching in, no one has to do a lot of work, and the job is done well before the half hour deadline. But now reverse this process; suppose that initially there were many volunteers but after a few moments people who thought they had the time to help remove chairs realized they had to leave (or suppose that each chair was a lot more trouble than anyone had realized at first: they are old, heavy, don't fold easily, etc.). Volunteers start to offer excuses and leave. As the number of volunteers decreases, the job becomes harder for those who remain, and they start to wish that they had not volunteered and now also begin to leave. At some point the remaining volunteers, working faster and faster to meet the half hour deadline, give up for they see they no longer can accomplish the job on time. Yet, had everyone stayed the job would have been completed in time. But given that volunteers would be leaving the job before it's finished, the remaining helpers feel that their decision to pitch in seems foolish.

If we turn to the largest social unit, the whole social system itself, we likewise see this same decision problem. For society to function, members must trust one another to "do their part." If there is a belief on anyone's part that enough of the others have given up their contributions toward the working of society and have become untrustworthy, then doubt and insecurity start to grow. Members begin to anticipate a breakdown in social services and a collapse of the networks of mutual reliance and mutual assistance that every society requires to be minimally successful. People stop contributing to the functioning of society as this breakdown in trust spreads. Clearly, it is no longer rational to continue contributing to a social system that seems no longer workable and looks like it's in the process of falling apart. (In an effort to highlight the importance of trust, we are putting to one side those parts of the social system that don't centrally depend on trust and decision, and so would continue to function if there were a breakdown of mutual trust, such as essential government functions, emergency services, police duties, and the like. But even here a minimal measure of trust is needed, unless society had a higher level of “enforcers” ensuring these essential services continued.)

Trust, then, is a very important bond between people. Without it, small groups of voluntary associates break down into individuals who are on their own, and small group efforts are abandoned; larger organizations and neighborhoods tend to dissolve into small groups, and local projects can’t be completed; and finally the social system itself can break down into local centers of power, and many large scale social goods cannot be achieved. On each level, a minimum amount of trust and trustworthiness is at the center of mutual reliance, at the center of each person’s expectation that others will be sufficiently reliable in fulfilling their part of the bargain, and at the center of everyone else’s belief that you, for example, will be responsible and committed to contribute to joint projects what is expected of you. On each level, the decision to participate and “do one’s part” is closely connected to a degree of trust and the expectations built on that degree of trust.

As important as trust is, however, it is a very fragile thing. The slightest hint that someone’s trust in others has weakened makes their trust in each other weaker, for unless every member does his or her part the group project, remember, can’t be successful. This breakdown in trust can quickly accelerate to the point where a joint effort is abandoned and each agent finds it best to do things on his or her own. Note that we are not talking about cases in which it is the failure of the effort that causes those involved to give up; the stag hunt, as we shall see, is a game that focuses attention on cases where it is the other way around: the expectation that those involved will give up can cause failure of the effort. The problem of having to trust others to hold up their part of the bargain (and likewise the problem of others trusting you to do your fair share), and knowing when to “jump ship” and go it alone, accomplishing things that you can do yourself, are important and delicate factors in many of our social interactions as well as on different levels of social interaction. Here is an example, a little model, of the importance and the fragility of trust.

Imagine that a few workers, trying to clear several heavy objects and a few light things from an area, have lifted a very heavy large object over their heads and must carry it a short distance and drop it. If just one worker does not use all his efforts, the object will fall, perhaps on one of the others crushing him; but if all use maximum effort they will succeed. Each is fully committed to the task and will carry his load providing, that is, he is sure the others are doing the same. As they struggle along, they continually assure each other that each is managing the load well. As they continue to strain with their load, however, each becomes increasingly aware that he must let go and jump out of the way if any other worker gives any sign that he will falter or is about to let go. As they struggle, the risk increases that one will falter. They trust each other less and less, for at any point one could let go in the heightened expectation that another will do so, and no one will let himself be crushed. Suddenly, one worker falters ever so slightly and makes a quick effort to regain position under the load. But this makes the others unsure and the load starts to tip. As they struggle to regain position, they also prepare to let go and jump out of the way. Their effort at holding up the load is now no longer maximum, each feels this, and they all abandon the effort at the same instance, each looking out for himself and "regretting" that the others might be crushed. The load crashes down, and each has saved his own skin in the nick of time. They give up this joint project, which now doesn't get done, and instead each only carries objects that he can lift alone, a less important task for clearing the area.

This little model of the importance and fragility of trust also highlights the main features of the stag hunt. Each agent can cooperate or defect. Each will cooperate to achieve a cooperative goal (remove the heavy object) that no agent can achieve alone, but only as long as the others do likewise. There is another goal (remove light objects), not as good as the cooperative goal, given that everyone is cooperating to achieve it, but a lot better than nothing (remove nothing). If there is a weakening of trust and other agents start to defect, defecting is better than cooperating.

Before we analyze and evaluate the stag hunt, let's look at a few more examples. The name "stag hunt" comes from the philosopher Jean Jacques Rousseau who first illustrated this game using a deer hunt. It is customary in the theory of rational choice to present this game using Rousseau's example.

Two hunters are hunting stag, which takes quite a bit of effort, but the likely result (though not guaranteed) is that they will eventually be successful and will eat well. Each cannot hunt stag alone, he will fail and go hungry. However, it is much easier to hunt hare, which each can do alone and is a sure thing. If both hunt hare, the result will be eating poorly, but if one hunts hare while the other is still after stag, the one does better. As they hunt stag, each trusts the other to cooperate and not (secretly) take off after hare; if one does, the other will also. As they hunt stag, periodically a hare happens by. For each hunter, it is tempting to abandon the stag hunt and take off after a hare, for the stag hunt requires trusting that the other hunter will stick to it and there is always the risk that the other will take off after a hare, making the one agent’s stag hunt efforts done in vain. What should each hunter do: hunt stag together (cooperate) or hare alone (defect)?

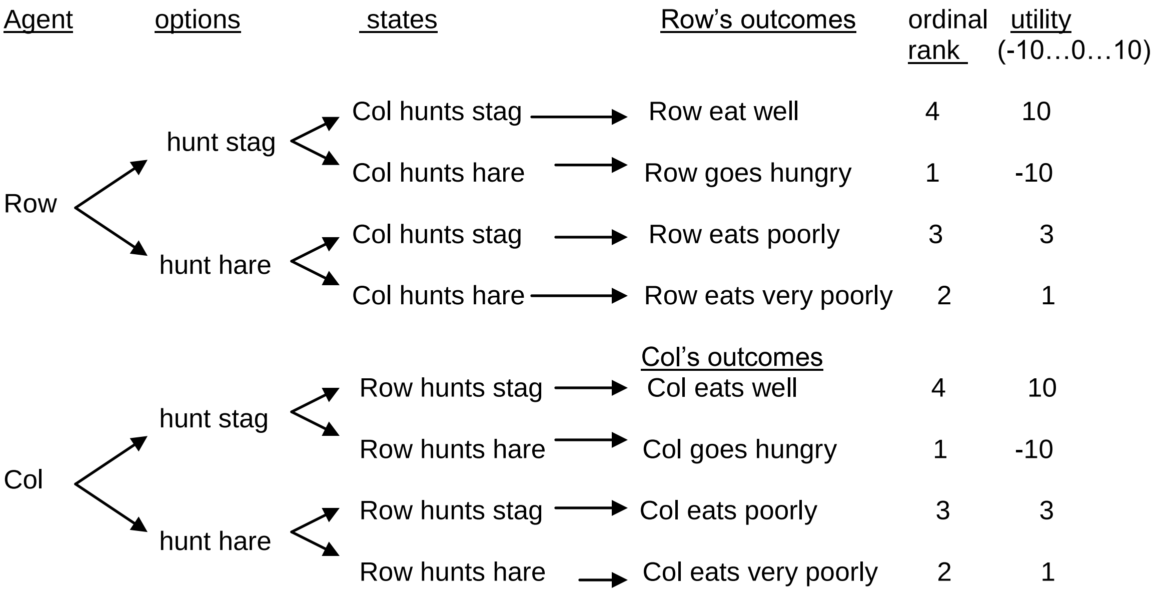
Here is an example of the stag hunt that often occurs during wartime. High school sweethearts are planning to get married but he is called to military service. He is a soldier in the war zone and will have to remain there for a tour of duty lasting several years. This couple has promised to stay together and remain faithful to their plans; their cooperative goal is to marry. However, after a long time apart things are becoming strained and difficult, and there are opportunities on each one’s part to start other relationships. They trust each other to stay together, but each realizes the growing risk that the war separation will take its toll on their feelings for each other. If one becomes involved with someone else, the other obviously will do likewise, and their future together will not happen.

Let’s look at an example involving international relations. Suppose two nations share a border with a war-torn third nation from which there is a massive refugee problem. If the two nations make a cooperative effort, they can jointly take in and help a large number of refugees from the third nation, but if the two nations act alone each can take in and help only a small number of refugees. They must act quickly, for the third-country refugees are beginning to die. Each nation accepts the good intentions of the other to cooperate on the refugee problem, but each realizes that this international-cooperation decision is officially up to the legislative powers in each nation’s government, and there is no assurance in either nation that the vote will favor cooperation. However, each nation does not need a legislative vote to take in and help a small number of refugees alone, it can immediately be done.

Here is an example of the stag hunt involving institutions. Imagine two small local colleges. Each has its history, traditions, and identity: one is an all-female college, and the other is an all-male college. If they continue to go it alone, they will remain small, not very well known, and will have periods of struggle. However, if they merge into a single co-ed college, the experts have assured each that the new college has a good chance of growing into a large, successful university. Merging, however, means neither can retain its history, traditions, or identity, and each must trust an entirely new administration to lead the proposed new college to success, as well as trust each other not to attempt to retain its independent identity or to reverse the decision to merge once the process has started and the new larger college faces challenges to establish itself. Remaining separate is better than merging with a partner who is not fully committed.

As a final example, picture a couple whose marriage is going through rough times. They have been married many years and have raised a family, but their enjoyment of each other's company seems to have declined, their interests have diverged, and neither feels the love they once experienced for each other. Meanwhile, each has achieved financial independence in a professional field. While they realize that there are still many important emotional and practical benefits to remaining married and want it to work, each is nevertheless considering divorce. They still trust each other not to abandon the marriage and start divorce action, but there is risk (each believes) that the other's assurance in these matters might not hold out. Each feels that divorce and complete independence is better than continued marriage to someone whose vows might have weakened and whose commitment could be doubtful.

Following tradition, we'll analyze the stag hunt using Rousseau's example, but you can easily substitute the specifics of any of the above examples. Each agent in Rousseau's example has the same goal: to acquire food to eat. (In the stag hunt, cooperation itself is not the goal, it's the means by which a common good is achieved.) To cooperate is to hunt stag in a joint effort; success here means a lot of food and eating well. To defect is to go after hare alone; this is guaranteed food, but of small quantity and poorer quality than stag. However, if only one agent hunts hare, there is more hare to eat than if both hunt hare.



Agent	options	states	Row's outcomes	ordinal rank	utility (-10…0…10)
Row	hunt stag	Col hunts stag	Row eat well	4	10
		Col hunts hare	Row goes hungry	1	-10
	hunt hare	Col hunts stag	Row eats poorly	3	3
		Col hunts hare	Row eats very poorly	2	1
			Col's outcomes		
Col	hunt stag	Row hunts stag	Col eats well	4	10
		Row hunts hare	Col goes hungry	1	-10
	hunt hare	Row hunts stag	Col eats poorly	3	3
		Row hunts hare	Col eats very poorly	2	1

We can see from this analysis that each agent desires mutual cooperation to a free ride, a free ride to mutual defection, and mutual defection to being a sucker. This preference order defines the game **stag hunt**. Now let's put this particular example of the stag hunt into matrix form and see if practical reasoning can discover a rational choice.

		Col: C1: cooperate (stag)	C2: defect (hare)
Row:	R1: cooperate (stag)	4, 4 (10, 10)	1, 3 (-10, 3)
	R2: defect (hare)	3, 1 (3, -10)	2, 2 (1, 1)

As you can see, neither Row nor Col has a dominant option. Suppose these agents reason by maximin; that is, what if each agent should "play it save" because the goal of acquiring food and eating is so very important? By maximin reasoning, the rational choice is (R2,C2) for the (1,1) outcome is clearly a saddle point. Also, this outcome is a Nash equilibrium; no agent could achieve as much or more of the goal by cooperating given that the other has defected; in fact, any agent who singly switched would lose the entire goal (go hungry in this example). It looks, then, as if the stag hunt has a rational choice solution: (R2,C2). However, I hope you are thinking: “not so fast! – isn’t it clearly more rational in the stag hunt for each agent to cooperate? In fact, isn’t (R2,C2) unacceptably sub-optimal, downright irrational, compared to mutual cooperation?”

This idea of “unacceptably sub-optimal” should be made into an explicit principle of practical reasoning, at this point, for it is an important principle that will come up again in the next potentially cooperative game we’ll be looking at, and will play a large role in the next chapter on bargaining and negotiations. The Italian economist and social theorist Vilfredo Pareto argued that it would be wrong for a society to accept an economic system of distribution that brought about a certain level of general welfare, if an alternative economic system of distribution was available to that society that would give everyone a higher level of welfare. “Wrong” is not understood in the moral sense (though it might also be morally wrong for a society to make such a social choice), it is more the economic sense of “wasteful” or “inefficient.” An economic system that results in a society being less well off is not as efficient in how goods get distributed as an alternative system that distributed those same goods resulting in the society being better off. If we add that social welfare is one of the goals of social organization, then accepting (choosing) an inefficient economic system when a more efficient one is available would be irrational for a society to do. An economic system, then, is said to be “Pareto efficient” or “Pareto optimal” if there is no such better alternative. In other words, if no one in a society can be made economically better off unless someone else in that society is made economically worse off, then the society’s economic system is at its Pareto optimal. In rational choice theory, Pareto’s idea is applied to potentially cooperative games as a principle of practical reasoning. An outcome said to be **Pareto sub-optimal** if the game has another outcome giving at least one player a higher payoff and no player a worse payoff. And if there is no such alternative outcome, an outcome is said to be **Pareto optimal**. Given that agents desire maximum goal achievement in game decisions, a Pareto sub-optimal outcome is unacceptable; it would be an irrational choice. Ideally, rational choices in potentially cooperative games should give agents Pareto optimal outcomes.

In the stag hunt (R2,C2) yields a Nash equilibrium outcome and in this respect they are rational choices. But (R2,C2) yields a Pareto sub-optimal outcome and in this respect should be unacceptable choices. But (R1,C1) are also rational choices; the mutual cooperation outcome (10, 10) is a Nash equilibrium outcome that is Pareto optimal. In a stag hunt, then, who in their right mind would switch (be the first to "jump ship") and singly give up the mutual cooperation outcome? Not only would it greatly reduce goal achievement for the agent who defected (going from utility 10 to utility 3), it would "punish" the other agent, who after all was being cooperative with the defector, with a plunge from her hope limit (full goal achievement: utility 10) to her security level (potential starvation: disutility -10). To take off after a hare, leaving your co-worker in the lurch, seems like the height of selfishness. It also looks like a betrayal of all those who rely on that agent to do his or her part of the work, dooming the collaborative project to failure. But most important for the theory of rational choice, it seems irrational – how could anyone decide not to cooperate? Let's consider this question in comparison with the two games we investigated in Chapter 11, the clash of wills and chicken.

In the stag hunt there are two Nash equilibrium outcomes. In this regard, the stag hunt is similar to the clash of wills and to chicken. But in the (strong) clash of wills and chicken each agent should choose different options; if one cooperates the other should defect, and if one defects the other should cooperate, so that one gets the free ride payoff and the other gets the sucker's payoff. Also, in each of these two games the two Nash equilibrium outcomes are equal (they are symmetric), each sums to the same utility, and each is Pareto optimal; there is equal "rational pull" or "rational attraction" toward each equilibrium outcome. In the stage hunt, however, each agent should choose the same option (defect if the other defects, cooperate if the other cooperates); in the two Nash equilibrium cells no agent gets the sucker's payoff. But, these two equilibrium outcomes are not equal; mutual cooperation is much better for both agents than the sub-optimal mutual defection outcome. The "rational pull" toward mutual cooperation ought to be much stronger than that of mutual defection. So, what's the problem in the stag hunt? Why isn't (R1,C1) the rational choice, plain and simple, and (R2,C2) irrational by the Pareto optimal principle? This is exactly where, in this decision problem, the worrisome issues of trust, risk, and assurance enter the picture. While choices (R1,C1) are justified by two methods of practical reasoning (Nash equilibrium and Pareto optimality), choices (R2,C2) are likewise justified by two methods (Nash equilibrium and maximin reasoning). And it turns out that maximin reasoning has special power in the stag hunt because of the issue of trust and its associated risk.

The problem with the mutual cooperation outcome is that trust is necessary to bring it about (if the agents are not yet hunting stag) and to sustain it (once they are hunting stag), but trust is fragile. Relying on other people involves risk, not so much because people are unreliable and untrustworthy out of a flaw in their character (though some people are, and in their case it is all-the-more rational to "play it safe" with self-protective maximin reasoning). Rather, mutual cooperation is risky in the stag hunt because of two factors. (1) It is human nature sometimes to falter when tasks are demanding, no matter how much determination and commitment a person might have. For example, an agent might get sick, or have a sincere change of heart, or suffer fatigue, or simply die, and as a result can't be expected to be unconditionally trustworthy. Did you ever have great plans, perhaps a vacation, fall apart because someone on whom the plans depended got sick or changed her mind at the last moment? (2) Trust between people who are involved in a joint undertaking tends to be reciprocal; that is, the levels of trust in agents are attuned to each other so that a change in one agent's level of trust causes a change in the other's levels of trust, which causes a change in the one's level, and so on. Trust is a bond between agents that continually readjusts in strength.

Now add these two factors together. If (1) it is part of the "human condition" that people can't always be expected to be completely trustworthy to fulfill their part of a collaborative project in spite of their good intentions, their promises and commitment, and determination (this is common knowledge), and (2) weakened trust tends to spreads to everyone in collaborative projects, then the mutual cooperative outcome, as attractive as it looks, might not be as practically rational as it appears. Trust is risky and risk, as we have seen in earlier chapters, discounts the value of the outcome. With risk, outcome utility becomes *expected* outcome utility. Defecting (hunting hare in our example) is an option whose outcome is only a small part of the goal (eating poorly but at least eating) – that's the downside. But it is an option that represents independence; all it takes, in theory, is self-reliance. It is in this regard risk free. Mutual cooperation (hunting stag) has a high likelihood of gaining the agent the full goal – that's the upside. But it is an option that represents dependency; an agent needs the help of others (as they need the agent's help). Mutual reliance can't be assured by practical reasoning, and so there is risk that others (or you!) will give up the project, making all your (their!) time and efforts given over to the collaborative venture in vain. Worse, the other agent might not even tell you that she has abandon the project, leaving you to find out on your own that you have been working alone on an important task that can't be done by one agent.

We are left, then, with the rationally unsatisfying, disappointing, sub-optimal, but safe maximin solution to the stag hunt: mutual defection. The rational choice appears to be to eat hare alone rather than trust that agents will be eating stag together. In the spirit of Rousseau, it seems that it is more rational to achieve small goals independently than to try for large goals in a condition of dependency on others. If this rational choice solution to the stag hunt is too hard to swallow, and if you think that the mutual cooperation outcome is too valuable to give up on, how might mutual cooperation be assured? Here are some possibilities to consider, but note that each one brings into the picture elements that are not methods of practical reasoning, and so mutual cooperation in these cases would not be justified or sustained by a rational choice. They involve non-rational additions to practical reasoning so that the mutual cooperative Nash equilibrium outcome is assured. What do you think of these possibilities? Can you think of any others?

1) Perhaps an agent could cooperate (hunt stag) only with cooperative family members or loved ones whose emotional ties to one another would never allow anyone to defect (take off after a hare).
2) Perhaps an agent could cooperate only with members of an organization, like a club or a street gang, all of whom have sworn to a loyalty oath never to leave a collaborative project until the goal has been achieved under pain of harsh punishment.
3) Perhaps an agent could cooperate only with others who have accepted an "enforcer of the project" – a boss or ruler or power that would assure that any agent who defected would be caught and forced to cooperate. In the case of international relations, a world power or international "enforcer" would be necessary, perhaps something like a greatly strengthened United Nations.

4) Perhaps an agent could cooperate only with agents all of whom have made a binding agreement, say a legal contract, not to leave the project until the goal is achieved or face a "breach of contract" lawsuit. In the example of the married couple considering divorce, some societies make divorce very difficult in an effort to assure mutual cooperation (the marriage union).

5) Perhaps an agent could cooperate only with others who come from a certain background or culture; that is, they have been brought up or have been taught to conform to social goals, to value community over individuality and contributions to society over self-interest, and to find meaning in life through service to others and collective efforts rather than through independence and self-reliance. (If this possibility is too weak, lets say that mutual cooperation is achieved by a vast social effort of indoctrination so that a society turns out people whose "trust" in one another and whose loyalty to the group and sense of patriotism makes defecting psychologically impossible to contemplate as an option. The loyalty and group cohesion brought about in military units might be a good example of this possibility, so that in battle very few soldiers ever desert their "brothers" or "sisters.")

6) Perhaps an agent could cooperate only with agents who are so deeply in each other's debt that they owe it to each other to cooperate (the debt might be financial, or exchanges of important favors, or even that they owe each other their very lives).

While these possibilities might assure mutual cooperation, note what has happened. In these scenarios the goal and the outcomes have been changed so that the defection option does not yield a lesser version of the goal having at least some value for the agents as does, for example, hunting hare in the original stag hunt game. Now defection is connected with outcomes having strong disutilities for the defecting agent, typically worse than the sucker's payoff: the threat of a lawsuit or severe punishment or the loss of loved ones and family bonds. In effect, mutual cooperation is practically assured in these six possibilities by adding non-rational elements to augment practical reasoning, and because of this the game has been changed from the stag hunt to a version of harmony. It is important to see that these possible ways of achieving mutual cooperation are not rational choices within the stag hunt resulting from the application of practical reasoning, instead they are external ways of changing the stag hunt into another game. In the end, the mutual defection Nash equilibrium outcome appears to be the rational choice for each agent in the stag hunt by maximin reasoning, even thought it is disappointing to see that it is not the Pareto optimal outcome. While it looks as if the outcome having maximum utility (mutual cooperation) should be easily achievable by rational agents using only methods of practical reasoning, the crucial requirement of trusting others and being assured of their cooperation makes this a shaky and perhaps overly optimistic ideal.

12.2 Prisoner's dilemma

If you have ever been taken advantage of or exploited, especially concerning an important matter, your experience will have taught you a painful lesson: protect yourself when you are vulnerable, if at all possible. This seems to be one of life's valuable lessons, often learned the hard way. Security and self-protection are important things we owe ourselves and should try to achieve. It seems only reasonable that agents should be concerned for their own well-being, especially when this seems to be under threat. This is the case for all types of individual agents: people, institutions, and of course nations. But what if the only way to protect yourself from being exploited is to exploit someone else (who is likewise trying to protect himself from being taken advantage of)? And suppose that this only invites retaliation, requiring all-the-more self-protection? The prisoner's dilemma is a game that explores this difficult position we sometimes find ourselves in: the position, that is, that to protect yourself you end up hurting yourself as well as others.

The prisoner's dilemma is the single most studied potentially cooperative 2-person game. It is rich in its power to represent a wide variety of human interactions, frustrating in its resistance to any easy or satisfying solution, and fascinating in the way it combines many of the interesting aspect and challenging tensions we have seen in the clash of wills, chicken, and the stag hunt. It has become the classic non-zero sum game for showing the difficulty of rationally achieving and maintaining cooperation. Let's look at some examples before analyzing and evaluating this decision problem, starting with a version of the original prisoner story that gave this game its name.

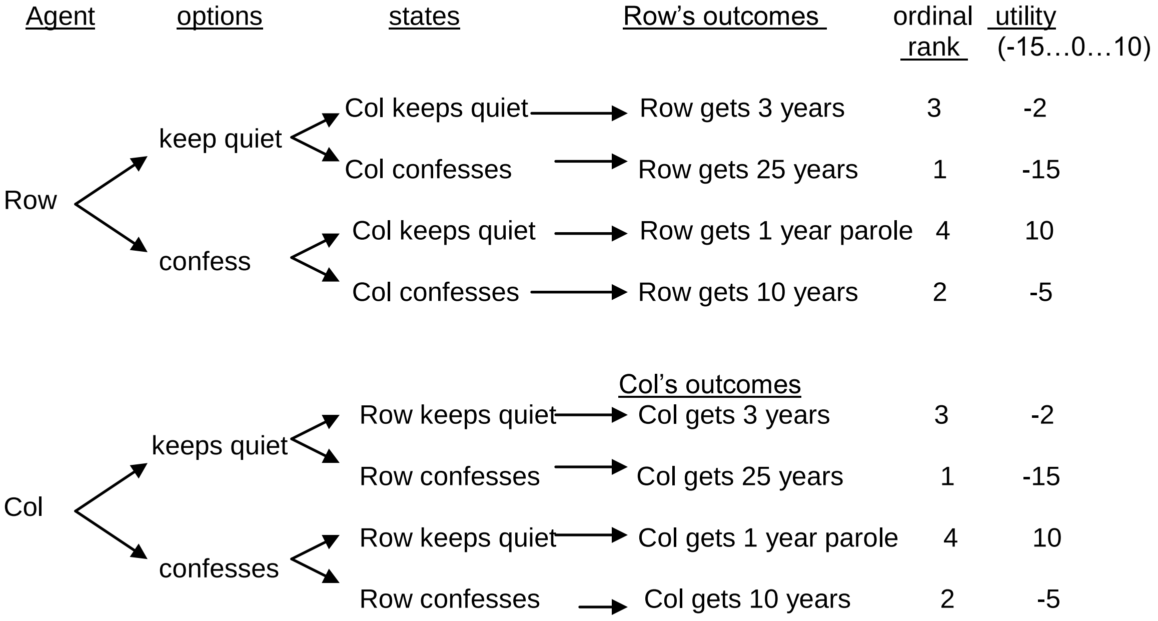
Imagine that two criminals are caught and held for armed robbery. The police separate them and each is offered the same deal (and each knows the other is offered it). If you confess that both you and your partner committed the crime and your partner refuses to confess, you will be rewarded with a light sentence of 1 year parole, no jail time, but your partner will get the maximum of 25 years for armed robbery using your confession to convict him. If you both confess, you'll each get 10 years in prison, reduced from the maximum 25 year sentence for helping with your joint conviction. However, if neither of you confesses, the best case the police have is a breaking and entering charge that will get you each 3 years in jail. You have one hour to decide what to do: confess (and turn your partner in) or keep quiet (cooperate with your partner). What should each prisoner do?

While we'll be looking at the prisoner's dilemma involving only two agents, here is an example with many people. Imagine a crowded nightclub; there is a great band and everyone is having a good time. Suddenly a fire breaks out and everyone must exit the building. If each person walks to an exit, lines up orderly, and takes their turn at the exits everyone will escape but many will inhale smoke and suffer minor burns. But if the people run for the exits to get out quickly, some will make it safely, but many will be trampled and injured and will suffer permanent lung injury from inhaling toxic smoke as well as disfigurement from burn injuries. In addition, many who are not so quick to the exits will end up dying in the fire and smoke because the exits will be jammed with people struggling to get out first. (Think of this scene not as fear and panic, but as a decision that each customer in the nightclub must make between an orderly emergency exit and one where each person rushes to beat others to the exits.) Of course, no one in the club wants to be the last one out. Each one thinks: if the others walk, I had better go a little faster to get ahead of them and assure my safety, but if the others rush I better rush for an exit even faster to assure my safety. (This example is based on a real event, a tragic fire in 2003 in the Station Concert Club, West Warwick, Rhode Island in which over 100 died. Survivors reported that most customers entered by the same few doors, and when the fire started – caused by the band's use of fire in their act – everyone rushed to get out quickly through the exit they had entered by, trampling many and trapping many in the smoke and flames.)

Here is an example from international relations. Two nations are historical rivals and each has developed for good reason a suspicion of the intentions of the other. They share as a common border a large river that is an important natural resource for each: the river provides fish for both food and industry for each nation. Technology and population growth have brought fishing rates to the point where the fish population can no longer support the food and economic requirements of both nations, but it can for one of them. The powerful fishing industry of each nation is lobbying hard to step up fishing so as to out-fish the rival nation before the fish population gets too low due to the other side's "rampant and reckless" fishing. (They argue: we must beat them to the fish or our fishing industries will go bankrupt and many of our people will be unemployed and sink into poverty). Meanwhile, environmentalists in each nation are counter-lobbying equally hard to reduce fishing drastically until the fish population rebounds, and thereafter limit fishing to a rate both nations can sustain without endangering the fish population. (They argue: the fish are more important than our tradition of over-reliance on our fishing industry or our rivalry with the other nation.) Each nation would prosper, provided the other nation reduced its fishing rates; but neither will reduce its fishing rate if the rival doesn't – this is completely unacceptable to each side. If each nation attempts to beat their rival to the fish by increasing fishing rates, the short-term gain will be far outweighed by the long-term loss of a traditional food source and a major industry. Each nation would be willing to adopt the environmentalists' plan, but only if the rival nation does likewise. However, each side has a well-founded suspicion that, should it reduce fishing, their rival will exploit this "generous unilateral cooperative gesture" by attempting secretly to increase fishing rates. In this climate of protecting vital national fishing interests and mutual suspicion, what should each nation do?

As a last example of the prisoner's dilemma, imagine a college class taking the final exam in a challenging course: let's say it is a practical reasoning course and the final exam is on non-zero sum games, with heavy emphasis on the prisoner's dilemma. Several students have been skipping class (this is not a course in their majors) and have fallen behind in studying this material. They now face the prospect of getting a low grade on the final exam, and maybe even failing – unless, that is, they cheat. These are all seniors who really need to pass the course to graduate. If they all cheat, however, the instructor will clearly discover the pattern and will fail them all. But if one, or just a few, cheat on the final exam, there is a good chance of not getting caught, passing the final and the course. However, those who don't cheat face a likely low grade, failing the final, and even failing the course; this would mean not graduating, having their post-graduate plans and having their job prospects fall apart. They can't admit to each other that each is considering cheating on the final exam, for someone might inform the instructor, and besides each realizes cheating is a moral failure and so would be embarrassed to admit such a plan. Thus, each must decide in secrete to cheat on the final exam or to take it honestly and risk failing the course. What should each of these students, who have fallen behind on the prisoner's dilemma material, do on their final exam: cheat or not?

It has become tradition in rational choice theory to present the prisoner's dilemma using the example that gave this game its name. We will analyze and evaluate this version, but you should try to do this with one or more of the other examples: two patrons deciding to walk or run to an exit to escape a fire, or two nations in rivalry over a shared decreasing vital resource, or two students contemplating cheating on an important final exam. For the prisoners, the goal is minimum jail time (we will understand this as a form of self-protection). Let's simply call one prisoner "Row" and the other one "Col". Cooperating in the prisoner's dilemma does not mean cooperating with the police by confessing; it is Row and Col cooperating with each other: that is, keeping quiet and not ratting on your partner by confessing to the crime. Defecting is to try to reduce your jail time by confessing and thereby turning against your partner.



Agent	options	states	Row's outcomes	ordinal rank	utility (-15…0…10)
Row	keep quiet	Col keeps quiet	Row gets 3 years	3	-2
		Col confesses	Row gets 25 years	1	-15
	confess	Col keeps quiet	Row gets 1 year parole	4	10
		Col confesses	Row gets 10 years	2	-5

			Col's outcomes		
Col	keeps quiet	Row keeps quiet	Col gets 3 years	3	-2
		Row confesses	Col gets 25 years	1	-15
	confesses	Row keeps quiet	Col gets 1 year parole	4	10
		Row confesses	Col gets 10 years	2	-5

We can see from this analysis that each agent prefers a free ride to mutual cooperation, mutual cooperation to mutual defection, and mutual defection to being a sucker. This preference order defines the game **prisoner's dilemma**. Now let's put this particular example of the prisoner's dilemma into an outcome matrix and see if practical reasoning can discover a rational choice.

		Col:	
		C1: cooperate (keep quiet)	C2: defect (confess)
Row:	R1: cooperate (keep quiet)	3, 3 (-2, -2)	1, 4 (-15, 10)
	R2: defect (confess)	4, 1 (10, -15)	2, 2 (-5, -5)

Look at this matrix carefully, it represents very clearly why the prisoner's dilemma is such a troubling interdependent decision problem. The mutual cooperation outcome gives each agent the 2nd best goal achievement (-2 disutility each), whereas mutual defection results in only the 3rd best (-5 disutility each). From the agents' point of view (though not from the police's) 10 years jail time is a lot worse than 3 years. So, because each agent desires the goal (better: deeply fears losing the goal since it would mean serious punishment in the form of jail time) each fears a lot more that both will choose confessing than that both refuse to confess. But each agent also fears 3 years jail time much more than no jail time, and so strongly desires to confess – providing, that is, her partner refuses to confess. The temptation to confess on condition that the other agent will keep quiet would be powerful, virtually irresistible, to anyone caught in a prisoner's dilemma. Of course, what each most fears is the maximum 25 year sentence; this is to be avoided above all. It is pretty clear, then, how the emotions of fear and desire will push and pull these prisoners trapped in this dilemma. Put yourself in their shoes, facing potential jail time; your emotions are powerful and there is no way to opt out. You have just an hour to make a decision.

What does practical reasoning reveal as the rational course of action? Notice that each agent has a strong dominant option; no matter what the other agent chooses, each agent is much better off in goal achievement defecting (confessing). If Col chooses C1, R2 gives Row a full 10 outcome utility as opposed to R1's -2. And if Col chooses C2, R2 gives Row -5 outcome disutility as opposed to R1's horrible -15. The same for Col; reasoning by dominance shows C2 to be in all possible cases better than C1. (R2,C2) is the clear rational choice by dominance.

How about maximin reasoning? Because agents in a prisoner's dilemma face such threatening possibilities to their well-being (e.g., jail time, death or disfigurement by burning, depletion of food resources, not graduating college(!) in the examples above), it would seem that self-protective maximin reasoning is especially appropriate. Again, (R2,C2) is the rational choice by the maximin method; it results in a clear saddle point outcome (-5, -5).

Notice, finally, that (R2,C2) results in the game's only Nash equilibrium outcome. No agent could possibly be rational switching to the other option, given that the other agent doesn't switch from the equilibrium option. The mutual cooperation outcome is clearly not an equilibrium point; just as we saw to be the case in chicken, in the prisoner's dilemma mutual cooperation is unstable. Each agent has very strong practical reason – so strong as to be overwhelming – to switch from mutual cooperation to the defect option; again, however, only on condition that the other agent stays with the cooperation option.

So, by all three principles of practical reasoning – that is, by Nash equilibrium outcomes, by maximin reasoning, and by dominance – the rational choice for each agent is to confess (defect). Given our assumption for interdependent decision problems that agents are equally rational, the result is the mutual defection outcome of 10 years jail time each.

At this point, I hope you are reacting as most people who learn about the prisoner's dilemma react: this can't be! It is a deeply unsatisfying solution. It seems, to use the principle introduced earlier, too *Pareto sub-optimal* concerning such a vital matter to be acceptable. As in the stag hunt, in the prisoner's dilemma you probably feel that if only the agents could cooperate they would be much better off, in this case a 3 year jail sentence as opposed to 10 years jail time. If you are thinking about the other examples above, say the terrible Station Concert Club fire, inhaling smoke and receiving minor burn injuries (the worse case possibility for mutual cooperation) are a lot better than living with permanent lung damage and disfigurement from burn scars (the worse case for mutual defection). How can it be rational, then, for each agent to defect so that they end up with the drastically sub-optimal mutual defection outcome?

In a prisoner's dilemma, it is rational for each agent to defect because (1) the maximin principle assures self-protection when much is at stake; an agent who does not cooperate cannot be exploited and be forced to live with the dreadful sucker's outcome. As bad as the mutual defection outcome might be, it is better than being a sucker in a prisoner's dilemma. The defection option avoids the worst of the worse outcomes. Go back to the Station Concert Club fire; the sucker's payoff is very likely death by fire. In the example of the prisoners it is 25 years in jail, the maximum sentence. Think about it: what would you do?

It is also rational to defect because (2) the free ride payoff is so tempting: it represents full goal achievement. The dominance principle assures the free ride payoff is not ruled out. The defection option keeps the best of the best outcomes among an agent's possible payoffs. If one agent were sure that the other agent would choose the cooperative option, all-the-more an agent rationally should defect and exploit this willingness to cooperate, *this irrational choice*, on the other's part. Again, reflect on this point: what would you do? Put yourself in the Club fire; would you move toward an exit a "little faster" than everyone else if they were walking in an orderly fashion, in order to make sure you got out of the burning building sooner rather than later? Suppose you were one of the prisoners: would you exploit your partner-in-crime's willingness to cooperation, his confessing, and by doing so reduce your jail time to zero? Once you appreciate these two considerations (namely, (1) how bad the sucker's payoff is in a prisoner's dilemma and (2) how attractive goal achievement is), put them together. Each agent knows how tempting it is for the other to exploit his cooperation, and how terrible the sucker's outcome would be. The only rational choice, then, is to defect. And so, the mutual defection outcome seems unavoidable as the rational choice solution in a prisoner's dilemma.

Still, it seems incredible that the cooperation option is not more rational than the defection option, for the mutual cooperation outcome clearly represents far better goal achievement than the mutual defection outcome. The latter, compared to the former payoff, looks unacceptably sub-optimal. And yet it is not easy to see how mutual cooperation can be achieved in a prisoner's dilemma. Take communication, for example. If practical reasoning in isolation from each other does not result in mutual cooperation, let's see what happens if we allow the prisoners to communicate with each other before making their decisions. Their communication, we may suppose, leads to an agreement not to confess. But if each prisoner merely gives his word, his promise or pledge, to keep quiet, why should they trust each other? If they don't, defection is rationally required. And if each agent actually believed the other's agreement to choose to cooperate, the temptation to defect is made stronger not weaker. Just as we saw in the game chicken, in a prisoner's dilemma recall that assurance that the other agent will choose the cooperative option gives an agent the strongest possible practical reason to defect. So, an agreement alone (one without a guarantee to back it up) is at best worthless, no better than empty words. At worst, it is nothing but a ploy to set the other agent up to be taken advantage of. An agent would be naïve and gullible, indeed, and would make herself much too vulnerable, to choose the cooperative option based only on an agreement, no matter how trustworthy and honest each believed the other to be. Too much is at stake. An agent that cooperated as a matter of faith in the other's good word might get lucky and actually achieve the mutual cooperation outcome, but it would be just that – luck – not a rational choice.

If an agreement alone on each agent's part to forfeit a greater individual gain (the free ride outcome) for a second best outcome (mutual cooperation) will not do it, what happens if we make the agreement binding? Suppose each agent can provide a guarantee that she will stick to the agreement. As we saw in the case of the clash of wills and in the stag hunt, an agreement is made binding by setting up physical, or psychological, or moral consequences for breaking it that the agent fears to an even greater degree than the agent fears the sucker's payoff. These "punishments" have to be worse than the sucker's payoff, and the desire to avoid then greater, or they will not have the power to be a guarantee that the agreement to cooperate won't be exploited. So, for example, let's imagine that the two prisoners are members of a gang that will kill any member who rats on a partner. Or, to take the Club fire example, let's suppose the government will impose a guaranteed death sentence on anyone who rushes to be among the first to escape a fire that killed people who couldn't make it to safety because they were trampled or found exits blocked with “free riders.” Or, suppose that people’s moral principles were so strong that they literally couldn’t live with themselves if they were free riders, or that their religious beliefs meant eternal damnation as punishment for choosing the free ride in a prisoner’s dilemma. Such severe consequences as these would have the power, let’s suppose, to make an agreement to cooperate binding.

But now look what has happened. To make the agreement binding, that is, to back it up with a guarantee that an agent’s cooperation will not be taken advantage of, we must introduce negative outcomes that are worse for the agents than the sucker’s payoff. This is no longer the prisoner’s dilemma. This is to change the game from the original prisoner’s dilemma form into something else, it is clearly not a rational choice that yields the mutual cooperation outcome in a prisoner’s dilemma. An agreement, then, won’t do the trick; if it is a binding agreement it is not a prisoner’s dilemma, and if the agreement is not binding it is powerless to bring about mutual cooperation.

This is also true for other possible ways of assuring cooperation. Suppose, for example, that the agents were loved ones or family members with whom they could never defect, or that the agents were indoctrinated by their education and culture always to cooperate, or that a powerful “enforcer” of cooperation existed. I’m sure that you can imagine other possibilities that would make both agents choose the cooperation option. None of these, however, is a method of practical reasoning that yields mutual cooperation as a rational choice in the prisoner’s dilemma. Instead, this is to resort to emotional bonds and social/political mechanisms that transform the original game into a very different decision problem. They are, in our special sense of the term, non-rational additions to practical reasoning, needed because the standard methods of practical reasoning result in a rational choice that agents find unacceptable.

12.3 The iterated prisoner’s dilemma

We have been looking at the **one-time prisoner’s dilemma** in which the goal represents an extremely important value for each agent’s well-being. In the examples used, there is little or no chance for the game to be repeated with one or both agents remaining the same. This form of prisoner’s dilemma, as you now know, has a rational choice solution, but it is one that is so troubling that many philosophers and other rational choice researchers who study this game treat it as unsolved and continue to seek a cooperative solution. But there seems to be no better rational choice short of giving the game up and transforming it into a more manageable decision problem.

Fortunately, this situation improves if we weaken the game. Let's relax the urgency of the goal, reduce the fear factor of goal loss, and make it a game that can be repeated. This is called the **iterated (weakened) prisoner's dilemma**. Here is an example that will serve for people as well as for nations as agents. Suppose two teenagers who don't live near each other (or two nations) collect cards; each wants to trade cards (goods) with the other as a way of improving their collection (resources). They must do this by mail (by shipping), each sending the desired items to the other; this is mutual cooperation. Each, naturally, would like to keep their own cards (goods) as well as own the other agent's cards (goods). In other words, each would like to have a desired card (goods) sent, but is very tempted not send her own card (goods) to the other, or perhaps send only a worthless card (damaged/defective goods). This is the free ride outcome. But neither wants to be fooled so as to end up minus a card (a shipment of good) with nothing or something worthless in return – the sucker's payoff. When the time comes to send the trading card (ship the goods), what should each agent do: sent their valuable item (cooperate) or not send it (defect)? (I will not frame this decision problem into a branching diagram or construct its outcome matrix, but for practice you might give it a try.)

Both teens (nations) have many cards (goods) to trade, so the exchange will be taking place repeatedly. The goal, of course, is to gain all the cards (goods) over time that each one desires to own. Once the initial round of exchange is finished, each agent – and this is an important point – will remember what the other agent did, and will use this information to decide what to do on the next trade. This means that after the 1st round an agent's **reputation** becomes an important factor in an iterated prisoner's dilemma. If each agent cooperates on the 1st round, each establishes a cooperator's reputation with the other which becomes stronger with each round in which they continue to decide to cooperate. But if one (or both) defects on the 1st exchange, the 2nd round is drastically affected.

Let's suppose, first, that there is mutual defection in the initial exchange; each side tried to sucker the other, as well as protect their valuable item from a form of "theft", by holding it back. Each now has a reputation with the other as a defector that can't be trusted to cooperate in a prisoner's dilemma. In effect, the exchange didn't happen and the status quo results. While there is no goal achievement on either part, at least nothing of value has been lost. There won't be future exchanges for it is now irrational to trust each other to cooperate on a replay. For an agent to defect on the 1st round of play in an iterated prisoner's dilemma, then, amounts to the agent treating the decision problem as a one-time prisoner's dilemma in which defection is the rational choice.

Now let's suppose that the 1st round exchange results in the free ride payoff for a defecting agent (the agent now has two cards or two shipments of goods: the one held back and the one received) and the sucker's outcome for the other agent (a trading card or shipment of goods sent and either nothing or something worthless in return for it). The cooperating agent has been deceived and will *want* the game to continue, or at least will want the free rider to <u>expect</u> future rounds, first in order to have a chance to get even, and second to be able to retaliate for having been exploited. How might retaliation happen? One possible way is simply by cutting off future trade, depriving the free rider of any further goal achievement. Another possibility is by finding a way to lure the free rider (who will now expect the other agent to retaliate) into further rounds and at some point get even by <u>unexpectedly</u> exploiting this former defector's present cooperation by defecting. So – and this is another important point – a willingness to **retaliate**, a refusal to "forgive and forget" or to let "bygones be bygone," must be part of an agent's reputation in an iterated prisoner's dilemma.

Here is a little story that wonderfully illustrates the importance of both reputation retaliation. A barber, who doesn't get paid very much for haircuts, has come to rely on tips to bolster his earnings. One day a customer comes in for a haircut, and the barber immediately pegs this person as a poor tipper. So the barber gives him a quick, very bad haircut. The customer pays for the haircut and then, to the barber's surprise, gives the barber a huge tip. Some time goes by and this customer comes back to the same barber for another haircut. This time the barber, who remembers the overly generous tip last time, spends a lot time and effort giving the customer a perfect haircut, the best the barber knows how to give. The customer pays and then tips the barber only 1 penny! Before the customer leaves, the barber asks for an explanation: “Last time I gave you a rotten haircut and you gave me an amazing tip. This time I go all out and give you a great haircut, and you only tip me a penny. What gives?” The customer answers: “*This* tip is for the last haircut, the last tip was for *this* haircut.” (What do you expect will happen when this customer returns a third time for a haircut?)

The iterated prisoner’s dilemma, then, is not a series of independent games involving the same agents and the same goal; it is better conceived as one multi-play game stretched out over time in which – starting with the second round – both history (the memory of past choices) and expectation (that there will be retaliation for defection) affects present and future decisions. Recall the gambler’s fallacy (Chapter 5.4) in which independent events were falsely taken to be connected by “memory”. In the iterated prisoner’s dilemma it is the other way around. Each round except the 1st <u>has</u> memory (in the form of reputations agents earn with each other), and it would be incorrect to treat each round of play as a series of independent events.

What, then, has the power to make mutual cooperation in an iterated prisoner’s dilemma more likely than not? Given that each agent has a goal that can only be achieved over time through repeated prisoner’s dilemma interactions, it is thought that three common knowledge conditions are necessary (the third is discussed in the next sub-section). (1) <u>The reputation condition</u> – each agent will remember the history of the other agent’s choices. (2) <u>The retaliation condition</u> – each agent believes that the other will somehow punish defection and get even by damaging future goal achievement. (Recall that these are common knowledge in the sense that both agents know these things about the other, and both know that both know these things about the other.) The idea, then, is that two rational agents who have this common knowledge and who find themselves in the first round of an iterated (weakened) prisoner’s dilemma will reason practically to choose the cooperation option and so achieve the mutual cooperation outcome, providing they can avoid the following problem.

12.3.1 The backward induction problem

It is important to see that cooperation in an iterated prisoner's dilemma very much depends on the special nature of the goal. Practical reasoning is able to discover that initial and continued cooperation is the rational choice because (1) the long-term sum of goal achievement by cooperation outweighs the short-term partial goal achievement that the free ride would bring the defector on the 1st (or an early) round, and (2) the potential goal loss in receiving the sucker's payoff is not so damaging to an agent as to be unsustainable. The iterated prisoner's dilemma is a much weakened version of the one-time version. In the one-time prisoner's dilemma the goal is typically of great value and urgency to the agent; its loss can be life threatening in some cases, and is not a goal to be achieved bit-by-bit over time by repeated decisions to cooperate. In the iterated case, an agent can afford (though prefers not) to be suckered in the short-term; this is generally not true in the one-time case.

However, the very same properties of the goal that allows for rational cooperation in the iterated prisoner's dilemma creates a serious problem if one or both agents know the exact number of times there will be an interaction. Suppose, to keep with our trading card example, 5 cards are involved and in each prisoner's dilemma iteration one card each is to be exchanged. (Use the same numbers for the case of two nations trading goods, if that's the example you are working with, so that there are to be 5 shipments of goods between each nation to complete the trade contract). The 1st round will establish a history (reputation), but notice that the 5th round has no future for it ends the game. This means, in effect, that history stops with the 4th round. In the last interaction, the 5th, neither agent need fear memory or retaliation, for there is no 6th round in which they can take effect. Reputation does not extend to a 6th play (there isn't any!) and so it can't function in the 5th to deter the free ride temptation that makes for such trouble in the one-time prisoner's dilemma. Thus, the 5th and last round becomes separated from the series and takes on the form of a one-time prisoner's dilemma in which it is rational not to cooperate.

With the 5th round gone the way of mutual defection, the 4th round becomes the last interaction in the series for cooperation to take place. But now the same practical reasoning will apply to it as applied to the 5th interaction. The same two conditions that tend to make cooperation rational in the series of interactions will not extend into the 4th round; neither agent needs a reputation for cooperation and neither need fear retaliation for defecting in the 4th round as a way to assure cooperation in the next (5th) round, for the next round of exchange is already doomed to mutual defection. So, now the 4th exchange breaks off from the series of prisoner's dilemmas and takes on the form of a one-time interaction. Each agent will find it rational to defect in the forth round. This makes the 3rd round the last iteration for cooperation to be rational, but … you see the problem. The whole original 5-play iterated prisoner's dilemma falls apart "backward" into a series of 5 independent one-time prisoner's dilemmas in each of which, starting with the last, defection is the rational choice. The cooperation option can't get started, whether it is 5, 10, or thousands of prisoner's dilemma iterations. So long as there is a last round the agents know about before they reach it, this last round of play transforms into a one-time interaction and this makes the next to last do the same. Iteration by iteration entire series collapses backward from the last to the first like a row of dominos. This is **the backward induction problem**; it makes cooperation irrational in an iterated prisoner's dilemma of known definite length to the agents.

To overcome the backward induction problem, to keep one multi-play (iterated) game from collapsing into a series of one-time independent games, a third necessary condition is needed for cooperation in an iterated (weakened) prisoner's dilemma. Along with (1) the reputation condition and (2) the retaliation condition, for cooperation to be likely (even if not guaranteed) there must be (3) an **ignorance** condition – the number of interactions or iterations of the game must be unknown to the agents. In other words, it must be common knowledge to the agents that they are in an iterated prisoner's dilemma of undetermined length, open-ended, extending indefinitely into the future. For example, the two girls who are trading cards might believe that there will be ongoing card trading (or for the two nations in a trade agreement, an ongoing exchange of goods). Even if this turns out to be false (say one of the girls loses interest at some point in trading cards, or one of the nations discovers a natural resource in its own territory and no longer desires the other nation's goods), the agents must be ignorant of when the series will end. As soon as they know, the backward induction problem kicks in and practical reasoning can no longer justify the cooperation option as the rational choice.

This ends our analysis and evaluation of the prisoner's dilemma, the last potentially cooperative game covered in this and the previous chapter. It is a troubling yet fascinating interdependent decision problem, as you can now appreciate, worthy of the vast amount of attention it has received. Once you become acquainted with this decision problem, you'll start to see one-time and iterated prisoner's dilemmas (as a threatening possibility if not actually taking place) wherever you see people trying to achieve cooperative interaction.

12.4 Summary: the sub-optimal outcome problem as a failure of practical reasoning

In Chapter 11 we saw that the clash of wills and chicken presented a challenge to practical reasoning: the equilibrium selection problem. No principle of rational choice, it seems, justifies a pair of choices – one for Row and one for Col – that selects one of two equally good outcomes. Recourse to "external" non-rational (but not irrational!) help in the form of flipping a coin or artful use of deception is required. It seems that practical reasoning, guided by the norms and methods of rational choice theory, reaches a limit in these two decision situations. With what problem do the stag hunt and the prisoner's dilemma challenge practical reasoning?

The stag hunt and the prisoner's dilemma are similar in that both seem to suffer from "too little." Each has a single rational choice solution by one or more standard methods of practical reasoning, but these solutions seem to fall so far short of what mutual cooperation would achieve of the goal as to be unacceptable to reasonable agents. Not only many rational choice theorists, but also the strong intuitions of non-experts, reject (or at least would like to reject) mutual defection, the rational choice in each case, as too sub-optimal, yielding only the third best (second worse) outcome out of four possibilities. Surely, they argue (and we all feel), humans can – and often do – do better than that. In these two important games, then, practical reasoning leaves agents with a **sub-optimal outcome problem**, and the benefits of cooperation are lost.

Consider again the rationalist-behaviorist controversy; we saw at the end of Chapter 11 that the equilibrium selection problem in both the clash of wills and chicken appears to strengthen the behaviorist position. The same can now be stated about the sub-optimal problem in the stag hunt and the prisoner’s dilemma. In light of these two problems, thinkers from a variety of disciplines that study human decision making have come to believe that perhaps rational choice theory is not just incomplete in offering ideals of practical reasoning, but has reached a dead end. After all, the behaviorist argues, in the long course of human evolution people have learned to cooperate to an amazing degree; the equilibrium selection and the sub-optimal problems have evidently not been a serious obstacle. There must be something about cooperative decisions that has allowed for this, something that rational choice theory has failed to discover and justify. Except for harmony (and perhaps the special cases of the weak clash of wills and the iterated weakened prisoner’s dilemma), practical reasoning and making rational choices appears to be an obstacle and not a path to human cooperation, at least for those cases of possible human interaction that fits any of these four games: the clash of wills, chicken, the stag hunt, and the prisoner’s dilemma.

The rationalist-behaviorist debate is a deep, complex, perennial philosophical controversy. We have seen just one version of it concerning the status of the methods and standards of practical reasoning as found within rational choice theory. In spite of this debate, however, for the purposes of this text we will continue to hold to the rationalist position. This means that we should try to satisfy the ideals and follow the methods of rational choice in our practical reasoning, and if we fail to do so we will look for flaws in ourselves instead of casting doubt on the theory.

EXERCISE:

1) Identify the potentially cooperative games (stag hunt or prisoner's dilemma) represented by the following matrices. (See if you can recognize the game from the matrix by trying to put yourself in each of the agent's shoes, rather than mechanically matching up the matrix with the forms in the chapter.) Which are symmetrical and which asymmetrical, and if asymmetrical which are Row-biased and which Col-biased?

a) example: Col

		C	D
Row	C	5, 5	-3, 0
	D	0, -3	-1, -1

stag hunt (symmetrical)

b) Col

		C	D
Row	C	4, 4	-9, 7
	D	9, -10	-8, -9

c) Col

		C	D
Row	C	.5, 1	-3, 10
	D	10, -3	0, 0

d) Col

		C	D
Row	C	25, 20	5, 10
	D	15, 1	7, 3

e) Col

		C	D
Row	C	0, 0	-6, 6
	D	6, -6	-2, -2

f) Col

		C	D
Row	C	5, 12	0, 7
	D	3, -5	1, 2

g) Col

		C	D
Row	C	-1, -1	-10, -2
	D	-2, -10	-3, -3

h) Col

		C	D
Row	C	-3, 1	-5, 8
	D	11, -2	-7, 0

i) Col

		C	D
Row	C	-5, -5	-15, 1
	D	1, -15	-10, -10

2) Here are four scenarios. Each is one of the potentially cooperative game covered in this chapter: stag hunt or prisoner's dilemma. (1) Analyze each into a branching diagram and game matrix; (2) identify the game by name, and (3) explain why the appropriate principles of practical reasoning justify a rational choice solution that appears unacceptable.

a) Roger and Carol are amateur musicians, he plays piano and she is a violinist. They are neighbors and for some time have been talking about playing music together and perhaps even performing together. In the past their plans for a musical evening have been frustrated, their demanding careers have interfered. At the last minute several times Roger had to cancel, leaving Carol with nothing to do for those evenings. Similarly, at other times at the last minute Carol had to cancel. Roger and Carol have now been asked to perform at the local high school featuring an evening of local talent; their goal is to do so giving the best performance. They discuss the possibilities. They can each prepare a difficult piece of music to perform together at the show, and this would surely be the high point of the evening. Or they can each prepare a simple piece to perform alone; this would make an adequate contribution to the talent show, but nothing as spectacular as performing together. The worse thing to happen, however, is for one to prepare their part of the difficult piece of music, only to have the other cancel at the last minute. Each, unfortunately, doesn't have the time to prepare both the difficult and the simple pieces of music. What should each prepare: the difficult music they can only perform together, or the simple music each can perform alone.

b) (With some changes to fit our purposes, this story was reported by the CBC, 4/8/2008). In southern California two families with adjourning properties each had the goal of conserving energy. One family (Solar) put up an array of solar panels that generated their electricity. Their neighbors (Shade) had planted rows of trees for shade, thereby reducing air-conditioning needs. A problem developed when the trees grew to a height where they provided full shade: they also blocked the full day's sunlight from the solar panels. After a series of requests and complains, counter-requests and counter-complaints – neither family agreeing to give up their energy conserving measures and each becoming angry at the other's interference – they grudgingly agreed to this compromise: Shade will cut down ½ their trees and Solar will take down ½ their solar panels. That is, they angrily agreed to reduce goal achievement by half. But each family secretly considered another option. Shade could keep all their trees and try to damage all the solar panels; this would solve the problem for Shade, if the destruction could be made to appear accidental, giving Shade full goal achievement. Likewise, Solar could keep all the solar panels working and kill off all of Shade's trees; this would solve the problem for Solar, if the damage to the trees could be made to appear the result of natural processes, giving Solar full goal achievement. What should each family decide, according to rational choice theory, given their goal of energy conservation: do what they agreed or opt for their secret plan?

c) Sally and Harry have just started dating. They live in an exciting city, but one with very high rents and precious few apartments available. Each is paying a large portion of salary toward rent for a very tiny studio apartment. Motivated by how much they are attracted to each other, they have begun to discuss giving up their separate apartments and renting a bigger apartment together. They agree that there are good reasons for deciding to do this: they desire to be together, their decision to do so would assure each of their commitment to one another and strengthen their relationship, both would save on rent, and an appealing apartment has just become available in Sally's building which must be grabbed right away or it will be gone very quickly. In their excitement at the prospects, strong emotions towards each other, and desire to assure each other that each was "serious" about their relationship, they decide to go ahead with this plan. The next day at work, however, the "glow" is lessening and each starts to have second thoughts. Friends point out to each that it seems a bad decision: they really don't know each other very well, they could not continue to live together if their relationship ended, one would have to move out and whoever did could not get their old apartment back and very likely wouldn't find anything to rent; this would jeopardize their ability to live and work in their city. These doubts made perfect sense, but they saddened Sally and Harry who now feel torn. The law does not allow them to rent their current apartment to someone else while they live together; each must either stick with their earlier decision to live together (if it works, this would result in a wonderful outcome), or must announce to the other that they are reversing their earlier decision. If both decide to reverse, this would surely weaken their new relationship but it might survive. The worse case, each believes, is to be emotionally committed to their earlier decision while the other announces they are reversing; it would not be possible for the committed one to continue dating the other after such a "rejection." What, according to rational choice theory, should Sally and Harry each decide: give renting together a try or stay in their separate apartments?

d) A terrorist group, the "People's Dagger," has been fighting for religious freedom. The National Army has been trying to eliminate the People's Dagger and their fundamentalist religion that they see as dangerous to the nation. After a bloody history of terrorist attacks and military counter-attacks, leaders of a terrorist group and heads of the National Army finally have negotiated a meeting to work out the possibility of a compromise peace. Both sides have the same goal ending the cycle of terrorist attacks and military counter-attacks. The People's Dagger has agreed to give up all terrorist activity and to disband if the nation will officially recognize and support their religion. The National Army has agreed to stop all efforts to eliminate the People's Dagger if they will no longer try to win converts to their fundamentalist religion. One option for each side is to attend the meeting in good faith and negotiate the best compromise within their agreements. This will yield for each their 2nd best goal achievement. However, each side has a 2nd option. The People's Dagger could go to the meeting as a ploy and use the occasion to try to kill the heads of the National Army; success would fatally weaken the National Army and would be a great gain for the People's Dagger's cause; full goal achievement but disaster for the National Army. Likewise, the National Army could go to the meeting as a ploy and take the opportunity to kill the People's Dagger's leaders: this would finish off the People's Dagger terrorism movement for good and mean full goal achievement for the National Army. Of course, if both sides take their second option, the "meeting" would turn into one more bloody battle between these two enemies and a continuation of the People's Dagger terrorist campaign. What option should each agent choose according to the principles of rational choice?

Sources and suggested reading:

The literature on non-zero sum games is huge, largely coming from the fields of economics, psychology, and related social sciences, in addition to philosophy. Try searching “prisoner’s dilemma”, for example, and you will be amazed at the number of hits. The main sources for this chapter are Davis (1983) Chapter 5, Luce and Raiffa (1957) Chapter 5, Mullen and Roth (2002) Chapter 8, Resnik (1987) Chapter 5, and Straffin (1993) Chapters 11 - 15. Davis is recommended for its non-technical presentation, Resnik for its presentation from a philosophical point of view, and Straffin for its mathematical clarity. Gauthier (1986) Chapter III and Schick (1997) Chapter 5 offer excellent introductions to games with a focus on practical reasoning and ethics. For specific games, see Skyrms’ (2004) exploration of the stag hunt from an evolutionary (non-rational choice) perspective, and Skyrms (1996) for an evolutionary approach to chicken, prisoner’s dilemma, and other cooperative decision problems. The prisoner’s dilemma has, by far, generated more literature in more disciplines than all the other games combined. Poundstone (1992) is a great read, not to be missed, and the historical background he provides is a real plus. If there is just thing further you’ll read on the prisoner’s dilemma, Axelrod (1984) gets my vote; it has become a classic. For an online source, Stanford University’s Encyclopedia of Philosophy (http://plato.stanford.edu/contents) has excellent current entries for game theory, evolutionary game theory, prisoner’s dilemma, etc. that focus primarily on philosophical issues rather than technical or applied (behavioral) problems, containing extensive bibliographies and links to related online resources.