Lecture Notes on
GAME THEORY

Brian Weatherson

2011

About this Document

These are teaching notes for a course on game theory in June-July 2011, at Arché, St Andrews. I've made some changes in response to comments received in that course. I'm particularly grateful to feedback from Josh Dever, Daniel Rothschild and Levi Spectre, but also to everyone else who attended any of the classes. Some of the material, particularly in the sixth part, needs serious revision, but I thought it useful to post the notes somewhat as they were delivered.

They are *not* original research. If they were, I'd have to cite sources for every (non-original) idea in here, and since most of the ideas aren't original, that would be a lot! I've relied on a lot of sources in putting this together, most notably:

- Wolfram Mathworld.
- Ben Polak's OpenYale Game Theory Course.
- *Playing for Real*, by Ken Binmore, Oxford University Press.

Reproduction

I'm more than happy for anyone who wants to use this in teaching, or for their own purposes, to copy and redistribute it. (Though note that it is still very much in draft form, and there are several things that I hope to improve.)

There are only three restrictions on reusing this document:

1. This can't be copied for commercial use; it must be freely distributed.
2. If you do modify it, the modification must also be freely distributed.
3. If you're using it for teaching purposes, email me at brian@weatherson.org to let me know.

Citation

Since this isn't a scholarly work, it shouldn't be cited in scholarly work.

Contents

Introduction to Games

Game theory is a slighttly oddly defined subject matter. A **game** is any decision problem where the outcome depends on the actions of more than one agent, as well as perhaps on other facts about the world. **Game Theory** is the study of what rational agents do in such situations. You might think that the way to figure that out would be to come up with a theory of how rational agents solve decision problems, i.e., figure out **Decision Theory**, and then apply it to the special case where the decision problem involves uncertainty about the behaviour of other rational agents. Indeed, that is really what I think. But historically the theory of games has diverged a little from the theory of decisions. And in any case, games are an interesting enough class of decision problems that they are worthy of attention because of their practical significance, even if they don't obviously form a natural kind.

Let's start with a very simple example of a game. Each player in the game is to choose a letter, A or B. After they make the choice, the player will be paired with a randomly chosen individual who has been faced with the very same choice. They will then get rewarded according to the following table.

- If they both choose A, they will both get £1
- If they both choose B, they will both get £3
- If one chooses A, and the other B, the one who chose A will get £5, and the one who chose B will get £0.

We can represent this information in a small table, as follows. (Where possible, we'll use uppercase letters for the choices on the rows, and lowercase letters for choices on the columns.)

	Choose a	Choose b
Choose A	£1, £1	£5, £0
Choose B	£0, £5	£3, £3

We represent one player, imaginatively called **Row**, or *R*, on the rows, and the other player, imaginatively called **Column**, or *C* on the columns. A cell of the table represents the outcomes, if *R* chose to be on that row, and *C* chose to be in that column. There are two monetary sums listed in each cell. We will put the row player's outcome first, and the column player's outcome second. You should verify that the table represents the same things as the text. (Actually you should do this for two reasons. One, it's good to make sure you understand what the tables are saying. Two, I'm really sloppy with details, so there's a far from zero chance I've got the table wrong.)

Now let's note a few distinctive features of this game.

- Whatever *C* does, *R* gets more money by choosing A. If *C* chooses a, then *R* gets £1 if she chooses A, and £0 if she chooses B; i.e., she gets more if she chooses A. And if *C* chooses b, then *R* gets £5 if she chooses A, and £3 if she chooses B; i.e., she gets more if she chooses A.
- Since the game is symmetric, that's true for *C* as well. Whatever *R* does, she gets more money if she chooses a.
- But the players collectively get the most money if they both choose B.

So doing what maximises the players' individual monetary rewards does not maximise, indeed it minimises, their collective monetary rewards.

I've been careful so far to distinguish two things: the monetary rewards each player gets, and what is best for each player. More elegantly, we need to distinguish the **outcomes** of a game from the **payoffs** of a game. The outcomes of the game are things we can easily physically describe: this player gets that much money, that player goes to jail, this other player becomes President of a failing Middle Eastern dictatorship, etc. The payoffs of a game describe how well off each player is with such an outcome. Without knowing much about the background of the players, we don't know much about the payoffs.

Let's make this explicit by looking at four ways in which the agents may value the outcomes of the game. The first way is that agents simply prefer that their monetary payoff is as high as possible. If that's the way the players value things, then the game looks as follows.

Game 1	Choose a	Choose b
Choose A	1, 1	5, 0
Choose B	0, 5	3, 3

Whenever we just put numbers in a table, we assume that they stand for **utils**. And we assume that players are constantly trying to maximise utils. We'll come back to this assumption presently. But first, let's note that it doesn't require that players only care about their own well-being. We could change the game, while keeping the outcome the same, if we imagine that R and C are parts of a commune, where all money is shared, and they both know this, so both players utility is given by how much money is added to the commune in a give outcome. That will give us the following game.

Game 2	Choose a	Choose b
Choose A	2, 2	5, 5
Choose B	5, 5	6, 6

In Game 1, the players face the following awkward circumstance. Each individual will be made **better off** by playing A rather than B, no matter what happens, but were they both to have played B, they would both be better off than if they'd both played A. That's not true in Game 2; here what is good for each player is good for the collective.

You might note that I've started numbering the games, and that I didn't number the initial description of the outcomes. There's a reason for this. Technically, we'll say that a game is specified by setting out what moves, or as we'll sometimes call them, *strategies* are available for each player, and what the *payoffs* are for each player, given the moves that they make. (And, perhaps, the state of the world; for now we're just looking at games where only moves matter for payoffs. And we're only looking for now at games where each player makes a simultaneous choice of strategy. We'll return to how general an account this is in a little while.) Specifying the outcome of a game in physical terms doesn't give us a unique game. We need to know more to get a genuine game specification.

There are yet more ways we could imagine the outcomes being mapped into a particular payoff matrix; i.e., a game. Imagine that the players have the following values. First, they care a lot about how much money goes to the two of them together. So the first determinant of their payoff is the sum of the money paid to each player. Second, they care a lot about agreement. If the two players play different strategies, that is equivalent to a cost of £5 to them. So here is the payoff table for players with those values.

Game 3	Choose a	Choose b
Choose A	2, 2	0, 0
Choose B	0, 0	6, 6

Something new happens in Game 3 which we haven't seen before. What is best for the players to do depends on what the other players do. In Game 1, each player was best off playing A, no matter what the other player did. In Game 2, each player was best off playing B, no matter what the other player did. But in Game 3, the best move for each player is to play what the other player does. If *R* plays A, then *C* gets 2 if she plays a, and 0 if she plays b. If *R* plays B, then *C* gets 6 if she plays b, and 0 if she plays a. There's no single best strategy for her, until she knows what *R* does.

We can mix and match these. Let's look at what the game is like if *R* has the egotistic preferences from Game 1, and *C* has the obsession with agreement of the players in Game 3.

Game 4	Choose a	Choose b
Choose A	1, 2	5, 0
Choose B	0, 0	3, 6

You should confirm this, but what I've attempted to do here is have the first number in each cell, i.e., *R*'s payoff, copy the matching cell in Game 1, and the second number in each cell, i.e., *C*'s payoff, copy the matching cell in Game 3. We will come back in a little while to what to say about Game 4, because it is more complicated than the other games we've seen to date. First, let's make three philosophical asides on what we've seen so far.

Prisoners' Dilemma

Game 1 is often called a **Prisoners' Dilemma**. There is perhaps some terminological confusion on this point, with some people using the term "Prisoners' Dilemma" to pick out any game whose *outcomes* are like those in the games we've seen so far, and some using it only to pick out games whose *payoffs* are like those in Game 1. Following what Simon Blackburn says in "Practical Tortoise Raising", I think it's not helpful to use the the term in the first way. So I'll only use it for games whose payoffs are like those in Game 1.

And what I mean by payoffs like those in Game 1 is the following pair of features.

- Each player is better off choosing A than B, no matter what the other player does.

- The players would both be better off if they both chose B rather than both chose A.

You might want to add a third condition, namely that the payoffs are symmetric. But just what that could *mean* is a little tricky. It's easy to compare *outcomes* of different players; it's much harder to compare *payoffs*. So we'll just leave it with these two conditions.

It is often very bad to have people in a Prisoners' Dilemma situation; everyone would be better off if they were out of it. Or so it might seem at first. Actually, what's really true is that the two players would be better off if they were out of the Prisoners' Dilemma situation. Third parties might stand to gain quite a lot from it. (If I'm paying out the money at the end of the game, I prefer that the players are in Game 1 to Game 2.) We'll come back to this point in a little. There are several ways we could try and escape a Prisoners' Dilemma. We'll mention four here, the first two of which we might naturally associate with Adam Smith.

The first way out is through **compassion**. If each of the players cares exactly as much about the welfare of the other player as they do about themselves, then we'll be in something like Game 2, not Game 1. Note though that there's a limit to how successful this method will be. There are variants of the Prisoners' Dilemma with arbitrarily many players, not just two. In these games, each player is better off if they choose A rather than B, no matter what the others do, but all players are better off if all players choose B rather than A. It stretches the limit of compassion to think we can in practice value each of these players's welfar equally to our own.

Moreover, even in the two player game, we need exact match of interests to avoid the possibility of a Prisoners' Dilemma. Let's say that *R* and *C* care about each other's welfare a large amount. In any game they play for money, each players' payoff is given by the number of pounds that player wins, plus 90% of the number of pounds the other player wins. Now let's assume they play a game with the following outcome structure.

	Choose a	Choose b
Choose A	£9.50, £9.50	£20, £0
Choose B	£0, £20	£10, £10

So we'll have the following payoff matrix.

Game 5	Choose a	Choose b
Choose A	18.05, 18.05	20, 18
Choose B	18, 20	19, 19

And that's still a Prisoners' Dilemma, even though the agents are very compassionate. So compassion can't do all the work. But probably none of the other 'solutions' can work unless compassion does some of the work. (That's partially why Adam Smith wrote the *Theory of Moral Sentiments* before going on to economic work; some moral sentiments are necessary for economic approaches to work.)

Our second way out is through **contract**. Let's say each party contracts with the other to choose B, and agrees to pay £ 2.50 to the other if they break the contract. Assuming that this contract will be enforced (and that the parties know this), here is what the outcome table now looks like.

	Choose a	Choose b
Choose A	£1, £1	£2.50, £2.50
Choose B	£2.50, £2	£3, £3

Now if we assume that the players just value money, those outcomes generate the following game.

Game 6	Choose a	Choose b
Choose A	1,1	2.5, 2.5
Choose B	2.5, 2.5	3,3

Interestingly, the game looks just like the original Prisoners' Dilemma as played between members of a commune. Basically, the existence of side contracts is enough to turn capitalists into communists.

A very closely related approach, one which is typically more efficient in games involving larger numbers of players, is to modify the outcomes, and hence the payoffs, with taxes. A striking modern example of this involves congestion charges in large cities. There are many circumstances where each person would prefer to drive somewhere than not, but if everyone drives, we're all worse off than if everyone took mass transit (or simply stayed home). The natural solution to this problem is simply to put a price on driving into the congested area. If the price is set at the right level, those who pay the charge are better off than if the charge was not there, since the amount they lose through the charge is gained back through the time they save.

In principle, we could always avoid Prisoners' Dilemma situations from arising through judicious use of taxes and charges. But it's hard to get the numbers right, and even harder to do the enforcement. So sometimes states will try to solve Prisoners' Dilemma situations with **regulation**. We see this in Beijing, for example, when they try to deal with congestion not by charging people money to enter the city, but by simply banning (certain classes of) people from driving into the city on given days. At a more abstract level, you might think of ethical prohibitions on 'free-riding' as being ways of morally regulating away certain options. If choosing B is simply ruled out, either by law or morality, there's clearly no Prisoners' Dilemma!

Having said that, the most important kind of regulation around here concerns making sure Prisoners' Dilemma situations survive, and are not contracted away. Let the two players be two firms in a duopoly; i.e., they are the only firms to provide a certain product. It is common for there to be only two firms in industries that require massive capital costs to startup, e.g., telecommunications or transport. In small towns (like St Andrews!) , it is common to have only two firms in more or less every sphere of economic life. In such cases there will usually be a big distance between the prices consumers are prepared to pay for the product, and the lowest price that the firm could provide the product and still turn a profit. Call these prices High and Low.

If the firms only care about maximising profit, then it looks like setting prices to High is like choosing B in Game 1, and setting prices to Low is like choosing A in that game. The two firms would be better off if each of them had High prices. But if one had High prices, the other would do better by undercutting them, and capturing (almost) all the market. And if both had Low prices, neither would be better off raising prices, because (almost) everyone would desert their company. So the firms face a Prisoners' Dilemma.

As Adam Smith observed, the usual way businesses deal with this is by agreeing to raise prices. More precisely, he says,

> People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices.

And that's not too surprising. There's a state where they are both better off than the state where they can compete. If by changing some of the payoffs they can make that state more likely to occur, then they will. And that's something that we should regulate away, if we want the benefits of market competition to accrue to consumers.

The final way to deal with a Prisoners' Dilemma is through **iteration**. But that's a big, complicated issue, and one that we'll come back to much later in these notes.

Payoffs and consequentialism

As we've stressed so far, there is a difference between outcomes and payoffs. An agent's payoff may be quite high, even if their outcome looks terrible, if the result of the game involves something they highly value. For instance, if one player values the wealth of the other player, an outcome that involves the other player ending up with lots of money will be one where the payoffs to both players are high.

In that respect the theory does not assume selfishness on the part of agents. It does assume that agents should try to get what they value, but that doesn't seem too big a constraint at first, assuming that agents are allowed to value anything. But in practice things are a little more complicated.

The model game theorists are using here is similar to the model that many ethicists, from G. E. Moore onward, have used to argue that any (plausible) ethical theory has a consequentialist form. To take one example, let's assume that we are virtue ethicists, and we think ethical considerations are 'trumps', and we are playing a game that goes from time t_0 to time t_1. Then we might say the payoff to any agent at t_1 is simply how virtuously they acted from t_0 to t_1. Since agents are supposed to be as virtuous as possible, this will give us, allegedly, the right evaluation of the agent's actions from t_0 to t_1.

Does this work in general? It certainly doesn't work as a theory of moral motivation, or indeed of any other kind of motivation. But consequentialism isn't really meant to be a theory of motivation. Utilitarians do not think that agents should aim to maximise utility *as such*. They think agents should do the things that, as a matter of fact, maximise utility. But a bigger worry is how this theory of value intersects with a theory of rational action under uncertainty. To settle this, we'd have to offer a theory of action under moral uncertainty, and we're not going to do *that* here. But we will note that there's a big issue here, and one that isn't easily settled by being liberal about what agents can value. (If you're interested in this topic, look up articles on whether moral theories can be *consequentialized*. The American spelling is because it is primarily Americans who are interested in this question, its Moorean origins notwithstanding!)

Knowledge and Dominance

Here's what looks like a very simple game.

Game 7	Choose a	Choose b
Choose A	20, 20	10, 1
Choose B	1,10	1,1

It seems clear that both players should choose A. After all, whatever the other player does, they are better off with A. And the two of them are collectively better off both choosing A, so any Prisoners' Dilemma related doubts that we had are not operational here.

But let me tell you a bit more about the background to this game. The payoffs are just payments in pounds. Each player values only their own winnings, and values each pound equally, so the function from outcomes to payoffs is easy. And it's really true that those are the payoffs the players will get if they choose either A or B. But the players don't know this. What they do know is that a fair coin is about to be tossed 10 times. They also know that if the coin comes down heads every time, then the payoffs are as above. Finally, they know that if the coin comes down tails even once, and either of them chooses A, then neither player will get any money. So the full table looks more like this.

Game 8	Choose a & 10 Heads	Choose a & at least 1 Tail	Choose b
Choose A & 10 Heads	20, 20	0, 0	10, 1
Choose A & at least 1 Tail	0, 0	0, 0	0, 0
Choose B	1,10	0,0	1,1

And now choosing A looks like a crazy gamble, since the odds are overwhelming that the coin will fall tails at least once. It seems then that Game 7 somehow misrepresents the situation facing the players. The table in Game 7 makes it look like it is best to choose A, but really in the situation facing the players, the smart move is to choose B.

What, though, is wrong with the representation of Game 7? It isn't that anything written on the table is *false*. Those really are the payouts the players will get, since the coin does, as a matter of fact, land heads 10 times in a row. What's wrong, at least on standard views, is that the players don't *know* that Game 7 represents the payoffs correctly. At a minimum, when we write a payoff matrix down, we assume that each player **knows** that the matrix is correct.

Sometimes, indeed often, we will make stronger assumptions than that. For instance, we'll almost always assume that each player knows that the other player knows the table is correct. And we'll often assume that each player knows that. And we'll often assume that each player knows that. And we'll often assume that each player knows that. And so on. But the basic assumption is that each player knows the table is correct.

Given that assumption, we infer that players should never play a **dominated** strategy. A dominated strategy is, roughly, a strategy such that some other strategy can do better, no matter how other things are. In other words, if a player knows that strategy s_1 will do better than s_2, then it is irrational for her to do s_2.

Those familiar with recent debates in epistemology will recognise this as a form of the Knowledge-Action Principle, which says that knowledge is sufficient for action. This principle plays a central role in work by John Hawthorne and Jason Stanley, and by Jeremy Fantl and Matthew McGrath. But it has also been the subject of some criticism. Orthodox game theory, and for that matter decision theory, incorporates this version of the Knowledge-Action principle, via the principle that dominated actions should not be chosen. Those who want to reject the Knowledge-Action principle will have to either do without orthodox game and decision theory, or find some other way to reinterpret game matricies so that the injunction against choosing dominated options makes sense.

It's important to be careful about what we mean by a dominated strategy. Here is a more careful definition.

Strong Domination A strategy s_1 *strongly dominates* strategy s_2 for player i iff for any combination of moves by other players, and states of the external world, playing s_1 provides a greater payoff than playing s_2, assuming other players make those moves, and the world is that way.

Strongly Dominated A strategy is strongly dominated iff some other strategy, available to the same player, strongly dominates it.

There is a potential scope ambiguity in the description of a strongly dominated strategy that it is important to be clear about. The claim is *not* that a strategy is strongly dominated if no matter what else happens, some strategy or other does better than it. It is that a strategy is dominated if some particular strategy does better in every circumstance. We can see the difference between these two ideas in the following game.

Game 9	l	r
U	3, 0	0, 0
M	2, 0	2, 0
D	0, 0	3, 0

Consider this game from R's perspective; who is choosing as always the rows. Her options are **U**p, **M**iddle and **D**own. C is choosing the columns; her choices are **l**eft or **r**ight. (I hope the ambiguity between r for *Right* and R for *Row* is not too confusing. It should be very hard in any given context to get them mixed up, and hopefully the convention we've adopted about cases will help.)

Notice that Middle is never the best outcome for R. If C chooses Left, R does best choosing Up. If C chooses Right, R does best choosing Down. But that does not mean Middle is dominated. Middle would only be dominated if one particular choice was better than it in both circumstances. And that's not true. Middle does better than Up in one circumstance (when C chooses Right) and does better than Down in another circumstance (when C chooses Left).

Indeed, there are situations where Middle might be uniquely rational. We need to say a bit more about expected utility theory to say this precisely, but consider what happens when R suspcts C is just going to flip a coin, and choose Left if it comes up Heads, and Right if it comes up Tails. (Since literally nothing is at stake for C in the game, this might be a reasonable hypothesis about what C will do.) Then it maximises R's **expected** return to choose Middle. We'll come back to this notion a lot.

So far we've talked about the notion of strong dominance. We also need a notion of **weak dominance**. Roughly, strategy s_1 weakly dominates strategy s_2 if s_1 can do better than s_2, and can't do worse. More formally,

Weak Domination A strategy s_1 *weak dominates* strategy s_2 for player i iff for some combination of moves by other players, and states of the external world, playing s_1 provides a greater payoff than playing s_2, assuming other players make those moves, and the world is that way, and for all combination of moves by other players, and states of the external world, playing s_1 provides at least as high a payoff as playing s_2, assuming other players make those moves, and the world is that way,

Weakly Dominated A strategy is weakly dominated iff some other strategy, available to the same player, weakly dominates it.

It does seem plausible that agents should prefer any strategy over an alternative that it weakly dominates. This leads to distinctive results in games like the following.

Game 10	a	b
A	1, 1	0, 0
B	0, 0	0, 0

In this game, choosing A does not strongly dominate choosing B for either player. The game is symmetric, so from now on we'll just analyse it from R's perspective. The reason choosing A does not strongly dominate is is that if C chooses b, then choosing A leads to no advantage. R gets 0 either way.

But choosing A does *weakly* dominate choosing B. A does better than B in one circumstance, namely when C chooses a, and never does worse. So a player who shuns weakly dominated options will always choose A rather than B in this game.

Iterated Dominance

A rational player, we've argued, won't choose dominated strategies. Now let's assume, as is often the case, that we're playing a game where each player knows that the other player is rational. In that case, the players will not only decline to play dominated strategies, they will decline to play strategies that only produce the best outcomes if the other player adopts a dominated strategy. This can be used to generate a prediction about what people will, or at least should, do in various game. We can see this going back to a variant of Prisoners' Dilemma from earlier on.

Game 4	Choose a	Choose b
Choose A	1, 2	5, 0
Choose B	0, 0	3, 6

If we look at things from *C*'s perspective, neither strategy is dominated. She wants to choose whatever *R* chooses. But if we look at things from *R*'s perspective, things are a little different. Here there is a strongly dominating strategy, namely choosing A. So *C* should really think to herself that there's no way *R*, who is rational, is going to choose B. Given that, the table really looks like this.

Game 4′	Choose a	Choose b
Choose A	1, 2	5, 0

I've put the prime there to indicate it is officially a different game. But really all I've done is delete a dominated strategy that the other player has. Now it is clear what *C* should do. In this 'reduced' game, the one with the dominated strategy deleted, there is a dominant strategy for *C*. It is choosing a. So *C* should choose a.

The reasoning here might have been a little convoluted, but the underlying idea is easy enough to express. *R* is better off choosing A, so she will. *C* wants to choose whatever *R* chooses. So *C* will choose a as well.

Let's go through a small variant of this game which might, after redescription, look fairly familiar.

Game 11	*l*	*r*
U	1, 1	1001, 0
D	0, 0	1000, 1

Just as in Game 4, R has a dominant strategy. It is to choose Up. (Again, I'm labelling just using the first letter of the description of the move.) And given that R will choose Up, the best thing for C to do is choose l. So it looks like we should end up in the top-left corner of the table, just like in Game 4.

Those of you who have taken some decision theory should recognise Game 11. It is just Newcomb's problem, with some assumptions about the payoffs. (If you don't know what Newcomb's Problem is, skip the next couple of paragraphs. We'll discuss Newcomb's problem in more detail later in the notes.) R in this case is the human player, who is usually the focus of attention in decision theory classes. Her payoffs here are just her payments in the usual statement of the game, divided by 1000. Up is her choosing both boxes, and Down is her choosing one box.

C is the demon. The demon isn't usually treated as a player in decision theoretic versions of the puzzle, but she clearly has views, and preferences. The demon wants to predict the move that the player makes. So we've represented her payoffs that way. Left is her predicting two boxes, Right is her predicting one box. And if she gets the prediction right, she gets a payoff of 1, if she gets it wrong, she gets a payoff of 0.

So Newcomb's problem is just a simple game, and it can be solved by noting that one player has a dominating strategy, and the other player, i.e., the demon, has a dominating strategy under the assumption that this dominating strategy is played.

We can use the idea of removing dominating strategies to illustrate some puzzling features of a couple of other games. I won't do tables for these games, because they'd be much too big. The first is a location game that has many applications.

Game 12

Two trucks have to choose where they will sell ice-cream on a particular beach. There are 11 locations to choose from, which we'll number 0, 1, . . . , 9, 10. Spot 0 is at the left end of the beach, Spot 10 is at the right end of the beach, and the other spots are equally spaced in between. There are 10 people at each location. Each of them will buy ice-cream. If one truck is closer, they will buy ice-cream from that truck. If two trucks are equally close, then 5 of them will buy ice-cream from one truck, and 5 from the other. Each truck aims to

> maximise the amount of ice-cream it sells. Where should the trucks end up?

Let's start by looking at a fragment of the payoff matrix. The payoffs are numbers of ice-creams sold. We'll call the players *R* for Row and *C* for column, as usual, and just use the number *n* for the strategy of choosing location *n*.

	0	1	2	3	4	...
0	55, 55	10, 100	15, 95	20, 90	25, 85	...
1	100, 10	55, 55	20, 90	25, 95	30, 80	...
...						

I'll leave it as an exercise to confirm that these numbers are indeed correct. But there's something already from the numbers that we can see. No matter what *C* selects, *R* is better off picking 1 than 0. If *C* picks 0 or 1, she is a lot better off; she sells 45 more ice-creams. And if *C* picks a higher number, she is a bit better off; she sells 5 more ice-creams. So picking 1 dominates picking 2.

Let's look at the opposite corner of the matrix.

	...	6	7	8	9	10
...						
9	...	30, 80	25, 85	20, 90	55, 55	100, 10
10	...	25, 85	20, 80	15, 95	10, 100	55, 55

Again, there should be a pattern. No matter what *C* does, *R* is better off picking 9 than 10. In most cases, this leads to selling 5 more ice-creams. If *C* also picks 9 or 10, the picking 9 gives *R* a big advantage. The argument here was obviously symmetric to the argument about picking 0 or picking 1, so I'll stop concentrating on what happens when both players select high numbers, and focus from here on the low numbers.

So there is a clear conclusion to be drawn from what we've said so far.

- Spot 0 is dominated by Spot 1, and Spot 10 is dominated by Spot 9. So if *R* is rational, she won't pick either of those spots. Since the game is symmetric, if *C* is rational, she won't pick either of those spots either.

Now let's turn to the comparison between Spot 1 and Spot 2. Again, we'll just look at a fragment of the matrix.

	0	1	2	3	4	5	...
1	100, 10	55, 55	20, 90	25, 95	30, 80	35, 75	...
2	95, 15	90, 20	55, 55	30, 80	35, 75	40, 70	...
...							

The pattern is also clear. If C selects any number above 2, then R sells 5 more ice-creams by picking 2 rather than 1. If C selects either 1 or 2, then R sells 35 more ice-creams by picking 2 rather than 1. But if C selects 0, then R sells 5 *fewer* ice-creams by picking 2 rather than 1. So picking 2 does *not* dominate picking 1.

But note the only circumstance when picking 2 is worse than picking 1 is if C picks 0. And picking 0 is, for C, a strongly dominated strategy. So picking 2 is sure to do better than picking 1 if C does not play a strongly dominated strategy. Assuming C is rational, we can represent this by deleting from the game matrix C's dominated options. Here's what the top left corner of the game matrix looks like when we do that.

	1	2	3	4	5	...
1	55, 55	20, 90	25, 95	30, 80	35, 75	...
2	90, 20	55, 55	30, 80	35, 75	40, 70	...
...						

And now it looks like picking 1 is dominated. This teaches us another lesson about the game.

- If we 'delete' the option of picking either 0 or 10 for C, then picking 2 dominates picking 1 for R. In other words, if R knows C is rational, and hence won't pick a dominated option, then picking 2 dominates picking 1 for R, relative to the space of epistemically possible moves in the game. For similar reasons, if R knows C is rational, then picking 8 dominates picking 9. And if C knows R is rational, then picking either 1 or 9 is dominated (by 2 and 8 respectively).

Summarising what we know so far,

- If the players are rational, they won't pick 0 or 10.
- If the players know the other player is rational, they also won't pick 1 or 9.

Let's continue down the matrix. For simplicity, we'll leave off columns 0 and 10, since they are dominated, and we have deleted those as possible options.

	1	2	3	4	5	...
2	90, 20	55, 55	30, 80	35, 75	40, 70	...
3	85, 25	80, 30	55, 55	40, 70	45, 65	...
...						

Picking 3 doesn't *quite* dominate picking 2. In most circumstances, *R* does better by picking 3 rather than 2. But she does a little worse if *C* picks 1. But wait! We just had an argument that *C* shouldn't pick 1. Or, at least, if *C* knows that *R* is rational, she shouldn't pick 1. Let's assume that *C* does know that *R* is rational, and *R* in turn knows that fact, so she can use it in reasoning. That means she knows *C* won't pick 1. So she can delete it from consideration too. And once she does, picking 3 dominates picking 2, relative to the reduced space of epistemically possible outcomes.

I haven't belaboured the point as much as in the previous paragraphs, but hopefully the following conclusion is clear.

- If the players know that the players know that the other player is rational, they won't pick 2 or 8.

And you can see where this is going. Once we rule out each player picking 2 or 8, then picking 4 dominates picking 3, so picking 3 should be ruled out. And picking 7 should be ruled out for symmetric reasons. But once 3 and 7 are ruled out, picking 5 dominates picking either 4 or 6. So both players will end up picking 5.

And that is, in the standard economic textbooks, a nice explanation of why we see so much 'clustering' of shops. (For instance, why so often there are several petrol stations on one corner, rather than spread over town.) Of course the full explanation of clustering is more common, but it is nice to see such a simple model deliver such a strong outcome.

This process is called the **Iterative Deletion of Dominated Strategies**. More precisely, what we've used here is the strategy of iteratively deleting **strongly** dominated strategies. This is a powerful technique for solving games that don't, at first glance, admit of any easy solution.

But it is worth reminding ourselves just how strong the assumptions that we used were. The standard terminology here can be a little confusing. After all, it isn't that picking 4 *dominates*, in any sense, picking 3. What's really true is that if we quantify over a restricted range of choices for C, then picking 4 is better for R than picking 3, no matter which choice *from that range*, C chooses. And that's a good reason to pick 4 rather than 3, provided that R knows that C will make a pick in that range. From that perspective, it's instructive to complete the list of lessons that we were compiling about the game.

- If the players are rational, they won't pick 0 or 10.
- If the players know the other player is rational, they also won't pick 1 or 9.
- If the players know that the players know that the other player is rational, they also won't pick 2 or 8.
- If the players know that the players know that the players know that the other player is rational, they won't pick 3 or 7.
- If the players know that the players know that the players know that the players know that the other player is rational, they also won't pick 4 or 6, i.e., they will pick 5

There are a lot of assumptions built in to all of this. It would be nice to have a way of summarising them. The standard approach traces back to David Lewis's *Convention*.

Common Knowledge In a game, it is common knowledge that p if each player knows it, each player knows that each player knows it, each player knows that each player knows that each player know it, and so on.

In many, but not all, games, we assume common knowledge of the rationality of the players. In Game 12, common knowledge of rationality makes picking 5 rationally mandatory.

There is a fun story that is usually told to illustrate the importance of common knowledge.

> **Slapville**
>
> In Slapville, it is culturally required to slap oneself if one is in public with a dirty face. Larry, Curly and Moe are in a room together, fortunately one without mirrors. Each of them has a dirty face, but they can't see their own faces, they can only see the other faces. And each face is dirty. Inspector Renault walks into the room and says,

> "I'm shocked! Someone in this room has a dirty face." After a long delay, Larry, Curly and Moe each slap themselves in the face (thereby getting dirty hands as well as dirty faces). Why?

One way to be puzzled by Slapville is to start with the concept of **mutual knowledge**. It is mutual knowledge that p if everyone in the game knows that p. In Slapville, it is mutual knowledge that someone has a dirty face. It is even, modulo Williamsonian concerns, mutual knowledge* that someone has a dirty face. (By S knows* that p, I mean S knows that p, and S knows that S knows p, and S knows that S knows that S knows that p, and so on.) So you might wonder what difference Renault's statement makes. After all, just like his namesake, he's just voicing something everyone already knew.

But it wasn't common knowledge that someone has a dirty face. Consider things from Larry's perspective. He knows someone has a dirty face. He can see Curly and Moe's dirty faces. And he knows that everyone knows that someone has a dirty face. He clearly knows it; he can see Curly and Moe. And Curly knows it; he can see Moe. And Moe knows it; he can see Curly.

But he doesn't know that everyone knows that everyone knows that someone has a dirty face. For all he knows, only Curly and Moe have dirty faces. If that's true, the only dirty face Curly knows about is Moe's. So for all Larry knows that Curly knows, only Moe has a dirty face. And if only Moe has a dirty face, then Moe doesn't know that someone has a dirty face. So for all Larry knows that Curly knows, Moe doesn't know that someone has a dirty face.

Or at least, that's the situation before Renault speaks. Once Renault speaks, it becomes *common knowledge* that someone has a dirty face. (Assume that it is common knowledge that Renault speaks the truth, at least when he is shocked.) Now let's trace back the consequences.

Consider again the world where only Moe has a dirty face. In that world, once Renault speaks, Moe slaps himself. That's because he learns that he has a dirty face by putting together the clean faces he can see with the fact that someone has a dirty face. (I've been assuming here that it is common knowledge that only Larry, Curly and Moe were in the room to start with. Hopefully that hasn't been too distracting, but it is crucial here.)

Now as a matter of fact, Moe does not immediately slap himself. That suffices to teach everyone something. In particular, it teaches them they were not in the world where only Moe has a dirty face. Of course, they each already knew that, but it is now clear to everyone that they all know it.

Consider next the world where only Curly and Moe have dirty faces. From Curly's perspective in that world, there are two possibilities. Either he and Moe have dirty faces, or only Moe has a dirty face. But we just ruled out that only Moe has a dirty face. So if we were in the world where only Curly and Moe have a dirty face, then Curly should slap himself.

But Curly doesn't slap himself yet. (I'll leave the question of precisely why he doesn't as an exercise; it should be clear given what we've said so far.) So that rules out the possibility that we're in the world where only Curly and Moe have dirty faces. But Larry knew to start with that we were either in the world where all of them have dirty faces, or in the world where only Curly and Moe have dirty faces. So they must be in the world where they all have dirty faces.

At this stage Larry realises this, and slaps himself in the face. At roughly the same time, Curly and Moe also slap themselves in the face. And it's all because of the difference between mutual knowledge and common knowledge.

Strong and Weak Dominance

The assumption of common knowledge of rationality is a really strong assumption though. The following game makes this very clear.

> **Game 13**
>
> Everyone in a large group selects an integer between 1 and 100 inclusive. The winner of the game is the person whose number is cloest to $^2/_3$ of the average of all of the numbers selected. That is, the payoff for the player who selects closest to $^2/_3$ of the average is 1. (If there is a tie between n players, their payoff is $^1/_n$.) The payoff for everyone else is 0.

This game can be played with any number of players, but we'll keep things simple by assuming there are just 10. This still gives us too big a game table. We need 10 dimensions, and 100^{10} cells. The latter is not too demanding; but a 10-dimensional representation is tricky on paper. So we'll just describe states of the game.

The first thing to note about the game is that a particular player, let's call her P, can't win if she selects a number between 68 and 100. That's because those numbers can't be 2/3 of the average unless the average is greater than 100. And, of course, the average can't be greater than 100. So those choices are dominated for P.

But we have to be rather careful here. What choice dominates picking, say, 70? We might say that 60 dominates it. After all, 60 could be the best possible play, while 70 could not. But in most circumstances, 60 and 70 will have the same payoff, namely 0. Unless the average is close to 90, or no one else picks around 60, P's payoff will be 0 whether she picks 60 or 70. And the same goes for any alternative to picking 70.

This is all to say that no alternative pick **strongly** dominates picking 70. But several picks do **weakly** dominate it. For instance, picking 64 does. Note that picking 70 can never do better than picking 64, because even if everyone else picks 100, if one player picks 64, the average will be 96.4, so 64 will be closest to 2/3 of the average. So any circumstance where 70 will be a winning play must be one where everyone else picks more than 70. But in those circumstances, picking 64 will win as well. Conversely, picking 64 could do better than picking 70. If everyone else picks 65, picking 64 will win, and picking 70 will lose. So 64 weakly dominates 70. And as we can see, all that really mattered for that argument was that 70 was always going to be higher than 2/3 of the average, so it would be weakly dominated by some numbers that could be closer to 2/3 of the average.

Again, let's list the lessons as we learn them.

- Any selection above 67 is weakly dominated.
- Since rational players do not play weakly dominated strategies, it is irrational to pick any number above 67.

We will, much later on, come back to the assumption that playing weakly dominated strategies is irrational. I think it is true, though it deserves a more careful treatment than we'll give it here. Let's just assume for now it is true.

Now we showed a way that P can win while playing 60. But it has to be said, that it isn't a particularly likely way. It requires the average of the selections to be nearly 90. And that requires a lot of other people to pick high numbers. That is, it requires other people to pick weakly dominated strategies. And that's not very plausible, assuming those other people are rational.

Let's assume, then, that P knows that the other players are rational, and hence will not choose weakly dominated strategies. So no other player will choose a number greater than 67. Then the average of what everyone picks can't be greater than 67. So 2/3 of the average can't be greater than 45. So once we remove the weakly dominated strategies, any selection greater than 45 can't be optimal (i.e., it must be considerably greater than 2/3 of the average), and we can give an argument similar to the above argument that it is weakly dominated.

As in the ice-cream game, the trick here is to delete dominated strategies. Once you do that, it is as if you are playing a different game. And in that game, more strategies are in turn dominated. That's because they are strategies that only made sense to play on the assumption that other people played dominated strategies. And, really, it isn't very plausible to assume that people will play dominated strategies. So we should delete the dominated strategies in this new game as well.

And once we do that, we'll find yet more strategies become dominated. Let's say we delete the strategies between 46 and 67. Now the most the average can be is 45. So 2/3 of the average can't be more than 30. So any pick greater than 30 can't be optimal, and so is weakly dominated, so should be deleted. But once those picks are deleted, the average can't be greater than 30, so 2/3 of the average can't be greater than 20, so any pick greater than 20 can't be optimal, and is weakly dominated, so should be deleted. And so on, until every pick greater than 1 is deleted. That's the next lesson from the game.

- The only strategy that survives the iterated deletion of weakly dominated strategies is to select 1.

So it seems rational players, who are playing the game with other rational players, should choose 1 right?

Not so fast! Here's a little tip for anyone playing this game in a large enough group. If you pick 1 you will lose with a probability more or less equal to 1. Just what number will win is harder to predict without knowing more about the group's features, but it won't be 1. Why not? Is it because there are irrational players in any group?

Not necessarily. What's really going on is that the assumptions needed to get to 1 are incredibly strong. Let's go through the argument for getting to 1 in some more detail.

- At the start, 2/3 of the average is at most 67.
- If everyone knows that, and is rational, 2/3 of the average is at most 45.

- If everyone knows that, and is rational, 2/3 of the average is at most 30.
- If everyone knows that, and is rational, 2/3 of the average is at most 20.
- If everyone knows that, and is rational, 2/3 of the average is at most 13.
- If everyone knows that, and is rational, 2/3 of the average is at most 9.
- If everyone knows that, and is rational, 2/3 of the average is at most 6.
- If everyone knows that, and is rational, 2/3 of the average is at most 4.
- If everyone knows that, and is rational, 2/3 of the average is at most 3.
- If everyone knows that, and is rational, 2/3 of the average is at most 2.
- If everyone knows that, and is rational, 2/3 of the average is at most 1.

Note that at every stage we need to make one more assumption about what the players know. By the end we've assumed that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone knows that everyone is rational. (There are 10 iterations of 'everyone knows that' in that sentence, in case you'd rather not count.) And that's really not a plausible assumption.

To put this in perspective, imagine a variant of the Slapville story where there aren't just 3 people, Larry, Curly and Moe, but 10 people. (Perhaps we add the 7 dwarves to the 3 stooges.) And they all have dirty faces. And none of them can see their own face, but Inspector Renault says that at least one of them has a dirty face. It is hard to imagine that this will make any difference at all. For one thing, the computations required to process the new common knowledge are horrendously difficult. (In a way, this is a denial that the players are *perfectly* rational, but it also brings out what a strong assumption that is already.) To assume that they can be done, and that everyone knows they can be done, and everyone knows that everyone knows they can be done, and so on for 8 more steps, is absurd.

Puzzles about Weak Domination

There is something very odd about choosing strategies that are strongly dominated. And from that, we can naturally deduce that we should delete strategies that are strongly dominated. The iterative deletion of strongly dominated strategies requires not just rationality, but common belief in rationality through as many iterations as there are steps of deletion. Sometimes that will be an implausible assumption, but it often seems reasonable enough in practice.

Weak domination, however, generates principles that are both more puzzling and harder to justify. Some people argue that weakly dominated strategies are perfectly rational to play in games like this one, which we might think of as Weak Prisoners' Dilemma.

Game 14	l	r
U	1, 1	100, 0
D	0, 100	100, 100

D is weakly dominated by U. But if R is confident that C will play r, then it may be rational to play D. And there's good reason to think that C will play r; the bottom-right corner is a good place for them to both end up. So you might think it is rational to play a weakly dominated strategy.

If we start iteratively deleting weakly dominated strategies, we get even less plausible outcomes.

Game 15	l	m	r
T	2, 2	100, 0	0, 90
M	0, 100	100, 100	100, 95
B	0, 95	95, 100	95, 95

Since B and r are weakly dominated by M and m respectively, we can delete them. And now we're back in a small variant of Game 14, where deleting weakly dominated strategies leads to the $\langle T, l\rangle$ equilibrium. But here it is even more attractive to play $\langle M, m\rangle$. For one thing, it is an equilibrium in an important sense that we'll talk more about later. The important sense is that given what the other player is doing, neither player can do better by changing. So if each player thinks the other player will play their half of $\langle M, m\rangle$, it makes sense for each player to play their half of $\langle M, m\rangle$. All that's true of the $\langle D, R\rangle$ equilibrium in Game 14. But in Game 14, the players couldn't do worse by changing their strategies if they are wrong about what the other players will play. Here they might. We'll have much more to say about this later, but there is an even stronger argument that $\langle M, m\rangle$ is a rational pair of plays in this game than there was that $\langle D, R\rangle$ was a rational pair of plays in Game 14. For roughly these reasons, Robert Stalnaker argues that there is a good sense of rationality (what he calls **perfect rationality**) which requires playing $\langle U, L\rangle$ in Game 14, but which is compatible with playing $\langle M, m\rangle$ in Game 15.

There are other puzzles with iteratively deleting weakly dominated strategies. Surprisingly, it can turn out that the order in which we delete strategies makes a big difference. This point is made in Elon Kohlberg and Jean-Francois Mertens's 1986 *Econometrica* paper "On the Strategic Stability of Equilibria". (Thanks to Daniel Rothschild for pointing this out to me.) Here is (a minor variation on) the example they use.

Game 16	l	r
T	2, 2	2, 2
M	1, 0	0, 1
B	0, 1	1, 0

Since B is dominated, it won't be chosen. But once we eliminate B, then for C, r (weakly) dominates l, so only r survives iterative deletion of weakly dominated strategies.

But wait! We could also reason as follows. Since M is dominated, it won't be chosen. But once we eliminate M, then for C, l (weakly) dominates r, so only l survives iterative deletion of weakly dominated strategies. What is going on?

Kohlberg and Mertens suggest that we should focus on strategies that survive *some* process of iterative deletion. Since for player 2, there is an iterative deletion path that l survives, and an iterative deletion path that r survives, then both strategies really survive iterative deletion.

You might be tempted by an alternative take on this example. Perhaps it was wrong to either delete M or to delete B. Perhaps we should say that when we are deleting strategies, the right thing to do is to delete *all* strategies that are dominated at a stroke. So we should simultaneously delete M and B, and then it will be clear that both L and R survive. This won't avoid the problem though, as we can see by a simple three player game.

Game 17

There are three players, 1, 2 and 3. They can each choose one of two options, which we'll label A and B. For player 1 and 2, the payoff structure is easy, they get 2 if they pick A, and 1 if they pick B. For player 3, it is a little more complicated. Player 3 gets:

- 2 if both players 1 and 2 pick A
- 0 if both players 1 and 2 pick B

- 1 if players 1 and 2 make opposite picks, and player 3 picks the same thing as player 1.
- 0 if players 1 and 2 make opposite picks, and player 3 picks the same thing as player 2.

One way we could analyse this is by saying that since *B* is dominated for both players 1 and 2, they won't pick it. And since player 3's choice doesn't matter if both player 1 and 2 pick *A*, then it doesn't matter what player 3 picks. But there are other ways we could go as well.

Since *B* is (strongly) dominated for player 1, we can rule it out. Now player 3 faces the following choice, assuming player 1 picks *A*. (We'll write player 3's choices on the rows, and player 2's on the columns, and put player 3's payoff first.)

	a	*b*
A	2, 2	1, 1
B	2, 2	0, 1

Now *A* weakly dominates *B*, so it is uniquely rational, we might think, for player 3 to pick *A*.

But wait! Since *b* is (strongly) dominated for player 2, we can rule it out. Now player 3 faces the following choice, assuming player 2 picks *A*. (We'll write player 3's choices on the rows, and player 1's on the columns, and put player 3's payoff first.)

	a	*b*
A	2, 2	0, 1
B	2, 2	1, 1

Now *B* weakly dominates *A*, so it is uniquely rational, we might think, for player 3 to pick *B*.

Now it looks like even we delete every dominated strategy that a player has when we get to that player in the analysis, the order in which we do the deletions still matters. Note though that none of the analysis we've just done seems to undermine the intuitive verdict that player 3 could rationally choose either *A* or *B*. She is going to get 2 whatever she chooses, since the other players will both choose *A*. So this doesn't undermine Kohlberg and Mertens's conclusion that if there is some path of strategy deletion that leads to a strategy being available and no more deletions being possible, then it is (for these purposes) rationally acceptable to choose that strategy.

Four Normative Statuses

We will spend a lot of time in these notes on various normative statuses that strategies in a game can have. We have, in effect, already seen four such statuses.

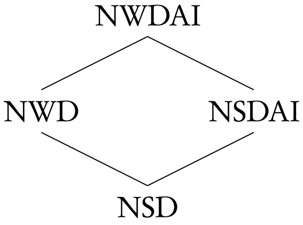
NSD That is, **N**ot **S**trongly **D**ominated.

NWD That is, **N**ot **W**eakly **D**ominated.

NSDAI That is, That is, **N**ot **S**trongly **D**ominated **A**fter **I**terations. In other words, it is still there after we repeatedly delete strongly dominated strategies until only undominated strategies remain.

NWDAI That is, That is, **N**ot **W**eakly **D**ominated **A**fter **I**terations. In other words, it is still there after we repeatedly delete weakly dominated strategies until only undominated strategies remain.

We can place these from weakest to strongest in the following lattice, with weakest being on the bottom.



That table actually encodes a lot of distinct claims. Let's go through them all, noting which ones are obvious, and proving the ones that aren't.

All NWD strategies are NSD

This follows from the fact that strong domination entails weak domination, so deleting weakly dominated strategeis will delete strongly dominated strategies.

All NWDAI strategies are NSDAI

This is true for more or less the same reason. Formally proving this is a little tricky, since you have to be careful about how the different iterative steps interact with the definitions of dominance, but it can be proven by induction on the number of iterative steps. We won't do the proof here, but you can show that any strategy that is deleted by step *n* of the process of iteratively deleting strongly dominated strategies will also be deleted by step *n* of the process of iteratively deleting weakly dominated strategies.

All NSDAI strategies are NSD

Obvious.

All NWDAI strategies are NWD

Also obvious.

Some NSD strategies are not NWD

In Game 13, the strategy of picking 90 is not strongly dominated, as we showed above, but is weakly dominated. So it is NSD, but not NWD.

Some NWD strategies are not NWDAI

In Game 13, the strategy of picking 60 is not weakly dominated, as we showed above, but as we also showed, it does not survive the iterative deletion process. Indeed, it is deleted at the second step.

Some NSD strategies are not NSDAI

In Game 12, the strategy of choosing location 1 is not strongly dominated. But it does not survive the iterative deletion process. Indeed, it is deleted at the second step.

Some NSDAI strategies are not NWDAI

In Game 13, no strategy is strongly dominated, so all strategies are NSDAI. But many strategies, indeed all but one, are not NWDAI.

Some NSDAI strategies are not NWD

Similarly in Game 13, some strategies, like choosing 90, aren't NWD. But as we just showed, they are NSDAI.

Some NWD strategies are not NSDAI

In Game 12, the only weakly dominated strategies are choosing 0 and 10. But the only NSDAI strategy is 5. So any other strategy is NWD but not NSDAI.

Games and Time

So far we've looked just at games where each player makes just one move, and they make it simultaneously. That might not feel like most games that you know about. It isn't how we play, for instance, chess. Instead, most games involve players making multiple moves, and these moves taking place in time. Here is one simple such game. It's not very interesting; you won't have fun games nights playing it. But it is useful to study.

> **Game 18**
>
> There are two players, who we'll call A and B. First A moves, then B, then finally A moves again. Each move involves announcing a number, 1 or 2. A wins if after the three moves, the numbers announced sum to 5. B wins otherwise.

This is a simple zero-sum game. The payoff is either 1 to A and 0 to B, or 1 to B and 0 to A. For simplicity, we'll describe this as A winning or B winning. We'll soon be interested in games with draws, which are a payoff of 1/2 to each player. But for now we're looking at games that someone wins.

Before we go on, it's worth thinking about how you would play this game from each player's perspective. The formal approach we'll eventually take is pretty similar, I think, to the way one thinks about the game.

Normal Form and Extensive Form

Figure 3.1 is the **extensive form** representation of Game 18. We will use $\mathcal{W}$ to denote that A wins, and $\mathcal{L}$ to denote that B wins.

To read what's going on here, start at the node that isn't filled in, which in this case is the node at the bottom of the tree. The first move is made by A. She chooses to play either 1 or 2. If she plays 1, we go to the left of the chart, and now B has a choice. She plays either 1 or 2, and now again A has a choice. As it turns out, A has the same choice to make whatever B does, though there is nothing essential to this. This is just a two-player game, though there is also nothing essential to this. We could easily have let it be the case that a third player moved at this point. Or we could have made it that who moved at this point depended on which move B had made. The extensive form representation allows for a lot of flexibility in this respect.

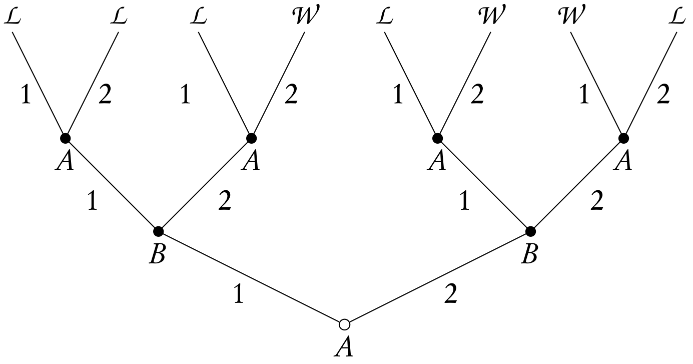


Figure 3.1: Game 18

At the top of the diagram is a record of who won. In the more general form, we would put the payoffs to each player here. But when we have a simple zero-sum game, it is easier to just record the payoffs to one player. You should check that the description of who wins and loses in each situation is right. In each case, *A* wins if 2 is played twice, and 1 once.

The extensive form representation form is very convenient for turn taking games. But you might think that it is much less convenient for games where players move simultaneously, as in Prisoners' Dilemma. But it turns out that we can represent that as well, through a small notational innovation. Figure 3.2 is the game tree for Prisoners' Dilemma.

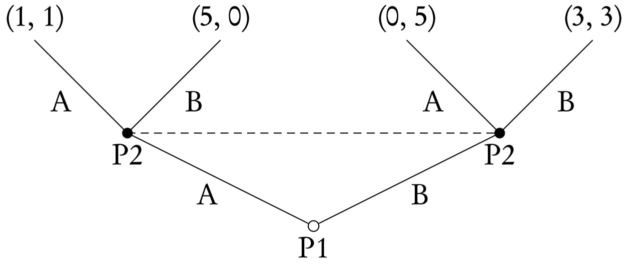


Figure 3.2: Prisoners' Dilemma

The crucial thing here is the dashed line between the two nodes where P2 (short for Player 2) moves. What this means is that P2 doesn't know which of these nodes she is at when she moves. We normally assume that a player knows what moves are available to her. (Interesting philosophical question, which hasn't been much explored: What happens when we drop this assumption?) So we normally only put this kind of dashed line in when it connects two nodes at which the same player moves, and the available moves are the same. When we put in notation like this, we say that the nodes that are connected form an **information set**. If we don't mark anything, we assume that a node is in a degenerate information set, one that only contains itself.

Strictly speaking, you could regard this tree as a representation of a game where Player 1 goes first, then Player 2 moves, but Player 2 does not know Player 1's move when she moves. But that would be reading, I think, too much into the symbolic representation. What's crucial is that we can represent simultaneous move games in extensive form.

There is a philosophical assumption in this notational convention that we might want to come back to later on. It is usual to use dashed lines, like I've done, or circles or ovals to represent that for all Player 2 knows, she is at one of the connected nodes. But drawing circles around a set of possibilities is only a good way to represent Player 2's uncertainty if we assume quite a lot about knowledge. In particular, it is only good if we assume that which nodes are epistemically open for Player 2 is independent of which node, within that group, she is at. In formal terms, it amounts to assuming that knowledge is an S5 modality. This isn't actually a true assumption, and it makes for some interesting complications to game theory if we drop it. But for now we'll follow orthodoxy and assume we can use representaions like these dashed lines to represent uncertainty.

These representations using graphs are knows as **extensive form** representations of games. The represenations using tables that we've used previously are known as **normal form** or **strategic form** representations. The idea is that any game really can be represented as game of one simultaneous move. The 'moves' the players make are the selections of **strategies**. A strategy in this sense is a plan for what to do in any circumstance whatsoever.

Let's do this for the Game 18. A strategy for *A* has three variables. It must specify what she does at the first move, what she does at the second move if *B* plays 1, and what she does at the second move if *B* plays 2. (We're assuming here that *A* will play what she decides to play at the first move. It's possible to drop that assumption, but it results in much more complexity.) So we'll describe A's strategy as $\alpha\beta\gamma$, where α is her move to begin with, β is what she does if B plays 1, and γ is what she does if B plays 2. A strategy for B needs to only have two variables: what to do if A plays 1, and what to do if A plays 2. So we'll notate her strategy as $\delta\epsilon$, where δ is what she does if A plays 1, and ϵ is what she does if A plays 2. So A has 8 possible strategies, and B has 4 possible strategies. Let's record the giant table listing the outcomes if thye play each of those strategies.

Game 18	11	12	21	22
111	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
112	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
121	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
122	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
211	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$
212	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
221	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$
222	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$

There is something quite dramatic about this representation. We can see what A should play. If her strategy is 221, then whatever strategy B plays, A wins. So she should play that; it is a (weakly) dominant strategy. This isn't completely obvious from the extended form graph.

Here's a related fact. Note that there are only 8 outcomes of the extended form game, but 32 cells in the table. Each outcome on the tree is represented by multiple cells of the table. Let's say we changed the game so that it finishes in a draw, represented by $\mathcal{D}$, if the numbers picked sum to 3. That just requires changing one thing on the graph; the $\mathcal{L}$ in the top-left corner has to be changed to a $\mathcal{D}$. But it requires making many changes to the table.

Game 18′	11	12	21	22
111	$\mathcal{D}$	$\mathcal{D}$	$\mathcal{L}$	$\mathcal{L}$
112	$\mathcal{D}$	$\mathcal{D}$	$\mathcal{W}$	$\mathcal{W}$
121	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
122	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{W}$
211	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$
212	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$	$\mathcal{L}$
221	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$	$\mathcal{W}$
222	$\mathcal{W}$	$\mathcal{L}$	$\mathcal{W}$	$\mathcal{L}$

In part because of this fact, i.e., because every change to the value assignment in the extensive form requires making many changes in the values on the normal form, it isn't a coincidence that there's a row containing nothing but $\mathcal{W}$. The following claim about our game can be proved.

> Assign $\mathcal{W}$ and $\mathcal{L}$ in any way you like to the eight outcomes of the extended form game. Then draw table that is the normal form representation of the game. It will either have a row containing nothing but $\mathcal{W}$, i.e., a winning strategy for *A*, or a column containing nothing but $\mathcal{L}$, i.e., a winning strategy for B.

We will prove this in the next section, but first we will look at how to 'solve' games like Game 18.

Backwards Induction

The way to think through games like Game 18 is by working from top to bottom. *A* moves last. Once we get to the point of the last move, there is no tactical decision making needed. *A* knows what payoff she gets from each move, and she simply will take the highest payoff (assuming rationality.)

So let's assume she does that. Let's assume, that is, that *A* does play her best strategy. Then we know three things about what will happen in the game.

- If *A* plays 1, and *B* plays 2, *A* will follow with 2.
- If *A* plays 2, and *B* plays 1, *A* will follow with 2.
- If *A* plays 2, and *B* plays 2, *A* will follow with 1.

Moreover, once we make this assumption there is, in effect, one fewer step in the game. Once *B* moves, the outcome is determined. So let's redraw the game using that assumption, and just listing payoffs after the second move. This will be Figure 3.3

Now we can make the same assumption about B. Assume that *B* will simply make her best move in this (reduced) game. Since *B* wins if *A* loses, *B*'s best move is to get to $\mathcal{L}$. This assumption then, gives us just one extra constraint.

- If *A* plays 1, *B* will follow with 1.

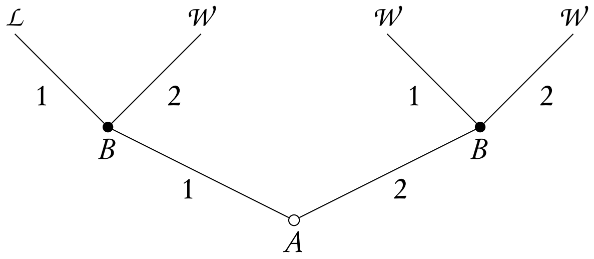


Figure 3.3: Game 18 with last move assumed

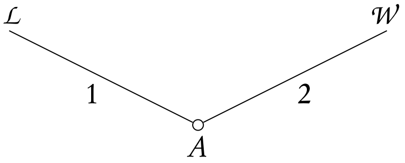


Figure 3.4: Game 18 with last two moves assumed

And, once again, we can replace *B*'s actual movement with a payoff of the game under the assumption that *B* makes the rational move. This gives us an even simpler representation of the game that we see in Figure 3.4.

And from this version of the game, we can draw two more conclusions, assuming *A* is rational.

- *A* will play 2 at the first move.
- *A* will win.

Let's put all of that together. We know *A* will start with 2, so her strategy will be of the form $2\beta\gamma$. We also know that *B* doesn't care which strategy she chooses at that point, so we can't make any further reductions. But we do know that if *A* plays 2 and *B* plays 1, *A* will follow with 2. So *A*'s strategy will be of the form 22γ. And we know that if *A* plays 2 and *B* plays 2, then *A* will play 1. So *A*'s strategy will be 221, as we saw on the table.

Note that, as in Game 13, the use of backwards induction here hides a multitude of assumptions. We have to assume each player is rational, and each player knows that, and each player knows that, and so on for at least as many iterations as there are steps in the game. If we didn't have those assumptions, it wouldn't be right to simply replace a huge tree with a single outcome. Moreover, we have to make those assumptions be very modally robust.

We can see this with a slight variant of the game. Let's say this time that the left two outcomes are $\mathcal{D}$. So the graph looks like Figure 3.5.

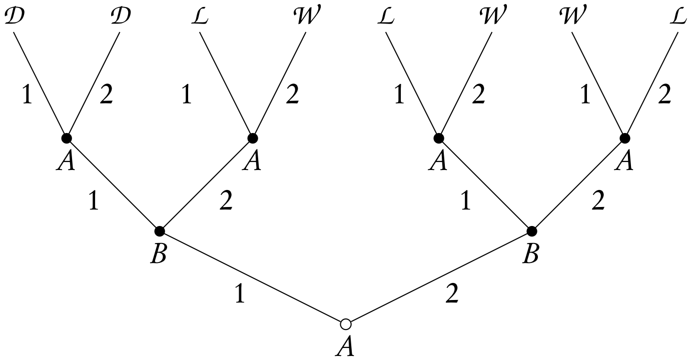


Figure 3.5: Game 18″

Now we assume that A makes the optimal move at the last step, so we can replace the top row of outcomes with the outcome that would happen if A moves optimally. This gets us Figure 3.6.

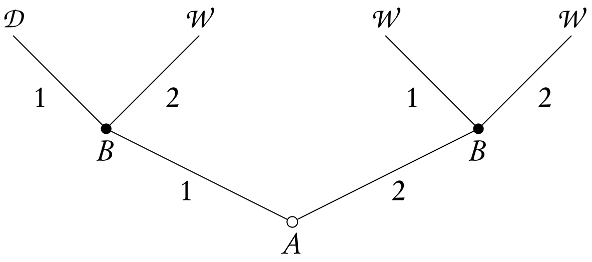


Figure 3.6: Game 18″ with last move assumed

Now assume that A plays 1 on the first round, then B has to move. From 3.6 it looks like B has an easy choice to make. If she plays 1, she gets a draw, if she plays 2, then *A* wins, i.e., she loses. Since drawing is better than losing, she should play 1 and take the draw.

But why think that playing 2 will lead to *A* winning? The argument that it did depending on assuming *A* is perfectly rational. And assuming *B* is in a position to make this choice, that seems like an unlikely assumption. After all, if *A* were perfectly rational, she'd have chosen 2, and given herself a chance to force a win.

Now you might think that even if *A* isn't perfectly rational, it still is crazy to leave her with an easy winning move. And that's probably a sufficient reason for *B* to accept the draw, i.e., play 1. But the argument that *B* should regard playing 2 as equivalent to choosing defeat seems mistaken. *B* knows that *A* isn't perfectly rational, and she shouldn't assume perfect rationality from here on out.

We will come back to this point, a lot, in subsequent discussions of backwards induction. Note that it is a point that doesn't really arise in the context of normal form games. There we might wonder about whether common knowledge of rationality is a legitimate assumption at the *start* of the game. But once we've settled that, we don't have a further issue to decide about whether it is still a legitimate assumption at later stages of the game.

Value of Games

Consider games with the following characteristics.

- *A* and *B* take turns making moves. We will call each point at which they make a move, or at which the game ends, a **node** of the game.
- At each move, each player knows what moves have been previously made.
- At each move, the players have only finitely many possible moves open.
- The players' preferences over the outcomes are opposite to one another. So if *A* prefers outcome o_1 to o_2, then *B* prefers o_2 to o_1, and if *A* is indifferent between o_1 and o_2, then *B* is indifferent between them as well.
- *A*'s preferences over the outcomes are complete; for any two outcomes, she either prefers the first, or prefers the second, or is indifferent between them.
- There is a finite limit to the total number of possible moves in the game.

The finiteness assumptions entail that there are only finitely many possible outcomes. So we can order the outcomes by *A*'s preferences. (Strictly speaking, we want to order sets of outcomes that are equivalence classes with respect to the relation that A is indifferent between them.) Assign the outcome A least prefers the value 0, the next outcome the value 1, and so on.

Now we recursively define the **value** of a node as follows.

- The value of a **terminal node**, i.e., a node at which the game ends, is the payoff at that node.
- The value of any node at which A makes a choice is the greatest value of the nodes between which A is choosing to move to.
- The value of any node at which B makes a choice is the least value of the nodes between which B is choosing to move to.

Finally, we say that the value of the game is the value of the initial node, i.e., the node at which the first choice is made. We can prove a number of things about the value of games. The proof will make crucial use of the notion of a **subgame**. In any extensive form game, a **subgame** of a game is the game we get by treating any perfect information node as the initial node of a game, and including the rest of the game 'downstream' from there.

By a perfect information node, I mean a node such that when it is reached, it is common knowledge that it is reached. That's true of all the nodes in most of the games we're looking at, but it isn't true in, for instance, the extensive form version of Prisoners' Dilemma we looked at. Nodes that are in non-degenerate information sets, i.e., information sets that contain other nodes, can't trigger subgames. That's because we typically assume that to play a game, players have to know what game they are playing.

Note that a subgame is really a game, just like a subset is a set. Once we're in the subgame, it doesn't matter a lot how we got there. Indeed, any game we represent is the consequence of some choices by the agent; they are all subgames of the game of life.

The value of a game is the value of one of the terminal nodes

> We prove this by induction on the length of games. (The length of a game is the *maximum* number of moves needed to reach a terminal node. We've only looked so far at games where every path takes the same number of moves to reach a conclusion, but that's not a compulsory feature of games.)

If the game has zero moves, then it has just one node, and its value is the value of that node. And that node is a terminal node, so the value is the value of a terminal node.

Now assume that the claim is true for any game of length k or less, and consider an arbitrary game of length $k + 1$ The first node of the game consists of a choice about which path to go down. So the value of the initial node is the value of one of the subsequent nodes. Once that choice is made, we are in a subgame of length k, no matter which choice is made. By the inductive hypothesis, the value of that subgame is the value of one of its terminal nodes. So the value of the game, which is the value of one of the immediate subsequent nodes to the initial node, is the value of one of its terminal nodes.

***A* can guarantee that the outcome of the game is at least the value of the game.**

Again, we prove this by induction on the length of games. It is trivial for games of length 0. So assume it is true for all games of length at most k, and we'll prove it for games of length $k+1$. The initial node of such a game is either a move by A or a move by B. We will consider these two cases separately.

Assume that A makes the first move. Then the value of the initial node is the maximum value of any immediate successor node. So A can select to go to the node with the same value as the value of the game. Then we're in a subgame of length k. By the inductive assumption, in that game A can guarantee that the outcome is at least the value of the subgame. And since the value of that node is the value of the subgame, so it is also the value of the initial node, i.e., the value of the initial game. So by choosing that node, and starting that subgame, A can guarantee that the outcome is at least the value of the game.

Now assume that B makes the first move. B can choose the node with the least value of the available choices. Then, as above, we'll be in a subgame in which (by the inductive hypothesis) B can guarantee has an outcome which is at most its value. That is, B can guarantee the outcome of the game is at most the value of the initial node of

> the subgame. And since *B* can guarantee that that subgame is played, *B* can guarantee that the game has an outcome of at most its value.

B **can guarantee that the outcome of the game is at most the value of the game.**

> The proof of this exactly parallels the previous proof, and the details are left as an exercise.

Let's note a couple of consequences of these theorems.

First, assume that the rationality of each player is common knowledge, and that it is also common knowledge that this will persist throughout the game. Then the kind of backwards induction argument we used is discussing Game 18 will show that the outcome of the game will be the value of the game. That's because if *A* is rational and knows this much about *B*, the outcome won't be lower than the value, and if *B* is rational and knows this much about *A*, the outcome won't be greater than the value.

Second, these theorems have many applications to real-life games. Both chess and checkers, for instance, satisfy the conditions we listed. The only condition that is not immediately obvious in each case is that the game ends in finite time. But the rules for draws in each game guarantee that is true. (Exercise: Prove this!) Since these games end with White win, Black win or draw, the value of the game must be one of those three outcomes.

In the case of checkers, we know what the value of the game is. It is a draw. This was proved by the Chinook project at the University of Alberta. We don't yet know what the value of chess is, but it is probably also a draw. Given how many possible moves there are in chess, and in principle how long games can go on, it is hard to believe that chess will be 'solved' any time soon. But advances in chess computers may be able to justify to everyone's satisfaction a particular solution, even if we can't prove that is the value of chess.

Best Responses

We have done as much as we can for now by merely thinking about dominance. We need to look at games where there is no dominant option. To do this requires taking a small detour into orthodox decision theory.

Crash Course in Probability Theory and Decision Theory

Probability

We'll start by going over the notion of probability that is used in decision theory. The best way to think about probability functions is by thinking about **measure** functions. Intuitively, a measure function is a non-negative **additive** function. That is, it is a function which only takes non-negative values, and where the value of a whole is equal to the sum of the value of the (non-overlapping) parts. In principle the domain of a measure function could be the size of any set, but we'll only consider functions with finite domains for now.

Such functions are frequently used in the real world. Consider, for instance, the functions from regions to their population and land mass. This table lists the (very approximate) population and land mass for England, Scotland, Wales and Northern Ireland.

	Population	Land Mass
	(Millions)	(Thousands of mi^2)
England	50	50
Scotland	5	30
Wales	3	8
Northern Ireland	2	5

The nice thing about tables like this is that they give you a lot of information implicitly. You don't need a separate table to tell you that the combined population of England and Scotland is 55 million, or that the combined land mass of England and Wales is 58,000 square miles. Those (approximate) facts follow from the facts listed on the table, as well as the fact that populations and land masses are additive. Once you know the population of the parts, you know the population of the whole.

The next notion we need is that of a **normalised** measure function. A normalisted measure is a measure where every function takes the value assigned to the largest whole is 1. We often use this notion when talking about proportions. Here is a table listing, again approximately, the portion of the UK population and land mass in each of the four countries.

	Population	Land Mass
	(Proportion)	(Proportion)
England	0.84	0.54
Scotland	0.08	0.32
Wales	0.05	0.09
Northern Ireland	0.03	0.05

The way we turn the regular measure function into a normalised measure is by dividing each value on the earlier table by the value of whole. So the land mass of England is (about) 50,000 mi^2, and the land mass of the UK is (about) 93,000 mi^2, and $^{50,000}/_{93,000}$ is about 0.54. So that's why we right 0.54 in the top right cell of the normalised measure table.

The last notion we need is that of a possibility space, or more formally, a **field** of propositions. Although these terms might be unfamiliar, the notion should be very familiar; it is just what we represent with a truth table. We will identify, for these purposes, propositions with sets of possible worlds. Then a field of propositions is a set of propositions such that whenever A and B are in the set, so are $\neg A$ and $A \wedge B$. This is, of course, just what we see with truth tables. Whenever you can represent two propositions on a truth table, you can represent their conjunction and each of their negations. Just like with truth-tables, we'll only consider finite fields. It's possible to extend this to cover any set-sized partition of possibility space, but we'll ignore such complications for now. (Philosophers sometimes write as if the restriction to set-sized partitions is unnecessary, and a probability function could assign a probability to any portion of possibility space. That would involve a serious deviation for mathematical orthodoxy, unless the space of possible worlds is set-sized.)

A **Probability function** is just a normalised measure on a possibility space. That is, it is a function satisfying these two conditions, where A and B are any propositions, and T is a logical truth.

- $0 \leq \Pr(A) \leq 1 = \Pr(T)$
- If A and B are disjoint, then $\Pr(A \vee B) = \Pr(A) + \Pr(B)$.

We normally suppose that any rational agent has **credences** over the space of (salient) epistemic possibilities, and these credences form a probability function.

In any finite field of propositions, there will be some **atoms**. These are propositions such that no stronger consistent proposition is in the field. (Some, but not all, infinite fields are atomic.) We will use these atoms a lot. I'll use the variable w to range over atoms, reminding us that for practical purposes, atoms behave a lot like *worlds*. They aren't worlds, of course; they only settle the truth values of a handful of propositions. But they are divisions of possibility space that are maximally precise for present purposes. That is, they are being treated as worlds. On a pragmatist approach to modality, there's no more to being a world than to be treated as a world in the right circumstances, and thinking about probability encourages, I think, such a pragmatist approach.

Conditional Probability

So far we've talked simply about the probability of various propositions. But sometimes we're not interested in the absolute probability of a proposition, we're interested in its **conditional** probability. That is, we're interested in the probability of the proposition *assuming* or *conditional on* some other proposition obtaining.

It isn't too hard to visualise how conditional probability works if we think of a possibility space simply as a truth table, so a measure ove a possibility space is simply a measure over on the truth table. Here's how we might represent a simple probability function, defined over the Boolean combinations of three propositions, if we're thinking about things this way.

p	q	r	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0.08
T	F	F	0.8
F	T	T	0.0002
F	T	F	0.001
F	F	T	0.01
F	F	F	0.1

If we assume that something , call it B is true, then we should 'zero out', i.e. assign probability 0, to all the possibilities where B doesn't obtain. We're now left with a measure over only the B-possibilities. The problem is that it isn't a normalised measure. The values will only sum to $\Pr(B)$, not to 1. We need to renormalise. So we divide by $\Pr(B)$ and we get a probability back. In a formula, we're left with

$$Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$$

Assume now that we're trying to find the conditional probability of p given q in the above table. We could do this in two different ways. First, we could set the probability of any line where q is false to 0. So we will get the following table.

p	q	r	
T	T	T	0.0008
T	T	F	0.008
T	F	T	0
T	F	F	0
F	T	T	0.0002
F	T	F	0.001
F	F	T	0
F	F	F	0

The numbers don't sum to 1 any more. They sum to 0.01. So we need to divide everything by 0.01. It's sometimes easier to conceptualise this as multiplying by $1/\Pr(q)$, i.e. by multiplying by 100. Then we'll end up with:

p	q	r	
T	T	T	0.08
T	T	F	0.8
T	F	T	0
T	F	F	0
F	T	T	0.02
F	T	F	0.1
F	F	T	0
F	F	F	0

And since p is true on the top two lines, the 'new' probability of p is 0.88. That is, the conditional probability of p given q is 0.88. As we were writing things above, $\Pr(p|q) = 0.88$.

Alternatively we could just use the formula given above. Just adding up rows gives us the following numbers.

$$\begin{aligned} \Pr(p \wedge q) &= 0.0008 + 0.008 = 0.0088 \\ \Pr(q) &= 0.0008 + 0.008 + 0.0002 + 0.001 = 0.01 \end{aligned}$$

Then we can apply the formula.

$$\begin{aligned} \Pr(p|q) &= \frac{\Pr(p \wedge q)}{\Pr(q)} \\ &= \frac{0.0088}{0.01} \\ &= 0.88 \end{aligned}$$

Bayes Theorem

This is a bit of a digression from what we're really focussing on in game theory, but it's important for understanding how conditional probabilities work. It is often easier to calculate conditional probabilities in the 'inverse' direction to what we are interested in. That is, if we want to know $\Pr(A|B)$, it might be much easier to discover $\Pr(B|A)$. In these cases, we use Bayes Theorem to get the right result. I'll state Bayes Theorem in two distinct ways, then show that the two ways are ultimately equivalent.

$$\begin{aligned} \Pr(A|B) &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \\ &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)Pr(A) + \Pr(B|\neg A)Pr(\neg A)} \end{aligned}$$

These are equivalent because $\Pr(B) = \Pr(B|A)Pr(A) + \Pr(B|\neg A)\Pr(\neg A)$. Since this is an independently interesting result, it's worth going through the proof of it. First note that

$$\begin{aligned}
\Pr(B|A)\Pr(A) &= \frac{\Pr(A \wedge B)}{\Pr(A)}\Pr(A) \\
&= \Pr(A \wedge B)
\end{aligned}$$

$$\begin{aligned}
\Pr(B|\neg A)\Pr(\neg A) &= \frac{\Pr(\neg A \wedge B)}{\Pr\neg(A)}\Pr(\neg A) \\
&= \Pr(\neg A \wedge B)
\end{aligned}$$

Adding those two together we get

$$\begin{aligned}
\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A) &= \Pr(A \wedge B) + \Pr(\neg A \wedge B) \\
&= \Pr((A \wedge B) \vee (\neg A \wedge B)) \\
&= \Pr(B)
\end{aligned}$$

The second line uses the fact that $A \wedge B$ and $\neg A \wedge B$ are inconsistent. And the third line uses the fact that $(A \wedge B) \vee (\neg A \wedge B)$ is equivalent to A. So we get a nice result, one that we'll have occasion to use a bit in what follows.

$$\Pr(B) = \Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)$$

So the two forms of Bayes Theorem are the same. We'll often find ourselves in a position to use the second form.

One kind of case where we have occasion to use Bayes Theorem is when we want to know how significant a test finding is. So imagine we're trying to decide whether the patient has disease D, and we're interested in how probable it is that the patient has the disease conditional on them returning a test that's positive for the disease. We also know the following background facts.

- In the relevant demographic group, 5% of patients have the disease.
- When a patient has the disease, the test returns a position result 80% of the time
- When a patient does not have the disease, the test returns a negative result 90% of the time

So in some sense, the test is fairly reliable. It usually returns a positive result when applied to disease carriers. And it usually returns a negative result when applied to non-carriers. But as we'll see when we apply Bayes Theorem, it is very unreliable in another sense. So let A be that the patient has the disease, and B be that the patient returns a positive test. We can use the above data to generate some 'prior' probabilities, i.e. probabilities that we use prior to getting information about the test.

- $\Pr(A) = 0.05$, and hence $\Pr(\neg A) = 0.95$
- $\Pr(B|A) = 0.8$
- $\Pr(B|\neg A) = 0.1$

Now we can apply Bayes theorem in its second form.

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\neg A)\Pr(\neg A)} \\
&= \frac{0.8 \times 0.05}{0.08 \times 0.05 + 0.1 \times 0.95} \\
&= \frac{0.04}{0.04 + 0.095} \\
&= \frac{0.04}{0.135} \\
&\approx 0.296
\end{aligned}$$

So in fact the probability of having the disease, conditional on having a positive test, is less than 0.3. So in that sense the test is quite unreliable.

This is actually a quite important point. The fact that the probability of B given A is quite high does not mean that the probability of A given B is equally high. By tweaking the percentages in the example I gave you, you can come up with cases where the probability of B given A is arbitrarily high, even 1, while the probability of A given B is arbitrarily low.

Confusing these two conditional probabilities is sometimes referred to as the *Prosecutors' fallacy*, though it's not clear how many actual prosecutors are guilty of it! The thought is that some prosecutors start with the premise that the probability of the defendant's blood (or DNA or whatever) matching the blood at the crime scene, conditional on the defendant being innocent, is 1 in a billion (or whatever it exactly is). They conclude that the probability of the defendant being innocent, conditional on their blood matching the crime scene, is about 1 in a billion. Because of derivations like the one we just saw, that is a clearly invalid move.

Expected Value

One very convenient notion to define using a probability function is that of an **expected value**. Say that a **random variable** is any function from possibilities to real numbers. We usually designate random variables with capital letters. So expressions like $X = x$, which you'll often see in probability textbooks, means that the random variable X takes real value x.

The expected value of a random variable is calculated using the following formula.

$$E(X) = \sum_{w} \Pr(w)X(w)$$

Remember that w is a variable that ranges over atoms. So the summation here is over the atoms in the field. $\Pr(w)$ is the probability of that atom obtaining, and $X(w)$ is the value of the random variable at w. Here's a worked example to make it clearer.

Let our random variable X be the number of children of the next person to walk through the door. Let's assume that the department contains 10 people; 2 of whom have three children, 3 of whom have two children, 3 of whom have one child, and 2 are childless. Each member of the department is equally likely to walk through the door, and no one else is going to. So the probability of any given member of the department walking through the door is 1/10. Then the expected value of X, i.e., $E(X)$, is:

$$\begin{gathered} 1/10 \times 3 + 1/10 \times 3 + \\ 1/10 \times 2 + 1/10 \times 2 + 1/10 \times 2 + \\ 1/10 \times 1 + 1/10 \times 1 + 1/10 \times 1 + \\ 1/10 \times 0 + 1/10 \times 0 = 1.5 \end{gathered}$$

Note that the expected value of X is *not* the value we 'expect', in the ordinary sense, X to take. We can be pretty sure that X's value will be an integer, so it won't be 1.5, for instance! That is to say, 'expected value' is a term of art.

As well as 'absolute' expected values, we can also define conditional expected values, as follows:

$$E(X|p) = \sum_{w} \Pr(w|p)X(w)$$

I won't go through the proof of this, but it turns out that conditional expected values satisfy a nice theorem. Let $\{p_1, \dots, p_n\}$ be a partition of possibility space. Then the following holds.

$$E(X) = \sum_{i=1}^{n} \Pr(p_i)E(X|p_i)$$

Here are a few of immediate corrollaries of that result.

$$\begin{aligned} E(X) - E(Y) &= E(X - Y) \\ &= \sum_{i=1}^{n} \Pr(p_i)(E(X|p_i) - E(Y|p_i)) \\ &= \sum_{i=1}^{n} \Pr(p_i)(E((X - Y)|p_i)) \end{aligned}$$

Often we know more about the difference between two expected values than we do about the absolute value of either. For instance, if we are comparing the expected values of taking and declining a small bet, we can figure out which action has a higher expected value without figuring out the expected value of either, which presumably depends on all sorts of facts about how our life might (probably) go. But as long as we can work out the difference in expected value between the two acts conditional on each p_i, and the probability of each p_i, we can work out at least which has a higher expected value, and by how much.

Orthodox Decision Theory

Given all that, it is fairly easy to summarise orthodox decision theory.

- Rational agents have a credence function over salient alternatives that is identical to some probability function Pr.
- Rational agents also have a real-valued utility function U over salient outcomes.
- The rational action for such an agent to do is the action that maximises **expected** utility, i.e., maximises the expected value of U relative to that function Pr.

Lexicographic Utility

Some authors suggest a small modification to this theory. This modification plays a prominent role in some recent work on game theory by Robert Stalnaker, so it will be useful to cover it now.

Let's say that the expected utility of two possible actions α and β are identical. But there are some possible states of the world that are live (in some salient sense), despite having probability 0. On every one of those states, α does better than β. Now how an action does on a probability 0 event makes no difference to its expected utility, so this feature doesn't increase α's expected utility. But it seems to be useful to know as a tie-breaker. Say that α is *barely better* than β iff they have the same expected utility, there are some live probability 0 events on which α is better than β, and no live probability 0 events on which β is better than α.

So we can define a new concept. Say that an action is **rational** iff there is no alternative action with a higher expected utility. And say that an action is **perfectly rational** iff it is rational, and there is no action that is barely better than it. Perfect rationality helps resolve some puzzles about infinite chains of events, and gives us a motivation for not choosing weakly dominated options.

Assume that a fair coin will be flipped a countable infinity of times. Let AH be the proposition that all the flips land heads, and EH the proposition that all the even numbered flips (i.e., the second, fourth etc.) land heads. An agent has the choice between α, which is accepting a gift of a bet that pays £1,000,000 if EH, β, which is accepting a gift of a bet that pays £1,000,000 if AH, and γ, which is to decline both gifts. All three options have an expected return of 0, since the probability of a coin landing heads on each of an infinity of flips is 0. But α is barely better than β, since if EH is true, it returns £1,000,000, while β still has an expected return of 0. And β is barely better than γ, since it returns £1,000,000 if AH, while γ still returns nothing.

On weak dominance, consider again Game 14.

Game 14	l	r
U	1, 1	100, 0
D	0, 100	100, 100

Assume that the row player thinks the probability that C will play r is 1. Then whether she plays U or D doesn't affect her rationality; she has an expected return of 100 either way. But it's a live possibility, in what seems to be the right sense, that C will play l. And if so, U does better than D. So U is barely better than D, and is the only perfectly rational play.

Best Responses

Let's return now to game theory, because we at last have the tools to illustrate a more subtle concept than domination.

Best Response A strategy s_i is a best response for player i iff there is some probability distribution Pr over the possible strategies of other players such that playing s_i maximises i's *expected* payoff, given Pr. (Note that we're using 'maximise' in such a way that it allows that other strategies do just as well; it just rules out other strategies doing better.)

Let's look at an example of this, using a game we have already seen.

Game 9	l	r
U	$3, 0$	$0, 0$
M	$2, 0$	$2, 0$
D	$0, 0$	$3, 0$

We will just look at things from the perspective of R, the player who chooses the row. What we want to show is that *all three* of the possible moves here are best responses.

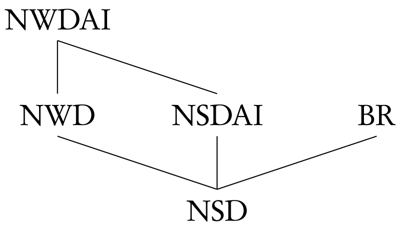
It is clear that U is a best response. Set $\Pr(l) = 1, \Pr(r) = 0$. Then $E(U) = 3, E(M) = 2, E(D) = 0$. It is also clear that D is a best response. Set $\Pr(l) = 0, \Pr(r) = 1$. Then $E(U) = 0, E(M) = 2, E(D) = 3$.

The striking thing is that M can also be a best response. Set $\Pr(l) = \Pr(r) = 1/2$. Then $E(U) = E(D) = 3/2$. But $E(M) = 2$, which is greater than $3/2$. So if R thinks it is equally likely that C will play either l or r, then R maximises expected utility by playing M. Of course, she doesn't maximise actual utility. Maximising actual utility requires making a gamble on which choice C will make. That isn't always wise; it might be best to take the safe option.

It's very plausible that agents should play best responses. If a move is not a best response, then there is no way that it maximises expected utility, no matter what one's views of the other players are. And one should maximise expected utility. So we have a new normative status to add to our list.

BR That is, is a **B**est **R**esponse.

Here is the updated graph of how these norms relate.



There are quite a few claims made in this graph; let's go through the proofs of them.

All BR strategies are NSD

> If s is strictly dominated by s', then the expected value of playing s' is greater than the expected value of playing s, no matter which strategy you think one's partners will play. So s is not a best response; s' is better. Hence any strategy that is a best response is not strictly dominated.

Some strategies are NWDAI, and hence NSDAI, NWD and NSD, but not BR

> Consider this variant on Game 9.

Game 19	l	r
U	3, 3	0, 0
M	1, 1	1, 1
D	0, 0	3, 3

Consider things from R's perspective. The probability that C will play l is x, for some x, and the probability that she will play r is $1-x$. So $E(U) = 3x, E(D) = 3(1-x)$ and $E(M) = 1$. If M is to be a best response, then we must have $E(M) \geq E(U)$, and $E(M) \geq E(D)$. The first condition entails that $1 \geq 3x$, i.e., $x \leq 1/3$. The second condition entails that $1 \geq 3(1-x)$, i.e., $x \geq 2/3$. But these two conditions can't both be satisfied, so M is not a best response under any circumstances.

But nor is M even weakly dominated. Since M sometimes does better than U, and sometimes does better than D, it is not dominated by either. Moreover, neither of C's strategies is weakly dominated by the other. So eliminating weakly dominated strategies removes no strategies whatsoever from the game. Hence M is NWDAI, and hence NSDAI, NWD and NSD, but it is not BR.

Some BR strategies are not NSDAI, and hence not NWDAI

Consider the following game, whose payoffs will be explained below.

Game 20	0	1	2	3
0	9, 9	6, 11.1	4, 9.2	2, 7.3
1	11.1, 6	7.1, 7.1	4.1, 9.2	2.1, 9.2
2	9.2, 4	9.2, 4.1	5.2, 5.2	2.2, 7.3
3	7.3, 2	7.3, 2.1	7.3, 2.2	3.3, 3.3

The idea is that R and C will 'play' three rounds of Prisoners' Dilemma. They have to specify their strategy in advance, and, to make things easy, they only have four options. Their strategy must be of the form "Tit-for-tat-minus-n", for $n \in \{0, 1, 2, 3\}$. The strategy "Tit-for-tat" in an iterated Prisoners' Dilemma is the strategy of playing B in the first game, then in any subsequent game, playing whatever one's partner played in the previous game. The strategy "Tit-for-tat-minus-n" is the strategy of playing Tit-for-tat until there are n moves to go, at which point one plays A the rest of the way.

To make sure various strategies are *strictly* dominated, I tinkered with the payoffs a little. Someone who plays Tit-for-tat-minus-n gets a bonus of $n/10$.

From now on, unless it makes too much ambiguity, we'll use n to refer to the strategy of playing Tit-for-Tat-minus n. Note that 0 is strictly dominated by both 1 and 2. If we assume that neither player will play 0, then 1 is dominated by 2. And if we assume that neither player will play 0 or 1, then 2 is dominated by 3. So iterative deletion of strictly dominated strategies leaves us with nothing but 3. That is, both players should play A from the start, if they want to play strategies that are NSDAI.

But also note that any strategy except 0 is BR. Given a probability of 1 that the other player will play 0, the best response is 1. Given a probability of 1 that the other player will play 1, the best response is 2. And given a probability of 1 that the other player will play 2 or 3, the best response is 3. So 1 and 2 are both BR, but not NSDAI, and hence not NWDAI.

Some BR strategies are not NWD

This one is a little odd, though it turns out to be important for a lot of subsequent analysis.

Game 10	l	r
U	1, 1	0, 0
D	0, 0	0, 0

This is a game of perfect co-operation; the players get the same payoffs in every case. And they both get 1 if we end up at $\langle U, l\rangle$, and 0 otherwise. Note that both D and R are weakly dominated. So $\langle U, l\rangle$ is the only NWD outcome of the game.

But both D and r are best responses. If C plays r, then it doesn't matter what R plays. Since a best response only has to be as good as the alternatives, not better, in that circumstance D is a best response. A similar argument shows that r is a best response. So we have two examples of strategies that are BR despite being weakly dominated.

Iterated Best Responses

Some best responses are pretty crazy when playing against a rational opponent. Consider the following game from R's perspective.

Game 21	L	r
U	5, 5	0, -5
D	0, 5	2, -5

In this game, D is a best response. It does best if C chooses r. But why on earth would C do that? C gets 5 for sure if she chooses L, and -5 for sure if she chooses r. Assuming the weakest possible rationality constraint on C, she won't choose a sure loss over a sure gain. So given that, C should choose U.

Of course, we could have shown that with just considering domination. Note that U is both NSDAI and NWDAI, while D has neither of these properties. The following example is a little more complex.

Game 22	l	m	r
U	1, 3	0, 1	1, 0
D	0, 0	1, 1	0, 3

In this game neither player has a dominated move. So just using domination techniques can't get us closer to solving the game. And, from R's perspective, thinking about best responses doesn't help either. U is a best response if C is going to play l or r with probability at least $1/2$, and D is best response if C is going to play m with probability at least $1/2$.

But note that m is not a best response for C. The argument for this is just the argument we used in Game 19. Now let's assume, when thinking from R's perspective, that C will play a best response. That is, we're assuming C will play either l or r. Given that, the best thing for R to do is to play U.

More carefully, if R knows that C is rational, and if R knows that rational agents always play best responses, then R has a compelling reason to play U. This suggests a new normative status.

BRBR That is, is a **B**est **R**esponse to a **B**est **R**esponse.

All BRBR strategies are BR

> This is obvious. All BRBR strategies are best responses to best responses, hence they are best responses to something, which by definition makes them BR.

Some BR strategies are not BRBR

> See Game 22. From *R*'s perspective, *D* is BR but not BRBR.

We can go on further. Think about Game 22 from C's perspective. Assuming R is rational, and knows C is rational, then R has a compelling reason to play U. And if R plays U, the best response for C is l. This doesn't mean that l is the only best response; r is also a best response. Nor does it mean l is the only best response to a best response; r is the best response to D, which as we showed above is a best response. What's really true is that l is the only best response to a best response to a best response.

That means that if C knows that R knows that C is rational, then C has a strong reason to play l. We could designate this with a new status, perhaps BRBRBR. But at this point it's best to simply iterate to infinity.

BRBRI That is, is a **B**est **R**esponse to a **B**est **R**esponse to a Best Response, and so on to **I**nfinity.

All BRBRI strategies are BRBR

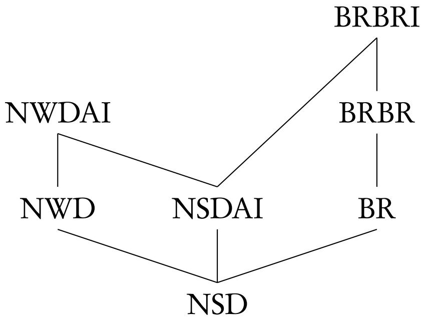
> This is obvious.

Some BRBR strategies are not BRBRI

> See Game 22. From C's perspective, *R* is BRBR but not BRBRI.

We say that p is **mutual knowledge** if every player in the game knows it. We say that p is **common knowledge** if everyone knows it, and everyone knows everyone knows it, and so on. It seems plausible that in any game where it is mutual knowledge that everyone is rational, agents should only play BRBR strategies. And at least there's a case to be made that in a game where it is common knowledge that the players are rational, players should play BRBRI strategies. (Though we will come back to this point repeatedly, especially in the context of extensive form games.)

Let's update our graph of different normative statuses.



We have already gone over some of the proofs of the claims in the graph; let's finish up this part by going over the rest.

All BRBRI strategies are NSDAI

> The proof of this is actually a little complicated. We will just go over the proof for a two-player game; generalising the proof introduces even more complications.
>
> Assume for *reductio* that it isn't true. That is, assume that a BRBRI strategy, call it s_0 is deleted in the process of iterative deletion of strongly dominated strategies. So s_0 is a best response to some strategy s_1 which is a best response to some strategy s_2 which is a best response to some strategy s_3 etc. (Since we're assuming games are finite, many of these s_i must be identical to each other.)
>
> Every such deletion happens at some round or other of deletion. Let the round that this BRBRI strategy is deleted be the n'th round. That is, given the strategies surviving after $n - 1$ rounds, some alternative strategy s' does better than s_0 no matter what alternative strategy is played. So s_1 must have been already deleted. That's because s_0 is a best response to s_1, and an inferior response to all surviving strategies. So s_1 is deleted in, at the latest, the $n - 1$'th round.
>
> Now consider the round at which s_1 is deleted. Some alternative to s_1 is a better response than s_1 to any strategy surviving to this stage of the iterative deletion process. But s_1 is, by hypothesis, a best response to s_2. So s_2 must have been antecedently deleted before this round.

That is, s_2 must have been deleted by at the latest the $n-2$'nd round of the iterative deletion process.

Generalising this argument, strategy s_k must be deleted by, at the latest, the $n-k$'th round of the iterative deletion process. But this is impossible. Consider, for instance, the case where $k = n+1$. Strategy s_{n+1} must be deleted by, at the latest, the -1'th round. But there is no -1'th round, so we get a contradiction, completing the *reductio*.

Some strategies which are BRBR are not NSDAI, and hence not NWDAI

See Game 20. Playing 1 is a best response (to playing 0). And playing 2 is a best response to playing 1, so it is BRBR. But iterative deletion of strongly dominated strategies deletes first 0, then 1, then 2. So playing 2 is not NSDAI.

Some strategies which are BRBRI, and hence BR, and not NWD

We again use Game 10. Playing D is a best response to playing R, which is a best response to playing D, which is a best response to playing R, etc. So each of these strategies is BRBRI. But both strategies are weakly dominated. So being BRBRI does not entail being NWD.

Nash Equilibrium

In a two-player game, for a strategy s_0 to be BRBRI, it must be the best response to some strategy s_1, which is the best response to some strategy s_2, which is the best response to some strategy s_3, etc. But we are assuming that there are only finitely many strategy choices available. So how can we get such an infinite chain going?

The answer, of course, is to repeat ourselves. As long as we get some kind of loop, we can extend such a chain forever, by keeping on circling around the loop. And the simplest loop will have just two steps in it. So consider any pair of strategies $\langle s_0, s_1\rangle$ such that s_0 is the best response to s_1, and s_1 is the best response to s_0. In that case, each strategy will be BRBRI, since we can run around that two-step 'loop' forever, with each stop on the loop being a strategy which is a best response to the strategy we previous stopped at.

When a pair of strategies fit together nicely like this, we say they are a **Nash Equilibrium**. More generally, in an n-player game, we use the following definition.

NE Some strategies $s_1, ..., s_n$ in an n-player game form a **N**ash **E**quilibrium iff for each i, s_i is a best response to the strategies played by the other $n-1$ players. That is, iff player i cannot do better by playing any alternative strategy to s_i, given that the other players are playing what they actually do.

The 'Nash' in Nash Equilibrium is in honour of John Nash, who developed the concept, and proved some striking mathematical results concerning it. (You may remember that Nash was the subject of a bio-pic some years back, 'A Beautiful Mind'. The movie required believing that Russell Crowe was a mad genius, and didn't properly deploy the notion of Nash Equilibrium, so maybe it is best if we keep our contact with Nash to the textbooks.)

The next few results are enough to let us place **NE** on our table of norms.

All NE strategies are BRBRI

> By hypothesis, if s is part of a NE pair, then there is some s' such that s is a best response to s', and s' is a best response to s. And that means that each of s and s' is BRBRI.

Some BRBRI are not NE

We'll prove a slight stronger result in a few pages, but for now we can get the result we need by looking at the following simple game.

Game 23	l	r
U	1, 0	0, 1
D	0, 1	1, 0

None of the (pure) strategies U, D, l or r are NE. That's because there's no NE pair we can make out of those four. And that's fairly obvious from the fact that whichever corner of the table we end up in, one of the players would have done better by swapping their strategy.

But note that each of the four strategies is BRBRI. U is a best response to l, which is a best response to D, which is a best response to r, which is a best response to U, which is ....

The point here should be clear once we think about how we got from the idea of BRBRI to the idea of NE. We wanted a 'loop' of strategies, such that each was a best response to the strategy before it. NE was what we got when we had a loop of length 2. But there are loops which are longer than that; for example, there are loops of length 4. And any loop is sufficient for the strategies on the loop to be BRBRI. And these strategies need not be NE.

Some NE strategies are not NWD, and hence not NWDAI

We again use Game 10.

Game 10	l	r
U	1, 1	0, 0
D	0, 0	0, 0

The pair $\langle D, r\rangle$ is a Nash Equilibrium pair. But neither D nor r survives the cutting out of weakly dominated strategies.

Some NWDAI strategies are not NE

This one is a little more complicated, but only a little. Basically, we take Game 23 and add a 'co-operative' strategy.

Game 24	*l*	*m*	*r*
U	4, 4	3, 3	3, 3
M	3, 3	5, 0	0, 5
D	3, 3	0, 5	5, 0

First, note that $\langle U, l\rangle$ is a Nash Equilibrium. Second, note that if *R* plays *M*, then *C*'s best response is to play *r*. And if *C* plays *r*, *R*'s best response is to play *D*, and so on. So *R* playing *M* can't be part of any Nash Equilibrium. If it were, there would have to be some response on *C*'s part to which *R* playing *M* was a best response, and there isn't. Similar arguments show that *R* playing *D* isn't part of any Nash Equilibrium. And similar arguments to that show that *C* playing either *m* or *r* can't be part of any Nash Equilibrium.

So $\langle U, l\rangle$ is the *only* (pure) Nash Equilibriumin the game. (I'll explain why I'm puttng 'pure' in occasionally very soon!) But eliminating dominated strategies gets us precisely nowhere, since as can be quickly verified, there are no dominated strategies in the game. So for each player, playing *M* is NWDAI, but is not NE.

Note that this example is also an argument that some BRBRI strategies are not NE, since for each player *M* is BRBRI. (Exercise: Prove this!)

Given all that, we can update the graph of normative statuses. The latest iteration is Figure 5.1.

Nash Equilibrium in Simultaneous Move Games

Let's return to Game 23. It looks at first like there won't be any Nash Equilibrium strategies in this game. That would be unfortunate; all of our statuses so far are exemplified by at least one strategy in each game.

But that would be too quick. It leaves out a large class of strategies. Game theorists say that as well as simply choosing to play *U*, or choosing to play *D*, *R* has another choice. She can play a **mixed strategy**. A mixed strategy is where

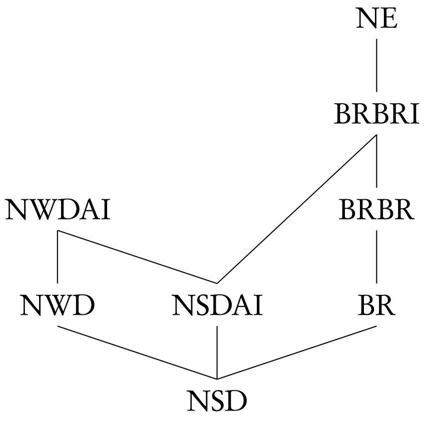


Figure 5.1: Normative Statuses

the player plays different pure strategies with different probabilities. We'll come back very soon to what we might possibly *mean* by 'probability' here, but for now let's explore the consequences for the existence of Nash Equilibria.

Let's assume that R plays U with probability $^1/_2$, and D with probability $^1/_2$. And similarly, assume that C plays l with probability $^1/_2$, and r with probability $^1/_2$. Can either player do better by deviating from these strategies?

Let's look at it first from C's perspective. If she plays l, then her expected return is given by the following equation.

$$\begin{aligned} E(L) &= \text{Prob that } R \text{ plays } U \times \text{Payoff of } \langle U, l\rangle \\ &\qquad + \text{Prob that } R \text{ plays } D \times \text{Payoff of } \langle D, l\rangle \\ &= {}^1/_2 \times 0 + {}^1/_2 \times 1 \\ &= {}^1/_2 \end{aligned}$$

And the expected return of playing r is given by the following equation.

$$\begin{aligned} E(R) &= \text{Prob that } R \text{ plays } U \times \text{Payoff of } \langle U, r\rangle \\ &\qquad + \text{Prob that } R \text{ plays } D \times \text{Payoff of } \langle D, r\rangle \\ &= {}^1/_2 \times 1 + {}^1/_2 \times 0 \\ &= {}^1/_2 \end{aligned}$$

Let M_x be the mixed strategy of playing L with probability x, and R with probability $1-x$. Then the expected value of M_x is given by the following equation.

$$\begin{aligned} E(M_x) &= \Pr(l)E(l) + \Pr(r)E(r) \\ &= {}^{x}/_{2} + {}^{1-x}/_{2} \\ &= {}^{1}/_{2} \end{aligned}$$

So whichever strategy C adopts, whether it is l, r or one of the continuum many values of M_x, she'll have an expected payout of $^1/_2$. That means that she can't do any better by playing any alternative to $M_{1/2}$. Of course, that's for the rather boring reason that any strategy is as good as any other at this point.

When we're discussing R's strategies, we'll say that M_x is the strategy of playing U with probability x, and D with probability $1-x$. A similar argument shows that given that C is playing $M_{1/2}$, all strategies are as good as each other for R. That means that the pair $\langle M_{1/2}, M_{1/2}\rangle$ is a Nash Equilibrium. Each player does as well as they can playing $M_{1/2}$ given that the other player is playing $M_{1/2}$. And that's the definition of a Nash Equilibrium.

It turns out that for any game with a finite number of choices for each player, there is always at least one Nash Equilibrium, if we include mixed strategies. The proof of this is beyond the scope of these notes, however.

Rather than using ad hoc naming conventions like M_x, it would be good to have better ways of referring to mixed strategies. I'll use the following (fairly standard) notation. If a player's choices for pure strategies are $s_1, s_2, ..., s_n$, then the vector $\langle x_1, x_2, ..., x_n\rangle$ will represent the mixed strategy of playing s_i with probability x_i. If the game is represented on a table, we'll let the first (i.e., leftmost) column be C's strategy s_1, the second column be her strategy s_2, and so on. And we'll let the first (i.e., highest) row be R's strategy s_1, the second row be her strategy s_2, and so on. This notation will need to get more complicated when we consider games in extensive form, but in fact we usually use mixed strategies for games displayed in strategic form, so this isn't a huge loss.

Mixtures and Dominance

Now that we have the notion of a mixed strategy, we need to revise a little bit about what we said about dominant strategies. Recall that we used the Game 19 to show that some strategies which were not dominated were nevertheless not best responses.

Game 19	l	r
U	3, 3	0, 0
M	1, 1	1, 1
D	0, 0	3, 3

Now compare the strategy M to the strategy $S = \langle 1/2, 0, 1/2 \rangle$. If C plays l, then M returns 1, but the mixed strategy has an expected return of 1.5. If C plays r, then M returns 1, but S has an expected return of 1.5. Indeed, if C plays any mixture of l and r, then M returns 1, but S still has an expected return of 1.5. In effect, M is a dominated strategy; it is dominated by S.

So when we are talking about whether a strategy is dominated, we have to distinguish the scope of the tacit quantifier. Remember, 'dominated' means 'dominated by something'. If we are only quantifying over pure strategies, then M is not dominated. If we are quantifying over mixed strategies as well, then M is dominated. At least some authors use 'dominated' with this more expansive quantifier domain in mind, so M is dominated because S dominates it. There is a nice advantage to doing this. (There are also costs; it is harder to tell just by looking whether strategies are dominated in this more expansive sense.) It means we can distinguish nicely between what I've called a Best Strategy, which is really a strategy that is not strongly dominated, and what we might call Perfect Strategies, which are strategies that are not weakly dominated in this sense. We'll come back to this when we look at Stalnaker's work in a few chapters.

What is a Mixed Strategy?

So far we've noted that there are some games that don't have pure strategy Nash Equilibria. And we said that if we expand the space of available strategies to include mixed strategies, then those games do have Nash Equilibria. In fact we've said, though not gone close to proving this, that all finite games have at least one Nash Equilibrium solution, once we allow that agents can play mixed strategies.

But what does it mean to 'play a mixed strategy'? As game theorists sometimes put it, how should be **interpret** talk of mixed strategies. It turns out the options here are very similar to the candidate 'interpretations' of probability. (See the SEP entry on interpretations of probability for more on this debate, if you're interested.) Basically the interpretations can be classified as either **metaphysical** or **epistemological**. We're going to start with one of the metaphysical interpretations, then look at a few epistemological interpretations, and finally return to some more mundane metaphysical interpretations.

The most obvious, and natural, interpretation uses objective chances. What it means to play the strategy $\langle x, 1-x\rangle$ is to grab some chance device that goes into one state with chance x, and another state with chance $1-x$, see which state it goes into, then play the s_1 if it is in the first state, and s_2 if it is in the second state. Consider, for instance, the game Rock, Paper, Scissors, here represented as **Game 25**.

Game 25	Rock	Paper	Scissors
Rock	0, 0	-1, 1	1, -1
Paper	1, -1	0, 0	-1, 1
Scissors	-1, 1	1, -1	0, 0

For each player, the equilibrium strategy is to play $\langle 1/3, 1/3, 1/3\rangle$. (Exercise: Verify this!) The chance interpretation of this mixed strategy is that the player takes some randomising device, say a die, that has a 1/3 chance of coming up in one of three states. Perhaps the player rolls the die and plays Rock if it lands 1 or 2, Paper if it lands 3 or 4, Scissors if it lands 5 or 6.

A slightly more elegant version of this strategy involves the game Matching Pennies. We've seen the formal version of this game before, but the informal version is fun as well. Basically each player reveals a penny, and Row wins if they are both heads up or both tails up, and Column wins if one coin is heads up and the other tails up. Apparently this was a source of some schoolyard amusement before students had things like Angry Birds, or football, to play. As I said, we've seen the game table before, though not with these labels.

Game 26	Heads	Tails
Heads	1, -1	-1, 1
Tails	-1, 1	1, -1

Again, the only equilibrium solution is for each player to play $\langle 1/2, 1/2\rangle$. And here the chance interpretation of this strategy is that each player plays by simply flipping their coin, and letting it land where it may.

But obviously it is very hard to procure a chance device on the spot for any mixed strategy one might choose to play. How should we interpret a mixed strategy then? The natural move is to opt for some kind of *epistemological* interpretation of mixtures.

One option is a straightforward subjectivist interpretation of the relevant probabilities. So Row plays a mixed strategy $\langle x, 1-x\rangle$ iff Column's subjective probability that Row is playing s_1 with is x, and her subjective probability that Row is playing s_2 is $1-x$. One does hear game theorists offer such subjective interpretations of mixed strategies, but actually they don't seem to make a lot of sense. For one thing, it's hard to say how Column's credences should be in any sense a *strategy* for Row, unless Row has Jedi mind-control powers. And if Row does have Jedi mind-control powers, then she shouldn't settle for any kind of mixed strategy equilibrium. In Rock, Paper, Scissors, for instance, she should follow the strategy of playing Rock and using her Jedi mind-control powers to get Column to think she's playing Paper.

Perhaps we can retain a subjectivist interpretation if we change who the subject is. Frank Arntzenius, in a recent paper called "No Regrets", offers a different kind of subjectivist interpretation. He says that an agent plays a mixed strategy $\langle x, 1-x\rangle$ if her credences at the end of rational deliberation are that she'll play s_1 with probability x, and s_2 with probability $1-x$. He admits there are oddities with this interpretation. In particular, he notes that it means that if you take some kind of equilibrium theory to be the theory of rationality (as he does), then our theory of rationality turns out to be a theory of what one should believe one will do, not what one will do. This feels a little like changing the subject.

Could some kind of objective version of an epistemological interpretation do any better? Perhaps we could say that to play $\langle x, 1-x\rangle$ is to act in such a way that the objectively rational thing to believe about the agent is that she'll play s_1 with probability x? Arguably, the objective chance interpretation is a version of this; given the Principal Principle, the rational thing to believe about a person who uses a randomising device that comes up s_1 with chance x is that they'll play s_1 with probability x. But beyond that, it is hard to see what advantage the view has. After all, if it is an available strategy to make rational people think that you're playing s_1 with probability x, in Rock, Paper, Scissors you should make them think you're likely playing Paper when you're actually playing Rock. So it's hard to see how the equilibrium solution is rational.

In Ken Binmore's decision theory textbook *Playing for Real*, he seems to endorse something like the objective epistemological interpretation of mixed strategies.

> Suppose that we deny Eve access to a randomizing device when she plays Matching Pennies with Adam. Is she now doomed to lose? Not if she knows her Shakespeare well! She can then make each choice of *head* or *tail* contingent on whether there is an odd or even number of speeches in the successive scenes of *Titus Andronicus*. Of course, Adam might in principle guess that this is what she is doing—but how likely is this? He would have to know her initial state of mind with a quite absurd precision in order to setle on such a hypothesis. Indeed, I don't know myself why I chose *Titus Andronicus* from all Shakespeare's plays ... To outguess me in such a manner, Adam would need to know my own mind better than I know it myself. (Binmore 2006, 185).

But why is the likelihood that Adam will figure out Eve's decision a component of Eve's *strategy*? Either Eve has some control over the probabilities Adam will assign to her making various choices, or she doesn't. If she doesn't, then she doesn't have the power to play any mixed strategy interpreted this way. If she does, she shouldn't use them to give Adam *true* beliefs about her likely decisions. So it's hard to see the advantages.

Perhaps then the detour through epistemological interpretations was a red herring. And this leads us back to two more metaphysical interpretations, both of them involving frequencies.

One of these interpretations is particularly important in biology, and we will return to this idea much later in the course. The idea is that a species plays a mixed strategy $\langle x, 1-x \rangle$ iff x of its population plays s_1, and $1-x$ of its population plays s_2. Once we're looking at population level 'strategies' we've obviously moved a little away from the focus on *rational* game playing that has been the focus of the course to date. I'm not sure that it even makes sense to assign agency to populations. (Probably many of you think something stronger, that it's clearly wrong to assign agency to populations. But I'm an extreme enough functionalist to think that populations might have beliefs, desires, intentions and the like. I don't want to make this part of the notes turn on my crazy functionalism though!) And presumably the way the population implements the policy of having this frequency distribution of strategies is by having some randomising device that sorts individual organisims into one of two types. So perhaps this isn't really an alternative to the objective chance interpretation either.

The other big alternative is hinted at in Binmore's discussion. Note that he refers to 'each choice of *head* or *tail*'. This implicates at least that there will be more than one. So what he's really interested in is the case where Adam and Eve play Matching Pennies repeatedly. In this case, we might say that playing a mixed strategy $\langle x, 1-x \rangle$ is playing s_1 in x of the repeated games, and playing s_2 in $1-x$ of the repeated games. (Actually, we might want something more specific than that, but saying just what we need is hard. See the SEP entry on "Chance versus Randomness" for more discussion of the difficulties here.)

But it's odd to bring repeated games in here. It looks suspiciously like changing the subject. What we care about is what we should do in this very game, not in a long sequence of games. Put another way, we should in the first instance be looking for a rule about what to do in a (known to be) one-shot game. What should we do in such a game? Considerations about what would happen were we to repeat the game seem irrelevant to that.

The issues about repeated games are complicated, and it is worth spending some time going over them directly.

Nash, Best Response and Repeated Games

Here's a hypothesis. I happen to think it's basically true, but for now all that matter is that it is a hypothesis. The hypothesis is that in any game where there is common knowledge of rationality, any strategy which is BRBRI (i.e., that is a best response to a best response to a best response ...) can be rationally played.

Now here's an objection to that hypothesis. Consider repeated games of Rock, Paper, Scissors. In any given game, playing Rock is BRBRI. That's because playing Rock is a best response to playing Scissors, which is a best response to playing Paper, which is a best response to playing Rock, and so on. But playing Rock repeatedly is dumb. If it weren't dumb, the following scene (from The Simpsons episode "The Front" (April 15, 1993)) wouldn't be funny.

> Lisa: Look, there's only one way to settle this. Rock-paper-scissors.
>
> Lisa's brain: Poor predictable Bart. Always takes 'rock'.
>
> Bart's brain: Good ol' 'rock'. Nuthin' beats that!
>
> Bart: Rock!
>
> Lisa: Paper.
>
> Bart: D'oh!

Since Bart's strategy is BRBRI, and is irrational, it follows that the hypothesis is false.

I think that objection is too quick, but it takes some time to say why. Let's think a bit about where Bart goes wrong. Imagine he and Lisa play Rock, Paper, Scissors (hereafter RPS) many many times. At first Lisa plays the mixed strategy $\langle 1/3, 1/3, 1/3 \rangle$, and Bart plays Rock. So each of them have an expected return of 0. By the 11þ round, Lisa has figured out what Bart is playing, and plays Paper, for an expected return of 1, while Bart has an expected return of -1.

Now let's think about when Bart goes wrong. To do this, I'm going to assume that Bart is rational. This is clearly false; Bart is obviously not playing rationally. But I want to see just where the assumption of rationality leads to a contradiction. Let's say we knew in advance nothing but that Bart was going to play Rock on the 11th round. Since we're still assuming Bart is rational, it follows that playing Rock is a best response given his credences over Lisa's moves. Without I hope loss of generality, let's assume Lisa is still playing $\langle 1/3, 1/3, 1/3 \rangle$. Then Bart's expected return is 0, like for any other strategy, so it's fine to play Rock.

But of course Bart doesn't just play Rock in the 11th round. He also plays in the previous ten rounds. And that means that Lisa won't still be playing $\langle 1/3, 1/3, 1/3 \rangle$ by the time we get to the 11th round. Instead, she'll be playing Paper. So Bart's expected return in the 11th round is not 0, it is (roughly) -1.

I think something follows from those reflections. When we added in information about Bart's play in the first ten rounds, Bart's expected return in the 11th round dropped. So I conclude that there was a long-term cost to his play in the first 10 rounds. When he played Rock all those times, his expected return *in that round* was 0. But he incurred a long-term cost by doing so. That long-term cost isn't properly represented in the matricies for each round of the game. When we include it, it no longer becomes clear that Rock is BRBRI in each round.

A natural question then is, what really is the payoff table for each round of RPS. The existing table isn't a payoff table for two reasons. First, it lists outcomes, not valuations of outcomes. And second, it only lists one kind of outcome, the short-run winnings, not the long-run consequences of any given strategy. What we really need is a valuation function over long-run consequences.

So what is the payoff table for each round of the game? I think that's just much too hard a question to answer. What we can say with some precision is what the short-run outcome table is for each round. And summing short-run outcomes gives us long-run outcomes, which hopefully will correlate in some sensible way with payoffs. But determining the long-run outcomes of any given strategy is much too hard.

And that's one reason to care about both mixed strategies, and the existence of equilibria. Sometimes the best we can do in specifying a game is to specify the short-run outcomes, and say that the long-run outcomes are sums of them. Now the following hypothesis is clearly false: In any long-run game, it is rationally permissible to, at each stage, play any strategy which would be BRBRI if the short-run outcomes were the total payoffs. The Bart and Lisa example refutes that hypothesis. But it doesn't refute the hypothesis that I started this section with, namely that any BRBRI strategy is rationally acceptable.

Games, Decisions and Equilibria

Let's start today with a different hypothesis. Today's hypothesis is that only Nash Equilibrium strategies are rationally permissible. This is something that many game theorists would endorse. But it's a striking thesis. It says that sometimes only mixed strategies are rationally permissible, since any pure strategy will not be in equilibrium. And this is surprising given what we said earlier about decision theory.

Remember that a game is just a decision problem where the relevant parts of the world include other rational actors. So we should think that decision theoretic considerations would apply. Indeed, the view that game theory and decision theory have radically different norms will have some odd consequences in situations where one is dealing with borderline rational agents, such as, perhaps, non-human primates or very young human children.

And yet, orthodox decision theory *never* says that mixed strategies are preferable to pure strategies. It might say that a mixed strategy is no worse than a pure strategy, but it will never say that the mixed strategy is better. We can see this quite easily. Assume our agent has two possible choices, s_1 and s_2. The expected value of s_1 is $E(s_1)$, and the expected value of s_2 is $E(s_2)$. Without loss of generality, assume $E(s_1) \geq E(s_2)$. (If that's not true, just relabel the strategies so it becomes true!) Let's work out the expected value of an arbitrary mixed strategy.

$$\begin{aligned}
E(\langle x, 1-x\rangle) &= xE(s_1) + (1-x)E(s_2) \\
&= xE(s_1) + (1-x)E(s_1) + (1-x)((E(s_2) - E(s_1)) \\
&\leq xE(s_1) + (1-x)E(s_1) \\
&= E(s_1)
\end{aligned}$$

The reasoning on the third line is that since $E(s_1) \geq E(s_2)$, $E(s_2) - E(s_1) \leq 0$. And since x is a probability, and hence in $[0,1]$, $1-x \geq 0$. So $(1-x)((E(s_2)-E(s_1)) \leq 0$, so $xE(s_1) + (1-x)E(s_1) + (1-x)((E(s_2) - E(s_1)) \leq xE(s_1) + (1-x)E(s_1)$.

So the mixed strategy can't do better than the better of the pure strategies. Indeed, it only does as well as s_1 when $(1-x)((E(s_2) - E(s_1)) = 0$, which means either $1 - x = 0$ or $E(s_2) - E(s_1)$. The first of these happens when $x = 1$, i.e., when the agent plays the pure strategy s_1. The second of these happens when

$E(s_2) = E(s_1)$, i.e., the two pure strategies have the same expected return, so it doesn't matter, from an expected utility perspective, which one we play. (We'll come back to this fact a lot next time.)

So if a mixed strategy never has a higher expected return than a pure strategy, how can it be better? To see what is really happening, we need to step back a bit and revisit some notions from decision theory. We'll need to rethink what we were saying about dominance, and as we do that, revisit what we were saying about expected value maximisation.

Dominance and Independence

Orthodox game theory endorses the following principle:

Dominance If choice c_1 does better than choice c_2 in every possible state, then c_1 is preferable to c_2.

We could derive that by showing that c_1 will have a higher expected utiility than c_2, but that would be partially missing the point. The reason we care about expected utility is because it endorses principles like Dominance.

But Dominance has some odd consequences. Jim Joyce suggests the following puzzle case:

> Suppose you have just parked in a seedy neighborhood when a man approaches and offers to "protect" your car from harm for £10. You recognize this as extortion and have heard that people who refuse "protection" invariably return to find their windshields smashed. Those who pay find their cars intact. You cannot park anywhere else because you are late for an important meeting. It costs £400 to replace a windshield. Should you buy "protection"? Dominance says that you should not. Since you would rather have the extra £10 both in the even that your windshield is smashed and in the event that it is not, Dominance tells you not to pay. (from pages 115-6 of Joyce, *The Foundations of Causal Decision Theory*.)

We can put this in a table to make the dominance argument that Joyce suggests clearer.

	Broken Windshield	Unbroken Windshield
Pay extortion	-£410	-£10
Don't pay	-£400	0

In each column, the number in the 'Don't pay' row is higher than the number in the 'Pay extortion' row. So it looks just like the case above where we said dominance gives a clear answer about what to do. But the conclusion is crazy. Here is how Joyce explains what goes wrong in the dominance argument.

> Of course, this is absurd. Your choice has a direct influence on the state of the world; refusing to pay makes it likely that your windshield will be smashed while paying makes this unlikely. The extortionist is a despicable person, but he has you over a barrel and investing a mere £10 now saves £400 down the line. You should pay now (and alert the police later).

This seems like a general principle we should endorse. We should define *states* as being, intuitively, independent of choices. The idea behind tables such as the one given here is that the outcome should depend on two factors - what you do and what the world does. If the 'states' are dependent on what choice you make, then we won't have successfully 'factorised' the dependence of outcomes into these two components.

There is a famous historical, or at least literary, example of the odd consequences of ignoring this dependence restriction. In *Henry V*, Shakespeare gives the title character the following little speech. The context is that the English are about to go to battle with the French at Agincourt, and they are heavily outnumbered. The king's cousin Westmoreland has said that he wishes they had more troops, and Henry strongly disagrees.

> What's he that wishes so?
> My cousin Westmoreland? No, my fair cousin;
> If we are marked to die, we are enough
> To do our country loss; and if to live,
> The fewer men, the greater share of honor.
> God's will! I pray thee, wish not one man more.

This isn't very persuasive reasoning, though I guess it worked in the play! And the reason that it isn't persuasive is that it takes the states to be *win* and *lose*, but these are dependent on what we're considering, namely whether it would be better to have more troops. It seems that we need some kind of restriction on dependent states. But what do we mean here by 'dependent'? It turns out that there are two quite distinct ways we can think about independence that lead to strikingly different responses.

First, we might treat independence as **probabilistic independence**. On this way of thinking about decision theory, we should only include states such that the probability of being in any one state is the same no matter what choice we make.

Second, we might treat independence as **causal independence**. On this way of thinking about decision theory, we should only include states such that which state we're in isn't caused (in whole or in part) by our decision.

Either strategy will be enough to deflect Joyce's argument, and King Henry's for that matter. If we don't pay, the probability of being in the broken windshield state is higher than if we do. And if we pay, this might cause us to be in one state (the unbroken windshield state) rather than another. So cases like Joyce's, or King Henry's, don't distinguish which independence criteria is appropriate. Fortunately, there are some cases that tease these two notions apart.

Newcomb's Problem

The most famous such case in philosophy is Newcomb's Problem. It was introduced to philosophers by Robert Nozick, who in turn credited it to William Newcomb, a physicist at Livermore. Here's how the puzzle is often presented. In front of you are two boxes, call them A and B. You call see that in box B there is £1000, but you cannot see what is in box A. You have a choice, but not perhaps the one you were expecting. Your first option is to take just box A, whose contents you do not know. Your other option is to take both box A and box B, with the extra £1000.

There is, as you may have guessed, a catch. A demon has predicted whether you will take just one box or take two boxes. The demon is very good at predicting these things – in the past she has made many similar predictions and been right every time. If the demon predicts that you will take both boxes, then she's put nothing in box A. If the demon predicts you will take just one box, she has put £1,000,000 in box A. So the table looks like this.

	Predicts 1 box	Predicts 2 boxes
Take 1 box	£1,000,000	£0
Take 2 boxes	£1,001,000	£1,000

Crucially, the demon has made a *prediction*. So there is no causal link between what choice is made, and what 'state' we're in. (The scare quotes there are because it isn't obvious that we should treat the demon's predictions as states, any more than King Henry should have treated winning and losing as states.) But there is a probabilistic connection. If you take 1 box, that's excellent evidence that the demon will have predicted you'll have taken one box. So it's more probable that the demon will have predicted one box if you selected one box, than if you selected two boxes.

Put more simply, the demon's predictions are causally, but not probabilistically, independent of your choices.

And that matters, because if the table above is a good representation of the decision problem, the obviously correct thing to do is to take both boxes. After all, taking both boxes does better no matter which state you are in. That is, taking both boxes does better no matter which state you are in, if the demon's predictions are genuinely states in the relevant sense.

Before going on, it's worth thinking about how *you* might respond to Newcomb's Problem, and why. In the few section, we'll look at some philosophically important responses.

Newcomb as a Game

Very occasionally, you hear people say that Newcomb's Problem relies on the idea that the demon is perfect, and since thisis impossible, the puzzle is dissolved. Really, neither of these claims is true. The puzzle arises even if the demon is very good. (As we'll see in a bit, the puzzle may arise if the demon is even marginally better than chance.) And it isn't obviously true that a perfect demon is impossible, or even incompatible with a free agent choosing. If determinism is true, and compatibilism is the true theory of free will, then the demon might be perfect, even if the agent is a free chooser.

Indeed, we can see the basic structure of the puzzle by considering Newcomb's Puzzle simply as a game. Here it is, with R as the agent, and C as the demon, and 1 util being £1000.

Game 11	l	r
U	1, 1	1001, 0
D	0, 0	1000, 1

You might think that as long as C is pretty good at figuring out what rational R will do, which game theorists will often assume is true, then the demon can be assumed do well at predicting the agent's move. But we have to be careful here. The demon is supposed to do two rather distinct things well.

- In the vast majority of cases where the agent takes both boxes, the demon predicts this.
- In the vast majority of cases where the agent takes one box, the demon predicts this.

The assumption that the demon is good in general at predicting what agents will do does not entail the conjunction of these claims. Imagine that the following hypotheses are all true. Rationality requires taking both boxes (since this looks like a dominating option). The demon assumes that most people are rational, so always predicts that they'll take both boxes. And, indeed, most people are rational and take both boxes. Then the demon will usually make accurate predictions, but the second claim isn't true. So it's important to be clear about what is being proposed about the demon, and how strong a claim it is. Still, I think it's a plausible claim.

When I was a graduate student back at Monash, I remember several discussions about how good a job any one of us could do as the demon. And I think it was pretty widely thought that if we had a few minutes to talk to the agent before predicting their decision, we could probably get to at least 80% accuracy, whether the agent was taking one box or two. That's enough to make the case intuitively problematic, and perfectly realistic. So the puzzle shouldn't be dismissed, as it often is by non-philosophers, as unrealistic.

Rule Following and Wealth

You might think that the following argument for taking one-box is compelling.

1. Most people who take one box end up rich, and most people who take two boxes don't.
2. It is better, at least in the context of this puzzle, to end up rich than not.

3. So you should do what the people who end up rich do.
4. So you should take one box.

The problem is that this argument over-generates. (I believe this point traces back to Allan Gibbard and William Harper, though I'm not sure about this.) The same argument implies that you should take just one box in the case where the demon makes the prediction, and then tells you what it is, before you choose. In other words, the argument implies that you should take one box in this game.

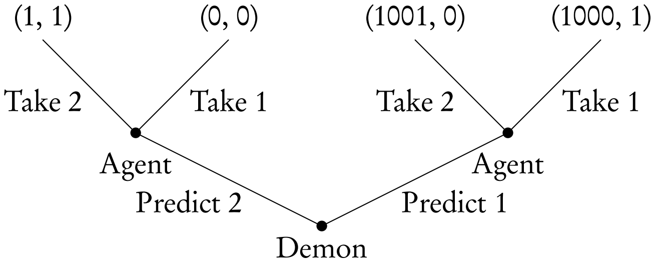


Figure 6.1: Transparent Newcomb's Problem

That seems remarkably implausible to me. So the argument that you should do what people who do well out of the game do also seems implausible. As David Lewis put it, this looks like a game where irrationality is rewarded. But that doesn't mean that one should become irrational; it just means the world has become a slightly odd place.

Evidential Decision Theory

Perhaps a better response to Newcomb like problems is evidential decision theory. Evidential decision theory says that we shouldn't take the probability of states as fixed. Rather, in our calculations of expected utility, we should use the probability of being in a state conditional on our having made a particular choice. So we should treat the value of a decision, which we'll write as $V(\varphi)$, as being given by the following rule, where the sum is over the members of some partition $\{p_1, \dots, p_n\}$

$$V(\varphi) = \sum_{i=1}^{n} \Pr(p_i|\varphi)V(\varphi \wedge p_i)$$

This obviously isn't a *reductive* definition; we assume that the value of each conjunction of proposition and action is given in advance.

Note that I didn't put any restrictions on the kind of partition mentioned. One of the nice features of this approach to decision problems is that any partition will do. Let's illustrate that by going through the Agincourt example. We'll make the following assumptions about value and probability. (In order to make the numbers easier to compare, I'll assume there is a 1/4 chance that the English will get more troops before the battle. This, like many other assumptions we're making about the case, is clearly unrealistic!)

First, the values of the four possible outcomes (to the English) are as follows:

- Winning with fewer troops is worth 11;
- Winning with more troops is worth 10;
- Losing with fewer troops is worth 1;
- Losing with more troops is worth 0

That matches up with King Henry's idea that whether you win or lose, it's better to do it with fewer troops. Now come the probabilities.

- The probability that the English will have fewer troops and win is 0.15.
- The probability that the English will have fewer troops and lose is 0.6.
- The probability that the English will have more troops and win is 0.15.
- The probability that the English will have more troops and lose is 0.1.

It's worth going through some conditional probabilities to get a cleaner sense of what these assumptions come to. I'll use 'win', 'lose', 'more' and 'fewer' with their obvious definitions from the English perspective, i.e, win or lose the battle, have more or fewer troops.

$$\begin{aligned}
\Pr(\text{win}|\text{more}) &= \frac{\Pr(\text{win} \wedge \text{more})}{\Pr(\text{more})} \\
&= \frac{\Pr(\text{win} \wedge \text{more})}{\Pr(\text{win} \wedge \text{more}) + \Pr(\text{lose} \wedge \text{more})} \\
&= \frac{0.15}{0.15 + 0.1} \\
&= 0.6
\end{aligned}$$

Since winning and losing are the only outcomes, it follows that Pr(lose|more) $= 0.4$. Now what happens if the English have fewer troops.

$$\begin{aligned}
\Pr(\text{win}|\text{fewer}) &= \frac{\Pr(\text{win} \wedge \text{fewer})}{\Pr(\text{fewer})} \\
&= \frac{\Pr(\text{win} \wedge \text{fewer})}{\Pr(\text{win} \wedge \text{fewer}) + \Pr(\text{lose} \wedge \text{fewer})} \\
&= \frac{0.15}{0.15 + 0.6} \\
&= 0.2
\end{aligned}$$

And again, since winning and losing are exclusive and exhaustive, it follows that $\Pr(\text{lose}|\text{fewer}) = 0.8$.

Given all this, let's work out the value of getting more troops. We'll use $\{\text{win}, \text{lose}\}$ as our partition.

$$\begin{aligned}
V(\text{more}) &= \Pr(\text{win}|\text{more})V(\text{win} \wedge \text{more}) + \Pr(\text{lose}|\text{more})V(\text{lose} \wedge \text{more}) \\
&= 0.6 \times 10 + 0.4 \times 0 \\
&= 6 \\
V(\text{fewer}) &= \Pr(\text{win}|\text{fewer})V(\text{win} \wedge \text{fewer}) + \Pr(\text{lose}|\text{fewer})V(\text{lose} \wedge \text{fewer}) \\
&= 0.2 \times 11 + 0.8 \times 1 \\
&= 3
\end{aligned}$$

And we get the intuitively correct response, that fighting with more troops is better than fighting with fewer troops.

Let's see what happens if we use a different partition. Let n be the proposition that the greater force will win. So it's equivalent to win $\leftrightarrow$ more. And now we'll redo the expected value calculation using $\{n, \neg n\}$ as the partition. I'll leave it as an exercise to figure out why it is true that $\Pr(n|\text{more}) = 0.6$, and $\Pr(n|\text{fewer}) = 0.8$.

$$\begin{aligned}
V(\text{more}) &= \Pr(n|\text{more})V(n \wedge \text{more}) + \Pr(\neg n|\text{more})V(\neg n \wedge \text{more}) \\
&= 0.6 \times 10 + 0.4 \times 0 \\
&= 6 \\
V(\text{fewer}) &= \Pr(n|\text{fewer})V(n \wedge \text{fewer}) + \Pr(\neg n|\text{fewer})V(\neg n \wedge \text{fewer}) \\
&= 0.8 \times 1 + 0.2 \times 11 \\
&= 3
\end{aligned}$$

And we get the same answer, for roughly the same reason. I won't go through the formal proof of this, but the general result is quite useful. If we're using this valuation function, it doesn't matter which partition of possibility space we use, we end up with the same value for each choice.

The decision theory that says we should make choices which maximise the value of this function V is known as *evidential decision theory*. It suggests that we should take one box in Newcomb's Problem, though we should take two boxes if we already know what the demon has predicted. The latter claim is, I hope, obvious. Here's the argument for the former, assuming that the demon is 99% reliable, whether the agent takes 1 or 2 boxes.

$$\begin{aligned}
V(\text{Take 2}) &= \text{Pr(Predict 2}|\text{Take 2)}V(\text{Predict 2} \wedge \text{Take 2})+ \\
&\quad \text{Pr(Predict 1}|\text{Take 2)}V(\text{Predict 1} \wedge \text{Take 2}) \\
&= 0.99 \times 1 + 0.01 \times 1001 \\
&= 11 \\
V(\text{Take 1}) &= \text{Pr(Predict 2}|\text{Take 1)}V(\text{Predict 2} \wedge \text{Take 1})+ \\
&\quad \text{Pr(Predict 1}|\text{Take 1)}V(\text{Predict 1} \wedge \text{Take 1}) \\
&= 0.01 \times 0 + 0.99 \times 1000 \\
&= 990
\end{aligned}$$

Since $990 > 11$, the theory says you should take just one box. Basically, it says that this case is just like the Agincourt case. Although you'll be better off taking one box whatever state you are in (win or lose in Agincourt, 1 or 2 boxes predicted in Newcomb), since there is a probabilistic connection between what you do and what state you're in, this isn't sufficient reason to act.

Many people find this an attractive account of Newcomb's Problem. But it seems to me (and to many others) that it wildly overgenerates predictions of when you should take actions that appear to be dominated. Consider the following fantasy story about the correlation between smoking and cancer.

> It's true that smokers get cancer at a higher rate than non-smokers. And that means that if you smoke, the probability that you'll get cancer will go up. That is, the *epistemic* probability that you will get cancer will go up; we now have more evidence that you will get cancern. But this isn't because smoking *causes* cancer. Rather, smoking and cancer have a common cause. The kind of people who are

> attracted to cigarettes are, unfortunately, disposed to get cancer. Perhaps this is because there is a gene that causes favourable attitudes towards cigarettes, and causes cancer.

Let's say that story is true. And let's say that our agent has a desire to smoke, but a stronger desire to avoid cancer. What should she do?

Well presumably she should smoke. It won't make her worse off. It will confirm that she is in the high risk pool for cancer, but if the story is true, then she is in that pool whether she smokes or not. So she may as well smoke.

This is hardly just an idiosyncratic view that people like me who think that you should take two boxes in Newcomb's Problem have. In the 1950s, some prominent scientists argued that the common cause story was plausibly true, or at least not ruled out by the evidence. (The most prominent such arguer was R. A. Fisher, the founder of modern statistics.) And the tobacco industry *loved* these claims, because they saw, rightly, that if people believed them they would keep smoking, while if people thought smoking caused cancer, they would stop.

But note that given Evidential Decision Theory, it doesn't seem to matter whether smoking causes cancer, or whether smoking and cancer merely have a common cause. Either way, smoking is evidence for cancer, just like taking both boxes is evidence that there's nothing in the opaque box. So either way, you shouldn't smoke. This is, most people think, wildly implausible, and that implausibility undermines Evidential Decision Theory.

There is, as you may imagine, more to be said here. One popular move made on behalf of Evidential Decision Theory at this point is the 'Tickle Defence'. It says that even if the story is true, smoking doesn't increase the probability of cancer *conditional on an observed desire to smoke*. There are critics of this move as well, and going into this would take us too far afield. Suffice to say that smoking cases cause problems for Evidential Decision Theorists, and many people think they cause unresolvable problems.

Causal Decision Theory

In response to these problems, a number of philosophers in the late 1970s and early 1980s developed **Causal Decision Theory**. The rough idea is that we start with a probability distribution over the causal structure of the world. Each element in this structure tells us what the causal consequences of particular choices will be. We don't know which of them is true; in some cases there may be a sense in which none of them are true *yet*. But we can assign a probability to each, and hence work out the probability of getting each outcome upon making each choice. We use that to generate an expected value of each choice, and then we maximise expected value.

There are disputes about how to implement this. I'm going to present the version due to Allan Gibbard and William Harper, which is the closest thing there is to an orthodox version of the theory. We'll use U for the utility of each choice, and write $A \mathrel{\Box\!\!\rightarrow} B$ for *If it were the case that A, it would be the case that B.*

$$U(\varphi) = \sum_{i=1}^{n} \Pr(\varphi \mathrel{\Box\!\!\rightarrow} p_i) U(\varphi \wedge p_i)$$

Again, this isn't a reductive account; we need utilities of more specific states to get the utilities of choices. But the hope is that those will be clear enough in any context that we can use the formula.

Let's see how this works in two cases, Newcomb's Problem and Joyce's car protection case.

In Newcomb's problem, it is natural to set the two causal hypotheses to be that the demon has put £1,000,000 in the opaque box, call that p_1, and that she has put nothing in the opaque box, call that p_2. Then since our choice doesn't make a causal difference to whether p_1 or p_2 is true, $\Pr(\varphi \mathrel{\Box\!\!\rightarrow} p_i)$ will just equal $Pr(p_i)$ no matter whether i is 1 or 2, or whatever choice φ is. The probability that there is a certain sum in the opaque box just is the probability that if we were to make a choice, there would be that amount in the opaque box. If we say that $\Pr(p_1) = x$, then we can work out the value of the two choices as follows.

$$\begin{aligned}
U(\text{Take 1 box}) &= \Pr(\text{Take 1 box} \mathrel{\Box\!\!\rightarrow} p_1) U(\text{Take 1 box} \wedge p_1) + \\
&\quad\ \Pr(\text{Take 1 box} \mathrel{\Box\!\!\rightarrow} p_2) U(\text{Take 1 box} \wedge p_2) \\
&= \Pr(p_1) U(\text{Take 1 box} \wedge p_1) + \Pr(p_2) U(\text{Take 1 box} \wedge p_2) \\
&= x \times U(\$1{,}000{,}000) + (1 - x) \times U(0) \\
U(\text{Take 2 boxes}) &= \Pr(\text{Take 2 boxes} \mathrel{\Box\!\!\rightarrow} p_1) U(\text{Take 2 boxes} \wedge p_1) + \\
&\quad\ \Pr(\text{Take 2 boxes} \mathrel{\Box\!\!\rightarrow} p_2) U(\text{Take 2 boxes} \wedge p_2) \\
&= \Pr(p_1) U(\text{Take 2 boxes} \wedge p_1) + \Pr(p_2) U(\text{Take 2 boxes} \wedge p_2) \\
&= x \times U(\$1{,}001{,}000) + (1 - x) \times U(\$1{,}000)
\end{aligned}$$

No matter what x is, as long as $U(£1{,}001{,}000) > U(£1{,}000{,}000)$, and $U(£1{,}000) > U(£0)$, then taking two boxes will be better than taking one box.

Now compare this to Joyce's example. As a reminder, here is the table of possible outcomes in that case.

	Broken Windshield	Unbroken Windshield
Pay extortion	-£410	-£10
Don't pay	-£400	0

Let's assume that a cost of £1 is worth -1 utils, and assume the following probabilities:

$$\begin{aligned}
\Pr(\text{Pay extortion} \,\Box\!\!\rightarrow \text{Broken Windshield}) &= 0.02\\
\Pr(\text{Pay extortion} \,\Box\!\!\rightarrow \text{Unbroken Windshield}) &= 0.98\\
\Pr(\text{Don't pay extortion} \,\Box\!\!\rightarrow \text{Broken Windshield}) &= 0.99\\
\Pr(\text{Don't pay extortion} \,\Box\!\!\rightarrow \text{Unbroken Windshield}) &= 0.01
\end{aligned}$$

Then we will have the following calculations of utilities.

$$\begin{aligned}
U(\text{Pay exortion}) &= 0.02 \times -410 + 0.98 \times -10\\
&= -8.2 + -9.8\\
&= -18\\
U(\text{Don't pay exortion}) &= 0.99 \times -400 + 0.01 \times 0\\
&= -396 + 0\\
&= -396
\end{aligned}$$

And clearly it is better to pay. That is, of course, the correct answer in this case.

As I mentioned, this isn't the only way to spell out Causal Decision Theory. David Lewis complains that the treatment of counterfactuals here is improper, for instance. But looking into those complications would take us too far afield. Instead we'll look at three other concerns one might have.

First, it's worth noting that our treatment of Newcomb's Problem left out a crucial fact. We proved that on a Causal Decision Theory, it would be better to take two boxes than one. But we proved that by proving a fact about comparative utilities, not by computing the actual utility of the two options. It is often non-trivial to do that. Once I've decided that I will take both boxes, presumably the probability that the demon will have put money in the opaque box will be low. But it's not clear that that's the probability I should be using. This might matter in some complicated variants of Newcomb's Problem. Let's say *S* has a choice between taking £2000 and playing Newcomb's Problem. And also say that *S* hasn't thought through what she would do were she to play Newcomb's Problem. Should she conclude that since she hasn't decided what she'll do, there is a substantial probability that she'll take only one box in Newcomb's Problem, so there is, even by Causal lights, a substantial probability that there will be money in the opaque box, so she should play rather than take the £2000? And if she should reason that way, is that a problem? That is, is it obvious that the ideas behind Causal Decision Theory should lead to taking the £2000 rather than playing? I'm not sure what to say about these cases, but I worry that there's a problem here.

Second, Causal Decision Theory leads to certain kinds of dilemmas. Consider the following story from Gibbard and Harper's paper introducing Causal Decision Theory.

> Consider the story of the man who met Death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, 'I AM COMING FOR YOU TOMORROW'. The terrified man that night bought a camel and rode to Aleppo. The next day, Death knocked on the door of the room where he was hiding, and said 'I HAVE COME FOR YOU'.
>
> 'But I thought you would be looking for me in Damascus', said the man.
>
> 'NOT AT ALL', said Death 'THAT IS WHY I WAS SURPRISED TO SEE YOU YESTERDAY. I KNEW THAT TODAY I WAS TO FIND YOU IN ALEPPO'.
>
> Now suppose the man knows the following. Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment

> with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...
>
> If... he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where Death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo.

If we note that Death has a preference for being wherever the man is, and the man has a preference for avoiding Death, then we can easily model this as a game. Indeed, we can model it as a familiar game, it is just Matching Pennies, i.e., Game 23. In this version of the game, Death plays Row and the man plays column, but we could easily reverse that.

Game 26	Damascus	Aleppo
Damascus	1, -1	-1, 1
Aleppo	-1, 1	1, -1

The response that Gibbard and Harper give to this game is that it is just a dilemma. Whatever the man plays, he'll regret it. Actually, that strikes me as the right thing to say, but not everyone is convinced. Reed Richter (in a 1984 *AJP* paper) argued that this was the wrong thing to say about *asymmetric* versions of the Death in Damascus case. Imagine that getting to Aleppo will cost a huge amount of money, and be incredibly painful. Then the table might look something like this:

Game 27	Damascus	Aleppo
Damascus	1, -1	-1, 0.5
Aleppo	-1, 1	1, -1.5

Again, whatever the man does, he will regret it, just like in the original Death in Damascus example. But it seems wrong to treat the two options available to the man symmetrically. After all, going to Aleppo is much worse for him. If forced to choose, some have argued, he should stay in Damascus.

Third, some people don't like Causal Decision Theory because it trucks in metaphysical notions like causation and counterfactuals. Richard Jeffrey was worried that Causal Decision Theory was too metaphysical, but he agreed that we should take both boxes in Newcomb's Problem. So he promoted a *Ratificationist* version of Evidential Decision Theory.

The idea behind ratificationism is that only *ratifiable* decisions are rationally allowed. A decision is ratifiable if it maximises expected utility conditional on that decision being taken. We can add a ratifiability clause to Evidential Decision Theory, as Jeffrey does, or to Causal Decision Theory, as (in effect) Frank Arntzenius has recently suggested.

If we add a ratifiability clause to Evidential Decision Theory, we get the result that rational agents should take both boxes. That's because only it is ratifiable. We computed earlier the expected utility of each choice according to Evidential Decision Theory, and concluded that the utility of taking just one box was higher. But now look what happens if we conditionalise on the hypothesis that we'll take just one box. (For simplicity, we'll again assume £1 is worth 1 util.) It is easy enough to see that taking both boxes is better.

$$\begin{aligned}
\text{Pr(Million in opaque box}|\text{Take one box)} &= 0.99 \therefore \\
V\text{(Take one box}|\text{Take one box)} &= 0.99 \times 1,000,000 + 0.01 \times 0 \\
&= 990,000 \\
V\text{(Take both box}|\text{Take one box)} &= 0.99 \times 1,001,000 + 0.01 \times 1,000 \\
&= 991,000
\end{aligned}$$

But there is something very odd about this way of putting things. It requires thinking about the expected value of an action conditional on something that entails the action is not taken. In Newcomb's Problem we can sort of make sense of this; we use the conditional assumption that we're taking one box to seed the probability of the demon doing certain actions, then we run the calculations from there. But I don't see any reason to think that we should, in general, be able to make sense of this notion.

A better approach, I think, is to mix ratificationism with Causal Decision Theory. (This isn't to say it's the right approach; I'm an old-fashioned Causal Decision Theorist. But it is *better*.) This lets us solve problems like the two-step Newcomb problem discussed earlier. Let's assume the demon is very very accurate; given that the player is choosing φ, the probability that the demon will predict φ is 0.9999. Now let's work through the values of taking each of the options. (Remember, the game is to take £2,000, or play Newcomb's Problem.

If the player does the latter, she must choose one box or two. And p_i is the proposition that the demon predicts that i boxes will be taken. We'll use T_i as shorthand for *Take i boxes*. And we'll assume, again that a pound is worth a util.)

$$\begin{aligned}
U(T_1|T_1) &= \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_1|T_1)U(T_1 \wedge p_1) + \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_2|T_1)U(T_1 \wedge p_2)\\
&= 0.9999 \times 1,000,000 + 0.0001 \times 0\\
&= 999,900\\
U(T_2|T_1) &= \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_1)U(T_2 \wedge p_1) + \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_2|T_1)U(T_2 \wedge p_2)\\
&= 0.9999 \times 1,001,000 + 0.0001 \times 1,000\\
&= 1,001,900\\
U(T_1|T_2) &= \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_1|T_2)U(T_1 \wedge p_1) + \Pr(T_1 \mathrel{\Box\!\!\rightarrow} p_2|T_2)U(T_1 \wedge p_2)\\
&= 0.0001 \times 1,000,000 + 0.9999 \times 0\\
&= 100\\
U(T_2|T_2) &= \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_2)U(T_2 \wedge p_1) + \Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_2|T_2)U(T_2 \wedge p_2)\\
&= 0.0001 \times 1,001,000 + 0.9999 \times 1,000\\
&= 1,100
\end{aligned}$$

The important thing to note about this calculation is that $\Pr(T_2 \mathrel{\Box\!\!\rightarrow} p_1|T_1)$ is very high, 0.9999 in our version of the game. What this says is that once we've assumed T_1, then the counterfactual $T_2 \mathrel{\Box\!\!\rightarrow} p_1$ is very very probable. That is, given that we're taking 1 box, it is very probable that if we had taken 2 boxes, there would still have been money in the opaque box. But that's just what Newcomb's problem suggests.

Note that neither T_1 nor T_2 is ratifiable. Given T_1, the player would be better with T_2. (The expected value of taking both boxes would be 1,001,900, as compared to an expected value of 999,900 for taking one box.) And given T_2, the player would be better with simply taking the £2,000 and not playing, since the expected payout of T_2 is a mere 1,100. But taking £2,000 is ratifiable. Given that the player is doing this, no other option is better. After all, if they are the kind of player who is moved by the reasoning that leads to taking the £2,000, then they are almost certainly two boxers, and so probably the opaque box would be empty. So the only ratifiable decision is to take £2,000. This ratificationist argument is, I think, intuitively plausible.

Note too that it is the very same conclusion that we reach through using a form of backward induction. Let's set up the game as a chart. We'll use P for the

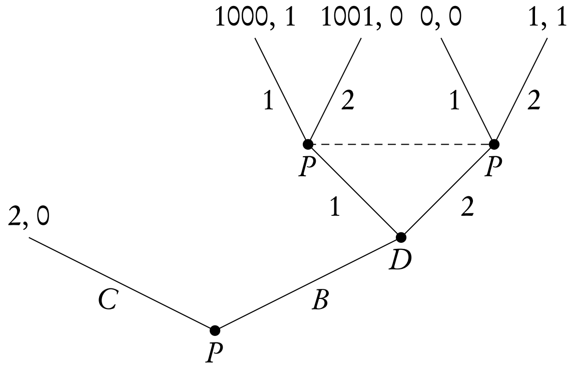


Figure 6.2: Game 28

player, and D for the demon. At move 1, P's choices are Cash or Box. After that, we'll use 1 and 2 for the obvious meanings; predictions for the demon, boxes for the players. The payoffs are player first, demon second. To make the numbers easier to read, we'll set 1 util as being worth £1,000. And we'll call the extended form of this **Game 28**.

Note crucially the dotted line between the player's choices in the top right. Although the player makes her choice *after* the demon does, the player doesn't know what the demon chooses. This makes backwards induction a little tricky; the demon can't assume that the player will make the best response to her choice, if the player doesn't know what the choice is.

But when we look at the numbers more closely, that worry dissipates. Whatever the demon does, the player is better off playing 2. So we should assume that the player will play 2. Given that the player will play 2, the demon is best off playing 2. And given that the demon is best off playing 2, the player at the initial node is best off playing C. So backwards induction leads to the same conclusion that ratificationist approaches do. And, as I said, this seems like an intuitive enough verdict, at least in this case.

Equilibrium and Ratification

A large number of fairly orthodox game theorists typically accept the following two propositions.

- It is strictly better to defect than co-operate in Prisoners' Dilemma, and in general one should always take strictly dominating options.
- In some games, such as Rock-Paper-Scissors, the best thing to do is to play a mixed strategy. In those games, playing a mixed strategy is preferably to playing any pure strategy.

In general, anyone who thinks there is something normatively significant in playing Nash Equilibrium strategies will have to accept something like those two claims. But if you look at the two biggest positions in philosophical decision theory, they are hard to reconcile. Evidential Decision Theorists may accept the second, but will reject the first. On Evidential Decision Theory, it may be best to accept a dominated option if, as in Newcomb's Problem, the states are evidentially connected to one's choice. And Causal Decision Theorists will accept the first, but not the second. On Causal Decision Theory, the expected value of a mixed strategy is just the (weighted) average value of the strategies being mixed, and the only rule is to maximise expected value, so the mixed strategy can't be preferable to each of the elements of the mixture.

Some of the tension is resolved if we add ratificationist decision theories to our menu of choices. In particular, a causal ratificationist might accept both of the bullet points above. It is obvious that they will accept the first. The co-operation option in Prisoners' Dilemma is both unratifiable and lower valued than the defection option. What's interesting is that, given one big extra assumption, they can accept the second as well.

The big extra assumption is that conditional on one's playing a strategy S, one should give probability 1 to the claim that the other player will do something that is in their best interests given that one is playing S. Let's apply that to Rock-Paper-Scissors. Conditional on playing Rock, one should give probability 1 to the proposition that the other player will play Paper. That is, one should give probability 0 to the proposition *If I were to play Rock, I would win*, while giving probability 1 to the proposition *If I were to play Scissors, I would win*. So conditional on playing Rock, the best thing to do, *from a causal perspective*, is to play Scissors.

So playing Rock is not ratifiable, and by similar reasoning, neither is playing Paper or Scissors. Does this mean nothing is ratifiable? Not at all; the mixed strategies might still be ratifiable. In particular, the mixed strategy where one plays each of the pure strategies with probability 1/3, is ratifiable. At least, it is ratifiable if we assume that causal ratificationism is the correct theory of rational choice. If one plays this mixed strategy, and the other player knows it, then every strategy the other player could play is equally valuable to them; they each have expected value 0. Given that each strategy is equally valuable, the other player could play any strategy that is rationally acceptable. Since we are assuming causal ratificationism, that means they could play any ratifiable strategy. But the only ratifiable strategy is the mixed strategy where one plays each of the pure strategies with probability 1/3. Conditional on the other player doing that, moving away from the mixed strategy has no advantages (though it also has no costs). So causal ratificationism, plus an assumption that the other player is an excellent mind reader, delivers the orthodox result.

There are other reasons to associate orthodoxy in game theory with causal ratificationism. Here, for instance, is Ken Binmore motivating the use of equilibrium concepts in a prominent game theory textbook (*Playing for Real*).

> Why should anyone care about Nash equilibria? There are at least two reasons. The first is that a game theory book can't authoratively point to a pair of strategies as the solution of a game unless it is a Nash equilibrium. Suppose, for example, that t weren't a best reply to s. [Player 2] would then reason that if [Player 1] follows the book's advice and plays s, then she would do better not to play t. But a book can't be authoritative on what is rational if rational people don't play as it predicts. (*Playing for Real*, 18-19)

It seems to me that this argument only makes sense if we assume some ratificationist theory of decision making. What Binmore is encouraging us to focus on here are strategies that are still rational strategies in situations where everyone believes that they are rational strategies. That's close, I think, to saying that we should only follow strategies that are best strategies conditional on being played.

There's a secondary argument Binmore is making here which I think is more misleading. The most the little argument he gives could show is that if a game has a unique solution, then that solution must be a Nash equilibrium. But it doesn't follow from that that there is anything special about Nash equilibrium as opposed, say, to strategies that are BRBRI. After all, if a game has a unique solution, each player will play a strategy that is BRBRI. And if each player has multiple BRBRI strategies, even if some of them are not part of any Nash equilibrium, it isn't clear why a book which said each BRBRI strategy was rational would be self-undermining. If I say that any pure or mixed strategy whatsoever could be rational in Rock-Paper-Scissors, and Player 1 believes me, and Player 2 knows this, Player 2 can't use that knowledge to undermine my advice.

But you might note that Binmore says there is a second reason. Here's what it is.

> Evolution provides a second reason why we should care about Nash equilibria. If the payoffs in a game correspond to how fit the players are, then adjustment processes that favour the more fit at the expense of the less fit will stop working when we get to a Nash equilibrium because all the survivors will then be as fit as it is possible to be in the circumstances. (*Playing for Real*, 19)

This seems to me like a very strong reason to care about Nash equilibrium in *repeated* games, like competition for food over time. The same is true in Rock-Paper-Scissors. It isn't true that Rock wins every time, and you shouldn't play Rock every time like Bart Simpson does. But that's because you'll give away your strategy. It doesn't show that a pure strategy of Rock is wrong in any given game of Rock-Paper-Scissors.

As we noted at the end of the last section, it is not clear that the standard representation of the payoffs in any given round of a repeated game are correct. Some strategies incur costs down the road that aren't properly reflected in the individual payoff matricies. But, as we also noticed, it is hard to do much about this. The best thing might be to just note that the payoffs aren't quite right, and look for equilibrium with respect to the not quite right payoffs.

In any case, we're going to want to come back to what Causal Decision Theory says about repeated games somewhat later, so it is best to set that aside for now. What I really have wanted to argue here was that ratificationism is necessary to motivate some of the core ideas of game theory.

So is ratificationism correct? I have my doubts, but I think it would be too much of a digression to go into them here. Instead, I want to focus on some of the interesting cases where we have to hunt a little for equilibrium, or where there are interesting factors that determine which equilibrium is chosen.

Finding and Choosing Equilibria

Finding Mixed Strategy Equilibria

Let's consider again the asymmetric version of Death in Damascus.

Game 27	Damascus	Aleppo
Damascus	1, -1	-1, 0.5
Aleppo	-1, 1	1, -1.5

Recall that Death is the Row player, and the Man is the Column player. The way we generated Game 27 was to take the basic structure of the game, which is Matching Pennies, and add in an 0.5 penalty for Man choosing Aleppo. It's an unpleasant journey from Damascus to Aleppo, particularly if you fear Death is at the other end.

There is still no pure strategy equilibrium in this game. Whatever Death plays, Man would prefer to play the other. And whatever Man plays, Death wants to play it. So there couldn't be a set of pure choices that they would both be happy with given that they know the other's play.

But the mixed strategy equilibrium that we looked at for Matching Pennies isn't an equilibrium either. We'll write $\langle x, y \rangle$ for the mixed strategy of going to Damascus with probability x, and going to Aleppo with probability y. Clearly we should have $x + y = 1$, but it will make the representation easier to use two variables here, rather than just writing $\langle x, 1 - x \rangle$ for the mixed strategies.

Given that representation, we can ask whether the state where each player plays $\langle 1/2, 1/2 \rangle$ is a Nash equilibrium. And, as you might guess, it is not. You might have guessed this because the game is not symmetric, so it would be odd if the equilibrium solution to the game is symmetric. But let's prove that it isn't an equilibrium. Assume that Death plays $\langle 1/2, 1/2 \rangle$. Then Man's expected return from staying in Damascus is:

$$1/2 \times -1 + 1/2 \times 1 = 0$$

while his return from going to Aleppo is

$$1/2 \times 0.5 + 1/2 \times -1.5 = -0.5$$

So if Death plays $\langle 1/2, 1/2\rangle$, Man is better off staying in Damascus than going to Aleppo. And if he's better off staying in Damascus that going to Aleppo, he's also better off staying in Damascus than playing some mixed strategy that gives some probability of going to Aleppo. In fact, the strategy $\langle x, y\rangle$ will have expected return $-y/2$, which is clearly worse than 0 when $y > 0$.

There's a general point here. The expected return of a mixed strategy is the weighted average of the returns of the pure strategies that make up the mixed strategy. In this example, for instance, if the expected value of staying in Damascus is d, and the expected value of going to Aleppo is a, the mixed strategy $\langle x, y\rangle$ will have expected value $xd + ya$. And since $x + y = 1$, the value of that will be strictly between a and d if $a \neq d$. On the other hand, if $a = d$, then $x + y = 1$ entails that $xd + ya = a = d$. So if $a = d$, then any mixed strategy will be just as good as any other, or indeed as either of the pure strategies. That implies that mixed strategies are candidates to be equilibrium points, since there is nothing to be gained by moving away from them.

This leads to an immediate, though somewhat counterintuitive, conclusion. Let's say we want to find strategies $\langle x_D, y_D\rangle$ for Death and $\langle x_M, y_M\rangle$ for Man that are in equilibrium. If the strategies are in equilibrium, then neither party can gain by moving away from them. And we just showed that that means that the expected return of Damascus must equal the expected return of Aleppo. So to find $\langle x_D, y_D\rangle$, we need to find values for x_D and y_D such that, given Man's values, staying in Damascus and leaving for Aleppo are equally valued. Note, and this is the slightly counterintuitive part, we don't need to look at *Death's* values. All that matters is that Death's strategy and Man's values together entail that the two options open to Man are equally valuable.

Given that Death is playing $\langle x_D, y_D\rangle$, we can work out the expected utility of Man's options fairly easily. (We'll occasionally appeal to the fact that $x_D + y_D = 1$.)

$$\begin{aligned}
U(\text{Damascus}) &= x_D \times -1 + y_D \times 1 \\
&= y_D - x_D \\
&= 1 - 2x_D \\
U(\text{Aleppo}) &= x_D \times 0.5 + y_D \times -1.5 \\
&= 0.5x_D - 1.5(1 - x_D) \\
&= 2x_D - 1.5
\end{aligned}$$

So there is equilibrium when $1 - 2x_D = 2x_D - 1.5$, i.e., when $x_D = {}^5/_8$. So any mixed strategy equilibrium will have to have Death playing $\langle {}^5/_8, {}^3/_8 \rangle$.

Now let's do the same calculation for Man's strategy. Given that Man is playing $\langle x_D, y_D \rangle$, we can work out the expected utility of Death's options. (Again, we'll occasionally appeal to the fact that $x_M + y_M = 1$.)

$$\begin{aligned}
U(\text{Damascus}) &= x_M \times 1 + y_M \times -1 \\
&= x_M - y_M \\
&= 2x_M - 1 \\
U(\text{Aleppo}) &= x_M \times -1 + y_M \times 1 \\
&= y_M - x_M \\
&= 1 - 2x_M
\end{aligned}$$

So there is equilibrium when $2x_M - 1 = 1 - 2x_M$, i.e., when $x_M = {}^1/_2$. So any mixed strategy equilibrium will have to have Man playing $\langle {}^1/_2, {}^1/_2 \rangle$. Indeed, we can work out that if Death plays $\langle {}^5/_8, {}^3/_8 \rangle$, and Man plays $\langle {}^1/_2, {}^1/_2 \rangle$, then any strategy for Death will have expected return 0, and any strategy for Man will have expected return of $-{}^1/_4$. So this pair is an equilibrium.

But note something very odd about what we just concluded. When we changed the payoffs for the two cities, we made it worse for *Man* to go to Aleppo. Intuitively, that should make Man more likely to stay in Damascus. But it turns out this isn't right, at least if the players play equilibrium strategies. The change to Man's payoffs doesn't change Man's strategy at all; he still plays $\langle {}^1/_2, {}^1/_2 \rangle$. What it does is change Death's strategy from $\langle {}^1/_2, {}^1/_2 \rangle$ to $\langle {}^5/_8, {}^3/_8 \rangle$.

Let's generalise this to a general recipe for finding equilibrium strategies in two player games with conflicting incentives. Assume we have the following very abstract form of a game:

Game 29	l	r
U	a_1, a_2	b_1, b_2
D	c_1, c_2	d_1, d_2

As usual, *R*ow chooses between *U*p and *D*own, while *C*olumn chooses between *l*eft and *r*ight. We will assume that R prefers the outcome to be on the northwest-southeast diagonal; that is, $a_1 > c_1$, and $d_1 > b_1$. And we'll assume that C prefers the other diagonal; that is, $c_2 > a_2$, and $b_2 > d_2$. We then have to find a pair of mixed strategies $\langle x_U, x_D \rangle$ and $\langle x_l, x_r \rangle$ that are in equilibrium. (We'll use x_A for the probability of playing A.)

What's crucial is that for each player, the expected value of each option is equal given what the other person plays. Let's compute them the expected value of playing U and D, given that C is playing $\langle x_l, x_r \rangle$.

$$\begin{aligned} U(U) &= x_l a_1 + x_r b_1 \\ U(D) &= x_l c_1 + x_r d_1 \end{aligned}$$

We get equilibrium when these two values are equal, and $x_l + x_r = 1$. So we can solve for x_l the following way:

$$\begin{aligned} & x_l a_1 + x_r b_1 = x_l c_1 + x_r d_1 \\ \Leftrightarrow\ & x_l a_1 - x_l c_1 = x_r d_1 - x_r b_1 \\ \Leftrightarrow\ & x_l (a_1 - c_1) = x_r (d_1 - b_1) \\ \Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} = x_r \\ \Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} = 1 - x_l \\ \Leftrightarrow\ & x_l \frac{a_1 - c_1}{d_1 - b_1} + x_l = 1 \\ \Leftrightarrow\ & x_l (\frac{a_1 - c_1}{d_1 - b_1} + 1) = 1 \\ \Leftrightarrow\ & x_l = \frac{1}{\frac{a_1 - c_1}{d_1 - b_1} + 1} \end{aligned}$$

I won't go through all the same steps, but a similar argument shows that

$$x_U = \frac{1}{\frac{b_2 - a_2}{c_2 - d_2} + 1}$$

I'll leave it as an exercise to confirm these answers are correct by working out the expected return of U, D, l and r if these strategies are played.

The crucial take-away lesson from this discussion is that to find a mixed strategy equilibrium, we look at the interaction between one player's mixture and the other player's payoffs. The idea is to set the probability for each move in such a way that even if the other player knew this, they wouldn't be able to improve their position, since any move would be just as good for them as any other.

I've focussed on the case of a game where each player has just two moves. When there are more than two moves available, things are a little more complicated, but only a little. We no longer need it to be the case that one player's mixed strategy must make *every* other strategy the other player has equally valuable. It only has to make every strategy that is part of the other player's mixture equally valuable. Consider, for instance, what happens if we expand our asymmetric Death in Damascus game to give Man the option of shooting himself.

Game 30	Damascus	Aleppo	Shoot
Damascus	1, -1	-1, 0.5	-1, -2
Aleppo	-1, 1	1, -1.5	-1, -2

The shooting option is no good for anyone; Death doesn't get to meet Man, and Man doesn't get to live the extra day. So if Death plays $\langle 5/8, 3/8 \rangle$, that will make Damascus and Aleppo equally valuable to Man, but Shooting will still have an expected return of -2, rather than the expected return of $-1/4$ that Damascus and Aleppo have. But that's consistent with Death's strategy being part of an equilibrium, since Man's strategy will be to play $\langle 1/2, 1/2, 0 \rangle$. Since Man isn't playing Shoot, it doesn't matter that Shoot is less valuable to him, given Death's move, than the two pure strategies.

Coordination Games

Many games have multiple equilibria. In such cases, it is interesting to work through the means by which one equilibrium rather than another may end up being chosen. Consider, for instance, the following three games.

Game 31	*a*	*b*
A	1, 1	0, 0
B	0, 0	1, 1

Game 32	*a*	*b*
A	2, 1	0, 0
B	0, 0	1, 2

Game 33	*a*	*b*
A	5, 5	0, 4
B	4, 0	2, 2

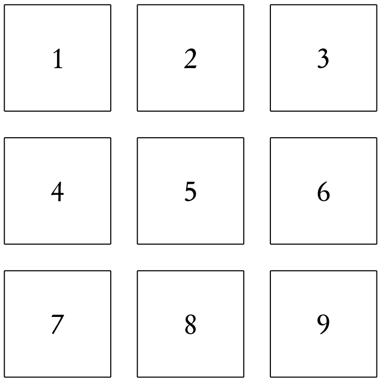
In each case, both $\langle A, a\rangle$ and $\langle B, b\rangle$ are Nash equilibria.

Game 31 is a purely cooperative game; the players have exactly the same payoffs in every outcome. It is a model for some real-world games. The two players, R and C, have to meet up, and it doesn't matter where. They could either go to location A or B. If they go to the same location, they are happy; if not, not.

In such an abstract presentation of game, it is hard to see how we could select an equilibrium out of the two possibilities. In practice, it often turns out not to be so hard. Thomas Schelling (in *The Strategy of Conflict*) noted that a lot of real-life versions of this game have what he called 'focal points'. (These are sometimes called Schelling points now, in his honour.) A focal point is a point that stands out from the other options in some salient respect, and it can be expected that other players will notice that it stands out. Here's a nice example from a recent paper by Christopher Potts (Interpretive Economy, Schelling Points, and evolutionary stability).

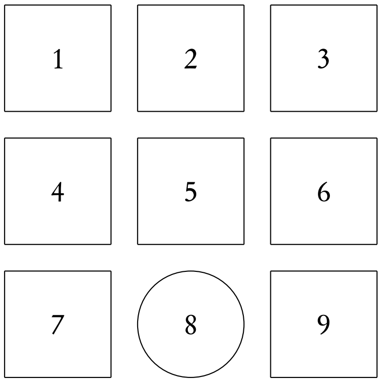
Game 34

A and *B* have to select, without communicating, one of the following nine figures. They each get a reward iff they select the same figure.



Game 35

A and *B* have to select, without communicating, one of the following nine figures. They each get a reward iff they select the same figure.



We could run experiments to test this, but intuitively, players will do better at Game 35 than at Game 34. That's because in Game 35, they have a focal point to select; one of the options stands out from the crowd.

Schelling tested this by asking people where they would go if they had to meet a stranger in a strange city, and had no way to communicate. The answers suggested that meetups would be much more common than you might expect. People in Paris would go to the Eiffel Tower; people in New York would go to (the information booth at) Grand Central Station, and so on. (This game may be easier if you don't actually live in the city in question. I live in New York and go to Grand Central about once a year; it would be far from the most obvious place I would select.)

Game 32 is often called 'Battle of the Sexes'. The real world example of it that is usually used is that *R* and *C* are a married couple trying to coordinate on a night out. But for whatever reason, they are at the stage where they simply have to go to one of two venues. *R* would prefer that they both go to *A*, while *C* would prefer that they both go to *B*. But the worst case scenario is that they go to different things.

Game 33 is often called 'Stag Hunt'. The idea behind the game goes back to Rousseau, although you can find similar examples in Hume. The general idea is that *A* is some kind of cooperative action, and *B* is a 'defecting' action. If players make the same selection, then they do well. If they both cooperate, they do better than if they both defect. But if they don't cooperate, it is better to defect than to cooperate.

Rousseau's example was of a group of hunters out hunting a stag, each of whom sees a hare going by. If they leave the group, they will catch the hare for sure, and have something to eat, but the group hunting the stag will be left with a much lower probability of stag catching. Hume was more explicit than Rousseau that these games come up with both two player and many player versions. Here is a nice many player version (from Ben Polak's OpenYale lectures).

> **Game 36**
>
> Everyone in the group has a choice between Investing and Not Investing. The payoff to anyone who doesn't invest is 0. If 90% or more of the group Invests, then everyone who Invests gains £1. If less than 90% of the group Invests, then everyone who invests loses £1.

Again, there are two Nash equilibria: everyone invests, and everyone does not invest. Hume's hypothesis was that in games like this, the cooperative move (in this case Investing) would be more likely in small groups than large groups.

Mixed Strategies in Multiple Equilibrium Games

All three of these games have mixed strategy equilibria as well as their pure strategy equilibria. For Game 31, the mixed strategy equilibrium is easy to find. Each player plays $\langle 1/2, 1/2 \rangle$. In that case the other player has an expected return of $1/2$ whatever they play, so it is an equilibrium.

Things are a little trickier in Game 32. To find the mixed strategy equilibrium there, we have to apply the lesson we learned earlier: find strategies that make the other player indifferent between their options. If R plays $\langle x, 1-x \rangle$, then C's expected return from playing a and b is:

$$\begin{aligned} U(a) &= x \times 1 + (1-x) \times 0 \\ &= x \\ U(b) &= x \times 0 + (1-x) \times 2 \\ &= 2(1-x) \end{aligned}$$

So there will be equilibrium only if $x = 2(1-x)$, i.e., only if $x = 2/3$. If that happens, C's expected return from any strategy will be $2/3$. A similar argument shows that we have equilibrium when C plays $\langle 1/3, 2/3 \rangle$, and in that case R's expected return from any strategy will be $2/3$. (Unlike in the asymmetric version of

Death in Damascus, here the equilibrium strategy is to bias one's mixture towards the result that one wants.)

We apply the same methodology in Game 33. So R will play $\langle x, 1-x\rangle$, and C's expected return from the two available pure strategies is the same. Those expected returns are:

$$\begin{aligned} U(a) &= x \times 5 + (1-x) \times 0 \\ &= 5x \\ U(b) &= x \times 4 + (1-x) \times 2 \\ &= 4x + 2 - 2x \\ &= 2 + 2x \end{aligned}$$

So there will be equilibrium only if $5x = 2 + 2x$, i.e., only if $x = 2/3$. If that happens, C's expected return from any strategy will be $10/3$. Since the game is completely symmetric, a very similar argument shows that if C plays $\langle 2/3, 1/3\rangle$, then R has the same payoff whatever she plays. So each player cooperating with probability $2/3$ is a Nash equilibrium.

In two of the cases, the mixed strategy equilibrium is worse for each player than the available pure strategy equilibria. In Game 33, the mixed equilibrium is better than the defecting equilibrium, but worse than the cooperating equilibrium. Say an equilibrium is **Pareto-preferred** to another equilibrium iff every player would prefer the first equilibrium to the second. An equilibrium is **Pareto-dominant** iff it is Pareto-preferred to all other equilibria. The cooperative equilibrium is Pareto-dominant in Game 33; neither of the other games have Pareto-dominant equiilbria.

Consider each of the above games from the perspective of a player who does not know what strategy their partner will play. Given a probability distribution over the other player's moves, we can work out which strategy has the higher expected return, or whether the various strategies have the same expected returns as each other. For any strategy S, and possible strategy S' of the other player, let f be a function that maps S into the set of probabilities x such that if the other player plays S' with probability x, S has the highest expected return. In two strategy games, we can remove the relativisation to S', since for the purposes we'll go on to, it doesn't matter which S' we use. In somewhat formal language,

$f(S)$ is the **basin of attraction** for S; it is the range of probability functions that points towards S being played.

Let m be the usual (Lesbegue) measure over intervals; all that you need to know about this measure is that for any interval $[x,y]$, $m([x,y]) = y - x$; i.e., it maps a continuous interval onto its length. Say $m(f(S))$ is the risk-attractiveness of S. Intuitively, this is a measure of how big a range of probability distributions over the other person's play is compatible with S being the best thing to play.

Say that a strategy S is risk-preferred iff $m(f(S))$ is larger than $m(f(S^*))$ for any alternative strategy S^* available to that agent. Say that an equilibrium is **risk-dominant** iff it consists of risk-preferred strategies for all players.

For simple two-player, two-option games like we've been looking at, all of this can be simplified a lot. An equilibrium is risk-dominant iff each of the moves in it are moves the players would make if they assigned probability 1/2 to each of the possible moves the other player could make.

Neither Game 31 nor Game 32 have a risk-dominant equilibrium. An asymmetric version of Game 31, such as this game, does.

Game 37	*a*	*b*
A	2, 2	0, 0
B	0, 0	1, 1

In this case $\langle A, a\rangle$ is both Pareto-dominant and risk-dominant. Given that, we might expect that both players would choose A. (What if an option is Pareto-dominated, and risk-dominated, but focal? Should it be chosen then? Perhaps; the Eiffel Tower isn't easy to get to, or particularly pleasant once you're there, but seems to be chosen in the Paris version of Schelling's meetup game because of its focality.)

The real interest of risk-dominance comes from Game 33, the Stag Hunt. In that game, cooperating is Pareto-dominant, but defecting is risk-dominant. The general class of Stag Hunt games is sometimes defined as the class of two-player, two-option games with two equilibria, one of which is Pareto-dominant, and the other of which is risk-dominant. By that definition, the most extreme Stag Hunt is a game we discussed earlier in the context of deleting weakly dominated strategies.

Game 14	l	r
T	1, 1	100, 0
B	0, 100	100, 100

The $\langle B, r\rangle$ equilibrium is obviously Pareto-dominant. But the $\langle T, l\rangle$ is *extremely* risk-dominant. Any probability distribution at all over what the other player might do, except for the distribution that assigns probability 1 to B/r, makes it better to play one's part of $\langle T, l\rangle$ than one's part of $\langle B, r\rangle$.

I won't go through all the variations here, in large part because I don't understand them all, but there are a number of ways of modifying the Stag Hunt so that risk-dominant strategies are preferred. Some of these are evolutionary; given certain rules about what kinds of mutations are possible, risk-dominant strategies will invariable evolve more successfully than Pareto-dominant strategies. And some of these involve uncertainty; given some uncertainty about the payoffs available to other players, risk-dominant strategies may be uniquely rational. But going the details of these results is beyond the scope of these notes.

Value of Communication

In all the games we've discussed to date, we have assumed that the players are not able to communicate before making their choices. Or, equivalently, we've assumed that the payoff structure is what it is after communication has taken place. If we relax that assumption, we need to think a bit about the kind of speech acts that can happen in communication.

Imagine that R and C are playing a Prisoners' Dilemma. There are two importantly different types of communication that might take place before play. First, they might promise really sincerely that they won't defect. Second, they might come to some arrangement whereby the person who defects will incur some costs. This could be by signing a contract promising to pay the other in the event of defection. Or it could be by one player making a plausible threat to punish the other for defection.

The second kind of communication can change the kind of game the players are playing. The first kind does not, at least not if the players do not regard promise breaking as a bad thing. That's because the second kind of communication, but not the first, can change the payoff structure the players face. If R and C each have to pay in the event of defecting, it might be that defecting no longer dominates cooperating, so the game is not really a Prisoners' Dilemma. But if they merely say that they will cooperate, and there is no cost to breaking their word, then the game still is a Prisoners' Dilemma.

Call any communication that does not change the payoff matrix **cheap talk**. In Prisoners' Dilemma, cheap talk seems to be useless. If the players are both rational, they will still both defect.

But it isn't always the case that cheap talk is useless. In a pure coordination game, like Game 34, cheap talk can be very useful. If a player says that they will play 7, then each player can be motivated to play 7 even if they have no interest in honesty. More precisely, assume that the hearer initially thinks that the speaker is just as likely to be lying as telling the truth when she says something about what she will do. So before she thinks too hard about it, she gives credence 0.5 to the speaker actually playing 7 when she says she will do so. But if there's a 50% chance the speaker will play 7, then it seems better to play 7 than anything else. In the absence of other information, the chance that the hearer will win when playing some number other than 7 will be much less than 50%; around 6% if she has equal credence in the speaker playing each of the other options. So the hearer should play 7. But if the speaker can reason through this, then she will play 7 as well. So her statement will be self-enforcing; by making the statement she gives herself a reason to make it true, even beyond any intrinsic value she assigns to being honest.

There is one step in that reasoning that goes by particularly fast. Just because we think it is in general just as likely that a speaker is lying as telling the truth, doesn't mean that we should think those things are equally likely *on this occasion*. If the speaker has a particular incentive to lie on this occasion, the fact that they are a truth-teller half the time is irrelevant. But in Game 34, they have no such incentive. In fact, they have an incentive to tell the truth, since truth-telling is the natural way to a good outcome for them in the game.

But this consideration is a problem in Battle of the Sexes. Assume that *R* says, "I'm going to play *A*, whatever you say you'll do." If *C* believes this, then she has a reason to play *a*, and that means *R* has a reason to do what she says. So you might think that Game 32 is like Game 34 as a game in which cheap talk makes a difference. But in fact the reasoning that we used in Game 34 breaks down a little. Here *R* has an incentive to make this speech independently of what they are planning to do. Unless we think *R* has a general policy of truth-telling, it seems speeches like this should be discounted, since *R*'s incentive to talk this way is independent of how the plan to play. And if R has a general policy of truth-telling, a policy they regard it as costly to break, this isn't really a case of cheap talk.

The same analysis seems to apply with even greater force in Game 33. There, R wants C to play a, whatever R is planning on playing. So she wants to give C a reason to play a. And saying that she'll play A would be, if believed, such a reason. So it seems we have a simpler explanation for why R says what she says, independent of what R plans to do. So I suspect that in both Game 32 and Game 33, this kind of cheap talk (i.e., solemn declarations of what one will play) is worthless.

But that's not all we can do when we communicate. We can also introduce new options. (Just whether this should be called genuinely cheap talk is perhaps a little dubious, but it seems to me to fall under the same general heading.) Assume that we modify Game 32 by allowing the two players to see the result of a fair coin flip. We introduce a new option for them each to play, which is to do A if the coin lands heads, and B if the coin lands tails. Call this option Z. The new game table looks like this. (Note that many of the payoffs are *expected* payoffs, not guarantees.)

Game 38	a	b	z
A	2, 1	0, 0	1, 0.5
B	0, 0	1, 2	0.5, 1
Z	1, 0.5	0.5, 1	1.5, 1.5

Note that $\langle Z, z\rangle$ is an equilibrium of the game. Indeed, it is a better equilibrium by far than the mixed strategy equilibrium that left each player with an expected return of 2/3. Not surprisingly, this kind of result is one that players with a chance to communicate often end up at.

Assume that R thinks that C will play A, B, z with probabilities $x, y, 1-x-y$. Then R's expected returns for her three strategies are:

$$\begin{aligned}
U(A) &= 2x + 0y + 1(1-x-y) \\
&= 1 + x - y \\
U(B) &= 0x + 1y + 0.5(1-x-y) \\
&= 0.5 - 0.5x + 0.5y \\
U(Z) &= 1x + 0.5y + 1.5(1-x-y) \\
&= 1.5 - 0.5x - y
\end{aligned}$$

A little algebra gives us the following inequalities.

$$U(A) > U(Z) \Leftrightarrow 3x > 1$$
$$U(A) > U(B) \Leftrightarrow x > 3y - 1$$
$$U(B) > U(Z) \Leftrightarrow 3y > 2$$

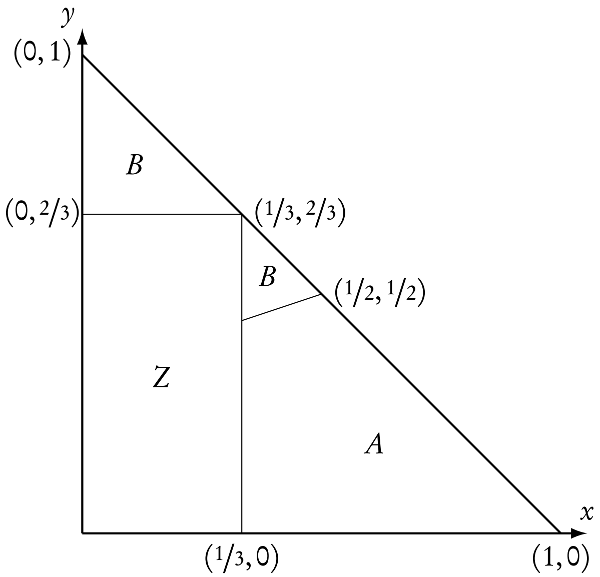
Putting these together we get the following results:

A is best iff $x > 1/3$ and $x > 3y - 1$

B is best iff $y > 2/3$ or ($x > 1/3$ and $3y - 1 > x$)

Z is best iff $x < 1/3$ and $y < 2/3$

Here is a graph showing where each of the three options has the highest utility.



A little geometry reveals that the area of the large rectangle where Z is best is $2/9$, the two triangles where B is best have area $1/18$ amd $1/54$ each, summing to $2/27$, and the remaining area in the triangle, the odd-shaped area where A is best, is therefore $1/6$. In other words, the largest region is the one where X is best. And that means, by definition, that Z is risk-preferred for R. A similar computation shows that z is risk-preferred for C. So we get the result that the newly available equilibrium is a risk-dominant equilibrium.

I'll leave the option of extending this analysis to Stag Hunt as an exercise for the interested reader.

Refining Nash Equilibria

In many games there are multiple Nash equilibria. Some of these equilibria seem highly unreasonable. For instance, the mixed equilibria in Battle of the Sexes has a lower payout to both players than either of the pure equilibria. In some coordination games, as we saw earlier, there are very poor outcomes that are still equilibria. For instance, in this game both *Aa* and *Bb* are equilibria, but we would expect players to play *Aa*, not *Bb*.

Game 39	a	b
A	1000, 1000	0, 0
B	0, 0	1, 1

Considerations such as these have pushed many game theorists to develop **refinements** of the concept of Nash equilibrium. A refinement is an equilibrium concept that is stronger than Nash equilibrium. We have already seen a couple of refinements of Nash equilibrium in the discussion of coordination games. Here are two such refinements. (The first name is attested in the literature, the second I made up because I couldn't find a standard name.)

Pareto Nash equilibrium is a Nash equilibrium that Pareto dominates every other Nash equilibrium.

Risk Nash equilibrium is a Nash equilibrium that risk dominates every other equilibrium.

We could also consider weak versions of these concepts. A weak Pareto Nash equilibrium is a Nash equilibrium that is not Pareto-dominated by any other equilibrium, and a weak risk Nash equilibrium is a Nash equilibrium that is not risk dominated by any other Nash equilibrium. Since we've already spent a fair amount of time on these concepts when we discussed coordination games, I won't go on more about them here.

We can also generate refinements of Nash equilibrium by conjoining dominance conditions to the definition of Nash equilibrium. For instance, the following two definitions are of concepts strictly stronger than Nash equilibrium.

Nash + Weak Dominance A Nash equilibrium that is not weakly dominated by another strategy.

Nash + Iterated Weak Dominance A Nash equilibrium that is not deleted by (some process of) iterative deletion of weakly dominated strategies.

In the following game, all three of Aa, Bb and Cc are Nash equilibria, but Cc is weakly dominated, and Bb does not survive iterated deletion of weakly dominated strategies.

Game 40	a	b	c
A	3, 3	2, 0	0, 0
B	0, 2	2, 2	1, 0
C	0, 0	0, 1	1, 1

But most refinements don't come from simply conjoining an independently motivated condition onto Nash equilibrium. We'll start with the most significant refinement, subgame perfect equilibrium.

Subgame Perfect Equilibrium

Consider the following little game. The two players, call them Player I and Player II, have a choice between two options, call them Good and Bad. The game is a sequential move game; first Player I moves then Player II moves. Each player gets 1 if they choose Good and 0 if they choose Bad. We will refer to this game as **Game 41**. Here is its game tree.

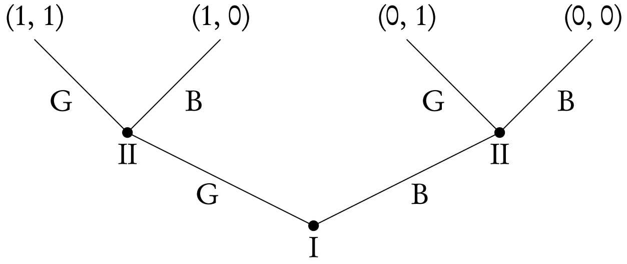


Figure 8.1: Game 41

A strategy for Player I in Game 41 is just a choice of one option, Good or Bad. A strategy for Player II is a little more complicated. She has to choose both what to do if Player I chooses Good, and what to do if Player II chooses Bad. We'll write her strategy as $\alpha\beta$, where α is what she does if Player I chooses Good, and β is what she does if Player II chooses Bad. (We will often use this kind of notation in what follows. Wherever it is potentially ambiguous, I'll try to explain it. But the notation is very common in works on game theory, and it is worth knowing.)

The most obvious Nash equilibrium of the game is that Player I chooses Good, and Player II chooses Good whatever Player I does. But there is another Nash equilibrium, as looking at the strategic form of the game reveals. We'll put Player I on the row and Player II on the column. (We'll also do this from now on unless there is a reason to do otherwise.)

Game 41	gg	gb	bg	bb
G	1, 1	**1, 1**	1, 0	1, 0
B	0, 1	0, 0	0, 1	0, 0

Look at the cell that I've bolded, where Player I plays Good, and Player II plays Good, Bad. That's a Nash equilibrium. Neither player can improve their outcome, given what the other player plays. But it is a very odd Nash equilibrium. It would be very odd for Player II to play this, since it risks getting 0 when they can guarantee getting 1.

It's true that Good, Bad is weakly dominated by Good, Good. But as we've already seen, and as we'll see in very similar examples to this soon, there are dangers in throwing out *all* weakly dominated strategies. Many people think that there is something else wrong with what Player II does here.

Consider the sub-game that starts with the right-hand decision node for Player II. That isn't a very interesting game; Player I has no choices, and Player II simply has a choice between 1 and 0. But it is a game. And note that Good, Bad is not a Nash equilibrium of that game. Indeed, it is a *strictly* dominated strategy in that game, since it involves taking 0 when 1 is freely available.

Say that a strategy pair is a **subgame perfect equilibrium** when it is a Nash equilibrium, and it is a Nash equilibrium of every sub-game of the game. The pair Good and Good, Bad is not subgame perfect, since it is not a Nash equilibrium of the right-hand subgame.

When we solve extensive form games by backwards induction, we not only find Nash equilibria, but subgame perfect equilibria. Solving this game by backwards induction would reveal that Player II would choose Good wherever she ends up, and then Player I will play Good at the first move. And the only subgame perfect equilibrium of the game is that Player I plays Good, and Player II plays Good, Good.

Forward Induction

In motivating subgame perfect equilibrium, we use the idea that players will suppose that future moves will be rational. We can also develop constraints based around the idea that past moves were rational. This kind of reasoning is called **forward induction**. A clear, and relatively uncontroversial, use of it is in this game by Cho and Kreps's paper "Signalling Games and Stable Equilibria" (QJE, 1987). We'll return to that paper in more detail below, but first we'll discuss **Game 42**.

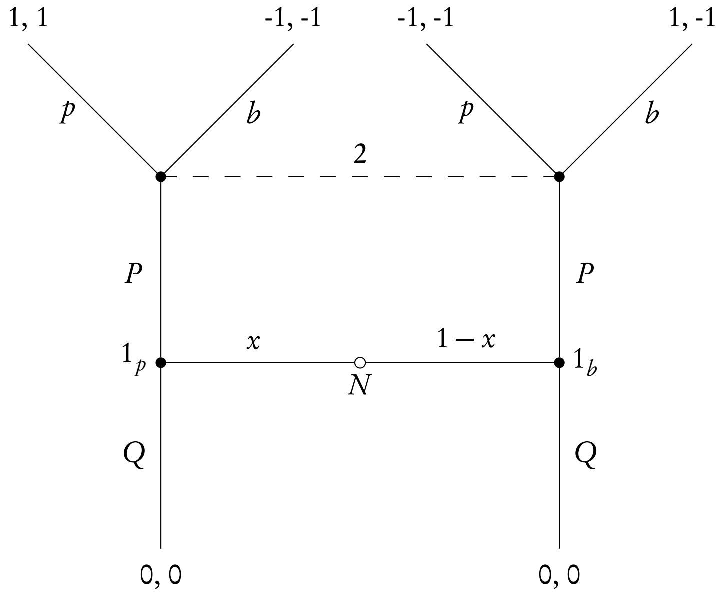


Figure 8.2: Game 42

This game takes a little more interpreting than some we've looked at previously. We use the convention that an empty disc indicates the initial node of the game. In this case, as in a few games we'll look at below, it is in the middle of the tree. The first move is made by Nature, denoted by N. Nature makes Player 1 playful with probability x, or bashful, with probability $1-x$. Player 1's personality type is revealed to Player 1, but not to Player 2. If she's playful, she moves to the node marked 1_p, if bashful to the node marked 1_b. Either way she has a choice about whether to play a guessing game with Player 2, where Player 2 has to guess her personality type. Player 1's payouts are as follows:

- If she doesn't play the guessing game, she gets 0.
- If she's playful, she gets 1 if Player 2 guesses correctly, and -1 if Player 2 guesses incorrectly.
- If she's bashful, she gets -1 either way.

Player 2's payouts are a little simpler.

- If the guessing game isn't played, she gets 0.
- If it is played and she guesses correctly, she gets 1.
- If it is played and she guesses wrongly, she gets -1.

The horizontal dashed line at the top indicates that if one of the upper nodes is reached, Player 2 doesn't know which node she is at. So we can't simply apply backward induction. Indeed, there aren't any subgames of this game, since there are no nodes that are neither initial nor terminal such that when they are reached, both players know they are there.

Player 1 has four possible strategies. She has to decide whether to Play or Quit both for when she is playful and when she is bashful. We'll write a strategy $\alpha\beta$ as the strategy of playing α if playful, and β if bashful. (We're going to repeat a lot of this notation when we get to signalling games, but it is worthwhile going over it a few times to be maximally clear.) Player 2 only has one choice and two possible strategies: if she gets to guess, she can guess $p$layful or $b$ashful. If we set out the strategic form of the game, we get the following expected payouts. (It's worth checking that you understand why these are the expected payouts for each strategy.)

Game 42	p	b
PP	$2x-1, 2x-1$	$-1, 1-2x$
PQ	x, x	$-x, -x$
QP	$x-1, x-1$	$x-1, 1-x$
QQ	$0, 0$	$0, 0$

Assuming $0 < x < 1$, it is easy to see that there are two pure Nash equilibria here: $\langle PQ, p\rangle$ and $\langle QQ, r\rangle$. But there is something very odd about the second equilibrium. Assume that both players are rational, and Player 2 actually gets to play. If Player 1 is bashful, then Quitting dominates *P*laying. So a rational bashful Player 1 wouldn't give Player 2 a chance to move. So if Player 2 gets a chance to move, Player 1 must be playful. And if Player 1 is playful, the best move for Player 2 is p. So by forward induction reasoning, Player 2 should play p. Moreover, Player 1 can figure all this out, so by backward induction reasoning she should play her best response to p, namely PQ.

We'll look at reasons for being sceptical of forward induction reasoning, or at least of some notable applications of it, next time. But at least in this case, it seems to deliver the right verdict: the players should get to the $\langle PQ, p\rangle$ equilibrium, not the $\langle QQ, r\rangle$ equilibrium.

Perfect Bayesian Equilibrium

The core idea behind subgame perfect equilibrium was that we wanted to eliminate equilibria that relied on 'incredible threats'. That is, there are some equilibria such that the first player makes a certain move only because if they make a different move, the second player to move would, on their current strategy, do something that makes things worse for both players. But there's no reason to think that the second player would actually do that.

The same kind of consideration can arise in games where there aren't any subgames. For instance, consider the following game, which we'll call **Game 43**. The game here is a little different to one we've seen so far. First Player 1 makes a move, either L, M or R. If her move is R, the game ends, with the (2, 2) payout. If she moves L or M, the game continues with a move by Player 2. But crucially, Player 2 does not know what move Player 1 made, so she does not know which node she is at. So there isn't a subgame here to start. Player 2 chooses l or r, and then the game ends.

There are two Nash equilibria to Game 43. The obvious equilibrium is Ll. In that equilibrium Player 1 gets her maximal payout, and Player 2 gets as much

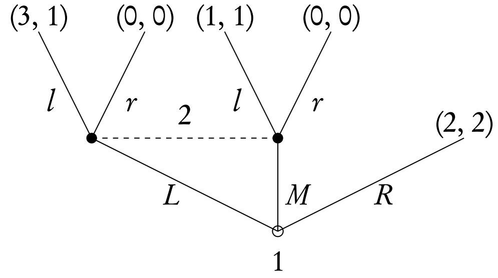


Figure 8.3: Game 43

as she can get given that the rightmost node is unobtainable. But there is another Nash equilibrium available, namely Rr. In that equilibrium, Player 2 gets her maximal payout, and Player 1 gets the most she can get given that Player 2 is playing r. But note what a strange equilibrium it is. It relies on the idea that Player 2 would play r were she to be given a move. But that is absurd. Once she gets a move, she has a straight choice between 1, if she plays l, and 0, if she plays r. So obviously she'll play l.

This is just like the examples that we used to motivate subgame perfect equilibrium, but that doesn't help us here. So we need a new concept. The core idea is that each player should be modelled as a Bayesian expected utility maximiser. More formally, the following constraints are put on players.

1. At each point in the game, each player has a probability distribution over where they are in the game. These probability distributions are correct about the other players' actions. That is, if a player is playing a strategy S, everyone has probability 1 that they are playing S. If S is a mixed strategy, this might involve having probabilities between 0 and 1 in propositions about which move the other player will make, but players have correct beliefs about other players' strategies.
2. No matter which node is reach, each player is disposed to maximise expected utility on arrival at that node.
3. When a player had a positive probability of arriving at a node, on reaching that node they update by conditionalisation.
4. When a player gave 0 probability to reaching a node (e.g., because the equilibrium being played did not include that node), they have some disposition or other to form a set of consistent beliefs at that node.

The last constraint is very weak, but it does enough to eliminate the equilibrium Rr. The constraint implies that when Player 2 moves, she must have some probability distribution Pr such that there's an x such that $\Pr(L) = x$ and $\Pr(M) = 1-x$. Whatever value x takes, the expected utility of l given Pr is 1, and the expected utility of r is 0. So being disposed to play r violates the second condition. So Rr is not a Perfect Bayesian equilibrium.

It's true, but I'm not going to prove it, that all Perfect Bayesian equilibria are Nash equilibria. It's also true that the converse does not hold, and this we have proved; Game 43 is an example.

Signalling Games

Concepts like Perfect Bayesian equilibrium are useful for the broad class of games known as signalling games. In a signalling game, Player 1 gets some information that is hidden from player 2. Many applications of these games involve the information being *de se* information, so the information that Player 1 gets is referred to as her *type*. But that's inessential; what is essential is that only one player gets this information. Player 1 then has a choice of move to make. There is a small loss of generality here, but we'll restrict our attention to games where Player 1's choices are independent of her type, i.e., of the information she receives. Player 2 sees Player 1's move (but not, remember, her type) and then has a choice of her own to make. Again with a small loss of generality, we'll restrict attention to cases where Player 2's available moves are independent of what Player 1 does. We'll start, in game **Game 44** with a signalling game where the parties' interests are perfectly aligned.

As above, we use an empty disc to signal the initial node of the game tree. In this case, it is the node in the centre of the tree. The first move is made by Nature, again denoted as N. Nature assigns a type to Player 1; that is, she makes some proposition true, and reveals it to Player 1. Call that proposition q. We'll say that Nature moves left if q is true, and right otherwise. We assume the probability (in some good sense of probability) of q is p, and this is known before the start of the game. After Nature moves, Player 1 has to choose Up or Down. Player 2 is shown Player 1's choice, but not Nature's move. That's the effect of the horizontal dashed lines. If the game reaches one of the upper nodes, Player 2 doesn't know which one it is, and if it reaches one of the lower nodes, again Player 2 doesn't know which it is. Then Player 2 has a make a choice, here simply denoted as $l$eft or $r$ight.

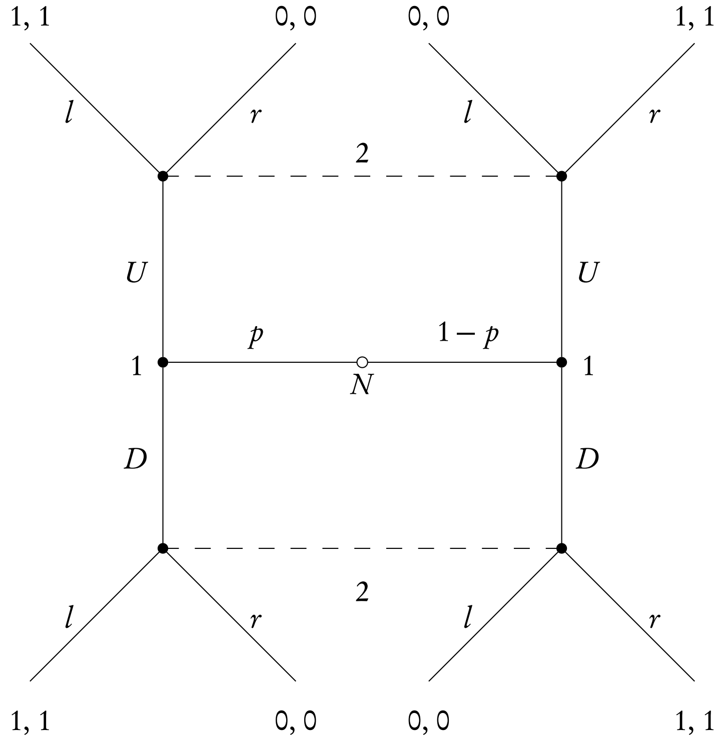


Figure 8.4: Game 44

In any game of this form, each player has a choice of four strategies. Player 1 has to choose what to do if q is true, and what to do if q is false. We'll write $\alpha\beta$ for the strategy of doing α if q, and β if $\neg q$. Since α could be U or D, and β could be U or D, there are four possible strategies. Player 2 has to choose what to do if U is played, and what to do if D is played. We'll write $\gamma\delta$ for the strategy of doing γ if U is played, and δ if D is played. Again, there are four possible choices. (If Player 2 knew what move Nature had made, there would be four degrees of freedom in her strategy choice, so she'd have 16 possible strategies. Fortunately, she doesn't have that much flexibility!)

Any game of this broad form is a signalling game. Signalling games differ in (a) the interpretation of the moves, and (b) the payoffs. Game 44 has a very simple payoff structure. Both players get 1 if Player 2 moves l iff q, and 0 otherwise. If we think of l as the formation of a belief that q, and r as the formation of the opposite belief, this becomes a simple communication game. The players get a payoff iff Player 2 ends up with a true belief about q, so Player 1 is trying to communicate to Player 2 whether q is true. This kind of simple communication game was Lewis used in *Convention* to show that game theoretic approaches could be fruitful in the study of meaning. The game is perfectly symmetric if $p = 1/2$; so as to introduce some asymmetries, I'll work through the case where $p = 3/5$.

Game 44 has a dizzying array of Nash equilibria, even given this asymmetry introducing assumption. They include the following:

- There are two **separating** equilibria, where what Player 1 does depends on what Nature does. These are $\langle UD, lr\rangle$ and $\langle DU, rl\rangle$. These are rather nice equilibria; both players are guaranteed to get their maximal payout.
- There are two **pooling** equilibria, where what Player 1 does is independent of what Nature does. These are $\langle UU, ll\rangle$ and $\langle DD, ll\rangle$. Given that she gets no information from Player 1, Player 2 may as well guess. And since $\Pr(q) > 1/2$, she should guess that q is true; i.e., she should play l. And given that Player 2 is going to guess, Player 1 may as well play anything. So these are also equilibria.
- And there are some **babbling** equilibria. For instance, there is the equilibrium where Player 1 plays either UU or DD with some probability r, and Player 2 plays ll.

Unfortunately, these are all Perfect Bayesian equilibria too. For the separating and babbling equilibria, it is easy to see that conditionalising on what Player 1 plays leads to Player 2 maximising expected utility by playing her part of the equilibrium. And for the pooling equilibria, as long as the probability of q stays above $1/2$ in any 'off-the-equilibrium-path' play (e.g., conditional on D in the $\langle UU, ll\rangle$ equilibrium), Player 2 maximises expected utility at every node.

That doesn't mean we have nothing to say. For one thing, the separating equilibria are Pareto-dominant in the sense that both players do better on those equilibria than they do on any other. So that's a non-coincidental reason to think that they will be the equilibria that arise. There are other refinements on Perfect Bayesian equilibria that are more narrowly tailored to signalling games. We'll introduce them by looking at a couple of famous signalling games.

Economists have been interested for several decades in games that give college a signalling function. So consider the following variant of the signalling game, **Game 45**. It has the following intended interpretation:

- Player 1 is a student and potential worker, Player 2 is an employer.
- The student is either bright or dull with probability p of being bright. Nature reveals the type to the student, but only the probability to the employer, so q is that the student is bright.
- The student has the choice of going to college (U) or the beach (D).
- The employer has the choice of hiring the student (l) or rejecting them (r).

In Game 45 we make the following extra assumptions about the payoffs.

- Being hired is worth 4 to the student.
- Going to college rather than the beach costs the bright student 1, and the dull student 5, since college is much harder work for dullards.
- The employer gets no benefit from hiring college students as such.
- Hiring a bright student pays the employer 1, hiring a dull student costs the employer 1, and rejections have no cost or benefit.

The resulting game tree is shown in 8.5. In this game, dull students never prefer to go to college, since even the lure of a job doesn't make up for the pain of actually having to study. So a rational strategy for Player 1 will never be of the form αU, since for dull students, college is a dominated option, being dominated by αD. But whether bright students should go to college is a trickier question. That is, it is trickier to say whether the right strategy for Player 1 is UD or DD, which are the two strategies consistent with eliminating the strictly dominated strategies. (Remember that strictly dominated strategies cannot be part of a Nash equilibrium.)

First, consider the case where $p < 1/2$. In that case, if Player 1 plays DD, then Player 2 gets $2p - 1$ from playing either ll or rl, and 0 from playing lr or rr. So she should play one of the latter two options. If she plays rr, then DD is a best response, since when employers aren't going to hire, students prefer to go to the beach. So there is one **pooling** equilibrium, namely $\langle DD, rr \rangle$. But what if Player 2 plays lr. Then Player 1's best response is UD, since bright students prefer college conditional on it leading to a job. So there is also a **separating** equilibrium, namely $\langle UD, lr \rangle$. The employer prefers that equilibrium, since her

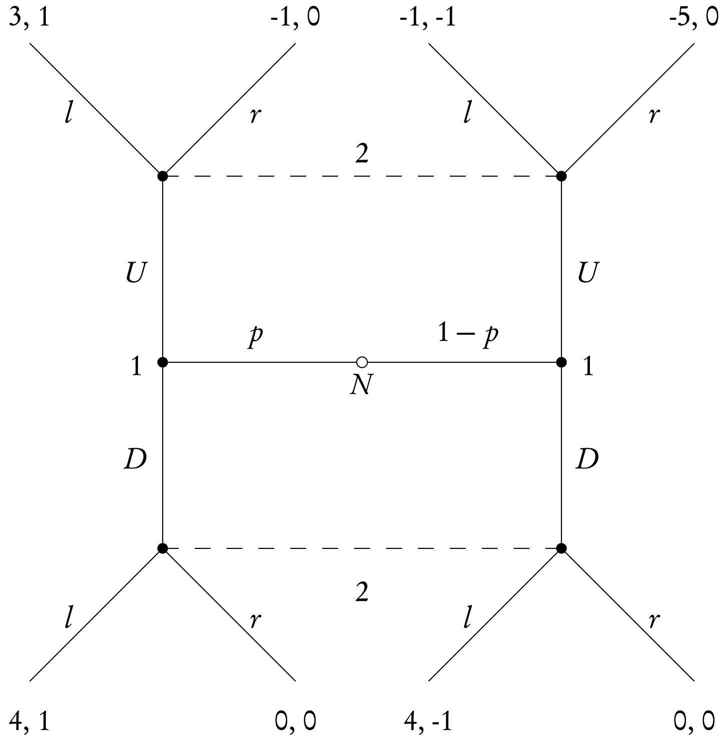


Figure 8.5: Game 45

payoff is now p rather than 0. And students prefer it, since their payoff is $3p$ rather than 0. So if we assume that people will end up at Pareto-dominant outcomes, we have reason to think that bright students, and only bright students, will go to college, and employers will hire college students, and only college students. And all this is true despite there being no advantage whatsoever to going to college in terms of how good an employee one will be.

Especially in popular treatments of the case, the existence of this kind of model can be used to motivate the idea that college *only* plays a signalling role. That is, some people argue that in the real world college does not make students more economically valuable, and the fact that college graduates have better employment rates, and better pay, can be explained by the signalling function of college. For what it's worth, I highly doubt that is the case. The wage premium one gets for going to college tends to *increase* as one gets further removed from college, although the further out from college you get, the less important a signal college participation is. One can try to explain this fact too on a pure signalling model of college's value, but frankly I think the assumptions needed to do so are heroic. The model is cute, but not really a picture of how actual college works.

So far we assumed that $p < 1/2$. If we drop that assumption, and assume instead that $p > 1/2$, the case becomes more complicated. Now if Player 1 plays DD, i.e., goes to the beach no matter what, Player 2's best response is still to hire them. But note that now a very odd equilibrium becomes available. The pair $\langle DD, rl \rangle$ is a Nash equilibrium, and, with the right assumptions, a Perfect Bayesian equilibrium. This pair says that Player 1 goes to the beach whatever her type, and Player 2 hires only beach goers.

This is a very odd strategy for Player 2 to adopt, but it is a little hard to say just why it is odd. It is clearly a Nash equilibrium. Given that Player 2 is playing rl, then clearly beach-going dominates college-going. And given that Player 1 is playing DD, playing rl gets as good a return as is available to Player 2, i.e., $2p - 1$. Could it also be a Perfect Bayesian equilibrium? It could, provided Player 2 has a rather odd update policy. Say that Player 2 thinks that if someone goes to college, they are a dullard with probability greater than $1/2$. That's consistent with what we've said; given that Player 1 is playing DD, the probability that a student goes to college is 0. So the conditional probability that a college-goer is bright is left open, and can be anything one likes in Perfect Bayesian equilibrium. So if Player 2 sets it to be, say, 0, then the rational reaction is to play rl.

But now note what an odd update strategy this is for Player 2. She has to assume that if someone deviates from the DD strategy, it is someone for whom the deviation is strictly dominated. Well, perhaps it isn't crazy to assume that someone who would deviate from an equilibrium isn't very bright, so maybe this isn't the oddest assumption in this particular context. But a few economists and game theorists have thought that we can put more substantive constraints on probabilities conditional on 'off-the-equilibrium-path' behaviour. One such constraint, is, roughly, that deviation shouldn't lead to playing a dominated strategy. This is the "**intuitive criterion**" of Cho and Kreps. In this game, all the criterion rules out is the odd pair $\langle DD, rl \rangle$. It doesn't rule out the very similiar pair $\langle DD, ll \rangle$. But the intuitive criterion makes more substantial constraints in other games. We'll close the discussion of signalling games with such an example, and a more careful statement of the criterion.

The tree in 8.6 represents **Game 46**, which is also a guessing game. The usual statement of it involves all sorts of unfortunate stereotypes about the type of men who have quiche and/or beer for breakfast, and I'm underwhelmed by it. So I'll run with an example that relies on different stereotypes.

A North American tourist is in a bar. 60% of the North Americans who pass through that bar are from Canada, the other 40% are from the US. (This is a little implausible for most parts of the world, but maybe it is very cold climate bar. In Cho and Kreps' version the split is 90/10 not 60/40, but all that matters is which side of 50/50 it is.) The tourist, call her Player 1, knows her nationality, although the barman doesn't. The tourist can ask for the bar TV to be turned to hockey or to baseball, and she knows once she does that the barman will guess at her nationality. (The barman might also try to rely on her accent, but she has a fairly neutral upper-Midwest/central Canada accent.) Here are the tourist's preferences.

- If she is American, she has a (weak) preference for watching baseball rather than hockey.
- If she is Canadian, she has a (weak) preference for watching hockey rather than baseball.
- Either way, she has a strong preference for being thought of as Canadian rather than American. This preference is considerably stronger than her preference over which sport to watch.

The barman's preferences are simpler; he prefers to make true guesses to false guesses about the tourist's nationality. All of this is common knowledge. So the decision tree is İn this tree, *B* means asking for baseball, *H* means asking for hockey, *a* means guessing the tourist is from the USA, *c* means guessing she is from Canada.

There are two Nash equilibria for this game. One is $\langle HH, ca\rangle$; everyone asks for hockey, and the barman guesses Canadian if hockey, American if baseball. It isn't too hard to check this is an equilibrium. The Canadian gets her best outcome, so it is clearly an equilibrium for her. The American gets the second-best outcome, but asking for baseball would lead to a worse outcome. And given that everyone asks for hockey, the best the barman can do is go with the prior probabilities, and that means guessing Canadian. It is easy to see how to extend this to a perfect Bayesian equilibrium; simply posit that conditional on baseball being asked for, the probability that the tourist is American is greater than $^1/_2$.

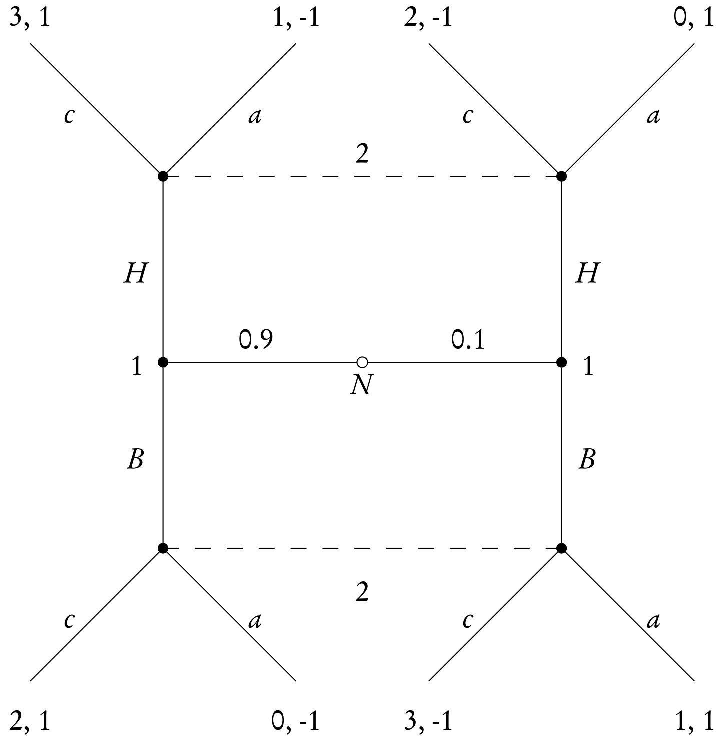


Figure 8.6: Game 46

The other equilibrium is rather odder. It is $\langle BB, ac\rangle$; everyone asks for baseball, and the barman guesses American if hockey, Canadian if baseball. Again, it isn't too hard to see how it is Nash equilibrium. The Canadian would rather have hockey, but not at the cost of being thought American, so she has no incentive to defect. The American gets her best outcome, so she has no incentive to defect. And the barman does as well as he can given that he gets no information out of the request, since everyone requests baseball.

Surprisingly, this could also be a perfect Bayesian equilibrium. This requires that the barman maximise utility at every node. The tricky thing is to ensure he maxmises utility at the node where hockey is chosen. This can be done provided that conditional on hockey being chosen, the probability of American rises to above $1/2$. Well, nothing we have said rules that out, so there *exists* a perfect

Bayesian equilibrium. But it doesn't seem very plausible. Why would the barman adopt just *this* updating disposition?

One of the active areas of research on signalling games is the development of formal rules to rule out intuitively crazy updating dispositions like this. (Small note for the methods folks: at least some of the economists working on this explicitly say that the aim is to develop formal criteria that capture clear intuitions about cases.) I'm just going to note one of the very early attempts in this area, Cho and Kreps' *Intuitive Criterion*.

Start with some folksy terminology that is (I think) novel to me. Let an outcome o of the game be any combination of moves by nature and the players. Call the players' moves collectively P and nature's move N. So an outcome is a pair $\langle P, N \rangle$. Divide the players into three types.

- The **happy** players in o are those such that given N, o maximises their possible returns. That is, for all possible P', their payoff in $\langle P, N \rangle$ is at least as large as their payoff in $\langle P', N \rangle$.
- The **content** players are those who are not happy, but who are such that for no strategy s' which is an alternative to their current strategy s, would they do better given what nature and the other players would do if they played s'.
- The **unhappy** players are those who are neither happy nor content.

The standard Nash equilibrium condition is that no player is unhappy. So assume we have a Nash equilibrium with only happy and content players. And assume it is also a Perfect Bayesian equilibrium. That is, for any 'off-equilibrium' outcome, each player would have some credence function such that their play maximises utility given that function.

Now add one constraint to those credence functions:

> **Intuitive Criterion - Weak Version**
>
> Assume a player has probability 1 that a strategy combination P will be played. Consider their credence function conditional on another player playing s', which is different to the strategy s they play in P. If it is consistent with everything else the player believes that s' could be played by a player who is content with the actual outcome $\langle P, N \rangle$, then give probability 1 to s' being played by a content player.

That is, if some deviation from equilibrium happens, give probability 1 to it being one of the content players, not one of the happy players, who makes the deviation.

That's enough to rule out the odd equilibrium for the baseball-hockey game. The only way that is a Perfect Bayesian equilibrium is if the barman responds to a hockey request by *increasing* the probability that the tourist is American. But in the odd equilibrium, the American is happy, and the Canadian is merely content. So the Intuitive Criterion says that the barman should give credence 1 to the hockey requester, i.e., the deviator from equilibrium, being Canadian. And if so, the barman won't respond to a hockey request by guessing the requestor is American, so the Canadian would prefer to request hockey. And that means the outcome is no longer an equilibrium, since the Canadian is no longer even content with requesting baseball.

In games where everyone has only two options, the weak version I've given is equivalent to what Cho and Kreps offers. The official version is a little more complicated. First some terminology of mine, then their terminology next.

- A player is **happy to have played** s rather than s' if the payoff for s in o is greater than any possible outcome for s' given N.
- A player is **content to have played** s rather than s' if they are not happy to have played s rather than s', but the payoff for s in o is as great as the payoff for s' given N and the other players' strategies.
- A player is **unhappy to have played** s rather than s' if they are neither happy nor content to have played s rather than s'.

If a player is happy to have played s rather than s', and s is part of some equilibrium outcome o, then Cho and Kreps say that s' is an **equilibrium-dominated** strategy. The full version of the Intuitive Criterion is then:

> **Intuitive Criterion - Full Version**
>
> Assume a player has probability 1 that a strategy combination P will be played. Consider their credence function conditional on another player playing s', which is different to the strategy s they play in P. If it is consistent with everything else the player believes that s' could be played by a player who is merely content to have played s rather than s', then give probability 1 to s' being played by someone who is

> content to have played s, rather than someone who is happy to have played s.

This refinement matters in games where players have three or more options. It might be that a player's options are s_1, s_2 and s_3, and their type is t_1 or t_2. In the Nash equilibrium in question, they play s_1. If they are type t_1, they are happy to have played s_1 rather than s_2, but merely content to have played s_1 rather than s_3. If they are type t_2, they are happy to have played s_1 rather than s_3, but merely content to have played s_1 rather than s_2. In my terminology above, both players are content rather than happy with the outcome. So the weak version of the Intuitive Criterion wouldn't put any restrictions on what we can say about them conditional on them playing, say s_2. But the full version does say something; it says other players should assign probability 1 to the player being type t_2 conditional on them playing s_2, since the alternative is that s_2 is played by a player who is happy they played s_1 rather than s_2. Similarly, it says that conditional on the player playing s_3, other players should assign probability 1 to their being of type t_1, since the alternative is to assign positive probability to a player deviating to a strategy they are happy not to play.

The Intuitive Criterion has been the subject of an enormous literature. Google Scholar lists nearly 2000 citations for the Cho and Kreps paper alone, and similar rules are discussed in other papers. So there are arguments that it is too weak and arguments too strong, and refinements in all sorts of directions. But in the interests of not getting too far outside my field of competence, we'll leave those. What I most wanted to stress was the *form* a refinement of Perfect Bayesian equilibrium would take. It is a constraint on the conditional credences of agents conditional on some probability 0 event happening. It's interesting that there are some plausible, even intuitive constraints; sometimes it seems the economic literature has investigated rationality under 0 probability evidence more than philosophers have!

Other Types of Constraint

We don't have the time, or perhaps expertise, to go into other kinds of constraints in as much detail, but I wanted to quickly mention two other ways in which game theorists have attempted to restrict Nash equilibrium.

The first is sometimes called **trembling-hand equilibrium**. The idea is that we should restrict our attention to those strategies that are utility maximising given a very very high credence that the other players will play the strategies they actually play, and some positive (but low) credence that they will play each other strategy. This is, I think very important to real-world applications, since it is very implausible that we should assign probability 1 to any claim about the other players, particularly claims of the form that they will play some equilibrium rather than another. (There is a connection here to the philosophical claim that rational credences are *regular*; that is, that they assign positive probability to anything possible.)

In normal form games, the main effect of this is to rule out strategies that are weakly dominated. Remember that there are many strategies that are equilibria that are weakly dominated, since equilibrium concepts typically only require that player can't do better given some other constraint. But if we have to assign positive probability to any alternative, then the weakly dominating strategy will get a utility boost from the alternative under which it is preferable.

Things get more complicated when we put a 'trembling-hand' constraint on solutions to extensive form games. The usual idea is that players should, at each node, assign a positive probability to each deviation from equilibrium play. This can end up being a rather tight constraint, especially when combined with such ideas as subgame-perfection.

The other kind of refinement I'll briefly discuss in **evolutionary stability**. This is, as the name suggests, a concept that arose out of game-theoretic models of evolution. As such, it only really applies to symmetric games. In such a game, we can write things like $U(s,s')$, meaning the payoff that one gets for playing s when the other player is playing s'. In an asymmetric game, we'd have to also specify which 'side' one was when playing s, but there's no need for that in a symmetric game.

We then say that a strategy is evolutionarily stable iff these two conditions hold.

$$\begin{aligned} &\forall t : (U(s,s) \geq U(t,s)) \\ &\forall t \neq s : (U(s,s) > U(t,s) \vee U(s,s) > U(t,t)) \end{aligned}$$

The idea is that a species playing s is immune to invasion if it satisfies these conditions. Any invader will play some alternative strategy t. Think then of a species playing *t* whose members have to play either against the existing strategy *s* with high probability, or against other members of their own species with low probability. The first clause says that the invader can't do better in the normal case, where they play with a dominant strategy. The second clause says that they do worse in one of the two cases that come up during the invasion, either playing the dominant strategy or playing their own kind. The effect is that the dominant strategy will do better, and the invader will die out.

For real-world applications we might often want to restrict the quantifier to biologically plausible strategies. And, of course, we'll want to be very careful with the specification of the game being played. But, I'm told, there are some nice applications of this concept in explanations of certain equilibrium outcomes. That's a topic for biology class though - not decision theory!

Backward Induction and Its Discontents

Problems with Backwards Induction

When each player only moves once, it is very plausible that each player should play their part of a subgame perfect equilibrium solution to the game. But this is less compelling when players have multiple moves. We can start to see why by looking at a game that Robert Stalnaker has used in a few places. Again, it is an extensive form game. We will call it **Game 47**.

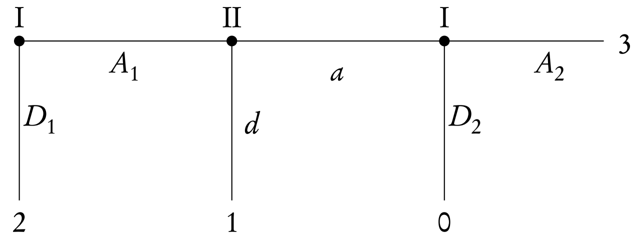


Figure 9.1: Game 47

The game starts in the upper-left corner. There are up to three moves, each of them Across or Down. As soon as one player moves Down, the game ends. It is a common interest game; each player gets the payout at the terminal node.

Since there is only one path to each decision node, a strategy merely has to consist of a set of plans for what to play if that node is reached. Alice's strategy will consist of two capitalised moves, and Bob's strategy will consist of one lower-case move.

If we apply backwards induction, we get that the unique solution of the game is $\langle A_1 A_2, a \rangle$. Alice would play A_2 if it gets that far. Given that, Bob would be better off playing a than d if he gets to play. And given that, Alice is better off playing A_1 than D_1.

But there are many other Nash equilibria of the game. One of them is $\langle D_1 A_2, d \rangle$. Given that Player I is playing D_1, Player II can play anything without changing her payout; it will be 2 come what may. Given that Player II is playing d, Player I is best off playing D_1 and taking 2, rather than just getting the 1 that comes from leaving Player II to make a play.

Could this equilibrium be one that rational players, who know each other to be rational, reach? Stalnaker argues that it could. Assume each player knows that the other player is rational, and is playing that strategy. Given what Player I knows, her strategy choice is clearly rational. She takes 2 rather than the 1 that she would get by playing A_1, and she is disposed to take 3 rather than 0 if she gets to the final decision node. So her actual choice is rational, and her disposition is to make another rational choice if we reach the end of the game.

Things are a little trickier for Player II. You might think it is impossible for rational Player II, who knows Player I to be rational, to move d. After all, if Player I is rational, then she'll play A_2, not D_2. And if she plays A_2, it is better to play a than d. So it looks hard to justify Player II's move. But looks can be deceiving. In fact there isn't anything wrong with Player II's move, as long as he has the right beliefs to justify it. It's very important to distinguish the following two conditionals.

- If Player I has the choice, she chooses A_2 over D_2.
- If Player I were to have the choice, she would choose A_2 over D_2.

Player II knows that the first, indicative, conditional is true. And indeed it is true. But he doesn't know that the second, subjunctive, conditional is true. After all, if Player I were to have the choice between A_2 and D_2, she would have, *irrationally*, chosen A_1 over D_1. And if she had chosen irrationally once, it's possible that she would choose irrationally again.

Here's an analogy that may help explain what's going on. The following set of beliefs is consistent.

- Any perfectly rational being gets all of the questions on their algebra exam right.
- Alice is perfectly rational.
- If Alice had got the second-hardest question on the algebra exam wrong, she would have got the hardest question on the algebra exam wrong as well.

Player II's beliefs are like that. He believes Player I is perfectly rational. He also believes that if Player I were to make an irrational move, she would continue to make irrational moves. That's consistent with belief in perfect rationality, and nothing about the game setup rules out such a belief. He also believes that playing

A_1 would be irrational. That's correct, given what Player I knows about Player II. Given all those beliefs, playing d, if he had the chance, would be rational.

Stalnaker argues that many game theorists have tacitly confused indicative and subjunctive conditionals in reasoning about games like Game 47. Let's look at some other games where similar reasoning takes place.

Money Burning Game

Elchanen Ben-Porath and Eddie Dekel suggested this variant to the Battle of the Sexes game. The variation is in two steps. First, we change the payouts for the basic game to the following. (Note that I'm using a lowercase letter for one of *R*'s options here; this is to distinguish the d of down from the D of Don't burn.)

	l	r
u	4, 1	0, 0
d	0, 0	1, 4

Then we give Player I the option of publicly burning 2 utils before playing this game. We will use D for Don't burn, and B for burn. So actually each player has four choices. Player I has to choose both D or B, then u or d. Player 2 has to choose whether to play l or r in each of the two possibilities: first, when D is played, second, when B is played. We'll write lr for the strategy of playing l if D, and r if B, and so on for the other strategies.

Game 48	ll	lr	rl	rr
Du	4, 1	4, 1	0, 0	0, 0
Dd	0, 0	0, 0	1, 4	1, 4
Bu	2, 1	-2, 0	2, 1	-2, 0
Bd	-2, 0	-1, 4	-2, 0	-1, 4

Now we can analyse this game a couple of ways. First, we can go through eliminating weakly dominated strategies. Note that Du strictly dominated Bd, so we can eliminate it. If Bd is out, then ll weakly dominates lr, and rl weakly dominates rr. So we can eliminate lr and rr. Now Bu weakly dominates Dd, relative to the remaining options, so we can eliminate it. Given that just Du and Bu remain, ll weakly dominates all options for Player II, so it is all that is left. And given that ll is all that remains, Du strongly dominates Bu. So the iterative deletion of weakly dominated strategies leaves us with $\langle Du, ll\rangle$ as the unique solution of the game.

Alternatively, we can think the players reason as follows. (As Stalnaker notes, this is an instance of the forward induction reasoning we discussed earlier.) Playing Bd is obviously irrational for Player I, since its maximum return is -1, and playing Du or Dd guarantees getting at least 0. So any rational player who plays B must be playing Bu. That is, if Player I is rational and plays B, she will play Bu. Moreover, this is common knowledge among the players. So if Player I plays B, she will recognise that Player II will know that she is playing Bu. And if Player II knows Player I is playing u after burning, she will play l, since that returns her 1 rather than 0. So Player I knows that playing B leads to the 2,1 state; i.e., it returns her 2. But now it is irrational to play Dd, since that gets at most 1, and playing B leads to a return of 2. So Player I will play Bu or Du. And since the reasoning that leads to this is common knowledge, Player II must play ll, since playing l is her best response to a play of u, and she knows Player I will play u. But if Player II is going to play ll, Player I doesn't need to burn; she can just play Du.

There is a crucial conditional in the middle of that argument; let's isolate it.

- If Player I is rational and plays B, she will play Bu.

But that's not what is needed to make the argument work. What Player I needs, and in fact needs to be common knowledge, is the following subjunctive.

- If Player I is rational then, if she were to play B, she would play Bu.

But in fact there's no reason to believe that. After all, if the forward induction argument is right, then it is *irrational* to burn the money. So if Player I were to play B, i.e., burn the money, then she would be doing something irrational. And the fact that Player I is actually rational is consistent with the counterfactual that if she were to do one irrational thing, it would be because she is following an irrational strategy. Note that on the assumption that Du is optimal, then if Player I finds herself having played B, it isn't clear that playing d is irrational; it isn't like it is dominated by u.

So it isn't true that if Player I were to burn the utils, Player II would know that she is playing Bu, and react accordingly by playing l. And if that's not true, then there's no reason to think that Bu is better than Dd. And if there's no reason to think that, there's no reason to be confident Player 2 will play ll.

Stalnaker goes further and provides a positive model where the players start with a common belief in rationality, and in what each other will do, but in which the players play Bu and rl. I'll leave it as an exercise to work out what counterfactuals the agents have to believe to make this rational, but suffice to say that it could be a perfectly rational solution.

There is another, relatively simple, equilibrium to the game. Player I will play Dd, and Player II will play rr. Player II is certain that Player I will play d, and will keep that belief even if Player I irrationally burns 2 utils to start with. Given that certainty, it is rational for Player I to play Dd, so Player II's belief is consistent with believing Player I to be rational. And since Player I is playing Dd, playing rr is perfectly rational for Player II; indeed it results in her best possible outcome. Moreover, given that Player II is playing rr, it makes sense for Player I to play d even if they were, irrationally, to burn the utils. So even though playing Bd would be irrational, since it is strictly dominated, if they were to (irrationally) play D, they *should* follow that by playing d. There's an important point about the scope of the claim that playing Bd is irrational. It *doesn't* imply that if the player were to irrationally play B, it would be irrational to follow with a play of d. If Player I was convinced that Player II was playing rr, then it would be rational to follow B with d.

Iterated Prisoners' Dilemma

Let's start with a slightly relabeled version of Game 1, where we use C and D for cooperate and defect.

Game 1	c	d
C	3, 3	0, 5
D	5, 0	1, 1

We'll call the players I and II, and use uppercase letters for Player I's strategies, and lower case letters for Player II's strategies. We'll assume that each player knows that the other players are perfectly rational in Stalnaker's sense.

There are a couple of arguments that the players must end up at the equilibrium where every player defects on every move. One of these is an argument by backwards induction.

At the last move of the game, defecting dominates cooperating. Both players know this. At the second-last move, you might have thought antecedently that there was some benefit to cooperating. After all, it might induce cooperation in the other player. But the only benefit one could get from cooperating was cooperation at the next (i.e., last) move. And no rational player will cooperate on the last move. So, if you're playing with a rational player, there's no benefit to cooperating on the second-last move. But if that's right, and everyone knows it, then there is no benefit to cooperating on the third-last move. The only benefit would be if that would induce cooperation, and it couldn't, since we just proved that any rational player would defect on the second-last move. And so on for any finite length game that you like.

Alternatively, we could look at the game from a strategic perspective. As we've seen in games like Game 48, sometimes there are Nash equilibria that don't appear in backwards induction reasoning. But this isn't (*I think*) the case in finitely iterated Prisoners' Dilemma. The only Nash equilibrium is that both players defect in every circumstance.

This is a little tricky to check since the strategic form of iterated Prisoners' Dilemma gets very complex very quickly. Let's just consider the three round version of the game. Already each player has 128 strategies to choose from. The choice of a strategy involves making 7 distinct choices:

1. What to do in the first round.
2. What to do in the second round if the other player cooperates in the first round.
3. What to do in the second round if the other player defects in the first round.
4. What to do in the third round if the other player cooperates in each of the first two rounds.
5. What to do in the third round if the other player cooperates in the first round then defects in the second round.
6. What to do in the third round if the other player defects in the first round then cooperates in the second round.
7. What to do in the third round is the other player defects in each fo the first two rounds.

Since these 7 choices are distinct, and the player has 2 choices at each point, there are $2^7 = 128$ possible strategies. So the strategic form of the table involves $128 \times 128 = 16384$ cells. Needless to say, we *won't* be putting that table here. (Though it isn't too hard to get a computer to draw the table for you. The four round game, where each player has 2^{15} choices, and there are over one billion cells in the decision table, requires more computing power!)

We'll write a strategy as $\langle x_1x_2x_3x_4x_5x_6x_7\rangle$, where x_i is 0 if the player's answer to the i'th question above is to cooperate, and 1 if it is to defect. So $\langle 0000000\rangle$ is the strategy of always cooperating, $\langle 1111111\rangle$ is the strategy of always defecting, $\langle 0010101\rangle$ is 'tit-for-tat', the strategy of cooperating on the first move, then copying the other player's previous move, and so on.

Let $\langle x_1x_2x_3x_4x_5x_6x_7\rangle$ be any strategy that doesn't always involve defection on the final round, i.e., a strategy where $x_4 + x_5 + x_6 + x_7 < 4$. It is easy enough to verify that such a strategy is weakly dominated by $\langle x_1x_2x_31111\rangle$. In some plays of the game, the defection on the final round leads to getting a better outcome. In other plays of the game, $\langle x_1x_2x_3x_4x_5x_6x_7\rangle$ and $\langle x_1x_2x_31111\rangle$ have the same play. For instance, $\langle 0001000\rangle$ will do just as well as $\langle 0001111\rangle$ if the opponent plays any strategy of the form $\langle 000x_4x_5x_6x_7\rangle$, but it can never do better than $\langle 0001111\rangle$. So if each player is perfectly rational, we can assume their strategy ends with 1111.

That cuts each player's choices down to 8. Let's do that table. We'll use C and c rather than 0, and D and d rather than 1, so it is easier to distinguish the two players' strategies. When we label the table, we'll leave off the trailing Ds and d's, since we assume players are defecting on the last round. We'll also leave off the 3 units the players each get in the last round. (In other words, this will look a lot like the table for a *two* round iterated Prisoners' Dilemma, but it is crucial that it is actually a three-round game.)

Game 49	ccc	ccd	cdc	cdd	dcc	dcd	ddc	ddd
CCC	6, 6	6, 6	3, 8	3, 8	3, 8	3, 8	0, 10	0, 10
CCD	6, 6	6, 6	3, 8	3, 8	5, 5	5, 5	1, 6	1, 6
CDC	6, 6	8, 3	4, 4	4, 4	3, 8	3, 8	0, 10	0, 10
CDD	6, 6	8, 3	4, 4	4, 4	5, 5	5, 5	1, 6	1, 6
DCC	6, 6	5, 5	8, 3	5, 5	4, 4	1, 6	4, 4	1, 6
DCD	6, 6	5, 5	8, 3	5, 5	6, 1	2, 2	6, 1	2, 2
DDC	6, 6	6, 1	10, 0	6, 1	4, 4	1, 6	4, 4	1, 6
DDD	6, 6	6, 1	10, 0	6, 1	6, 1	2, 2	6, 1	2, 2

In this game CCC is *strongly* dominated by CDD, and DCC is strongly dominated by DDD. Similarly ccc and dcc are strongly dominated. So let's delete them.

Game 49′	ccd	cdc	cdd	dcd	ddc	ddd
CCD	6, 6	3, 8	3, 8	5, 5	1, 6	1, 6
CDC	8, 3	4, 4	4, 4	3, 8	0, 10	0, 10
CDD	8, 3	4, 4	4, 4	5, 5	1, 6	1, 6
DCD	5, 5	8, 3	5, 5	2, 2	6, 1	2, 2
DDC	6, 1	10, 0	6, 1	1, 6	4, 4	1, 6
DDD	6, 1	10, 0	6, 1	2, 2	6, 1	2, 2

Note that each of these is a best response. In particular,

- CCD is a best response to dcd.
- CDC is a best response to ccd.
- CDD is a best response to ccd and dcd.
- DCD is a best response to ddc.
- DDC is a best response to cdc and cdd.
- DDD is a best response to cdc, cdd, ddc and ddd.

Now the only Nash equilibrium of that table is the bottom-right corner. But there are plenty of strategies that, in a strategic version of the game, would be consistent with common belief in perfect rationality. The players could, for instance, play CDC and cdc, thinking that the other players are playing ccd and CCD respectively. Those beliefs would be false, but they wouldn't be signs that the other players are irrational, or that they are thinking the other players are irrational.

But you might suspect that the strategic form of the game and the extensive form are crucially different. The very fact that there are Nash equilibria that are not subgame perfect equilibria suggests that there are tighter constraints on what can be played in an extensive game consistent with rationality and belief in rationality. Stalnaker argues, however, that this isn't right. In particular, any strategy that is (given the right beliefs) perfectly rational in the strategic form of the game is also (given the right beliefs and belief updating dispositions) perfectly rational in the extensive form. We'll illustrate this by working more slowly through the argument that the game play $\langle CDC, cdc\rangle$ is consistent with common belief in perfect rationality.

The first thing you might worry about is that it isn't clear that CDC is perfectly rational, since it is looks to be weakly dominated by CDD, and perfect rationality is inconsistent with playing weakly dominated strategies. But in fact

CDC isnt weakly dominated by CDD. It's true that on the six columns represented here, CDC never does better than CDD. But remember that each of these strategies is short for a *three*-round strategy; they are really short for $CDCDDDD$ and $CDDDDDD$. And $CDCDDDD$ is not weakly dominated by $CDDDDDD$; it does better, for example, against $dcccddd$. That is the strategy is defecting the first round, cooperating the second, then defecting on the third round unless the other player has cooperated on each of the first two rounds. Since CDC does cooperate each of the first two rounds, it gets the advantage of defecting against a cooperator on the final round, and ends up with 8 points, whereas CDD merely ends up with 6.

But why would we think Player II might play $dcccddd$? After all, it is itself a weakly dominated strategy, and perfectly rational beings (like Player II) don't play weakly dominated strategies. But we're not actually thinking Player II *will* play that. Remember, Player I's assumption is that Player II will play $ccdddddd$, which is not a weakly dominated strategy. (Indeed, it is tit-for-tat-minus-one, which is a well-known strategy.) What Player I also thinks is that if she's wrong about what Player II will play, then it is possible that Player II is not actually perfectly rational. That's consistent with believing Player II is actually perfectly rational. The assumption of common belief in perfect rationality is not sufficient for the assumption that one should believe that the other player is perfectly rational *no matter what surprises happen*. Indeed, that assumption is barely coherent; some assumptions are inconsistent with perfect rationality.

One might be tempted by a weaker assumption. Perhaps a player should hold on to the belief that the other player is perfectly rational unless they get evidence that is *inconsistent* with that belief. But it isn't clear what could motivate that. In general, when we are surprised, we have to give up something that isn't required by the surprise. If we antecedently believe $p \wedge q$, and learn $\neg(p \wedge q)$, then what we've learned is inconsistent with neither p nor q, but we must give up one of those beliefs. Similarly here, it seems we must be prepared to give up a belief in the perfect rationality of the other player in some circumstances when that is not entailed by our surprise. (Note that even if you don't buy the argument of this paragraph, there still isn't a reason to think that perfectly rational players can't play CDD. But analysing that possibility is beyond the scope of these notes.)

What happens in the extensive form version of the game? Well, each player cooperates, thinking this will induce cooperation in the other, then each player is surprised by a defection at round two, then at the last round they both defect because it is a one-shot Prisoners' Dilemma. Nothing seems irrational there. We might wonder why neither defected at round one. Well, if they believed that the other player was playing tit-for-tat-minus-one, then it is better to cooperate at round one (and collect 8 points over the first two rounds) than to defect (and collect at most 6). And playing tit-for-tat-minus-one is rational if the other person is going to defect on the first and third rounds, and play tit-for-tat on round two. So as long as Player I thinks that Player II thinks that Player I thinks that she is going to defect on the first and third rounds, and play tit-for-tat on round two, then her cooperation at round one is rational, and consistent with believing that the other player is rational.

But note that we had to attribute an odd belief to Player II. We had to assume that Player II is *wrong* about what Player I will play. It turns out, at least for Prisoners' Dilemma, that this is crucial. If both players are certain that both players are perfectly rational, and have no false beliefs, then the only strategy that can be rationalised is permanent defection. The proof (which I'm leaving out) is in Stalnaker's "Knowledge, Belief and Counterfactual Reasoning in Games".

This *isn't* because the no false beliefs principle suffices for backwards induction reasoning in general. In Game 47, we can have a model of the game where both players are perfectly rational, and have correct beliefs about what the other player will do, and both those things are common belief, and yet the backwards induction solution is not played. In that game A believes, truly, that B will play d, and B believes, truly, that A will play A_1D_2. And each of these moves is optimal given (true!) beliefs about the other player's strategy.

But Prisoners' Dilemma is special. It isn't that in a game played between players who know each other to have no false beliefs and be perfectly rational that we must end up at the bottom-right corner of the strategic table. But we must end up in a game where every player defects every time. It could be that Player I thinks Player II is playing either dcd or ddd, with ddd being much more, and on that basis she decides to play dcd. And Player II could have the converse beliefs. So we'll end up with the play being $\langle DCD, dcd \rangle$. But of course that means each player will defect on the first two rounds, and then again on the last round since they are perfectly rational. In such a case the requirement that each player have no false beliefs won't even be enough to get us to Nash equilibrium since $\langle DCD, dcd \rangle$ is not a Nash equilibrium. But it is enough to get permanent defection.