GRESHAM COLLEGE
Founded 1597

When Maths Doesn't Work: What we learn from the Prisoners' Dilemma
Transcript

Date: Monday, 16 February 2015 - 6:00PM

Location: Barnard's Inn Hall

16 February 2015

When Maths Doesn't Work: What We Learn from the Prisoners' Dilemma

Dr Tony Mann

Good evening.

This is the second of my three lectures this term on paradoxes in mathematics and computing. In March I shall be talking about a recently-discovered paradox about mathematical games, discovered by a quantum physicist, which astonished mathematicians. This month my lecture is about another mathematical paradox – an example, called the Prisoners' Dilemma, which at one time presented a serious problem for the relatively new mathematical subject of game theory, and how (with the help of computers) our understanding has developed so that instead of being an embarrassment, this example is now at the core of our research into fundamental aspects of human life like co-operation, reputation and trust.

For me, this is a very personal story, not because I have personally carried out research in this area, but because it has fascinated me ever since I first came across it. And so I intend to present this story in a very personal way: I'm going to tell you how I first came across this paradox, and how my view of it has changed over the course of my life.

I first came across game theory when I was seventeen. I was working for the engineering company Ferranti before going to university (writing systems software in machine code – a fascinating and valuable opportunity) and the department had what it grandly called a library – which was simply a shelf of old maths books which someone had donated. One of them was called *Theory of Games and Economic Behaviour*, by John von Neumann and Oskar Morgenstern. At the time I had no idea of the importance of this 1944 book, nor of the eminence of its authors. But I loved playing chess and bridge, so I naturally picked this book out to read in my lunch break.

But I was disappointed. The book doesn't help much with games like chess. To play the perfect game of chess all one has to do is list all the possible opening moves, then for each one list all possible responses, and all possible responses to these, and so on. There are only finitely many possible games, and at any stage one simply chooses a move which always leads to a win, if there is one, or if not, to a draw if that is possible. There is little mathematical interest in this analysis, which is just brute force calculation, and, of course, this is not a very practical solution!

Nor does the book help with bridge, which it describes as a two-player game (each partnership being regarded a single player, albeit one without full information since neither half of the partnership knows what cards their partner holds). But since the most fascinating part of bridge for me was the bidding systems that allow the two halves of a partnership to communicate their information, viewing bridge in this way seems to discard what makes the game particularly interesting.

The book contained a lot about poker, which I didn't play, and it seemed to be making analogies between poker and economic decisions, but I found it hard to follow, so I cast it aside, preferring to spend my lunch hour playing a number-guessing game that someone had programmed into our mini-computer.

So I was initially disappointed by the book by von Neumann and Morgenstern, not realising how significant it was nor how the mathematics it introduced to the world, Game Theory, would fascinate me for the next forty years (so far). But my failure to appreciate it didn't stop me, a year or so later, attending a lecture on Game Theory given to our undergraduate mathematics society by the eminent number theorist (and bridge international and strong chess player) Sir Peter Swinnerton-Dyer.

I found Sir Peter's lecture utterly absorbing (perhaps because I had been primed by my attempt to read the von Neumann-Morgenstern book). He took us through the basic ideas of game theory. I don't remember many of his specific examples, but it was something along the following lines.

We are interested in “games” between two (or sometimes more) players, where the outcome depends on choices made simultaneously by the players. So these are not games like chess, where the players play alternately and each has perfect information about the current position, but rather games like “Rock, Paper, Scissors” where players make simultaneous choices, knowing what the opponent’s options are, but not knowing which choice they will make. And the term “game” includes any scenario in which players interact in this way – so negotiation, trading, and decision-making and all sorts of non-recreational situations are classed as “games” and are covered by this analysis.

In such a game we can draw up a table showing the relationship between the choices and the outcome. My table will have a row for each of my possible choices or actions, and a column for each of my opponent’s options. The entry in the table will show the result for me corresponding to these choices by me and by my opponent. So for Rock, Paper, Scissors, the table would look like this (remember the outcomes are from my perspective, not my opponent’s):

	Opponent’s choice		
My choice	Rock	Paper	Scissors
Rock	Draw	Lose	Win
Paper	Win	Draw	Lose
Scissors	Lose	Win	Draw

My opponent will have a similar table showing the results from their point of view.

Another example might be the decisions made when a penalty kick is taken in a football match. The striker decides (in a simplified model) to place the ball to the left or right. The goalkeeper can choose to go left or right, or to wait to see where the striker aims, in which case he has less chance of saving a well-struck penalty but more chance of saving a weak one. The likelihoods of scoring might be as follows – the number in each cell is the probability that the striker will score (these numbers are made up for illustrative purposes only and are not based on research!)

	Goalkeeper’s choice		
Striker’s choice	Goes left	Goes right	Waits
Left	0.95	0.4	0.7
Right	0.4	0.95	0.7

These figures mean that if the goalkeeper guesses the right way to go, he is likely to save the penalty, but if he guesses wrong the striker will almost certainly score (he might hit the post or bar or shoot wide, so it’s not quite a certainty).

In both these “games”, if either player is able to predict correctly what the other will do, then they have a big advantage. If I know my opponent always chooses paper, I will choose scissors and win. If the goalkeeper knows that a player always shoots right, then by going left he can maximise his chance of saving the penalty: if the striker knows that the goalkeeper always goes right, then by shooting right he can maximise his chance of scoring. It’s intuitively obvious, therefore, that both players should vary their tactics. But how? If my opponent always chooses rock, then scissors, then paper, then rock again, and so on, then I can spot the pattern and win every time. So how should the players vary their tactics? Any pattern might be worked out by the opponent, so a random strategy of some sort is necessary, and the mathematics of game theory allows one to work out how often this random strategy should choose each option.

How does this maths work? Let’s analyse the Rock, Paper, Scissors example. Now, if I am optimistic, I may think that I can outguess my opponent and predict what they are going to do. But do I have any cause for optimism? Mathematicians are naturally pessimistic. If we assume the worst, so that, if it is possible for them to do so, our opponents will always play optimally to exploit our strategies, then what we want to do is to find the

strategy which they can exploit least effectively. In other words, we want to find the strategy that, whatever our opponent may do, provides the best result for us. Being pessimists, we might assume that we are facing a possible overall loss, and we want to minimise the maximum possible expected loss.

Let's assign the reasonable values of +1 for a win, 0 for a draw and -1 for a loss, so our table for Rock, Paper, Scissors now looks like this:

My choice	Opponent's choice		
	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

Suppose that I decide to play each of the three options with probabilities ρ for rock, π for paper and σ for scissors (where the sum of these three probabilities is 1). What can my opponent do?

Well, suppose of my three probabilities that ρ is the biggest, so that I play rock more often than either of the others. Then what happens if my opponent plays paper on every round? We will win when we choose scissors – probability σ – and lose when we choose rock – probability ρ – and we will draw the remainder. Our expected result is 1xσ – 1xρ, and since we assumed ρ was the biggest of the three probabilities, this quantity is negative. On average we will lose.

What this analysis tells us is that if our strategy allows them to do so, our opponent can ensure a negative outcome for us by choosing to play every time the antidote to our most common choice. The only way we can achieve a non-negative expected outcome is by choosing each option with equal likelihood, in which case our expected outcome is 0, whatever our opponent does! (This requires selecting each play with probability 1/3, so I might choose by throwing a dice, hidden from my opponent, and playing rock if it comes up 1 or 2, paper if it is 3 or 4, and scissors if it is 5 or 6.) So in the sense of game theory this is our optimal strategy.

We could find a similar optimal strategy for the striker and one for the goalkeeper in my football penalty example. For a fuller analysis than mine, which takes into account that right-footed players do better shooting to the left and details like that, see http://williamspaniel.com/2014/06/12/the-game-theory-of-soccer-penalty-kicks/.

(Incidentally, computers are possibly better than humans both at choosing options at random and at spotting patterns in their opponent's choices. If you want to try your hand at rock, paper, scissors against an experienced computer, go to http://www.nytimes.com/interactive/science/rock-paper-scissors.html.)

Rock, paper, scissors can be extended – the French game "pierre, papier, ciseaux, puits" adds a well into which rock and scissors fall and lose, while paper covers the well and wins. This removes the symmetry and makes the game somewhat more interesting mathematically: rock and scissors each beat only one of the other three, while well and paper beat two. Since well obtains the same results against scissors and paper as rock does, and well beats rock, why would one ever choose rock? And if neither player ever chooses rock, we are essentially back to the original three-weapon game "well, paper, scissors". A German version adds bull, which drinks the well, eats paper, but is stabbed by scissors and crushed by rock.

Those of you who watch *The Big Bang Theory* will be familiar with another extended version of the game, "rock-paper-scissors-lizard-Spock", created by Sam Kass and Karen Bryla and referencing *Star Trek*. Here Spock smashes scissors and vaporizes rock, so Spock defeats scissors and rock; but he is poisoned by lizard and disproven by paper so lizard and paper defeat Spock. Similar logic defines the outcomes for lizard.

This one appears to be symmetric, and looks very similar to the original three-weapon game, but with less likelihood of a draw. However there is a twist: some players have difficulty making the Spock symbol – the Vulcan salute with the thumb separate, the fore and middle fingers together, and the ring finger and pinkie together! So this physical consideration will change the mathematics!

One common feature of all of the games I have mentioned so far is that, what one player gains, the other player loses. If the striker scores, the goalkeeper concedes. If I choose Spock and my opponent chooses rock, I win and they lose. We call these "zero-sum" games – if we add up the outcomes for the two players we get zero, since one player's winnings exactly balance out the other's loss. But not all games in Game Theory are like this.

Sir Peter Swinnerton-Dyer finished his lecture by describing a game called the Prisoners' Dilemma. Here is the scenario as I remember it from his talk.

It is set in the Wild West. Two cowboys are arrested by the sheriff who is seeking to punish someone for a bank robbery. For the purposes of the discussion it doesn't matter whether our cowboys are guilty or not, and the sheriff doesn't care either. But the sheriff has no evidence, so to get a conviction he needs one of the cowboys to implicate the other. He arrests them both, puts them separately in solitary confinement, and offers each of them the following deal.

"If neither of you admit to the crime, then I will pin some other offence on you and you will both go to jail for two years. But if you admit that you did it together, and your partner denies it, then I'll use your evidence to convict him and he will go to jail for 20 years while, in return for your evidence, you can be out after six months. Similarly, if he admits it and you don't, he'll get six months and you'll get twenty years. If you both admit to the crime, then your confession isn't worth so much to me and you'll both go to prison for fifteen years."

Now both cowboys know that the other is being offered the same deal, but they cannot communicate with each other. What should they do? We can consider this a game, in the general sense, between the two cowboys, in which the options for each are to Co-operate (with each other, not with the sheriff) by refusing to admit to the robbery, or to Defect on the other by admitting to the robbery and implicating the other. And we can draw up a table showing the pay-offs:

	My partner's choice	
My choice	C (Co-operate)	D (Defect)
C (Co-operate)	-2	-20
D (Defect)	-0.5	-15

Here the entries represent the time to be served in prison, in years, and they are negative because game theorists like to find the maximum payoff: so expressing the prison terms as negative numbers means that higher numbers represent better outcomes. Note that this is not a zero-sum game: the sum of the pay-offs can range from -4 (if both co-operate) to -30 (if both defect).

Suppose I am in this position. What should I do? If I were a hardened criminal, I wouldn't betray my partner and I could rely on them not to betray me. But what if I am a mathematician? Let's apply some game theory. If we look at this table, we can see that every outcome in the second row (D) is better for me than the corresponding outcome in the first row (C). In the technical jargon, we say D "dominates" C. In other words, whatever my partner does, I will do better if I defect than if I co-operate. I don't know what my partner is going to do, but whatever they do, I will get a shorter sentence if I defect. Logically, it's a no-brainer. So our mathematical cowboy will defect, and so, applying the same logic, will their mathematical partner.

This is rather unfortunate. Our two mathematicians, following impeccable mathematical logic, both defect, and get fifteen years in prison each, while in the next town, two cowboys in the same position, who had the good fortune to be untrained in mathematics, both co-operate and get out after two years. This isn't a good advert for the value of mathematics – in this instance mathematics has got the two of us an extra thirteen years in jail!

Swinnerton-Dyer finished his lecture (as my perhaps not very reliable memory has it) by saying that this paradox was the end of game theory. As a branch of mathematics, it couldn't survive this disastrous paradox. Applying mathematics should lead to the best outcome, not the worst! For non-zero-sum games, the mathematics failed, and mathematicians abandoned this previously promising subject. It was a very anti-climactic ending to an otherwise inspiring lecture. (I now wonder if, in fact, Sir Peter was quite as negative about game theory as my memory records.)

The Prisoners' Dilemma fascinated me. Following this lecture I started seeing Prisoners' Dilemmas everywhere. At that time of high inflation and industrial unrest, some politicians were accusing trade unionists of irrationality in making high pay demands that would, it was suggested, fuel inflation and leave them worse off than if they had moderated their pay claim. But it seemed to me that even if the dubious claimed connection between pay and inflation were true, the unions were in a Prisoners' Dilemma situation and "defecting" by seeking a big pay increase was the mathematically correct behaviour. Another example was a scare over vaccination against whooping cough. Some people believed there was a small risk of adverse reaction to whooping cough vaccine. If every-one else has been vaccinated, then there will be nobody I can catch the disease from, and I can avoid the risk of side-effects by not being vaccinated. But if a lot of people think that way, then rates of vaccination will fall and we will all be at greater risk – indeed there were two whooping-cough epidemics in the 1970s following a decline in vaccination, showing how the Prisoners' Dilemma could have serious consequences in real life. These examples are over-simplified, but the depressing point is that the mathematics appears to show that even if everyone makes a sensible choice we can't always achieve the outcome that is the best possible for everyone.

I later discovered that these are variations of a scenario called "The Tragedy of the Commons". Herders were allowed to graze their cows or sheep on common land. By grazing one more animal, a herder derived benefit while the cost is shared by all. Logically, each herder was incentivised to graze as many animals as possible. There was nothing to be gained by an individual herder by not grazing every animal they could, and something to be lost. The inevitable tragic consequence was that the land was over-grazed and the common resource was lost. This, too, is a form of Prisoners' Dilemma.

So what does the Prisoners' Dilemma tell us? Perhaps that maths doesn't work. It might appear also to indicate that, left to themselves, people will not do what is necessary for the common good: if they behave logically, selfishness will rule and possible benefits are lost. Do we need a nanny state to dictate what we must do to ensure the best outcomes for everyone?

There are many other examples of the Prisoners' Dilemma, and I'm going to present two, of rather different degrees of seriousness.

First, the nuclear arms race in the Cold War. We have two superpowers. Each can choose to devote enormous resources to developing vast arsenals of nuclear weapons or to devote the money, research and labour to hospitals and schools and other good things. We can draw up a table:

	The other superpower's choice	
Our choice	Use resources for other things	Expand nuclear arsenal
Use resources for other things	Military stalemate: lots of social benefits	Military disadvantage
Expand nuclear arsenal	Military advantage	Military stalemate: less

		social benefit

Since being militarily weaker than the other superpower would be disastrous, and being stronger would be politically useful, the entries in the bottom row of the matrix all dominate the row above, and the rational choice for the superpower leader is to build more nuclear weapons. Of course, the leader of the other superpower, being equally rational, makes the same choice. The inevitable result is a world in which resources always go into swords rather than ploughshares.

The second example comes from the world of opera – specifically Puccini's *Tosca.* In the crucial scene of the opera, the heroine, the primadonna Tosca, is negotiating with the evil police chief Scarpia for the life of her lover the rebel Mario, who has been sentenced to death. They make an agreement: Tosca will sleep with Scarpia if he arranges that the firing squad will use fake bullets so that Mario can play dead and then make his escape. Tosca is prepared to make this sacrifice to save the life of her lover.

However each party has an opportunity to defect on this agreement. Scarpia can write an order which does not specify fake bullets, while Tosca, after Scarpia has sent off his order, can stab Scarpia to death with a convenient knife. Here's the pay-off table for this game, from Tosca's viewpoint:

	Scarpia's choice	
Tosca's choice	Co-operate (fake bullets)	Defect (real bullets)
Co-operate (don't kill Scarpia)	Mario is free but Tosca has to have sex with Scarpia – OK outcome	Mario is dead and Tosca has to have sex with Scarpia – worst outcome
Defect (kill Scarpia)	Mario is free and Tosca doesn't have sex with Scarpia – excellent outcome	Mario is dead but Tosca doesn't have to have sex with Scarpia – bad outcome

And here's the corresponding table from Scarpia's perspective:

	Tosca's choice	
Scarpia's choice	Co-operate (don't kill Scarpia)	Defect (kill Scarpia)
Co-operate (fake bullets)	Scarpia lives and has sex with Tosca but frees Mario – OK outcome	Scarpia is dead and Mario is free – worst possible outcome
Defect (real bullets)	Scarpia lives and has sex with Tosca, and Mario is killed – excellent outcome	Scarpia is dead but at least Mario is dead too – pretty bad outcome

Now, each of these tables reveals a characteristic Prisoners' Dilemma. For both Scarpia and Tosca, each cell in the Defect row is better than the corresponding one in the Co-operate row. Mathematically, neither has any rational option but to defect.

What happens in the opera? Sadly, both characters turn out to be good mathematicians and they both defect –

Tosca kills Scarpia and then finds, in the moving finale, that the chief of police has reneged on the deal and the firing squad used real bullets. Furthermore, this isn't a one-off choice: every time I listen to my CD of *Tosca* both Tosca and Scarpia make exactly the same decisions. And it could all have ended happily if only they had both co-operated! (Mind you, it wouldn't have been a very interesting opera.)

Now, the first of these examples may offer genuine insights into the economics of the Cold War arms race, but does game theory really help us understand the opera? The essence of the Prisoners' Dilemma is that both parties know that the other is in the same position and are aware of the options available to the other. But in *Tosca* that isn't the case. The heroine's happiness as she watches her lover being shot with what she wrongly thinks are fake bullets doesn't suggest that she has considered the possibility that Scarpia might go back on his word. And if Scarpia had considered that Tosca might stab him, he quite possibly wouldn't have left the knife on his desk for her to use. So this example (which is due to the eminent mathematical psychologist Anatol Rapoport, of whom we will hear more shortly) isn't to be taken too seriously (and I'm sure Rapoport didn't intend that it should). Incidentally, Rapaport, rather ironically for a leading game theorist, was apparently very good at chess but not so at poker.

Game theory may have been initially treated with suspicion by mathematicians because of the paradox of the Prisoners' Dilemma, as I think I remember being told in the undergraduate lecture, but it was taken up by economists: indeed twelve game theorists have won the Nobel Prize for Economics. It has been used by Steven J. Brams to analyse the behaviour of politicians, characters in literature and even the relationship between God and his creations, and recently Michael Suk-Young Chwe has argued (not, for me, entirely convincingly) that Jane Austen's novels are a systematic exploration of game theory, 150 years before von Neumann and Morgenstern wrote their book. And between 2007 and 2009 a version of the Prisoners' Dilemma featured on the TV game show *Golden Balls*, hosted by Jasper Carrott: two contestants had to choose whether to "split" or "steal" their jackpot. If they both split, each got half: if one chose to steal and the other split, the stealer took home all, and if both chose to steal, they both took home nothing. We can draw up the table:

	Split	Steal
Split	Half of Jackpot	0
Steal	Total Jackpot	0

We see that the "Steal" choice dominates "Split" – whatever the other player does, the outcome is better if we Steal. But the same logic applies to the other player, so if we both choose rationally, we both go home with nothing, when we could have had half the jackpot each!

There are some interesting clips of this on youtube which are well worth watching: see https://www.youtube.com/watch?v=p3Uos2fzIJ0

and (my favourite) https://www.youtube.com/watch?v=S0qjK3TWZE8. It's interesting to see how people behave in a Prisoners' Dilemma situation in real life (or at least in a TV game show).

One particularly interesting application of game theory came in biology. Evolutionary biologists like John Maynard Smith were using the theory to understand animal behaviour. Here's an example. Suppose we have two birds competing for food. They can fight for a piece of food (worth perhaps 50 units in some scale of avian benefit), or they can use some other method such as a competitive display (costing each bird 10 units) following which the loser leaves the food to the winner. If there is a fight, then the loser suffers injury costing 100 units.

We consider two possible behaviours. "Hawks" will be prepared to fight for any piece of food. "Doves" will always yield to a hawk, while two doves will display, with the winner getting the food. (The terms "hawk" and "dove" refer to the behaviour rather than to the species of bird.) Here's the table of expected pay-offs:

	Hawk	Dove
Hawk	– 25 (= ½ x 50 – ½ x 100)	+ 50
Dove	0	+ 15 (= ½ x 50 + ½ x 0) – 10

In this example, notice that neither choice dominates the other. In fact a colony consisting entirely of doves will do well, but they are vulnerable: if a few hawks join the colony then they will get more than their fair share of the food, with very little risk of injury, since almost all their competitors will be doves. So a colony which is all doves is potentially unstable. But we find there is an *evolutionarily stable strategy*, which in this case arises when about 58% of the birds are hawks (or alternatively each bird behaves randomly as a hawk or a dove with a probability of about 58% that it is hawkish in any one interaction). If this is the situation, then no incomers with a different strategy can exploit the colony. Game theory has provided a significant insight.

So game theory was useful in biology, but for me the Prisoners' Dilemma remained frustrating. And then in May 1983 I read Douglas Hofstadter's "Metamagical Themas" column in *Scientific American*, about the work of the political scientist Robert Axelrod, and suddenly a whole rich new understanding of the Prisoners' Dilemma opened up.

What Axelrod had done was to explore repeated Prisoners' Dilemmas. He used the power of the computer to see what happened over a series of games. Axelrod invited people to submit computer programmes with strategies to be applied during a long series of games against the same opponent. These strategies could be based on the opponent's previous choices. The winning strategy was submitted by Anatol Rapoport, who we mentioned earlier, and was called "Tit for Tat" (TFT). It's very simple: TFT co-operates on the first round, and thereafter does whatever its opponent did on the previous round. Axelrod told people about the result of the first tournament, and then arranged another computer tournament. Even though people knew TFT had won the first tournament and tried to devise strategies to defeat it, TFT won again.

Furthermore, Axelrod carried out computer tournaments which modelled natural selection: successful programmes generated more copies of themselves, whereas unsuccessful programmes died out. Tit for Tat was phenomenally successful in this scenario too, driving out all competitors. This is the more remarkable, because in a series of games against the same opponent, TFT can never outscore its opponent! But TFT is exceptionally successful in stimulating mutually rewarding behaviour in its opponents: while other strategies may succeed in exploiting some opponents, when they come up against each other they tend to end up with low-scoring series of mutual defections while TFT is picking up points through mutual co-operation.

Axelrod (who was collaborating with the evolutionary biologist W.D. Hamilton) identified four traits which contributed to TFT's success. TFT is *nice* (it is never the first to defect in a series of games against the same opponent). It is *forgiving* (it doesn't maintain a grudge: if you defect against it, it retaliates on the next round but after that, if you go back to co-operating, your defection is forgotten). TFT is *retaliatory* (if you defect, it will immediately defect back). And it is *clear* (it is easy for the opponent to work out what it is doing).

Axelrod presented his tournaments as a solution to a major problem in evolutionary theory. Humans (and other animals) are often remarkably unselfish. We do favours for people we don't know and who aren't related to us. We pass over opportunities to take advantage of others. We seem to have evolved to be reasonably altruistic. But evolutionary theory appears to suggest that selfish traits should be favoured by evolution, while altruism

towards strangers appears to have no evolutionary benefit. So how does altruism arise? Axelrod's Prisoners' Dilemma tournaments presented a model in which altruistic behaviour – that is, the Tit for Tat strategy – could take advantage of the non-zero-sum rewards to be more successful in evolutionary terms than more selfish alternative strategies. And Axelrod presented examples showing the Tit for Tat strategy apparently in action, both in the natural world and in human war and sport.

This *Scientific American* article seemed to me to be rather wonderful. The one-time mathematical paradox was now a significant factor in the solution of a major scientific problem. And, unlike the Tragedy of the Commons and similar analyses, it was showing the value of nice, co-operative behaviour! Maths was showing that it pays off to be unselfish! I remember reading that article and walking through London feeling that something that I had been puzzling over for many years now suddenly made sense.

So this fascinating example in game theory, rather than showing that mathematics prevents individuals from arriving at the mutually best outcome, and dictating that the Tragedy of the Commons is a recurrent motif of human existence, now shows us that altruism works and that co-operation can evolve naturally rather than having to be dictated by oppressive state control. What a happy outcome!

Except, of course, that it is rather more complicated than that, as I realised when I read the wonderful popular books of Matt Ridley, *The Red Queen*, and especially *The Origins of Virtue* which came out in 1996. Ridley is a remarkable science writer: he is extremely well-informed about an enormous range of current research and he presents fascinating and thought-provoking ideas very lucidly. Although I often disagree with his political views, I find his science writing very convincing and he has introduced me to whole fields of fascinating work. It was only recently that I discovered that writing wasn't his only career: he was chairman of the bank Northern Rock when in 2007 it became the first bank in 150 years to suffer a run, with customers queuing to withdraw their deposits. But happily Ridley is continuing to write excellent popular science books, most recently *The Rational Optimist*.

Ridley strongly challenges the version of the Tragedy of the Commons that I gave above. He argues that, contrary to the pessimistic conclusion I drew above, arrangements like common grazing worked extremely well. The community managed the land, and fear of acquiring a reputation for selfishness, or simple unwritten rules, generally prevented over-grazing. Only when society and communities changed did over-grazing occur. Legislation is not always needed to maintain the common good in a Tragedy of the Commons scenario.

Ridley's books present research which investigates various ramifications of the use of the Prisoners' Dilemma in modelling co-operation. For example, what happens when one has imperfect information about one's opponent's choices, or if a player makes a mistake? If I am playing the Prisoner's Dilemma game repeatedly against the same opponent, and we are both playing Tit for Tat, then we will both prosper. But what happens if I mistakenly think my opponent defected on the last turn? Then I defect, they will defect in turn, I will retaliate against their new defection, and we end up in an endless cycle of each alternately exploiting and being exploited by the other, resulting in us both being much worse off than if we had co-operated throughout. Similarly, if one of us decides to experiment by defecting on one turn, or if one accidentally makes the wrong choice, then we both lose out for the rest of time. So where information is imperfect, or players make mistakes, Tit for Tat is no longer such a successful strategy.

Unfortunately it is not always the case that everyone being nice to each other leads to the best outcome for all, or that one can never gain by treating people badly. In game theory as in life, there are situations in which co-operating with some people and defecting against others can lead to better outcomes for the exploiter. This mathematics can possibly help us understand human behaviour, but it doesn't tell us what we should or should not do.

The mathematics of co-operation and related matters like trust and reputation has become a major area of research. We are more ready to do a favour for X, who has a reputation for altruism and donates generously to charity, than for Y, who we know to have treated people badly in the past: these issues are now a major part of research into co-operation. The Prisoners' Dilemma, and computer tournaments and simulation, are an important tool in this work, which provides insights into politics, negotiation, and how to obtain mutual benefits while minimising the opportunities for free-loaders.

For example, why do we get angry? Anger over trivial matters appears to be a gross over-reaction. If my neighbour takes my parking place, the cost to me is small. My getting angry and punching him risks my getting prosecuted or hurt by his retaliation, both of which are much more costly to me than having an extra five metres to walk from my car to my front door. But the possibility of my losing my temper serves as a deterrent. My neighbour daren't use my parking space for fear of being punched – however irrational it would be for me to do so – because he knows that I may not be able to control my temper. If we behave in what appears to be a strictly rational way, people can take advantage of us. Perhaps some games players and sportsmen deliberately cultivate a reputation for losing their temper because if an opponent thinks you are irrational, their decisions must allow for your potential illogical reactions.

Another useful lesson from game theory is that co-operation can be fostered by breaking a big, high-risk game down into a series of lower-risk ones. When I meet someone who might become a friend, rather than exchanging important secrets when we first meet, it's better that the relationship develops gradually, so that, at each small step, we have much less to lose if the other turns out to be untrustworthy. So game theory can help us in everyday life.

But you may perhaps wonder, as I sometimes do, whether we need game theory to tell us this! Thinking about game theory over the years, there have been times when I have wondered whether the insights we gain are not sometimes rather banal – simply putting in mathematical language what we already know about obvious human behaviour. Furthermore, there have been times when I have wondered whether it is appropriate to apply game theory to human behaviour and relationships, or whether the theory trivialises what is important about being human. Does casting Jane Austen's novels as lessons in game theory do full justice to literary classics?

But my doubts have been allayed by Martin Nowak, author of an important recent book on the science of co-operation. Nowak shows how the Prisoners' Dilemma is still at the heart of computer modelling and simulation, which are helping us better understand the factors underlying co-operation (like how we establish trust and the role played by reputation). One interesting result is that it turns out that, while in the traditional applied mathematics of engineering and physics, systems usually end up in a steady state in which everything is at equilibrium, that is not true in social systems. In computer models of communities which start off being highly altruistic, then there are opportunities for more selfish individuals to prosper and the society, over time, becomes more selfish. But then altruism builds up again, and the community swings backwards and forwards, having periods of relative altruism and periods of comparative selfishness. The modelling suggests that these cycles, rather than a steady state, might be the natural state of society.

Nowak argues that the insights we gain from our research may provide the only hope for avoiding global disaster through climate change. Crudely, we have a potential analogy of the Tragedy of the Commons. No single nation can act on their own to prevent climate change, and any individual nation can get an economic advantage by doing less than the others. A politician who commits their country to make more than their fair share of sacrifices may be a visionary potential saviour of the world, but this isn't a platform which is likely to get them elected to a position where they could actually put their plans into practice. But the game theory of co-operation has the potential to help nations around the world collaborate on major challenges like climate change, and may therefore hold the key to the future of the world.

So in this lecture I have tried to share my personal journey regarding the Prisoners' Dilemma. I first met it as an alarming paradox which seemed to put into question an otherwise exciting branch of mathematics. But then, Axelrod's work showed that the Prisoners' Dilemma was not just an annoying paradox but rather was at the heart of the explanation of the evolution of altruism: it solved a significant puzzle in evolutionary theory. It offers insights into the plots of opera and fiction, and it is at the heart of research which improves our understanding of human behaviour. By helping us understand how to promote co-operation between nations, it may even have a profound effect on the future of our planet.

My lifetime fascination with the Prisoners' Dilemma is perfectly summarised by the words of Martin Nowak. "There's still so much more left to find out. We have only explored a small subset of this extraordinary game … Our analysis of how to solve the Dilemma will never be completed. This Dilemma has no end."

But this lecture does have an end. Thank you for listening.

Suggested Further Reading:

John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behaviour* (Princeton, 60th anniversary edition, 2004)

Steven J. Brams, *Game Theory and the Humanities: Bridging Two Worlds* (MIT Press, 2011)

Robert Axelrod, *The Evolution of Co-operation* (Penguin 1990; Basic 2009)

Matt Ridley, *The Origins of Virtue* (Penguin, 1997)

Martin Nowak with Roger Highfield: *Super Cooperators: Evolution, Altruism and Human Behaviour, or Why We Need Each Other to Succeed* (Canongate, 2011)