

## APPENDIX A

### The Theory of Games

The Theory of Games helps us understand our reasoning when we make decisions involving more than one person. It shows why I need to take account of other people's decision-making as well as my own, why what has happened in the past is relevant as well as what may happen in the future, and why my values need to develop to encompass our common good and not just my own individual good.

In the Theory of Games each decision-maker, or "player", has a number of choices, yielding a large number of "outcomes" according to the choices made by himself and other players. Thus if there are four players each with three possible courses of action, there will be 81 (*i.e.*  $3 \times 3 \times 3 \times 3$ ) possible outcomes. Each outcome is evaluated by each player according to his system of values, and the value he assigns to it is called his "pay-off". The pay-off is normally expressed in numerical terms, with the suggestion that we are dealing with the cardinal, interpersonal utilities that utilitarians believe in, but there is no need to assume that they are always cardinal and interpersonal; for most purposes it is enough that each player can decide his order of priorities as between the various outcomes that may result from his and others' choices. The outcomes are evaluated differently by the different players whose actions brought them about.

The Theory of Games enables us to characterize cooperative activities as opposed to purely competitive ones. In a competition there are necessarily losers as well as winners. They are "Zero-Sum Games" since my gain is your loss. In cooperative activities, however, there need be no losers, since by collaborating we both do better than we would have done on our own. These are "Non-zero Sum Games". Many are of a simple unproblematic sort: there is one outcome which is better from every player's point of view, and so each has a good reason of choosing to act so as to bring it about. But some pose problems for those who construe rationality in terms of maximising one's own pay-off.

The Rule of the Road shows the importance of conventions—"Coordination Norms"—in enabling players in a many-person game

to concert their decisions so as to secure outcomes that they all prefer. In driving, in communicating, in dancing and in many other social activities, we need to coordinate our actions with one another, so as to concert our efforts and avoid collisions. Schematically we represent two motorists, Mr Knight and M. Chevalier approaching each other, and needing to move over in order not to run into each other, by the matrix (with Mr Knight's pay-offs in top right of each outcome, and M.Chevalier's in bottom right):

**The Rule of the Road**

	Mr Knight goes right	Mr Knight goes left
M. Chevalier <i>va à droite</i>	<p style="text-align: right;">5</p> <p style="text-align: center;">each passes other safely</p> <p>5</p>	<p style="text-align: right;">0</p> <p style="text-align: center;">collision</p> <p>0</p>
M. Chevalier <i>va à gauche</i>	<p style="text-align: right;">0</p> <p style="text-align: center;">collision</p> <p>0</p>	<p style="text-align: right;">5</p> <p style="text-align: center;">each passes other safely</p> <p>5</p>

Theory of Games: Table 1

Provided both go right, or both go left, they will pass each other safely: what is essential is that they do not each decide what he, on his own, thinks best, but both abide by some convention, or rule, or law, or mutual agreement. That is to say, I should not attempt to do whatever seems to me to be productive of the best consequences, but should reliably act in the way that other people expect me to act. I should drive on the left and not cut corners, give way when the other driver has the right of way, and press forward when I have, so that other drivers know where they are with me, and can plan their own movements accordingly. There is a necessary imperfection of information about the future actions of free agents in the absence of publicly avowed rules: norm-observance—deontology—is the key to coordination. A simple maximising strategy is impossible, and each player must keep in step with others, usually by means of their all abiding by some relevant convention. Whatever the apparent attractions of consequentialism for the sin-

gle operator, they are shown to be illusory, even by consequentialist standards, once the agent sees himself to be not a solipsistic loner, but one person among many, each needing to recognise others as initiators of action with minds of their own whose decisions can be anticipated only if they adhere to well-known rules.

In the Battle of the Sexes He and She want to spend their holiday together, but He would prefer to go mountaineering in the Alps, whereas She would rather they both spent it sunbathing by the sea. The matrix is:

**The Battle of the Sexes**

	She goes to Alps	She goes to the sea
He goes to Alps	<p style="text-align: right;">8</p> <p>lovely for him: good for her</p> <p>10</p>	<p style="text-align: right;">4</p> <p>“wish you were here too”</p> <p>4</p>
He goes to sea	<p style="text-align: right;">0</p> <p>beastly for him: beastly for her</p> <p>0</p>	<p style="text-align: right;">0</p> <p>good for him: lovely for her</p> <p>8</p>

Theory of Games: Table 2

Since for either of them the second best is so much better than the third or fourth alternatives, it would pay either to settle for that if the very best appeared unattainable. And therefore it would pay the other to make it seem so. If She can throw a fit of hysterics and say she cannot abide the Alps and will not go there at any price, then He, if he is reasonable, will abandon his hopes of an Alpine holiday, and settle for the sea, which he would like twice as much—8—as solitary mountaineering. But equally He may see that the moment has come to take a firm masculine line, and let the little woman face up to the realities of the situation, and either come along with him or go her separate way. And if once it becomes clear that this is the choice, She will have no option but to cave in, and buy a knapsack instead of a new bikini. It is thus irrational to be guided only by the pay-offs of the outcomes that are available at any one time, because that enables the other to manipulate one's

choices. If I am to retain my autonomy, I cannot be altogether a direct consequentialist. Once you know that I am guided by consequences alone, you can induce me to do whatever you want by rigging the situation in such a way that by the time I come to make a decision the least bad outcome available to me is to fall in with your plans. Rationality, rather, requires that we extend our consideration over time as well as person.

The Battle of the Sexes shows the importance not of other persons but of other times. If we are to avoid being manipulated by unscrupulous fixers, we need long-term assessments, and a guarantee of not discounting the past as being merely water under the bridge. We cannot alter the past, but we can still assess it and take it into account, and thus free ourselves from being at the mercy of anyone who can rig the outcomes at one particular time. In the Theory of Games it is often an advantage to be able to bind oneself absolutely, or equivalently to rule out certain options absolutely. The strategy of Mutually Assured Destruction only worked provided both sides believed that the other was not governed solely by consequentialist considerations, and really would retaliate if attacked, even though there would be then no advantage in doing so. In order to reinforce this expectation, mechanical devices were constructed which in the event of a nuclear attack would operate automatically without the possibility of being switched off by any consequentialist survivors. In a less grisly way the whole logic of making and keeping promises is to ensure that some actions of an agent need not be altered simply by reason of factors which had been future becoming, by the effluxion of time, past. If we discount all past considerations we not only lay ourselves open to manipulation, but give only a partial account of the context in which our decisions are made, and from which they obtain their significance. I cannot be coherently oriented towards the future alone once I recognise that all my futures will one day be past.

If it were not for existence of some transferable token of value, economic transactions would mostly be instances of the Battle of the Sexes: most of the benefit would accrue to one of the parties, and the other would have the choice of either cutting off his nose to spite the other's face or of letting the other get away with the lion's share of the cooperative cake. If, however, there are not just two stark alternatives but an almost continuous range of intermediate courses of action, the claim by the one party that his offer is the

only one available, and that the other must either take it or leave it, becomes implausible, and the other can counter with an offer which is more plausible as a final offer, and which the first party would be evidently foolish to turn down out of hand. Bargaining becomes possible, and a refusal to bargain unacceptable.

The Prisoners' Dilemma was first discerned by Protagoras, and greatly impressed Plato, and later Hobbes, who made it the cornerstone of his argument for Leviathan. In its modern form it is due to A.W. Tucker. He considers two prisoners, Bill Sykes and Kevin Slob, held incommunicado, who have jointly committed a serious crime. The prosecution, however, does not have sufficient evidence to convict either of them, and they know it. But it does have evidence to convict each of them of a less serious crime, say tax-evasion, for which the penalty is six months imprisonment. The prosecution then suggests some plea-bargaining to each: if he will confess to the major crime, and give evidence so as to secure the conviction of the other, he will be pardoned for both the major and the minor crime. If he confesses, and the other confesses too, both will receive a suitably reduced sentence for having pleaded guilty, say five years. If he does not confess, but is convicted on the evidence of the other, then he will receive the full sentence of ten years. The prosecution lets each prisoner know that it has made the same proposition to the other. Each prisoner then has a strong incentive to confess: for if the other confesses too, he would get ten years unless he did, while if the other does not confess, he will get off scot-free, instead of doing six months for the minor offence. So, if they act according to their individual scale of values, they will both confess. But by so doing they will both end up worse off than if they both kept silent. If they both kept silent, they would each receive only six months for the minor offence; but by both confessing, they receive the five years for having pleaded guilty to the major crime. The matrix is given on the opposite page.

There are many Prisoners' Dilemmas in real life: tax-evasion, fare-dodging, stealing, are all familiar instances, where, other things being equal, it would seem like a good idea oneself to do them, but a very bad idea to have other people doing them too. Hence the need, argued for in Chapter 9 (§9.3), for laws backed by the sanctions of a State wielding coercive power. The importance of the Prisoners' Dilemma, however, lies not only in its showing the need for the State, but in its revealing the inadequacy of static

**The Prisoners' Dilemma**

	Sykes keeps silent	Sykes confesses
Slob keeps silent	-1 Both jailed for tax -1	0 Sykes let off: maximum jail for Slob -10
Slob confesses	-10 Slob let off: with maximum jail for Sykes 0	-5 Both jailed reduced sentences -5

Theory of Games: Table 3

ascriptions of value to individuals. For there is a sense in which it is obviously in the prisoners' interests not to confess, and this rationality the static schema employed by the Theory of Games occludes. This point is often missed, because the prisoners are *ex hypothesi* wrongdoers, and hence presumed to be selfish. If only people were unselfish, and put others before self, then, so the argument runs, all would be well: the prisoners would not confess, the tax-payer would pay his taxes, the traveller buy his ticket, and nobody would ever wrong his neighbour. That all would not be well, however, is evident once we consider the dilemma of the altruistic couple where He tries to maximise Her pay-off, and She His, with the result they both end up with something they neither want. Thus He might be keen on cars, and She on food. If He mends the car and She cooks, they have a good lunch, followed by a drive in the country. If He helps Her cook, instead of messing about in the garage, they have an absolutely super lunch, though no drive in the country. If, on the other hand, She helps Him mend the car, the car will go like greased lightning, but they will have to eat in a Transport Cafe. But if they each insist on doing what the other wants, He will try His hand in the kitchen, while She will wriggle under the car, and the result will be an indifferent lunch followed by a mediocre drive, much worse for both of them than if each had acted non-altruistically. The matrix is:

**The Altruists' Dilemma**

	She cooks	She helps Him mend the car
He mends the car	5 good lunch, followed by pleasant drive 5	0 record journey, with meal in Transport Cafe 10
He helps her cook	10 super lunch, but no drive  0	1 indifferent lunch, followed by mediocre drive 1

Theory of Games: Table 4

The Altruists' Dilemma is the mirror image of the Prisoners' Dilemma, and shows that the trouble lies not in one's being concerned to maximise one's own pay-off, but in being tied to just one pay-off throughout. In practice we are able to resolve or surmount the Prisoners' Dilemma because we modify our original preferences in the light of what we come to know about others', and are not confined to a single occasion. I conjugate over persons, and knowing what you want, see that we shall both be better off if we follow a cooperative strategy, and for that reason come to want it. Although other things being equal, I want to get off scot-free, and prefer a short prison sentence to a longer one, I do not want to let down my confederate. I identify with him, and begin to take his interests to heart, and consider what is best for us jointly, rather than for just me individually. I may not do so completely, and make his interests mine, as the utilitarians urge, but I do so enough to alter the balance of advantage so as to favour the cooperative strategy. Of course, in so doing, I make myself vulnerable to being let down by him; but in real life few situations are evidently and certainly one-off, and anyone who lets me down on one occasion will forfeit my trust thereafter. In the long run I shall do worse if I let people down in order to maximise my own pay-off on each occasion than if I respond to each person as he did to me the last time we met, and

give those I have not met before the benefit of the doubt and trusting them to behave decently. Being reasonable seems reasonable once we conjugate over persons, and proves to be the best policy once we conjugate over time too. A completely static and purely individualist approach is inadequate and demonstrably irrational: if we are to be rational we must take the values of others into consideration as well as our own, and must be prepared to change our priorities in the light of them.

Each of these arguments is a *reductio ad absurdum*. We start by assuming, as the classical economists did, that rationality can be defined in terms of maximising future pay-offs, and then show that even within its own terms, such a definition is self-contradictory. The Rule of the Road shows that it is better to keep to the rules than to try, as the Act Utilitarians counsel, to perform the act that will have the best consequences: each of us should recognise that he is not the only pebble on the beach, that it is not for him to choose which course of events shall occur, and that often the best he can do is to fit in with what other people are likely to do. The Battle of the Sexes shows that it is irrational to have regard only to future outcomes; an agent has a past as well as a future, and should make up his mind what he is going to do with regard to what he has decided in the past as well as what will ensue in the future. The Prisoners' Dilemma shows that he should take into account not only the existence but the interests and ideals of other people, and that it is irrational to ignore the collective point of view. Contrary to the static, solipsistic, future-oriented, exclusively individualistic standpoint of the classical economists, we are forced, by thinking about these three cases, to recognise that rationality is dynamic, leading us to take a longer temporal and wider personal view of what is involved in the decisions we are called on to take.