None of this is meant to imply that fairness is somehow irrational. On the contrary, it seems to me to be the most important of the conventions that humans use to resolve equilibrium selection problems in everyday coordination games. But rather than regarding fairness as a substitute for compromises reached by rational bargaining, John Rawls's *Theory of Justice* makes rational bargaining the foundation stone of his definition of a fair outcome. Rawls identifies a fair deal with the agreement that Alice and Bob would reach if they were to bargain behind a 'veil of ignorance' that concealed their identity during the bargaining. Neither Alice nor Bob would then wish to disadvantage anyone, because they might themselves turn out to be the disadvantaged party.

I have devoted a substantial chunk of my life using game theory to examine the implications of Rawls's definition. Why does it strike us as reasonable? Does it lead to a utilitarian outcome as claimed by Harsanyi, or an egalitarian outcome as claimed by Rawls? However, life is too short to explain why I think Rawls defended a sound intuition with a wrong argument.

Forming coalitions

How can we apply what we have learned about how two rational people bargain to the bargaining that takes place when coalitions form? Von Neumann and Morgenstern proposed the simplest toy model in which coalitions matter.

Alice, Bob, and Carol are to play Divide-the-Dollar. Who gets how much is determined by majority voting. Any coalition of two players can therefore dispose of the dollar as they choose. But which coalition will form? Who will be the odd man out? How will the dollar be divided?

Outside options

Alice's outside option when bargaining with Bob is the most she can get elsewhere if their negotiations break down altogether.

Labour economists continue to make the error of identifying the *status quo* payoffs with the players' outside options when using the Nash bargaining solution to predict the outcome of wage negotiations. For example, if Bob will become unemployed if he fails to come to an agreement with Alice, then his *status quo* payoff is taken to be the level of social benefit.

To see why it is usually a mistake to use the Nash bargaining solution in this way, it is necessary to modify the Rubinstein bargaining model so that Alice and Bob always have the opportunity to take up their outside option after refusing an offer. It then becomes obvious that the outside options are relevant to the bargaining outcome only to the extent that we should discard all payoff pairs from the agreement set that assign somebody less than their outside option. The *status quo* needs to be identified with the payoffs the players receive *while* negotiating. For example, if Alice and Bob are seeking to negotiate the end of a strike, then their *status quo* payoffs are their respective incomes *during* the strike.

In order for it to be right to identify the *status quo* payoffs with the players' outside options, any breakdown in the negotiations needs be forced rather than voluntary. To model such a forced breakdown in Rubinstein's model, one can introduce a chance move that ends the negotiations with some small probability after each refusal. This would correspond to the case in which any delay in reaching an agreement might result in the surplus over which Alice and Bob are bargaining being stolen by a third party.

Odd-Man-Out

Our three-player version of Divide-the-Dollar can be regarded as three two-player bargaining problems to which we can apply Nash's cooperative bargaining theory. When two players bargain about how they will split the dollar should they agree to form a coalition on how to vote, their outside options are the deals that each would reach if they were to bargain with the odd-man-out instead.

It follows that Alice must expect the same payoff if she succeeds in forming a coalition with Bob as when she succeeds in forming a coalition with Carol – otherwise one of the potential agreements would require her to accept less than her outside option in that situation. Together with the Coase theorem, this fact ties down the three possible deals. In the case when the players are all risk neutral, we are led to the unsurprising conclusion that the coalition which forms will split the dollar fifty-fifty, leaving the odd-man-out with nothing.

The symmetry of the problem makes it impossible to say which of the three possible coalitions will form. However, the following noncooperative model breaks the symmetry by requiring that Alice, Bob, and Carol rotate in making payoff demands. When it is your turn to move, you may either accept any demand that has been made previously or else make a new demand of your own. The unique subgame-perfect equilibrium predicts that the very first opportunity to form a coalition will be seized by Alice and Bob. In order that their shares of the dollar approximate our cooperative prediction, the time interval between successive demands needs to be very small.

Core

What can be said about how coalitions form in more general situations? One proposal is that we should reject a payoff profile as a possible solution outcome if some coalition can object to it on the grounds that it is able to enforce an alternative payoff profile that all its members prefer. The set of all payoff profiles to which no such objection can be found is called the *core* of a cooperative game.

Economists like the idea because the core of a large enough market game approximates what will happen if buyers and sellers trade at whatever prices equate supply and demand. However, applying the idea to Odd-Man-Out in the case when all the players are risk neutral isn't very encouraging.

We have seen that one possible solution outcome in Odd-Man-Out is for Alice and Bob to form a coalition on the understanding that they will vote to split the dollar so that each gets 50 cents. But this outcome can't be in the core, because Bob and Carol can object that they are able to enforce an outcome that they both prefer by voting to split the dollar so that Bob gets 51 cents and Carol gets 49 cents. Since similar reasoning can be used to exclude any payoff profile whatever, the core of Odd-Man-Out is empty.

Condorcet paradox

The Marquis de Condorcet was an idealistic French revolutionary who discovered a similar problem when exploring possible voting systems. If Alice and Bob form a coalition that disadvantages Carol, she will offer whoever will listen a little more than they are currently getting. If Bob takes up Carol's offer and abandons Alice, then Alice will become the disadvantaged party, with an incentive to offer Carol a little more than she is currently getting. If Carol agrees, Bob will then approach Alice. And so on.

The results in real life can be devastating. For example, the border between England and Wales where I live was a battlefield for centuries. Powerful lords on the English side supposedly guarded the border or marches against raids by the Welsh tribes, but warfare was actually continuous as the Welsh, the King of England, and the local Marcher Lord shifted alliances to combine against whichever of the three was currently most powerful.

Condorcet's life didn't work out any better than the victims of the unstable social systems whose mechanics he succeeded in identifying. He had hoped to create a utopia by mathematical reasoning, but was sentenced to the guillotine instead.

Stable sets

Von Neumann and Morgenstern understood that Bob would be unwise to listen to Carol in Odd-Man-Out when she explains that he can get 51 cents by joining a coalition with her rather the 50 cents that Alice has promised him. If it is a good idea to dump Alice when he is approached by Carol, then it will be a good idea for Carol to dump him when she is approached by Alice.

To capture this idea, Von Neumann and Morgenstern invented a notion that is nowadays called a *stable set*. They argued that objections which aren't themselves possible solution outcomes should be ignored. Anything outside a stable set is still excluded because an objection from within the stable set can be found, but something inside a stable set need only be immune from objections within the stable set.

Their chief example was Odd-Man-Out when the players are all risk neutral. One stable set consists of the three possible outcomes in which the dollar is divided equally between two of the players. However, there are lots of other stable sets. For example, the set of all of outcomes in which Carol gets 25 cents and the rest of the dollar is split in all possible ways between Alice and Bob is stable.

It isn't easy to make sense of these new stable sets. Other game theorists disagree, but I think their appearance simply shows that the idea of a stable set isn't precise enough. So there are sometimes too many stable sets – but this is the least of our troubles. William Lucas found a cooperative game with many players that has no stable sets at all, and so there are also sometimes too few stable sets.

Shapley value

I was once summoned urgently to London to explain what the French government was talking about when it suggested that the costs of a proposed tunnel under the English Channel be allocated to countries in the European Union using the Shapley value. The latter is the brainchild of Lloyd Shapley, who was another of the brilliant group of graduate students who studied mathematics alongside John Nash at Princeton.

Shapley followed Nash's example by proposing a set of assumptions that define a unique prediction for the outcome of a cooperative game. However, unlike Nash, his assumptions apply not just to bargaining games with only two players, but to any cooperative game with 'transferable utility'. The leading case of interest is when the players are all risk neutral and the payoffs are measured in dollars. It can then be argued that everything that matters about a coalition is what I shall call the value of the coalition – the largest number of dollars that it can guarantee is available to be shared out among its members. These payoffs include any 'side payments' necessary to buy the loyalty of any member of the coalition who might think the grass looks greener elsewhere.

For example, in Odd-Man-Out, the value of each coalition with two players is one dollar. The value of the grand coalition of all three players is also one dollar. The value of a coalition with only one player is zero. The empty coalition with no players also has value zero.

The easiest way to find the Shapley value makes it explicit that it is intended as an *average* over all the possible ways that coalitions might form. Start with the empty coalition and add players until you get to the grand coalition. When Alice is added to a coalition, write down her marginal contribution to the coalition – the amount by which her inclusion increases the value of the coalition. The payoff assigned to Alice by the Shapley value is then the average of all her marginal contributions taken over all the possible ways in which the grand coalition can be assembled one player at a time.

Odd-Man-Out has three players, and so there are six ways of ordering the players: ABC, ACB, BAC, BCA, CAB, CBA. Alice's marginal contributions are respectively: 0, 0, 1, 0, 1, 0. So the Shapley value assigns Alice a payoff of 1/3 of a dollar, which is what we argued she would get on average in the previous section on coalitions.

How useful is the Shapley value? I think there is no doubt of its relevance to cost-sharing exercises of the type proposed by the French government, but it doesn't fare too well when tested by the Nash program. Like much else in game theory, there remains a great deal about coalition formation that we do not yet understand.

Chapter 10
Puzzles and paradoxes

Feedback phenomena and human intuition are uncomfortable bedfellows. When people dislike where an equilibrium argument takes them, it is therefore unsurprising that they invent simpler arguments that lead to more palatable conclusions. However, the first principle of rational thought is never to allow your preferences to influence your beliefs.

Fallacies of the Prisoner’s Dilemma

The fact that both players would be better off if they didn’t play their equilibrium strategies in the Prisoner’s Dilemma is said to be a paradox of rationality that requires resolution.

Categorical imperative

In colloquial language, Immanuel Kant’s categorical imperative says that it is rational to do what you wish everybody would do. If this were true, it would be rational to cooperate in the Prisoner’s Dilemma. But wishful thinking is never rational. It is a constant source of amazement to me that Kant is never held to account for proposing a rationality principle without giving any reasons why we should take it seriously.

Fallacy of the twins

Two rational people facing the same problem will necessarily choose the same action. So Alice and Bob will either both play *hawk* or both play *dove* in the Prisoner's Dilemma. Since Alice prefers the outcome (*dove*, *dove*) to (*hawk*, *hawk*), she should therefore choose *dove*.

The fallacy is attractive because it would be correct if Alice and Bob were genetically identical twins, and we were talking about what genetically determined behaviour best promotes biological fitness (see Kin selection in Chapter 8). But the relevant game wouldn't then be the Prisoner's Dilemma; it would be a game with only one player.

As is commonplace when looking at fallacies of the Prisoner's Dilemma, we are offered a correct analysis of the wrong game. The Prisoner's Dilemma is a two-player game in which Alice and Bob choose their strategies *independently*. The twins fallacy wrongly assumes that Bob will make the same choice as Alice whatever strategy she chooses. This can't be right, because Bob is supposedly rational and one of his two choices is irrational.

One can modify the assumptions of the fallacy so that Alice and Bob's strategies coincide only with some sufficiently high probability. The story told to justify such a correlation in their behaviour often kicks up enough dust to obscure the fact that any correlation at all implies that Alice and Bob aren't choosing independently. But if they don't choose independently, they aren't playing the Prisoner's Dilemma. Even if Alice and Bob's information were correlated, as hypothesized in Aumann's notion of a correlated equilibrium, they still wouldn't play *hawk*, because *hawk* is strongly dominated whatever the players may learn about other matters.

Myth of the wasted vote

A version of the twins fallacy is routinely trotted out at election time, when pundits argue that 'every vote counts' (see Mixed Nash equilibria, Chapter 2). If a wasted vote is one that doesn't affect the outcome of the election, then the only time that your vote can count is when only one vote separates the winner and the runner-up. If they are separated by two or more votes, then a change in your vote would make no difference at all to who is elected. However, an election for a seat in a national assembly is almost never settled by a margin of only one vote.

Here is a hypothetical example of an election even closer than the actual race between Bush and Gore in the United States in 2000. A reliable opinion poll says that the voters in a pivotal state who have made up their minds are split 51% to 49% in favour of Bush. The probability that a floating voter will go for Bush is just enough to ensure that he will beat Gore by 500 votes on average. Things look so close that Alice decides to vote. What are the chances that her vote will count – that the result would have been different if she had stayed home and watched the television?

With one million voters of whom 5% are undecided, Alice's vote would count only once in every 8,000 years, even if the same freakish circumstances were repeated every four years. But they won't be. The chances that the votes cast by floaters will almost balance those cast by the decided voters are infinitesimal. If the floaters in our example voted for Bush with the same frequency as the rest of the population, Alice's vote would count only once in every 20 billion billion years. No wonder no state has ever been decided by a single vote in a presidential election!

Naive folk imagine that to accept this argument is to precipitate the downfall of democracy. We are therefore told that you are wrong to count only the effect of your vote alone – you should instead count the total number of votes cast by all those people who think and feel as you think and feel, and hence will vote as you vote. If you have 10,000 such soul mates or *twins*, your vote wouldn't then be wasted, because the probability that an election will be decided by a margin of 10,000 votes or less is often very high. This argument is faulty for the same reason that the twins fallacy fails in the Prisoner's Dilemma. There may be large numbers of people who think and feel like you, but their decisions on whether to go out and vote won't change if you stay home and watch the television.

Critics sometimes accuse game theorists of a lack of public spirit in exposing this fallacy, but they are wrong to think that democracy would fall apart if people were encouraged to think about the realities of the election process. Cheering at a football game is a useful analogy. Few cheers would be raised if what people were trying to do by cheering was to increase the general noise level in the stadium. No single voice can make an appreciable difference to how much noise is being made when a crowd of people is cheering. But nobody cheers at a football game because they want to increase the general noise level. They shout words of wisdom and advice at their team even when they are at home in front of a television set.

The same goes for voting. You are kidding yourself if you vote because your vote has a significant chance of being pivotal. But it makes perfectly good sense to vote for the same reason that football fans yell advice at their teams. And, just as it is more satisfying to shout good advice rather than bad, so many game theorists think that you get most out of participating in an election by voting *as though* you were going to be the pivotal voter, even though you know the probability of one vote making a difference is too small to matter. A Kantian would assume that everyone is similarly strategic, but I prefer to use opinion polls when guessing the most likely way a tie might arise.

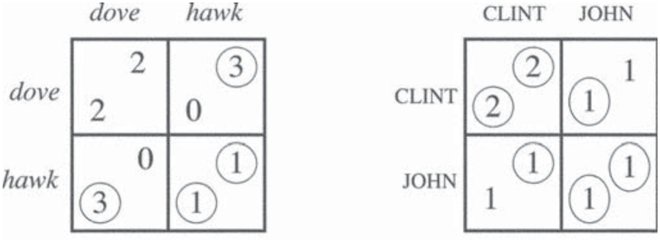
For example, Ralph Nader was the green candidate in the presidential election when Bush just beat Gore. I am hot on green issues, but I wouldn't have voted for Nader, because if there had been a tie, it would almost certainly have been between Bush and Gore. In Europe, such strategic voting will sometimes result in your voting for a minor party. The same pundits who tell you that every vote counts will also tell you that such a strategic vote is a wasted vote. But they can't be allowed to have it both ways!

Transparent disposition fallacy

This fallacy asks us to believe two doubtful propositions. The first is that rational people have the willpower to commit themselves in advance to playing games in a particular way. The second is that other people can read our body language well enough to know when we are telling the truth. If we truthfully claim that we have made an irrevocable commitment, we will therefore be believed.

If these propositions were correct, our world would certainly be very different! Charles Darwin's *Expression of the Emotions* would be wrong in denying that our involuntary facial muscles make it impossible to conceal our emotional state, and so actors would be out of a job. Politicians would be incorruptible. Poker would be impossible to play. Rationality would be a defence against drug addiction. However, the logic of game theory would still apply.

As an example, consider two possible mental dispositions called CLINT and JOHN. The former is a retaliating strategy named after the character played by Clint Eastwood in the spaghetti westerns (see Evolution of Cooperation, Chapter 8). The latter commemorates a hilarious movie I once saw in which John Wayne played the part of Genghis Khan. To choose the disposition JOHN is to advertise that you have committed yourself to play *hawk* in the Prisoner's Dilemma no matter what. To choose the disposition



Prisoner's Dilemma

Film Star Game

34. Transparent disposition fallacy

CLINT is to advertise that you are committed to play *dove* in the Prisoner's Dilemma if and only if your opponent is advertising the same commitment. Otherwise you play *hawk*.

If Alice and Bob are allowed to commit themselves transparently to one of these two dispositions, they won't be playing the Prisoner's Dilemma any more; they will be playing the Film Star Game of Figure 34 in which the players' strategies are CLINT and JOHN. If both players choose CLINT in the Film Star Game, they are then committed to playing *dove* in the Prisoner's Dilemma; otherwise they are committed to playing *hawk*.

As the circled payoffs show, CLINT is a (weakly) dominant strategy in the Film Star Game. So if Alice and Bob choose CLINT, they will be playing a Nash equilibrium that results in their cooperating in the Prisoner's Dilemma. Advocates of the transparent disposition fallacy think that this shows that cooperation is rational in the Prisoner's Dilemma. It would be nice if they were right that real-life games are really all Film Star Games of some kind – especially if one could choose to be Adam Smith or Charles Darwin rather than John Wayne or Clint Eastwood. But even then it wouldn't follow that rationality requires cooperating in the Prisoner's Dilemma. The argument shows only that it is rational to play CLINT in the Film Star Game.

Newcomb’s paradox

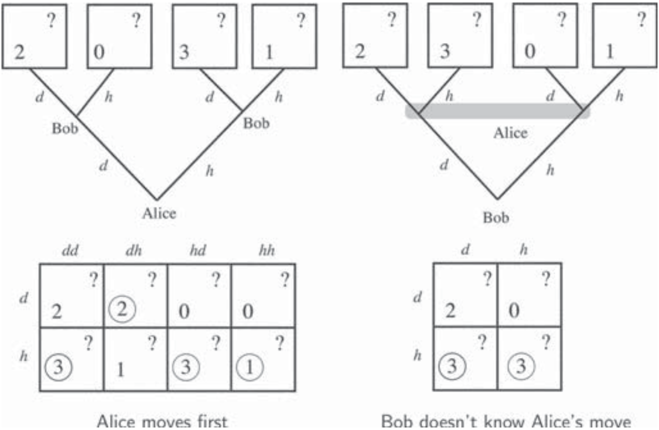
Two boxes possibly have money inside. Alice is free to take either the first box or both boxes. If she cares only for money, what should she do? This seems an easy problem. If *dove* represents taking only the first box and *hawk* represents taking both boxes, then Alice should choose *hawk* because she then gets at least as much money as with *dove*.

However, there is a catch. It is certain that the second box contains one dollar. The first box contains either two dollars or nothing. The decision about whether there should be money in the first box is made by Bob, who knows Alice so well that he is always able to make a perfect prediction of what she will do. Like Alice, he has two choices, *dove* and *hawk*. His dovelike choice is to put two dollars in the first box. His hawkish choice is to put nothing in the first box. His motivation is to catch Alice out. He therefore plays *dove* if he predicts that Alice will choose *dove*. He plays *hawk* if he predicts that Alice will choose *hawk*.

Choosing *hawk* doesn’t look so good for Alice now. If she chooses *hawk*, Bob predicts her choice and puts nothing in the first box, so that Alice gets only the single dollar in the second box. But if Alice chooses *dove*, Bob will predict her choice and put two dollars in the first box for her to pick up.

The Harvard philosopher Robert Nozick created a craze in his profession (aptly described as Newcombmania) by claiming that Newcomb’s paradox shows you can sometimes maximize your payoff by playing a strongly dominated strategy. He could equally well have argued that it shows 2 + 2 = 5, since anything can be deduced from a contradiction. The contradiction in Newcomb’s paradox consists in assuming the existence of a game in which:

1. Alice moves after Bob.
2. Bob knows Alice’s choice.
3. Alice has more than one choice.



	dd	dh	hd	hh
d	2 ?	② ?	0 ?	0 ?
h	③ ?	1 ?	③ ?	① ?

Alice moves first

	d	h
d	2 ?	0 ?
h	③ ?	③ ?

Bob doesn't know Alice's move

35. Two attempts to satisfy Newcomb's requirements. The information set in the right-hand game indicates that Alice doesn't know Bob's prediction. The payoff tables underneath each game tree are the relevant strategic forms

Figure 35 shows two attempts to create such a game without being specific about Bob's payoffs; the game on the left fails item 1 on the list, and that on the right fails item 2. We can satisfy both item 1 and item 2 by offering Alice only one choice in the right-hand game, but then we fall foul of item 3.

When arguing that Alice must play *dove* to maximize her payoff, Nozick assumes that Bob will play *dh* in the left-hand game. That is to say, Bob will predict *d* when Alice plays *d* and *h* when she plays *h*. However, Alice's strategy *d* isn't dominated in the left-hand game. To argue that Alice's strategy *d* is dominated, one has to appeal to the right-hand game. But it isn't paradoxical that Alice might play differently in different games.

One can muddy the waters by giving up the requirement that Bob can predict Alice's behaviour *perfectly*. We can then create a game in which the three requirements of Newcomb's paradox are satisfied by introducing chance moves into the right-hand game that remove Alice's opportunity to choose differently from Bob some of the time. But no amount of juggling with the parameters will make it optimal to play a strongly dominated strategy!

Surprise test paradox

The British telecom auction that raised $35 billion has been mentioned several times. Everybody was surprised at this enormous amount – except for the media experts, who finally got the figure roughly right by predicting a bigger number whenever the bidding in the auction falsified their previous prediction. Everybody can see the fraud perpetrated by the media experts on the public in this story, but the fraud isn't so easily detected when it appears in one of the many versions of the surprise test paradox, through which most people first learn of backward induction.

Alice is a teacher who tells her class that they are to be given a test one day next week, but the day on which the test is given will come as a surprise. Bob is a pupil who works backward through the days of the coming school week. If Alice hasn't set the test by the time school is over on Thursday, Bob figures that she will then have no choice but to set the test on Friday – this being the last day of the school week. So if the test were given on Friday, Bob wouldn't be surprised. Bob therefore deduces that Alice can't plan to give the test on Friday. But this means that the test must be given on Monday, Tuesday, Wednesday, or Thursday. Having reached this conclusion, Bob now applies the backward induction argument again to eliminate Thursday as a possible day for the test. Once Thursday has been eliminated, he is then in a position to eliminate Wednesday. Once he has eliminated all the days of the school week by this method, he sighs with relief and makes no attempt to study over the weekend. But then Alice takes him by surprise by setting the test first thing on Monday morning!

This isn't really a paradox at all, because Bob shouldn't have been so quick to sigh with relief. If the backward induction argument is correct, then Alice's two statements are inconsistent, and so at least one of them must be wrong. But why should Bob assume that the wrong statement is that a test will be given, and not that the test will come as a surprise? This observation is usually brushed aside, because what people really want to hear about is whether the backward induction argument is right. But what they should be asking is whether backward induction has been applied to the right game.

In the game that people imagine is being analysed, Eve chooses one of five days on which to hold the test, and Bob predicts which of the five days she will choose. If his prediction is wrong, then he will be taken by surprise. The solution of this version of Matching Pennies is that Alice and Bob both choose each day with equal probability. Bob is then surprised four times out of five.

This isn't the conclusion we reached before, because the surprise test paradox applies backward induction to a game in which Bob is always allowed to predict that the test will be today, even though he may have wrongly predicted that it was going to take place yesterday. In this bizarre game, Bob's optimal strategy is therefore to predict Monday on Monday, Tuesday on Tuesday, Wednesday on Wednesday, Thursday on Thursday, and Friday on Friday. No wonder Bob is never surprised by having the test occur on a day he didn't predict!

The surprise test paradox has circulated ever since I can remember. Occasionally it gets a new airing in newspapers and magazines. It has even been the object of learned articles in philosophical journals. The confusion persists because people fail to ask the right questions. One of the major virtues of adopting a systematic formalism in game theory is that asking the correct questions becomes automatic. You then don't need to be a genius like Von Neumann to stay on the right track. His formalism does the thinking for you.

Common knowledge

Why do we attach so much importance to eye contact? I think the reason is that something becomes common knowledge only if it is implied by an event that couldn't have occurred without everybody knowing it. For example, if Alice and Bob observe each other observing that Carol has a dirty face, then it becomes common knowledge between Alice and Bob that Carol has a dirty face. Similarly, when two people look each other in the eye, it becomes common knowledge between them that they are aware of each other as individuals.

Three old ladies

Alice, Beatrice, and Carol are three respectable ladies at a midwestern county fair. Each has a dirty face, but nobody is blushing, although a respectable lady who was conscious of appearing in public with a dirty face would surely do so. It follows

36. Three midwestern ladies

that none of the ladies knows that her own face is dirty, although each can clearly see the dirty faces of the others.

Midwestern clergymen always tell the truth, and so the ladies pay close attention when a local minister announces that one of the ladies has a dirty face. After his announcement, one of the ladies blushes. How come? Didn't the minister simply tell the ladies something they knew already?

To understand what the minister added to what the ladies already knew, we need to look at the chain of reasoning that leads to the conclusion that at least one of the ladies must blush. If neither Beatrice nor Carol blushes, Alice would reason as follows:

> *Alice:* Suppose that my face were clean. Then Beatrice would reason as follows:
>
> > *Beatrice:* I see that Alice's face is clean. Suppose that my face were also clean. Then Carol would reason as follows:
> >
> > > *Carol:* I see that Alice and Beatrice's faces are clean. If my face were clean, nobody's face would be dirty. But the minister's announcement proves otherwise. So my face is dirty, and I must blush.
> >
> > *Beatrice:* Since Carol hasn't blushed, my face is dirty. So I must blush.
>
> *Alice:* Since Beatrice hasn't blushed, my face is dirty. So I must blush.

So what did the minister add to what the ladies already knew? For Alice's reasoning to work, she needed to know that Beatrice knows that Carol knows that Alice and Beatrice know that someone has a dirty face. All these knowings became possible only after the minister's announcement makes it common knowledge that someone has a dirty face. It is then not only true that Alice, Beatrice, and Carol know that one of them has a dirty face;

they all know that they all know that they all know that they know it.

A coordination paradox

Is a magnificent beard necessary to make advances in interactive epistemology? The only evidence I have to offer is that the bearded Princeton philosopher David Lewis shares the credit for recognizing the importance of common knowledge in game theory with the equally hirsute Bob Aumann. But what are we to make of Lewis's claim that a convention can't be operational unless it is common knowledge that the players are planning to use it?

For something to become common knowledge, we need an equivalent of the tactless clergyman in the story of the three midwestern ladies. But no such clergyman is usually to be found. Nearly all the conventions we use in daily life therefore fail Lewis's test. So how come they seem to work so well?

Computer scientists worried about the implications for distributed systems illustrate the problem by telling a story about two Byzantine generals trying to coordinate an attack on an enemy army that lies in a valley between them, but I prefer a less dramatic example.

Alice and Bob want to get together tomorrow in New York. Alice emails the suggestion that they meet at Grand Central Station at noon. Bob emails a confirmation. This exchange would be adequate for most of us, but Lewis would object that the agreement isn't common knowledge because Bob doesn't know that Alice received his confirmation. She should therefore email to confirm that she received his confirmation. Bob should then email a confirmation of her confirmation, and so on. Since there is always a small probability that an email message won't be received, their attempt to agree on a convention will never become common knowledge.

But why should a convention have to be common knowledge to be operational? Ariel Rubinstein studied this question by analysing a new Email Game in which Alice and Bob's Meeting Game is replaced by the Stag Hunt Game of Chapter 4. The default convention is for Alice and Bob to play *dove* in the Stag Hunt Game, but every so often the labels of both their strategies get reversed, so that choosing *dove* will result in *hawk* actually being played. Only Alice observes when this happens. She sends an email message to Bob saying that they should play *hawk* on this occasion rather than *dove*. He automatically sends a confirmation. She automatically sends a confirmation of his confirmation, and so on.

A strategy in the Email Game says whether *dove* or *hawk* should be played depending on the number of messages a player has received. We can then short-circuit the common knowledge question by asking whether there is a Nash equilibrium of the Email Game in which Alice and Bob always succeed in coordinating on the equilibrium they both prefer in the Stag Hunt Game. Rubinstein's answer seems to confirm Lewis's intuition. The only Nash equilibrium in the Email Game in which Alice and Bob play *dove* when no message is sent requires that they *always* play *dove* no matter how many messages they may receive.

However, the picture changes when we allow Alice and Bob to choose whether or not to send or receive messages. The modified Email Game then has many Nash equilibria, the most pleasant of which requires that both players play *hawk* whenever Alice proposes doing so and Bob says OK – as when friends agree to meet in a coffee shop. But there are other Nash equilibria in which the players settle on *hawk* only after a long exchange of confirmations of confirmations. Hosts of polite dinner parties suffer from such equilibria when their guests start moving with glacial slowness towards the door at the end of the evening, stopping every inch or so in order that the host and the guest can repeatedly assure each other that departing at this time is socially acceptable to both sides.

The common-sense conclusion is that conventions don't need to be common knowledge to work. Most conventions are established by the forces of cultural evolution. Sometimes evolutionary stability considerations make it possible to eliminate some Nash equilibria. In the modified Email Game, one might hope that such considerations would eventually eliminate the equilibria that generate 'long goodbyes' after dinner parties, but the prognosis isn't good. Ironically, only Rubinstein's equilibrium, in which Alice and Bob play *dove* no matter what happens, fails to pass an appropriate evolutionary stability test.

Monty Hall problem

Alice is a contestant in an old quiz show run by Monty Hall. She must choose from three boxes, only one of which contains a prize. Monty knows which box contains the prize, but Alice doesn't. After she chooses *Box 2*, Monty opens one of the other boxes that he knows to be empty. Alice then has the opportunity to change her mind about her choice of box. What should she do?

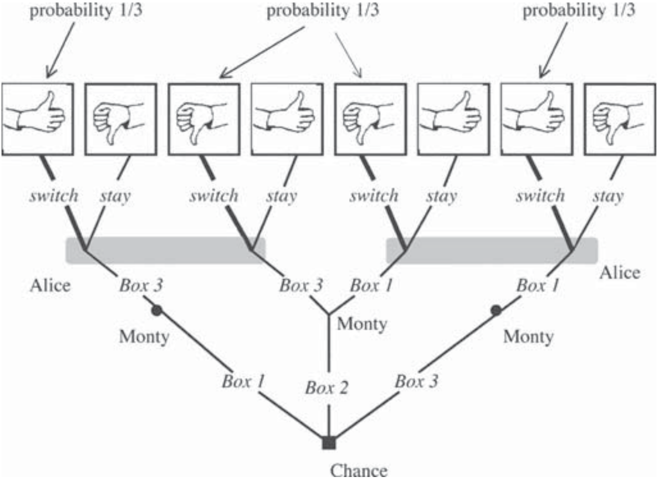
People usually say it doesn't matter. They reason that Alice's probability of winning when she chose *Box 2* was 1/3 because there was then an equal chance of the prize being in any of the three boxes. After another box is shown to be empty, the probability that *Box 2* contains the prize goes up to 1/2, because there is now an equal chance that the prize is in one of the two unopened boxes. If Alice switches boxes, her probability of winning will therefore still be 1/2. So why bother changing?

Marilyn Vos Savant apparently has the highest IQ ever recorded. When she explained in *Parade* magazine that Alice should always switch boxes, various self-appointed mathematical gurus laughed her to scorn, but she was right.

The probability that the prize is either in *Box 1* or *Box 3* is 2/3. If she switches to whichever of these boxes isn't opened, Alice will therefore win with probability 2/3.

This argument is deceptively easy. Even top mathematicians sometimes fail to see why Monty's action conveys so much information to Alice. After all, it wouldn't have conveyed any useful information at all if he had opened a box at random that just happened to be empty – but he deliberately chose a box that he knew to be empty.

However, you don't need to have the highest IQ ever recorded to get the answer right if you are willing to let Von Neumann do your thinking for you. Figure 37 shows the game that Alice and Monty



37. The Monty Hall Game. Only Alice's payoffs are shown. The chance move is shown as a square. Alice's information sets show that she doesn't know which box contains the prize, but she does know which box Monty opens. Her switching choice is thickened. The figure shows that whatever strategy Monty chooses, Alice wins with probability 2/3 if she switches

are playing. It doesn't matter what Monty's payoffs are, but we might as well assume that he wants Alice to lose. A chance move first puts the prize in one of the boxes. Monty then decides whether to open *Box 1* or *Box 3*. (He only has a genuine choice when the prize is actually in *Box 2*.) Alice then chooses whether to stay with *Box 2* or to switch to whichever of *Box 1* or *Box 3* Monty didn't open.

There is now no need to think at all. If Alice always switches, the figure makes it impossible not to recognize that she wins when the prize is in *Box 1* or *Box 3* and loses when the prize is in *Box 2*. So she wins with probability 2/3.

References and further reading

Chapter 1

Ken Binmore, *Playing for Real* (New York: Oxford University Press, 2007). This textbook on game theory is light on mathematics.

Ken Binmore, *Natural Justice* (New York: Oxford University Press, 2005). Why game theory matters in ethics.

Colin Camerer, *Behavioral Game Theory* (Princeton: Princeton University Press, 2003). Some aspects of game theory work well in the laboratory, and some don't. This book surveys the evidence, and looks at possible psychological explanations of deviations from the theory.

John Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1972). Rawls is often said to be the greatest moral philosopher of the 20th century. He refers to the maximin principle as the difference principle when proposing it as a rational substitute for maximizing average utility.

John Maynard Smith, *Evolution and the Theory of Games* (Cambridge: Cambridge University Press, 1982). This beautiful book introduced game theory to biology.

Barry Nalebuff and Avinash Dixit, *Thinking Strategically* (New York: Norton, 1991). A book-club choice, it contains many examples of game theory in action, both in business and in everyday life.

Sylvia Nasar, *A Beautiful Mind* (New York: Simon and Schuster, 1998). A best-selling biography of John Nash.

Alvin Roth and John Kagel, *Handbook of Experimental Game Theory* (Princeton: Princeton University Press, 1995). The survey by John Ledyard documents the immense amount of data supporting the claim that experienced subjects seldom cooperate in the Prisoner's Dilemma.

John Von Neumann and Oskar Morgenstern, *The Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944). Not a best-seller. Their theory of utility appears in an appendix.

Chapter 2

Robert Aumann, *Lectures on Game Theory* (Boulder, CO: Westview Press Underground Classics in Economics, 1989). The classroom notes of one of the great game theorists.

Ken Binmore, *Does Game Theory Work?* (Cambridge, MA: MIT Press, 2007). This book includes my own experiment on zero-sum games and references to others.

Steve Heine, *John von Neumann and Norbert Wiener* (Cambridge, MA: MIT Press, 1982). I write 'Von Neumann' rather than 'von Neumann' because one gets into trouble in some parts of the German-speaking world for according him the title that his father bought from the Hungarian government.

J. D. Williams, *The Compleat Strategyst* (New York: Dover, 1954). A delightful collection of simple two-person, zero-sum games.

Chapter 3

Robert Aumann, 'Interactive Epistemology', *International Journal of Game Theory*, 28 (1999): 263–314.

Martin Gardner, *Mathematical Diversions* (Chicago: University of Chicago Press, 1966) and *Hexaflexagons* (Chicago: University of Chicago Press, 1988). These books gather together many delightful games and brainteasers from the author's long-standing column in *Scientific American*.

Robert Gibbons, *Game Theory for Applied Economists* (Princeton: Princeton University Press, 1992). An unfussy introduction to game theory, with an orthodox treatment of refinements.

David Lewis, *Counterfactuals* (Cambridge, MA: Harvard University Press, 1973).

Larry Samuelson, *Evolutionary Games and Equilibrium Selection* (Cambridge, MA: MIT Press, 1997). This includes our paper on the replicator dynamics in the Ultimatum Game.

Chapter 4

Steven Brams, *Superior Beings: If They Exist, How Would We Know? Game Theoretic Implications in Omniscience, Omnipotence, Immortality and Comprehensibility* (New York: Springer Verlag, 1983).

John Harsanyi and Reinhard Selten, *A General Theory of Equilibrium Selection in Games* (Cambridge, MA: MIT Press, 1988).

David Hume, *A Treatise of Human Nature* (Oxford: Clarendon Press, 1978; first published 1739). Arguably the greatest work of philosophy ever.

David Lewis, *Conventions* (Princeton: Princeton University Press, 1969).

Thomas Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960). Schelling once bravely told a large audience of game theorists that game theory had contributed nothing whatever to the theory of focal points – except perhaps the idea of a payoff table!

Thomas Schelling, *Micromotives and Macrobehavior* (New York: Norton, 1978). Schelling's Solitaire and a lot more.

Brian Skyrms, *The Stag Hunt and the Evolution of the Social Structure* (Cambridge: Cambridge University Press, 2003).

Peyton Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton: Princeton University Press, 1998).

Chapter 5

Bob Axelrod, *Evolution of Cooperation* (New York: Basic Books, 1984). This book sold the world on the idea that reciprocity matters.

'Review of *The Complexity of Cooperation* by Ken Binmore', *Journal of Artificial Societies*, http://jasss.soc.surrey.ac.uk/1/1/review1.html. The book is a sequel to Axelrod's *Evolution of Cooperation*; the review assesses his reiterated claims for TIT-FOR-TAT. See also Karl Sigmund's *Games of Life* (Chapter 8 below).

Joe Heinrich *et al.* (eds), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (New York: Oxford University Press, 2004). An attempt to refute the repeated-game explanation of social norms that backfired. The paper by the anthropologist Jean Ensminger is particularly instructive.

George Mailath and Larry Samuelson, *Repeated Games and Reputations: Long-Term Relationships* (New York: Oxford University Press, 2006). Folk theorems with imperfect monitoring for mathematicians.

Bob Trivers, *Social Evolution* (Menlo Park, CA: Cummings, 1985). Reciprocity and much else in animal societies.

Chapter 6

Helena Cronin, *The Ant and the Peacock* (Cambridge: Cambridge University Press, 1991).

John Harsanyi, *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations* (Cambridge: Cambridge University Press, 1977).

Roger Myerson, *Game Theory: Analysis of Conflict* (Cambridge, MA: Harvard University Press, 1991).

Chapter 7

Ken Binmore and Paul Klemperer, 'The Biggest Auction Ever: The Sale of British 3G Licences', *Economic Journal*, 112 (2002): C74–C96.

R. Cassady, *Auctions and Auctioneering* (Berkeley, CA: University of California Press, 1967). Lots of good stories.

Paul Klemperer, *Auctions: Theory and Practice* (Princeton: Princeton University Press, 2004).

Paul Milgrom, *Putting Auction Theory to Work* (Cambridge: Cambridge University Press, 2004).

Chapter 8

John Alcock, *The Triumph of Sociobiology* (Oxford: Oxford University Press, 2001). Sociobiologists aren't the intellectual fascists they have been painted. Aside from offering wonderful examples of real sociobiology in action, this book lays bare the dishonest campaign of vilification directed at Edward Wilson and his followers by Gould, Lewontin, and other politically motivated polemicists.

Ken Binmore and Larry Samuelson, 'Evolutionary Stability in Repeated Games Played by Finite Automata', *Journal of Economic Theory*, 57 (1992): 278–305.

Richard Dawkins, *The Selfish Gene* (Oxford: Oxford University Press, 1976). One of the great works of popular science.

Peter Hammerstein, *Genetic and Cultural Evolution of Cooperation* (Cambridge, MA: MIT Press, 2003).

William Hamilton, *The Narrow Roads of Geneland* (Oxford: Oxford University Press, 1995). A collection of some of Bill Hamilton's path-breaking papers in evolutionary biology. The papers themselves are not easy reading for the general reader, but the linking remarks are a fascinating social commentary on how it was to be a graduate student in the old days, doing work so original that the academic establishment was unable to appreciate its value.

John Maynard Smith, *Evolution and the Theory of Games* (Cambridge: Cambridge University Press, 1984). Many wonderful examples.

Karl Sigmund, *Games of Life: Explorations in Ecology, Evolution and Behaviour* (Harmondsworth: Penguin Books, 1993). Among other delights, this book reports on some of the author's computer simulations with Martin Nowack. Their name for TIT-FOR-TAT is PAVLOV (see Chapter 5).

James Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA* (New York: Touchstone, 1968).

Vero Wynne-Edwards, *Animal Dispersion in Relation to Social Behaviour* (Edinburgh: Oliver and Boyd, 1962).

Chapter 9

Ken Binmore, *Playing for Real* (New York: Oxford University Press, 2007). Four chapters are devoted to bargaining issues.

Ken Binmore, *Natural Justice* (New York: Oxford University Press, 2005). This book explains why I side with Rawls rather than Harsanyi on the implications of using the original position to make fairness judgements.

Roger Fisher *et al.*, *Getting to Yes* (London: Houghton Mifflin, 1992). This best-seller argues that good bargaining consists of insisting on a fair deal. Thinking strategically is dismissed as a dirty trick!

Howard Raiffa, *The Art and Science of Negotiation* (Cambridge, MA: Harvard University Press, 1982).

Chapter 10

Ken Binmore, *Playing Fair: Game Theory and the Social Contract I* (Cambridge, MA: MIT Press, 1995). Chapter 3 discusses more fallacies of the Prisoner's Dilemma that circulate in the philosophical literature.

Bob Frank, *Passions with Reason* (New York: Norton, 1988). An economist makes a case for the transparent disposition fallacy.

David Lewis, *Conventions: A Philosophical Study* (Cambridge, MA: Harvard University Press, 1969).

J. E. Littlewood, *Mathematical Miscellany* (Cambridge: Cambridge University Press, 1953). I was a schoolboy when I first came across the paradox of three old ladies in this popular work by one of the great mathematicians.

Index

LOGIC

A Very Short Introduction

Graham Priest

Logic is often perceived as an esoteric subject, having little to do with the rest of philosophy, and even less to do with real life. In this lively and accessible introduction, Graham Priest shows how wrong this conception is. He explores the philosophical roots of the subject, explaining how modern formal logic deals with issues ranging from the existence of God and the reality of time to paradoxes of self-reference, change, and probability. Along the way, the book explains the basic ideas of formal logic in simple, non-technical terms, as well as the philosophical pressures to which these have responded. This is a book for anyone who has ever been puzzled by a piece of reasoning.

‘a delightful and engaging introduction to the basic concepts of logic. Whilst not shirking the problems, Priest always manages to keep his discussion accessible and instructive.’

Adrian Moore, St Hugh’s College, Oxford

‘an excellent way to whet the appetite for logic. . . . Even if you read no other book on modern logic but this one, you will come away with a deeper and broader grasp of the *raison d’être* for logic.’

Chris Mortensen, University of Adelaide

www.oup.co.uk/isbn/0-19-289320-3

DARWIN

A Very Short Introduction

Jonathan Howard

Darwin's theory of evolution, which implied that our ancestors were apes, caused a furore in the scientific world and beyond when *The Origin of Species* was published in 1859. Arguments still rage about the implications of his evolutionary theory, and scepticism about the value of Darwin's contribution to knowledge is widespread. In this analysis of Darwin's major insights and arguments, Jonathan Howard reasserts the importance of Darwin's work for the development of modern science and culture.

> 'Jonathan Howard has produced an intellectual *tour de force*, a classic in the genre of popular scientific exposition which will still be read in fifty years' time'
>
> ***Times Literary Supplement***

www.oup.co.uk/isbn/0-19-285454-2