Part IV
Using Formal Models

10

Choice Models

Katie Steele

1. Introduction

Adequate explanation and prediction of human behaviour often requires understanding the beliefs and values that motivate action. In this way, we gain a deeper understanding of the behaviour in question, beyond just noting that it has some regularity. For instance, the hypothesis that Dave goes to church weekly is prima facie less informative than the hypothesis that Dave goes to church primarily because he is in love with another member of the congregation. Likewise, the hypothesis that many voluntary public goods projects fail is prima facie less informative than the hypothesis that certain projects fail because individuals do not place positive value on their own participation and believe they can free ride on the efforts of others. In short, the social sciences involve special kinds of models that track our notions, based on common sense or 'folk' psychology, of the causes of human behaviour. These are models that depict the choices/behaviour of persons as resulting from what philosophers call, as we learn in Chapter 5, the persons' 'intentional attitudes'—their beliefs and values.

Formal choice models are used to represent these subjective beliefs and values in a concise way. The field is commonly divided into three domains: *decision theory, social choice theory,* and *game theory*. Decision theory concerns the intentional attitudes and choices of a single person or *agent*. This is of intrinsic interest in psychology and cognitive science. In fact, decision theory has wide reach as it is a building block of the other two domains of choice models as well, which concern groups of agents. *Social choice theory* models a group of agents who must reconcile their attitudes in order to act as a single agent with 'shared' attitudes (think of voters who must settle on a choice of leader); social choice theory is employed in political science and

public economics, amongst other areas. Finally, *game theory* models strategic interaction amongst individuals—all choose their own course of action, but they strategize because the combination of choices affects the outcomes of all involved. Game theory is used extensively in (micro)economics, and increasingly in sociology and other social science disciplines.

All of these domains of choice theory—decision, social choice, and game theory—are also used to explore normative questions, i.e. questions about how things *should* be done, rationally and/or morally speaking. Philosophers and other normative theorists appeal to choice theory to answer questions like 'What constitutes *rational* choice attitudes?' or 'What group attitudes *justly* represent the attitudes of its constituent members?' or 'How should public institutions be designed to deliver *adequate* outcomes, given the attitudes that citizens are likely to have and the strategic behaviour they are likely to engage in?' These are all interesting questions, but they are not the focus in this chapter; here I am primarily interested in the use of choice models in the empirical sciences. I concentrate just on decision and game theory, in the interests of space, and also because social choice theory arguably has more normative than descriptive applications. In any case, we shall see, however, that the normative and descriptive perspectives on choice theory are intimately linked, for better and worse reasons.

The chapter is organized as follows. In section 2, the basics of individual decision and game models are introduced. I will then, in section 3, address criticisms of the deployment of these models in the social sciences. My general position is that, while there is ample scope for criticizing particular applications of choice models, this should not be taken as an argument against choice modelling in the social sciences *tout court*.

2. Introducing Choice Models for the Social Sciences

This section introduces the basic notation for choice models in the social sciences. I consider, in turn, individual decision and game models. The two are not entirely distinct. It will come as no surprise that game models involve individuals. But game theoretic considerations also impact on individual decisions, so neither is obviously the more general choice model. To some extent, it is simply the case that the two kinds of models are useful in different contexts.

2.1 *Individual Decision Models*

I shall first introduce individual decision models. The main components of these models are *prospects* and *preferences*. Prospects may be further divided

into *acts, states,* and *outcomes*. In this way, the complexity of an agent's attitudes towards the world is reduced to a relatively simple model.

In short, the agent has preferences over prospects. That is, there are various 'items' or 'states of affairs' (the prospects) that the agent compares and ranks in terms of which ones he/she likes better than others (the preferences). The core prospects are acts, which, as the name suggests, are things that an agent can do of their own volition, like 'pick the orange from the fruit bowl', or 'go to the park', or, at a slightly grander level, 'pursue graduate studies in psychoanalysis'.

Let us denote a *weak preference* for act A_i over act A_j as $A_i \geq A_j$: either the agent *strictly prefers* A_i to A_j (written $A_i > A_j$), or the agent is *indifferent* between A_i and A_j (written $A_i \sim A_j$). Acts may be decomposed in terms of states of the world and outcomes. The states of the world are the possible (mutually exclusive) ways the agent thinks the world may be; this is the locus of the agent's uncertainty about the world. The uncertainty at issue may concern a very local and mundane issue like the result of a coin toss (where the states are 'heads' and 'tails'), or, for more worldly decisions, the uncertainty may concern the economic prospects of a country (where e.g. the states may be 'permitted to stay in the Eurozone with large debt repayments until 2020', and so on). Associated with each act and state is an outcome, which is what will come about, at least from the point of view of the agent, if the act is performed and the state in question is the true state.

The relationship between acts, states, and outcomes is best appreciated by considering the typical tabular representation of a decision problem, as per Table 10.1.

For instance, we might be interested in Mary, who is very conscientious and prudent; we want to know what she will do when confronted with various health insurance options. The acts $A_1 \ldots A_r$ are possible insurance packages she could purchase. The states $S_1 \ldots S_n$ represent the relevant possible scenarios that Mary envisages. For example, S_1 might be the state where she has only minor ailments during the next ten years, while S_2 is the state where she has some significant but not critical health problem, like reduced eyesight. The outcomes $O_{1,1} \ldots O_{r,n}$ are the various possible outcomes that Mary

Table 10.1 Tabular representation of decision problem

	S_1	S_2	…	S_n
A_1	$O_{1,1}$	$O_{1,2}$	…	$O_{1,n}$
A_2	$O_{2,1}$	$O_{2,2}$	…	$O_{2,n}$
…	…	…	…	…
A_r	$O_{r,1}$	$O_{r,2}$	…	$O_{r,n}$

anticipates, given the insurance benefits and her health prospects. For example, $O_{2,2}$ might be the outcome that Mary loses much vision yet her insurance gives her free access to consultations and glasses, etc.

The agent has preferences over the acts of the decision problem of interest, but we presumably need to work out/predict what these preferences are. This is done by making assumptions, perhaps based on past observations, about the agent's beliefs and desires. Shortly I shall expand on the basis for these assumptions; for now, note that, standardly, the agent's beliefs about the states of the world are represented by a probability function, *Pr*, over these states, and the agent's relative desires for the basic outcomes are represented by an *interval-valued* utility function, *U*, over these outcomes (to be explained shortly). For example, the strength of Mary's belief that she will have only minor ailments in the next ten years might be 0.7, and in this case, by the rules of probability, the strength of her belief in the contrary is 0.3. A utility function is *interval-valued* or *cardinal* just in case the differences between the utilities are meaningful in the following way. If, say, the decision outcome for minor ailments under package A_1 has utility 10, the outcome under package A_2 has utility 4 and the outcome under package A_3 has utility 1, then Mary prefers the A_1 outcome to the A_2 outcome twice as much as she prefers the A_2 outcome to the A_3 outcome. By contrast, if Mary's utility function were only *ordinal*, then the differences in utilities would be meaningless and all we could infer from the specified values is that Mary prefers the A_1 outcome to the A_2 outcome to the A_3 outcome.[1]

The standard calculus for deciding between options or acts is the *expected utility principle*, which claims that a (rational) agent's preference ranking tracks expected utility, or the average utility for a prospect given the probabilities and utilities for each of the possible outcomes, such that:

$A_i \geq A_j$ IF and only IF

$$\sum_n \Pr(S_n) \times U(O_{i,n}) \geq \sum_n \Pr(S_n) \times U(O_{j,n})$$

In other words, one act is preferred to another act just in case, for the first act, the sum over all the states of the world of the probability of the state multiplied by the utility of the outcome of the act in that state is greater than this

[1] Note that an interval-valued utility function is more formally described as a utility function that is *unique up to positive linear transformation*. In general, the uniqueness conditions associated with a mathematical measure tell us what information is given by the measure. To say that the utility function is *unique up to positive linear transformation* is to say that, if e.g. chocolate ice-cream has utility 5 while vanilla has utility 2, then this very same information is represented by utilities $5m + c$ and $2m + c$, for any positive m and any c. In other words, no conclusions should be based on the utility values 5 and 2 that do not also hold for utility values $5m + c$ and $2m + c$.

same sum for the second act. In the case of indifference, the sums for the acts are equivalent. Note that the utility function, U, must be interval-valued or cardinal, otherwise it is not meaningful to multiple utilities with probabilities and then sum these terms.

If we can characterize Mary in terms of her probabilities for states and her cardinal utilities for outcomes, then we can work out her preferences for health insurance packages, provided we also assume she is an expected utility maximizer. But one might well ask what justifies characterizing Mary in this way. Indeed, this is a significant part of the scientist's task. Recall Mary's example utility function, where the outcome for minor ailments under package A_1 has utility 10, the outcome under package A_2 has utility 4, and the outcome under package A_3 has utility 1. How could we know that this is Mary's utility function? That is, how could we know that she prefers the A_1 outcome to the A_2 outcome twice as much as she prefers the A_2 outcome to the A_3 outcome?

In the 1920s Frank Ramsey proposed an ingenious way to determine, at least in theory, a person's utility function over outcomes *and* their probability function over states of the world. It involves asking the person to rank a large number of bets. One must also assume that the person is an expected utility maximizer, i.e. they rank bets according to the rule just stated. Sometime later in the 1940s von Neumann and Morgenstern developed a similar method for measuring a person's utility; on their account we need to know the person's preference ranking over a large number of lotteries. By way of example, here is a very simple construction of a cardinal utility scale for the three holiday destinations Rome, Paris, and Barcelona. Assume the agent has the following preference ranking over the destinations: Rome > Barcelona > Paris. Now there is some lottery over Rome and Paris with objective probability p, i.e. a lottery that has chance p that the agent will go to Rome, and chance $1 - p$ that the agent will go to Paris, such that the agent is indifferent between this lottery and Barcelona. The probability p for this lottery is the utility for Barcelona, if the utility for Paris is set at 0 and the utility for Rome at 1, since then the expected utility calculations are in keeping with the claim that the agent is indifferent between Barcelona and the lottery. (Note that the expected utility for the lottery is $p \times 1 + (1 - p) \times 0 = p$). We can check that this makes sense by considering a couple of cases. If going to Barcelona is only marginally worse than going to Rome, then the agent would require a very high probability for Rome to be indifferent between Barcelona and the lottery. After all, the lottery might land the agent in Paris. In this case, the utility for Barcelona matches this high probability. On the other hand, if Barcelona is just a bit better than Paris, then the agent would require only a small probability for Rome to be indifferent between Barcelona and the lottery; this small probability is then the utility for Barcelona. So we see that lotteries

enable a utility representation of an agent's preferences over prospects. We could use this construction of the agent's utility function (in addition to the probabilistic measure of their beliefs) to predict the agent's preferences in other more complex decision problems. In Mary's case, we might test how she ranks a number of lotteries involving her basic health outcomes, and so determine her utility function over these outcomes. We could then use this utility function to predict what she would choose in the 'real-life' decision of choosing a medical insurance option.

Even if a person's utility and probability functions may be determined from their preferences over bets/lotteries, there is still the large question of how to elicit these preferences. It is difficult to ascertain a person's preferences over just a few prospects, let alone the many required for detailed measures of belief and desire. The key question is: what sort of data serve as evidence for an agent's current preferences? The answer depends on how preference is to be understood, and in particular, what is the presumed relationship between preference and choice behaviour, and between preference and other psychological traits. I can only touch the surface of this issue here. An extreme yet surprisingly mainstream position, particularly in economics, is the so-called theory of *revealed preference*, which holds that preference simply *is* choice, or perhaps choice disposition, rather than some deeper psychological attitude. In that case, we cannot look further than a person's choice behaviour in order to determine their preferences because these preferences are constituted by the choice behaviour and have no deeper grounding. It follows that decision models are reduced to mere descriptions of an agent's choices, where the probability and utility functions are just mathematical constructs and not properly representative of belief and desire at all.

The revealed preference interpretation offers little by way of explanation and prediction of choice behaviour. It does not have the resources to express what *particular* features of a prospect an agent finds (un)desirable and therefore (un)choiceworthy. There is thus no way to determine whether some current/future choice situation is relevantly similar to a past choice situation, such that one would expect an agent to behave similarly. In short, revealed preference theory has no inductive (predictive) power. Moreover, the view does not countenance an agent having irrational preferences or choosing contrary to their preferences, due to impulsiveness or weakness of will.

The more fruitful position is surely that preference is distinct from, but plays some kind of motivating role with respect to, choice behaviour. As such, choice behaviour is still important *evidence* for preferences. This might be choice behaviour in the laboratory, or in the wild, so to speak. Indeed, one could yet argue that choice behaviour and preferences are inextricably linked, whether on metaphysical or evidential grounds or both. For instance, some hold that preference, which is a psychological attitude, completely

determines choice behaviour, even if it is not identical to it. This is to say that it is impossible for a person to choose contrary to their preferences, and so phenomena like weakness of will would need some alternative explanation. Others maintain a 'causal gap' between preference and choice, filled by strength of will, habits, urges, instincts, and the like. Either of these views may or may not be coupled with the evidential claim that choice behaviour is the *only* reliable evidence for preferences. Even if that were true, experimenters must do much inferential work in deciphering choice behaviour if they want to construct appropriate utility functions for agents—the experimenters must make assumptions about how the agents perceive the options available to them, whether the agents have sufficiently reflected on their choices, and what are the agents' relevant background beliefs. Alternatively, or in addition to observing choice behaviour, experimenters could ask agents outright what their preferences are over some prospects, or ask agents various questions that are deemed relevant to their preferences. Different techniques will be more or less successful in different contexts; indeed, the appropriate way to conceive of and elicit preferences in particular settings is a topic of debate amongst social scientists.

2.2 *Game Models*

Game models are intended to capture the special situation of strategic interaction between agents. 'Strategic' here does not necessarily mean conniving or in some way dishonest; it simply means that, when choosing a course of action, agents consider what others are likely to do and how this bears on the possible outcomes. There are many such situations in social life, from political manoeuvring during an election campaign to simple coordination problems like choosing to drive on one side of the road rather than the other. This section serves to introduce and interpret game models and to outline the standard ways of 'solving' games; section 3 discusses the usefulness of game models in the social sciences.

There are some basic components to the *description* of a game. First, there are the *players*—in the simplest case, two players. These players have a set of available *acts* or *strategies*, and we ultimately want to know which of these the players each choose. The rules or procedures of the game specify the *outcomes* for each player, given each possible combination of players' strategies. Players have preferences over these outcomes, represented by a utility function that may be cardinal or ordinal. (Recall that for the former kind of utility function, unlike the latter kind, the distances between utility values are significant and represent strength of preference.) As already alluded to, the players may well be entirely self-regarding, but they need not be. As with individual decisions, in terms of preferences, agents in a game come in many stripes and colours.

Table 10.2 A literal prisoners' dilemma

	Cooperate (confess)	Defect (don't confess)
Cooperate	(light sentence, light sentence)	(very harsh sentence, go free)
Defect	(go free, very harsh sentence)	(harsh sentence, harsh sentence)

Table 10.2 depicts a *normal-form* (tabular) game model for two players; let us call the players 'row' and 'column'. The table shows that both players have two possible strategies: 'cooperate' and 'defect'. (The strategies of the 'row' player can be read off the rows of the table, while the strategies of the 'column' player can be read off the columns.)

The table describes a well-known type of game; it in fact depicts the original narrative used to illustrate this type of game. The story goes like this: there are two prisoners accused of a crime. The prisoners are told that if they each cooperate (i.e. confess), they will both get a light sentence. If they each rather defect (i.e. fail to confess), they will both get a harsh sentence. If one confesses but the other does not, the one that confesses will get a very harsh sentence while the other goes free. The convention is for the utility payoff to 'row' to be specified first in the appropriate cell of the table, followed by the utility payoff to 'column'. Take the case where both players cooperate. That is, 'row' plays 'cooperate' and 'column' plays 'cooperate' as well. Here they both get light sentences, which for each of them is the second-best possible outcome. In the case that 'row' cooperates whereas 'column' defects, 'row' gets the worst outcome—a very harsh sentence—while 'column' gets the best outcome of going free. The reverse situation occurs if 'column' cooperates while 'row' does not. The paired outcomes associated with each pair of strategies constitute the rules of the game.

I have here described a literal prisoners' dilemma, but the reason Table 10.2 describes what is known as a *Prisoners' Dilemma* (PD) game, in the technical sense, is simply the pattern of outcome utilities associated with the pairs of strategies. This pattern can be seen more clearly if we replace the literal outcomes with utilities as in Table 10.3. To characterize the PD game, the utilities need only be ordinal, i.e. only the ordering of the utilities matters and not the differences in their values. One can check that the ordering of the utility values reflects the preference ordering of the outcomes in Table 10.2.

There are many social scenarios involving two players (and many more involving more than two players, if we were to generalize) that have the pattern of outcomes given in Table 10.3. The situation is one where free-riding, or defecting when the other cooperates, gives the most utility for the defector and the least utility for the cooperator. It is better that both cooperate, however, than both defect. An infamous example of the PD, at least according

to some, is the arms race of the Cold War, where cooperating is pursuing disarmament and defecting is continuing to stockpile weapons. Of course, it is questionable whether the PD model accurately characterizes the Cold War scenario, but it is not implausible that the players' ordinal utilities (players here being national governments) were in line with those in Table 10.3. Other prime examples of PD games concern environmental protection. Here the agents may be national governments with a common interest in, say, global fish stocks or forestation or climate stability, or, at the other end of the spectrum, they may be individual citizens with a common interest in local air or river pollution. These are PD games just in case all prefer a sufficiently clean environment where everyone does 'their bit', to a polluted environment, and yet all players most prefer that the others refrain from polluting while they pollute and still enjoy a sufficiently clean environment. (It must also be the case that all players least prefer doing their bit when others don't do theirs, presumably because this makes the do-gooders' efforts futile.) Here again, it is questionable whether any particular environmental protection problem does in fact have the form of a PD game rather than another type of game. For one thing, it may be that polluting while others preserve the environment is not most preferred, perhaps because all players value helping the cause, or because even a small amount of pollution is damaging enough to be undesirable. In short, whether or not a scenario can be represented as a PD game as per Table 10.3 depends on the players' preferences, and also the way the world is, both in terms of natural laws and social institutions, as these structures dictate the rules of the game.

Beyond the description of a game, there is the *solution concept*, which specifies how the agents eventually settle on their respective strategies and associated payoffs. The central solution concept in game theory is the *Nash equilibrium*, formulated by its namesake John Nash in the early 1950s in his graduate thesis. A Nash equilibrium is a combination of strategies such that each player's strategy is the *best response* to the other player's strategies. That is, each player's strategy affords them the highest possible utility given what the other players do. A simple way to identify Nash equilibria in a two-player game model is to circle the row player's maximum values for each column (i.e. row's best response to each of the strategies that column might choose), and likewise circle the column player's maximum values for each row. Remember

Table 10.3 The general prisoners' dilemma

	Cooperate	Defect
Cooperate	(3,3)	(–1,4)
Defect	(4,–1)	(0,0)

that, for each strategy pair, the value for row is given first in the table cell, followed by the value for column. The Nash equilibria are the cells that have both row and column utility payoffs circled.

The PD game has only one Nash equilibrium: the strategy combination where both players defect, which yields the utility payoffs (0, 0) by the representation in Table 10.3. Note that if 'row' chooses to defect, then 'column' does best by also defecting, thereby gaining utility 0 rather than –1. Likewise, if 'column' chooses to defect, 'row' does better by also defecting, thereby gaining utility 0 rather than –1. That is why both players defecting is a Nash equilibrium; neither player can do better by switching to another strategy. None of the other strategy pairs are Nash equilibria. Consider the potential case where both 'row' and 'column' cooperate. This is not a Nash equilibrium because either player would do better by switching to defection, thereby gaining utility 4 instead of 3. In fact, the combination of both players defecting is a particularly robust kind of Nash equilibrium because it is a *dominant solution* for both players: for each player, the defect strategy is best, not just on the assumption that the other player is also playing their Nash equilibrium strategy, but *no matter what the other player does*. Even if 'row' is irrational and does not choose to defect, it is still better for 'column' to defect, and vice versa.

In a Nash equilibrium (including the special case of a dominant solution), no player can do better by *unilaterally* switching to a different strategy. It is possible, however, that all players could do better by *simultaneously* switching to a different strategy, i.e. a Nash equilibrium is not necessarily *Pareto optimal*. A Pareto optimal strategy set is optimal in this sense: there is no other strategy set such that shifting to this strategy set would result in at least one player being better off and no player being worse off. The tragedy of the Prisoners' Dilemma is that the one Nash equilibrium (here dominant solution) is the *only* strategy combination that is not Pareto optimal. Indeed, *both* players would be better off if they both cooperated. (Even the strategy pairs where one player defects and the other cooperates are Pareto optimal, because a shift to another strategy pair would result in at least one player—the one getting the maximum utility of 4—becoming worse off.) This is why the Prisoners' Dilemma is said to illustrate the 'failure of collective rationality'.

One might wonder what is so special about Nash equilibria, especially if they can in a sense yield inferior outcomes, as per the Prisoners' Dilemma. The logic is as follows: when reasoning about what strategy to opt for, an agent knows the set-up of the game, including the utilities of all players, and they also know that other players know the set-up of the game, and that these other players know that the agent knows they know the set-up of the game, and so on. Some claim that it is this *common knowledge assumption*—the cycle of 'I know you know I know you know etc.' reasoning—that leads to the

privileging of Nash equilibria, which are joint best responses to each other's strategies. Others disagree that common knowledge leads players to reason their way to Nash equilibria; there is simply an additional assumption that the players each believe their opponents will choose the/a Nash equilibrium strategy, perhaps due to the history of plays of the game. In any case, game theory rests on the assumption that players make 'intelligent' choices that lead them to settle on Nash equilibria. Either the players reason intelligently all the way to Nash equilibria, in which case it is necessary that they have rational preferences and are aware of the game description, or, at the other end of the spectrum, the players do not reason at all but are subject to environmental or 'selective' pressures that favour players who at least *act* as if they intelligently choose Nash equilibria after repetitions of the game.

Note that it may be desirable to treat the selective-pressure/evolutionary interpretation of Nash solutions explicitly. Indeed, this has given rise to *evolutionary game theory*, which finds application in biology as well as social science. The games in these models can even be 'played' by unconscious entities like bacteria that pursue strategies in an automated fashion. Strategies are replicated according to their success. The solution concepts of evolutionary game theory are in many ways similar to those of regular game theory, but there are differences, and of course the set-up and interpretation of the models differ (Alexander 2009). This chapter does not explore evolutionary game theory, but rather sticks to games with genuinely intelligent players.

Nash equilibria are central in this way: it is widely regarded a *necessary* condition of a game solution that it be a Nash equilibrium. Some games have multiple Nash equilibria, however, and this raises the question of what are *sufficient* conditions for something to count as *the* game solution. Game theorists refer to this as the 'refinement program'.

A number of games with multiple Nash equilibria arguably feature in social life. In many such cases, the question of what is the unique equilibrium 'solution' of the game is pressing. For example, the game in Table 10.4 represents a typical coordination problem; provided agents play the same strategies, the outcome is good. The game is commonly known as the 'driving game' because the choice of two players of whether to drive on the left or the right side of an isolated country road is a simple coordination problem that fits the pattern. We see that the top left and the bottom right entries in the table, corresponding to both players choosing 'left' and both choosing 'right' respectively, result in each player receiving (ordinal) utility 1. Both of these strategy combinations are Nash equilibria: if one player opts for 'left' then the best response of the other player is also 'left', and likewise if one player opts for 'right', then the best response of the other player is also 'right'. Note that these two Nash equilibria are also Pareto optimal. The question then arises: what if the players shoot for different Nash equilibria?

Table 10.4 Multiple Nash equilibria: the driving game

	Left	Right
Left	(1,1)	(–1,–1)
Right	(–1,–1)	(1,1)

This will yield bad outcomes. If one player plumps for the 'left' Nash equilibrium while the other player plumps for the 'right' Nash equilibrium, then we are in the top right or bottom left entries of the table, where the players receive -1 utility apiece. So in this game, it is important that the players aim for the same Nash equilibrium. (In other games, it may not be important that the players aim for the same Nash equilibrium, if their utility payoffs are in any case unaffected.)

The question of what Nash equilibrium players will/should aim for arguably depends on principles that go beyond game theory proper, like which equilibrium is more salient for all players. For instance, in a literal driving game on an isolated country road, presumably the salient equilibrium is the one that accords with the driving rules of the nearest municipality, whether this be 'always drive on the left' or 'always drive on the right'. Note that there are many coordination problems that have the formal structure of a driving game, as given in Table 10.4, and the conditions for salience will differ in these different applications. For instance, consider the case where two complementary NGOs can achieve good outcomes if they each target the same community, but there will be no good result if they divide their efforts. How should they each proceed? Presumably, if the NGOs were able to come to an agreement on which equilibrium to aim for (i.e. which community to direct their efforts at), this equilibrium would be very salient indeed. The NGOs would have no reason to deviate from such an agreement. Other game situations, however, may not lend themselves to straightforward negotiations regarding multiple equilibria; a major obstacle is when different equilibria are better for different players.

Indeed, the 'driving game' is just one example of a simple game structure with multiple Nash equilibria. There are other similar kinds of games, and moreover, these sorts of games are arguably more prevalent in the social world than the tragic Prisoners' Dilemma. Indeed, as we can see in Chapter 11, Cristina Bicchieri argues as much, by appeal to social norms; in brief, she claims that we internalize social norms, like 'one must do one's bit for community projects rather than free-riding on the efforts of others', such that satisfaction and thereby utility is gained from obeying the said norms, or otherwise lost by disobeying the norms. This phenomenon of *norm-internalization* serves to

'convert' a scenario that would otherwise be a PD game, for instance, into a more benign game with better collective equilibria. Of course, for multiple equilibria, there remains the issue raised of whether agents can coordinate effectively.

There are many more facets to game theory, and its applications are much richer in variety than have been touched upon here. For instance, there are game models for any number of players, not just two. There are game models that distinguish the order in which players choose their strategies, and the amount of information they each have about the prior choices of the other players. There are special considerations that apply to repeated plays of the same game. And, as mentioned earlier, there is the burgeoning area of evolutionary game theory, which increasingly finds application in the biological and social sciences, and which can accommodate players that merely *act* as if they are intelligent upon repetitions of a game. These are just some of the further dimensions to game theory. The most powerful contributions of game theory to the social sciences are arguably the more basic ideas, however, which I have introduced here.

3. Ideal Type and its Discontents

Having introduced formal choice models for the social sciences, we now consider criticisms of their deployment. The critics have two main targets: the idealized characterisation of individuals in choice models, and the fact that individuals and not societies are the primary building blocks of models of social interaction. I discuss these in turn. My general point is that standard criticisms of choice models tend to overshoot; one must bear in mind the general limitations of scientific modelling and also appreciate that choice models come in a variety of forms.

3.1 *Standard Decision Theory: Unfalsifiable or Simply False?*

There is a temptation to reject wholesale the project of generalizing and predicting human motivation and behaviour. But this is surely too pessimistic and casts doubt on the social sciences as a whole. Indeed, Max Weber (1864–1920), claimed to be one of the 'principal architects of modern social science', started from a more optimistic position. Weber argued that the social scientist's aim, unlike the historian's, is not to represent the vagaries in attitudes of some particular persons, but rather to represent attitudes and behaviour that can be generalized across people and/or across time and which therefore have explanatory value. To use Weber's terminology, the social scientist explains and predicts on the basis of *ideal types*.

According to Weber himself, a key ideal type for the social sciences is the agent who rationally furthers her own ends, or who, in other words, has rational preferences over prospects. Weber's claim then is that, when studying human behaviour and social interaction, it is useful to depict agents as making choices on the basis of rational preferences. This assumption is indeed borne out in the social sciences, and particularly in economics; while we know that people sometimes act irrationally or have inconsistent preferences, it is assumed that such cases of irrational behaviour are temporary deviations that in some sense 'cancel out' in large groups or in the long run. Consequently, the ideal type in economics as well as other disciplines is commonly in line with the standard theory of rational choice outlined in the previous section—expected utility (EU) theory. As such, EU theory is the locus of criticism with respect to formal choice models in the social sciences. Oddly, however, EU theory is criticized on two opposing fronts: some argue that the theory is unfalsifiable, i.e. that it cannot be proven wrong and so is vacuous, while others argue that it is outright false. How could both of these views have gained purchase? I claim that it is due to different understandings of the flexibility of choice models.

Start with the charge that EU theory is descriptively false, even as an approximation. Consider standard economic models—typically, the outcomes that agents are supposed to care about are bundles of consumer goods, or money, and the more of these goods the agent procures for him or herself, the better. In other words, it is standardly assumed that the only relevant distinction between decision outcomes is the amount of goods or money that will accrue to the agent. So, for example, $50 is always worth the same, no matter how this amount was procured, whether by hard labour or by cheating one's friend, and two sacks of potatoes are better than one, regardless of these products' provenance. It is easy to see that a real agent, and a perfectly rational one at that, may not be well described by an expected utility model constructed in this way. To begin with, the agent may not be indifferent between outcomes that are supposedly identical, like $50 procured by hard labour and $50 procured by cheating.

This point is perhaps more vivid when it comes to game models. To give an example: a classic game for characterizing and testing the bargaining behaviour of agents is the 'Ultimatum Game'. In this game, there is some pot of goods, often just a sum of money that needs to be divided between the two players. The rules of the game are as follows. One player—the dealer—selects a split of 'the pot', and the other player—the receiver—decides whether or not to accept the split. So, for instance, the dealer might recommend a 50:50 split (the fair strategy), or she might recommend, say, an 80:20 split (an unfair strategy), which is to say that she keeps 80 per cent of the pot and gives 20 per cent to the receiver. If the receiver accepts, then each gets what the dealer

Table 10.5 Ultimatum game with monetary outcomes

	Accept	Reject
Fair	($5, $5)	($0, $0)
Unfair	($8, $2)	($0, $0)

decreed, but if the receiver rejects the offer, neither gets anything. A simple version of this game, where the dealer has just the two strategies stated and where the pot is $10, is given in Table 10.5.

The outcomes of the game are specified in monetary amounts, and moreover, these amounts are assumed to track the agents' utilities. This is typical of game models in economics. But the Nash solution of ($8, $2) is not very compelling in this case, and indeed laboratory experiments and real-world observations suggest that agents in this kind of scenario overwhelmingly settle on the 'fair/yes' strategy pair, i.e. the dealer offers 50:50 and the receiver accepts. Furthermore, it seems entirely rational for these agents to reject the apparent Nash solution: they care not just about money but also about whether justice has been done.

There are further limitations of EU models that distinguish outcomes only in terms of personal holdings of goods/money. Besides having altruistic tendencies, many agents are sensitive to the menu of available options, and to what might have happened if they chose differently or if things turned out differently. So, for instance, an agent's preference for relaxing at home over a luxury holiday might switch depending on whether a third option of an adventure holiday is available. (The latter might make the 'stay-at-home' option seem like the timid choice rather than the modest choice.) To give a different example: losing a gamble and receiving nothing may be better or worse depending on whether one might otherwise have won $10 or $1,000 dollars. A decision problem developed by the late French economist Maurice Allais received much attention for exposing such attitudes of regret and its effect on agents' evaluations of outcomes. These sorts of preferences are not consistent with an EU model that distinguishes outcomes only in terms of money/material goods.

Is expected utility theory then false, both descriptively and normatively? There is an obvious argument to the contrary: EU models must simply be sufficiently detailed such that outcomes include all properties that agents care about, including the well-being of others and unrealized possibilities that inspire feelings of regret/envy/self-critique. (Decision theorists use the term *comprehensive outcomes*.) Then there are no obvious conflicts between EU theory and reasonable choice behaviour. For instance, returning to the

Table 10.6 Ultimatum game with comprehensive outcomes

	Yes	No
Fair	$(5 + \delta, 5 + \delta)$	(0, 0)
Unfair	$(8 - \theta, 2 - \theta)$	(0, 0)

Ultimatum Game, the monetary amounts apparently do not represent typical agents' utilities for outcomes. The 'right' characterization of the game is plausibly as per Table 10.6. The δ and θ amounts here represent the utility and disutility associated with fair play and unfair play respectively. In this case the 'fair/yes' strategy pair may well be a Nash equilibrium. Likewise, if outcomes in decisions include the properties that inspire regret or risk attitudes, then EU theory may well capture common choice behaviour amongst risky options.

This brings us to the other critique of expected utility theory: that it is unfalsifiable. The charge here is that EU theory is not false precisely because it *cannot* be proven false. Perhaps in avoiding one problem we run into another: if outcomes are so comprehensive that they include everything that an agent cares about, it would seem that EU and game models could be fitted to any choice behaviour and social interaction whatsoever. This would suggest that the theories in fact have no empirical/descriptive content; there is no substance to the claim that an agent is an expected utility maximizer, or chooses Nash equilibrium strategies, because the agent could not fail to have preferences of this sort.

For instance, I mentioned the phenomenon of regret that Maurice Allais saw as a challenge to EU theory, at least from the descriptive point of view. Allais showed that people seem to value possible outcomes differently depending on what are the other possible outcomes that they might have received instead. To give a new example, consider someone who values the outcome 'win travel scholarship' differently, depending on whether the other possible outcome was 'win nothing' or 'win travel scholarship plus living expenses'. These different evaluations of the same outcome, depending on what are the other chancy outcomes, are not consistent with EU theory. There is, however, a way to accommodate such regret phenomena within EU theory. At the extreme, we can simply include in the description of an outcome what else might have, but did not, occur. The following could then be regarded as two distinct outcomes: 'win travel scholarship when the other possibility was nothing' and 'win travel scholarship when the other possibility was a better scholarship'. Distinct outcomes can of course be evaluated differently by the lights of EU theory. But at what cost do we 'save EU theory' in this way? If, whenever an agent appears to care about more than just the *expected utility* of

options, the attitudes in question, such as regret, can simply be accounted for by enlarging the description of outcomes, it is unclear whether maximizing expected utility has any substantial meaning. Indeed, a number of theorists have debated this issue. A similar worry applies to game models. If, whenever agents appear to strategize at odds with Nash equilibria, we can simply redefine the game so that the utility functions track everything the agents seem to care about (as per my treatment of the Ultimatum Game), the theory is apparently unfalsifiable, i.e. it cannot be proven wrong.

So we see that the question of whether EU/game theory is false or unfalsifiable depends on the level of detail in which we may describe decision outcomes. The trouble is, it is very difficult to justify any particular level of detail, or any particular list of properties that may be used to distinguish the outcomes. This is a very pressing question for normative inquiry—it would be worrying indeed if our theories of rational choice were wrong or vacuous. The question is not so relevant for empirical uses of choice models, however, and this is our interest here. Note that there is an important difference between EU *theory* being unfalsifiable, and any particular EU *model* being unfalsifiable. In the former case, the key issue is whether it is only *rational* preferences/choice behaviour that can be represented by *some* EU model, or whether in fact *all* preferences/choice behaviour can be represented by *some* EU model. The arguments in the literature and my comments concerning regret attitudes are directed at this issue. This need not concern us as social scientists however. The important question for us is whether *particular* EU/game models that describe acts/states/outcomes in a specific way are adequate or inadequate for given descriptive, explanatory, or predictive tasks. For instance, does an EU model with utility increasing linearly with net family income adequately describe the impact on Bangladeshi families of a microfinance scheme? To give another example: is it useful, for the purposes of explaining international climate action to date and predicting future prospects for cooperation, to characterize national governments as playing a Prisoners' Dilemma game with respect to controlling their carbon dioxide emissions?

It is worth noting that it may well be empirically useful to treat EU *theory* as unfalsifiable. Francesco Guala has recently (2008) argued that EU theory is often used as a measuring instrument; we assume that an agent maximizes EU, and this allows us to measure her preferences in a given context and thereby characterize her with a particular EU *model*. The utility function (and/or the belief function) is determined by what model best 'fits' observed choice behaviour. Indeed, I noted earlier that using betting behaviour to infer an agent's probabilistic beliefs and/or utilities for outcomes has a long pedigree in decision theory that can be traced at least to Ramsey. Game theory can similarly serve as a measuring instrument. Under the assumption that agents maximize expected utility and opt for Nash equilibria, if both players

settle on the 50:50 split in the Ultimatum Game, for instance, we can infer that they care about fairness and not just monetary outcomes. In this way we come to understand the agents' psychology and we have principled reasons for including fairness properties, over and above monetary outcomes, in models intended for further predictive purposes.

There may be good reasons, however, to restrict the content of choice-model outcomes. Perhaps the model must achieve a certain amount of generality, or perhaps, if the model is to be practically viable, the identifying properties of outcomes must be measurable by independent means. This would presumably rule out subtle risk and fairness attitudes. In some applications, at least, it is surely the case that the best balance between simplicity and predictive power involves distinguishing outcomes only in terms of money/material goods.

In the case of crude outcomes, it should not be taken for granted that agents are best represented as expected utility maximizers. Indeed, once we acknowledge some slippage between empirical choice models and rational choice theory, the question arises as to whether non-EU models may be more adequate for at least some empirical purposes. One might argue that the ideal type in the social sciences need not be the rational type. Alternatively, one might argue that the ideal type *is* the rational type, and yet, *when model outcomes are crudely described*, the rational type need not correspond to expected utility maximization. The properties of preferences that are linked to maximizing expected utility theory are plausibly self-evident requirements of rationality *only if* the preferences are interpreted as expressing an agent's all-things-considered attitudes of approval. But this interpretation rests on outcomes 'containing' everything that an agent cares about. This is not typical of outcomes in empirical models—they are generally cruder or more restricted than this—and thus the corresponding preferences must be interpreted differently, and need not conform to the expected utility principle.

So-called *behavioural* decision theories may be understood in this light. These are proposed alternatives to EU theory, and include prospect theory, regret theory, satisficing theory, and various choice *heuristics* or shortcuts (see, for instance, Kahneman et al. 1982). Prospect theory, for example, deviates from EU theory in this way. Instead of assessing acts in terms of the sum of the probability of each state multiplied by the utility of the outcome associated with that state, the probability term (which, recall, represents belief) is transformed or converted according to the agent's risk attitudes. The comparison of two acts *Ai* and *Aj* can then be written as follows:

$A_i \geq A_j$ IF and only IF

$$\sum_n r(\Pr(S_n)) \times U(O_{i,n}) \geq \sum_n r(\Pr(S_n)) \times U(O_{j,n})$$

This is very similar to the expected utility rule described earlier, except that the utility of outcomes is multiplied by $r(\Pr(S_n))$ instead of simply $\Pr(S_n)$, where r stands for the agent's (personal) risk function and is subject to the constraint that $r(0) = 0$ and $r(1) = 1$. The idea is that people tend to overweight or give extra importance to low probabilities, and underweight or give lesser importance to high probabilities. So, for instance, a probabilistic belief of 0.01 (that, say, one's house will burn down) might be converted by the r function to 0.10, i.e. $r(0.01) = 0.10$, and a probabilistic belief of 0.99 (that, say, one will win a million dollars) might be converted by the r function to 0.90, i.e. $r(0.99) = 0.90$.

There are various criticisms of basic prospect theory that have led to a refined version known as 'cumulative' prospect theory. But these details need not concern us here. The point is simply that prospect theory is one amongst a number of behavioural or *non-expected utility* decision theories. These theories are varied in form and motivation. Moreover, they are understood in different ways. Some consider their favoured behavioural decision theory to be a challenge to EU theory as an account of *rational* choice. Others simply consider these decision theories to be better suited to empirical tasks because they represent the irrational agents that we mostly are. There is, however, a third way. The theories may, in specific contexts, be better for empirical tasks because they represent rational choice given a limited or crude interpretation of outcomes. Prospect theory, for instance, may be considered useful because it can accommodate varying risk attitudes and yet still confine the description of outcomes to monetary amounts or material goods.

3.2 *Game Theory: The Reductionist Critique*

Some of these points have bearing on a further criticism of game theory: that it is inappropriately *reductionist* in its representation of social interaction, given that the basic units of game models are individuals, and specifically, their preferences. Indeed, game theory is paradigmatic of *methodological individualism*, a term first used to describe the Weberian programme of explaining group-level phenomena in terms of the attitudes and behaviour of the individuals who constitute the relevant groups. It is a form of explanation whereby groups are *reduced* to their constituent individuals. As we see in Chapter 6, some argue, however, that depicting individuals as basic is grossly misleading, given the way individuals themselves are shaped by social norms and shared institutions. In other words, the charge is that game theory has the arrow of explanation going the wrong way—from individual attitudes to social arrangements rather than from social arrangements to individual attitudes.

There is both a negative and a positive response to this charge. Let us begin with the negative or defensive line, and that is that the critics are confusing

their targets. They are arguing against *atomism*, where individuals are conceived as asocial entities that interact but do not influence each other's psychology. But atomism neither implies nor is implied by methodological individualism. To begin with, as already emphasized, there is no reason to take a narrow view with respect to the outcomes that individuals care about; the agents in a game model may be sensitive to each other's plights and to all sorts of social factors that go beyond material goods. Furthermore, the provenance of agents' attitudes is left open; it may well be that agents' preferences were socially conditioned or in other words simply reflect the dominant attitudes of the group.

More generally, the existing social and institutional arrangements act as constraints on individuals, and play a corresponding role in game models. Beyond the agents' preferences, the prevailing social setting determines what strategies are actually available to agents, and the sorts of outcomes that result from combinations of strategy choices. For instance, the institutional setting of scientific publishing constrains scientists' options with respect to project funding and submission of work, and also influences the outcomes associated with the various players' choices of projects and submissions. When choosing an avenue of work, the scientist must take into account the competition of her peers and also their expected assessment of her work, as this, together with the procedures of journals and grant agencies, partially determines her outcomes. The social setting, in other words, shapes the 'rules of the game', and this is expressed in the parameters of a game model. In other words, game theory need not be blind to the fact that individuals are embedded in a social/institutional context!

The critic might nonetheless push further and argue that the real explanatory task is to explain these social forces and institutions that constrain individuals, and not simply to posit them in a model. To some extent, however, this argument misses the mark. A scientific model can only represent/explain some limited aspect of the world. In one model, particular social norms or institutions may serve as boundary conditions or constraining assumptions, but this is not to say that the emergence and/or persistence of these institutions cannot be explained via a different game model. Presumably, in such a model, 'higher-order' or more basic institutions will serve as the background constraints. Returning to our scientific publishing example, a different game model could be used to explain the procedures of the publishing business that served as the institutional context in the model alluded to; the players in the new model might be competing journals or grant agencies, rather than scientists themselves. The 'higher order' institutions will presumably amount to the broader economic and educational setting which shapes the possibilities and outcomes for the scientific journals.

The appeal to different orders of game models goes some way towards answering the critic's charge that game theory does not explain what we most

want to understand: social forces and institutions. There remains the problem, however, of explaining the mechanisms by which social forces shape individual preferences. We can describe ever richer layers of game models, but these models are still underpinned, at bottom, by fixed individual preferences. Indeed, this must be acknowledged as a limit of standard game theory (if not evolutionary game theory); it does not elucidate the process of preference change, since a basic ingredient of any game model is the set of players with their fixed preferences over prospects.

So much for the negative argument—what game models need not be. In short, they need not assume asocial agents. The positive argument is that the insights afforded by game models can be very valuable. In fact, Weber claimed that individualist explanation is what distinguishes the social sciences from other sciences: a key role of the social sciences is to offer an interpretative explanation of action in terms of *subjective attitudes*, and only individuals, as opposed to groups, are the locus of subjective attitudes. The basic idea is that we do not properly understand what people do unless we understand *why* they do it, and to answer such questions we must inevitably appeal to the beliefs and desires of individuals.

Furthermore, some striking cases highlight that we ignore individual attitudes at our peril. We need not look further than games like the Prisoners' Dilemma, introduced earlier. This game shows that, even if there is a collective goal or a common preferred outcome within a group, this outcome may nonetheless not be realized due to the structure of individual incentives. (Recall for the Prisoners' Dilemma that both agents value the fruits of cooperation, but, alas, both have an incentive to 'free-ride'.) The contemporary social and political theorist Jon Elster makes much of these game-theoretic insights in arguing for the acceptance of methodological individualism. If a common goal is salient in a group, we may make hasty predictions that the goal will be achieved, when a closer examination at the level of individuals would reveal otherwise. Elster (1985) appeals to a striking example in political science to make this point vivid: Marxist theory. He argues that Marxists appealed to the general will of the proletariat as the predicted driver of social change and failed to appreciate the more subtle motivations of individual workers. These individuals may well be sympathetic to the end-result of revolutionary change but nonetheless have an incentive to free-ride on the revolutionary efforts of others.

Some decades earlier in the twentieth century, and prior to the popularization of games like the Prisoners' Dilemma, the economist Friedrich von Hayek made similar cautionary remarks about the realization of social ideals. Hayek did not stress free-riding so much as the different frames of motivation of individual citizens versus policy-makers. While policy-makers may be keenly aware of society-level variables like inflation, citizens going about

their regular lives do not generally respond to these factors but rather have a more local set of concerns. The upshot is that the connection between macroeconomic variables and individual behaviour is not straightforward. The best-laid plans for the collective may be difficult to orchestrate at the individual level. (This is, in a sense, the flip-side of the *invisible hand* phenomenon, where unorchestrated individual activity leads to the appearance of planning at the collective level.)

Hayek's message is that we should be sceptical of social planning initiatives. We need not accept his conclusion, but we should take heed of Hayek's worries. Indeed, the conclusion one might draw is that social planning proposals may be better understood through careful game-theoretic modelling!

4. Concluding Remarks

One major conclusion to draw from this chapter is that we should not ask whether choice models, *in general*, are true/false or unfalsifiable in the social sciences, but rather whether *particular* choice models are useful. To be useful, a model must be true enough, and it must yield insights or predictions that justify the trouble of modelling in the first place. Furthermore, the question of how we can *permissibly* construct choice models, and, in particular, define outcomes, is misguided in the empirical setting, even if the question is important in the normative setting. What are the appropriate properties of decision outcomes in empirical models simply depends on what facilitates adequate explanation and/or prediction.

A consequence of these points is that we cannot take for granted that the best empirical models are expected utility models. Having said that, the expected utility principle has a powerful simplicity, and it underlies much analysis in game theory, so should not be dismissed too hastily.

References

Alexander, J. M. (2009). 'Evolutionary Game Theory', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Fall 2009 Edition) <http://plato.stanford.edu/archives/fall2009/entries/game-evolutionary>.

Elster, J. (1985). *Making Sense of Marx: Studies in Marxism and Social Theory*. Cambridge: Cambridge University Press.

Guala, F. (2008). 'Paradigmatic Experiments: The Ultimatum Game from Testing to Measurement Device', *Philosophy of Science*, 75: 658–69.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Weber, M. (1978). *Economy and Society*, ed. G. Roth and C. Wittich. Berkeley, CA: University of California Press.

Further Readings

Binmore, K. (2007). *Game Theory: A Very Short Introduction*. Oxford: Oxford University Press.

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty and Time*. Oxford and Cambridge: Basil Blackwell.

Hausman, D. M., and McPherson, M. S. (1996). *Economic Analysis and Moral Philosophy*. Cambridge and New York: Cambridge University Press.

Heath, J. (2011). 'Methodological Individualism', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2011 Edition) <http://plato.stanford.edu/archives/spr2011/entries/methodological-individualism>.

Resnik, M. D. (1987). *Choices: An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press.