Chapter 1
Preliminaries

1.1 The Logic of Game Theory

Human actions can be made sense of in various ways. Combining the dichotomy of understanding and explanation on the one hand, and that of individualism and holism on the other, we can explain actions as expressing meaning, as governed by rules, as fulfilling functions in larger systems, or as being based on individual reasons.[1]

Which of these four modes applies to the theory of games? This question suggests that there is only one way to use game theory in the social sciences, which may not be too plausible given the creativity of the social scientist working in any of the four frameworks and applying game theory in a way she judges productive. Nonetheless, if we take the declared aims of non-cooperative game theory seriously, we are almost automatically led to think of game theory as conforming to the belief–desire framework of action explanation. Game theory explains actions in terms of the reasons agents have to carry them out.[2] While this may not be the only way to make sense of game theory, it is certainly a central sense of game theory. The belief–desire framework forms the basis of this book.

This chapter first deepens our understanding of the contrast between decision theory and game theory, and shows the ways in which to give a precise description of the differences between the two theories in epistemic terms. Subsequently, I will consider normal form and extensive form game-playing in order to distinguish, as their relevant elements, the possible actions a player can choose to perform, her preference ordering and rationality principle, the action eventually performed, and, finally, the beliefs about—with recursion intended—all five ingredients. For extensive games, a one-shot interpretation and a many-moment interpretation will be set apart.

---

[1] Martin Hollis, *The Philosophy of Social Science: An Introduction* (Cambridge: Cambridge University Press, 1994).

[2] The belief–desire framework has most notably been defended by Donald Davidson, *Essays on Actions and Events* (Oxford: Clarendon Press, 1980).

B. de Bruin, *Explaining Games*, Synthese Library 346,
DOI 10.1007/978-1-4020-9906-9_1,

1.1.1 Decision Theory and Game Theory

It is standard to account for the differences between decision and game theory in terms of the number of players, and to note that where decision theory is concerned with one player who has to act in a situation of certainty, uncertainty or risk, game theory is concerned with several players who interact strategically. The number of players, however, is not the crucial factor here. Sometimes, for instance, decision theory is held as the study of individuals playing against a second player, nature. Rather, the difference is that decision theory involves only one agent with beliefs and desires—nature does not have beliefs or desires—and that game-theoretic agents all have beliefs and desires.[3]

This distinctive feature is unequivocally mirrored in the way games against nature and games against opponents with beliefs and desires are represented. The entries of the decision matrix are pairs of real numbers in game theory, but only single real numbers in decision theory. Nonetheless, what these differences lead to—what game-theoretic agents are supposed to do with their beliefs and desires—remains to be examined.

1.1.1.1 The Ban on Exogenous Information

John von Neumann and Oskar Morgenstern describe the function of preferences as follows:

> Every participant can determine the variables which describe his own actions but not those of the others. Nevertheless those 'alien' variables cannot, from his point of view, be described by statistical assumptions. This is because the others are guided, just as he himself, by rational principles—whatever that may mean—and no *modus procedendi* can be correct which does not attempt to understand those principles and the interactions of the conflicting interests of all participants.[4]

That being the case, the difference between decision and game theory is not so much that opponents have beliefs and desires, but rather that any player has to consider her opponents as having such beliefs and desires. The right way to think about your opponent is the way that you think about yourself, as a rational being acting on beliefs and desires, not in the way that you think of the weather or radioactive decay.

One of the most striking consequences of this view—a consequence that recurs throughout the entire book—is that in order to find out what your opponents will do, the only information a game-theoretic agent is supposed to use is her beliefs and desires. Statistical data or any other exogenously based data are out. To conceive of your opponents' agency as guided by rationality means that numerous epistemic

[3] In an alternative vocabulary this is the distinction between *parametric* and *strategic* choice situations. Decision theory is also called *rational choice theory*. A classic reference is R. Duncan Luce and Howard Raiffa, *Games and Decisions: Introduction and Critical Survey* (New York: Wiley, 1957).

[4] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1944), 11.

policies that are entirely natural in meteorology or nuclear physics are out. This is quite literally a *ban on exogenous information*. Decision-theoretic models typically have individuals base their probabilistic beliefs on statistics and other outside data. For instance, in order to choose with the help of decision theory whether to build a nuclear plant, statistical data about nuclear catastrophes are essential. Von Neumann and Morgenstern's call to choose one's strategy on the basis of reasoning processes that do not transcend the game structure, but only refer to the rationality principles and the utility functions of the players, excludes such forms of information, however.

In general, various ways to form beliefs about what opponents will play are available. Players may consult statistical surveys, they may call upon experience, they may read historical texts, or they may watch videos of previous games. The ban on exogenous information rules all this out, though, and so the question is: how can game-theoretic agents, if the ban is genuinely enforced, ever obtain information about the prospective strategic choices of their opponents?

To show how merely endogenous information can form the basis of substantial belief formation and lead players to select certain strategies and to exclude others is, in fact, the main research goal of non-cooperative game theory. Up until the 1990s, game-theorists thought that solution concepts that refine the Nash equilibrium would accomplish this goal; but nowadays the Epistemic Programmeis considered the most likely source of insight, particularly by way of mathematical results called *epistemic characterisation theorems*.[5] Strictly obeying the ban on exogenous information, the main idea is that I can derive sensible and precise predictions of my opponents' prospective strategic behaviour from the assumption that they are as rational as I am.[6]

1.1.1.2 Epistemic Characterisation Theorems

In order to get some feel for the general form of epistemic characterisation theorems, let me first examine the logic of decision-theoretic modelling, and then contrast it with the logic of game-theoretic modelling. Consider the decision problem shown in Figure 1.1 where agent i has to choose between actions i_1, i_2 and i_3, and where the possible states of nature are ω_1, ω_2 and ω_3. If i does not have any information about what state of nature will obtain, she may decide to act on a number of divergent and mutually inconsistent principles. She may decide to maximise the minimal payoff or utility, to minimise the maximum possible regret (risk), to use Hurwicz's pessimism-optimism index criterion, or she may use a principle of decision under uncertainty

[5] While 'Decision-Theoretic Foundations Programme', 'Epistemic Foundations Programme' or 'Interactive Epistemology' are good candidates, the terminology of 'Epistemic Programme' seems to be gaining popularity in the literature (Adam Brandenburger, personal communication).

[6] It may be worth pointing out that the ban on exogenous information still enjoys wide support. Adam Brandenburger and Amanda Friedenberg, 'Intrinsic Correlation in Games', *Journal of Economic Theory*, 141 (2008), 30 consider as endogenous the beliefs of the players in so far as they are derived from a hierarchy of common beliefs, emphasising that it is standard in the Epistemic Programme to consider these variables as part of the description of the game.

	ω_1	ω_2	ω_3
i_1	1	0	1
i_2	5	0	3
i_3	3	1	2

Fig. 1.1

based on Laplace's criterion of insufficient reason. But it is clear that she will never choose her first strategy—unless she wishes to lose—as it is *strictly* (or *strongly*) *dominated* by i_3.

Equally clear is that if she possessed information to the effect that the first or third state of nature will obtain, she would choose her second strategy and if she believed that the second state is the actual world, she would choose i_3. In easily understood formalism, this reasoning is of the form

$$(i\text{'s utility} \wedge i\text{'s rationality} \wedge \Box_i(\text{possible states of nature})) \rightarrow i\text{'s actions},$$

where the $\Box_i$ is used for i's beliefs. The sentence attests that i's actions ensue from her payoffs, rationality and beliefs about possible states of nature.[7]

	2_1	2_2	2_3
1_1	(1,5)	(0,1)	(1,3)
1_2	(5,0)	(0,0)	(3,1)
1_3	(3,3)	(1,1)	(2,2)

Fig. 1.2

Turning to game theory proper, consider the game shown in Figure 1.2. It is a normal form game between two players 1 and 2. Player 1 has the payoffs agent i had in the above decision problem, and as a result she will not play her first strategy

[7] I adopt the game-theoretic convention that actions and objects of beliefs are thought of extensionally. It is crucial to note the plural in the consequent, because rationality principles do not always fix unique actions.

here either. Furthermore, it is plain that her choice between her second and her third strategy ought to depend on what she believes about the prospective choices of her opponent. If he chooses 2_2, then 1_3 is the best choice for her; if he chooses something else, then she should play 1_2—unless, again, she wishes to lose.

The decision-theoretic analysis only proceeds if the agent possesses exogenous, statistical information about nature's prospective moves. How does the Epistemic Programme circumvent the use of such data? In the present case, player 1 can further develop her beliefs about her opponent in purely endogenous ways, that is, only referring to beliefs, utilities and rationalities. To see this, consider what player 1 can do with the extra information about player 2's utility function. Player 1 observes that playing 2_3 is always better for player 2 than playing 2_2, and this kind of reasoning helps her to decide upon her own action; playing 1_2 is the best choice if player 2 does not play 2_2. Nor is this all, because player 2 only avoids playing dominated strategies if he does not wish to lose. For that reason, player 1 also has to have information about her opponent's rationality; she should know that he maximises expected utility. In formalism, the assumptions allowing me to conclude that player 1 believes that player 2 does not play 2_2 and that therefore player 1 plays her second strategy are fully captured in the antecedent of

$$(1\text{'s utility} \wedge 1\text{'s rationality} \wedge \Box_1(2\text{'s utility}) \wedge \Box_1(2\text{'s rationality})) \rightarrow 1\text{'s actions}.$$

It is natural to ask whether similar reasoning can be used to determine player 2's choice of strategy. The analogue of the above assumptions,

$$(2\text{'s utility} \wedge 2\text{'s rationality} \wedge \Box_2(1\text{'s utility}) \wedge \Box_2(1\text{'s rationality})) \rightarrow 2\text{'s actions},$$

is insufficient, though. Being rational, player 2 will not play 2_2; it is strictly dominated by 2_3. Strategy 2_3, moreover, is a better choice against 1_2, but 2_1 is better against 1_1 and 1_3. Since player 2 believes 1 to be rational, he will exclude her from playing 1_1. That does not help him too much, however, as long as he does not have more information about whether 1 will play 1_2 or 1_3. Unfortunately for player 2, there is no way to obtain more information on the basis of the epistemic setting described in the antecedent of the above implication. While player 1 considers 1_1 to be a bad strategy no matter what, her opinion about 1_2 and 1_3 depends on what she believes that player 2 will do. But as long as player 2 has no information about player 1's beliefs about what 2 will do, player 2 has no basis for beliefs about which of the two strategies 1_2 and 1_3 player 1 plays.

With more elaborate, yet still exclusively endogenous information, progress can be made, though. To see this, suppose that

$$\begin{aligned} 2\text{'s utility} \wedge 2\text{'s rationality} \wedge \Box_2(1\text{'s utility}) \wedge \Box_2(1\text{'s rationality}) \wedge \\ \Box_2\Box_1(2\text{'s utility}) \wedge \Box_2\Box_1(2\text{'s rationality}). \end{aligned}$$

Player 2 believes that 1 is rational and that 1 has a utility function as shown in Figure 1.2. From this, player 2 concludes that 1 will not play 1_1. Player 2 believes, in addition, that 1 believes that 2 is rational and that 2 has payoffs as stipulated

in the game matrix. A rational player possessing such preferences never plays 2_2, and therefore player 2 believes that 1 believes that 2 will not play 2_2. Player 2 then observes that—provided 1 is rational and has the utility function the matrix details—the best response of 1 to the belief that 2_2 will not be played is playing 1_2. Consequently, player 2 believes that 1 will play 1_2. He has made his beliefs about his opponent more precise, and this allows him, in particular, to decide between 2_1 and 2_3. Believing that his opponent will play her second action, he chooses 2_3. That is, a precise description of player 2's choice of strategy can be derived from the more elaborate epistemic assumption.

The general form of epistemic characterisation results emerges. What I have established is something of the form

$$\begin{aligned}(\mathbf{u}_1 = \dots) \wedge \mathbf{rat}_1 \wedge (\mathbf{u}_2 = \dots) \wedge \mathbf{rat}_2 \wedge \\ \Box_1(\mathbf{u}_2 = \dots) \wedge \Box_2(\mathbf{u}_1 = \dots) \wedge \Box_1\Box_2(\mathbf{u}_1 = \dots) \wedge \Box_2\Box_1(\mathbf{u}_2 = \dots) \wedge \\ \Box_1(\mathbf{rat}_2) \wedge \Box_2(\mathbf{rat}_1) \wedge \Box_1\Box_2(\mathbf{rat}_1) \wedge \Box_2\Box_1(\mathbf{rat}_2) \rightarrow \mathbf{1}_2 \wedge \mathbf{2}_3,\end{aligned}$$

where $\mathbf{u}_i = \dots$ abbreviates a complete description of i's utility function, $\mathbf{rat}_i$ means that i is an expected utility maximiser (is rational), $\mathbf{i}_k$ means the player i plays her kth strategy. Anticipating the discussion in Chapter 2, this is a particular instance of the epistemic characterisation theorem to the effect that common true belief about rationality and utility entails that players choose strategies that survive the *iterated elimination* of strictly dominated strategies. Epistemic characterisation results, in short, are sentences of the form

$$\varphi(\mathbf{rat}_1, \mathbf{rat}_2, \mathbf{u}_1, \mathbf{u}_2) \rightarrow \text{actions},$$

where φ is a formula in which epistemic operators $\Box_1$ and $\Box_2$ may be used, nested arbitrarily deeply. The statement about actions in the consequent is a statement of the form $\mathbf{1}_k \wedge \mathbf{2}_l$ describing a situation in which 1 plays her kth strategy and 2 his lth, or a (finite) disjunction of such statements $\bigvee_{k,l}(\mathbf{1}_k \wedge \mathbf{2}_l)$, if the antecedent epistemic conditions are insufficient to attribute performing one single action to each player.[8]

1.1.2 Normal Form Games

Different assumptions about beliefs, desires and rationality principles epistemically characterise different game-theoretic solution concepts.[9] Similar assumptions appear in almost every characterisation result, and the aim of this section is to expli-

[8] A consequence of non-uniqueness is that a full explanation of the specific action performed by an agent cannot always be given in entirely game-theoretic terms. Decision and game theory explain that the action actually chosen lies in some set of possible actions. But neither theory can always account for why one action is chosen rather than another.

[9] The treatment of these issues has benefited from detailed written comments by Wlodek Rabinowicz, for which I am very grateful.

cate them. While it is entirely false that no models could be developed for situations in which these assumptions are not satisfied, without them epistemic characterisation results would often make less sense to the Epistemic Programme in game theory—but, as will become clear, there are important exceptions. Evolutionary, behavioural, stochastic and cooperative game theory all provide ample space for relaxing the assumptions—and some non-cooperative game theorists may also wish to resist adopting the viewpoint promoted by the Epistemic Programme—but this does not contradict my argument in this section, because the aims of those researchers are crucially different from ours.[10]

To start with, players of a normal form game can choose from a set of possible actions, in most models finite, and never containing fewer than two elements. Modelling a situation as a decision or game theorist means ascribing to the agents weak total preference orderings over all possible outcomes of the game. Ordinal orderings are often sufficient in the Epistemic Programme, but the full force of the von Neumann–Morgenstern axioms is needed whenever preference orderings are to be uniquely represented by means of utility functions modulo linear transformations; this is the content of Theorem A.1 (see the Appendix A).[11] Games are almost never specified using preference orderings. Utility functions are the standard.

Even though preferences are essential, they are insufficient to give a full motivation for actions; they do not on their own constitute reasons to act, but stand in need of at least a principle telling the agent what to do with her preferences. Decisions under certainty or uncertainty allow for a fair number of different principles. Decisions under risk involve the maximisation of expected utility, and this is the principle employed in game theory, too.[12] In the Epistemic Programme, no agent ever acts without a principle of rationality.

It follows that acting is rationally choosing an action from a non-empty, non-singleton choice set. If agents did not choose, they would dawdle and leave the game unfinished, and if agents chose more than one action, they would be spoilsports and ruin the game. Only if all players choose precisely one action can the game reach an outcome. In normal form games, a related condition is often phrased by stipulating that players play simultaneously, which means that players have no chance to observe what their opponents do. They act in ignorance of what their opponents choose—notwithstanding the fact that they may have very accurate beliefs that would predict their opponents' actions. This requirement could be envisaged by players who make their choice in private first, handing their choice over to

[10] The stance here is logical, not epistemological. No critical evaluation of the plausibility of the assumptions is carried out; only an investigation into the logic underlying epistemic characterisation results.

[11] A weak total ordering is a reflexive linear ordering. To be precise, the ordering ranges over the lotteries composed of the possible outcomes of the decision problem or game, satisfying, in addition, conditions of monotonicity, substitutability, continuity and reduction, i.e., the von Neumann–Morgenstern axioms. See, e.g., Luce and Raiffa, op. cit. 23–31. For an alternative rendering that has become the standard in the Epistemic Programme, see the Appendix A.

[12] If a player's preference ordering is an ordinal one, or if her beliefs are not of the Kolmogorov form, not all of the mathematical details of expected utility maximisation are needed in full to make a player play, and a simpler definition of rationality can be used.

an objective umpire who reveals the choices—and thereby announces the achieved outcome—as soon as she has received them all.

Accompanied by highly specific utility functions, rationality principles are still often powerless if a player does not have beliefs. As I have suggested, fully pinning down the strategic choice of a game-theoretic agent involves quite complex reasoning with nested epistemic operators involving the beliefs of one player about those of another. In a typical situation, a player possesses beliefs about her and her opponents' possible actions, about her and her opponents' preference orderings, about her and her opponents' rationality principles, about her and her opponents' actions to be performed, and—with recursion intended to generate the necessary hierarchy of beliefs—about her and her opponents' beliefs.

Before turning to these five kinds of beliefs in more detail, it is important to realise that what applies to preferences also applies to beliefs: they take, in game theory, a specific kind of object. While in most concrete examples the beliefs as well as the preferences can be stated in everyday language, beliefs are regularly—but not always—taken to be probability measures over outcomes. They have to satisfy the Kolmogorov axioms to ensure, most importantly, that summing up probabilities is allowed whenever the probability of the disjunction of two independent events will be calculated. More intricate models are needed when we are interested in the ways players would change their beliefs if they learned that their current beliefs were wrong. Plain standard probability theory does not tell us much about how to apply Bayes' Rule to null events. Belief revision theory, by contrast, sorts out more (theoretically) rational from less rational ways to update and correct one's epistemic state. In the characterisation of several game-theoretic solution concepts, the Epistemic Programme has employed this theory to describe how the players' dispositions to revise beliefs influence the outcome of the game.

Of the five ingredients, the first is clear. Without beliefs about what to choose between, players cannot make a choice, and without beliefs about what outcomes may result, the players' choices of action cannot be guided by their opinions about the desirability of possible outcomes. Information about possible actions suffices to that end, because the set of possible outcomes is entirely determined by the possible actions of the players.

Second, players ought to have correct beliefs about their own preferences—or at least, beliefs that are sufficiently correct to guide decision making. If players believe they do not have preference orderings, preferences cannot be considered as reasons for their actions; and if they believe they have certain preferences, but their true preferences are very different, it is hard to tell whether they inform their agency. In fact, it seems as though completely incorrect beliefs about preferences are incoherent—at least for the purposes of decision and game theory. If players believe they have preferences that are different from those stipulated by the model, and act on those beliefs, then there is much to recommend that the theorist substitute the preference ordering in the model with the believed ordering. In the absence of any beliefs about preferences, only unconscious motivation would make sense, and

without beliefs about preferences that are not at least approximately correct, there is no theory of games at all.[13]

This point is important, in particular if various degrees of autonomous choice are to be distinguished. Unconscious motivation has to be rejected, because players who are unconscious of their own preferences can hardly be thought to play a game. Of course, there is nothing incoherent about describing a person as gradually getting to know her real preferences and making conscious what she was previously unconscious of, and this may not be completely impossible in decision theory. In fact, someone's actions may be neatly modelled as maximin actions in some model ascribing preferences to the agent she first refused to acknowledge as her own. As the theorist continues modelling, however, we find that the model describes her behaviour correctly more often than not, and this prompts the agent to accept the model's specification of her preferences as correct. The models, one could say, explained her actions correctly, but it took a while before the agent herself became aware of that.

Yet this does not make sense in non-cooperative game theory as long as the perspective of the Epistemic Programme is adopted. A similar scenario where someone's actions are modelled as iteratively strictly undominated, for instance, may indeed turn out to describe her behaviour adequately, and the agent may indeed acknowledge that this reveals her true preferences. It is doubtful, however, whether a real explanation of her actions would be given, because the epistemic conditions of iterated strict dominance were not satisfied; these conditions involve common true belief about rationality and utility, and this was lacking as the modelled preferences were different from the actual preferences—the beliefs were common but not true. The distinction between players' *real* and *believed* preferences, in other words, is rejected, because at most only one preference ordering can play a motivating role, and the motivating preference ordering would need to appear in the game-theoretic model. Moreover, while not denying the conceptual possibility of players entertaining the belief that they do not possess any preference orderings, I reject the relevance of such possibilities to decision and game theory. If such beliefs are false, the preference ordering probably provides unconscious reasons for action; and if they true, there is just nothing decision and game theory can do.

Nor does it make sense, for game-theoretic purposes, to talk about players who are wholly uncertain about their preferences. If players are so uncertain about their preferences as to make it impossible to decide upon an action, then they just lack preference orderings. If, on the other hand, they can still decide on actions, they are, for game-theoretic purposes, just players with a particular preference ordering—or a range of possible orderings—that inspires the performance of one particular action. To summarise, as Adam Brandenburger and Robert Aumann write in their seminal paper on the epistemic characterisation of the Nash equilibrium, 'knowl-

[13] Players may be ignorant of their future preferences in extensive games under the many-moment interpretation.

edge of one's own payoff function may be considered tautologous'.[14] Without this assumption, the Epistemic Programme would not fully flourish.

Players not only need beliefs about their own preferences, they also need beliefs about the rationality principle that they use. If I explain the behaviour of a certain agent in terms of maximin, but she sees herself as solving a maximisation problem corresponding to the utility function and the probabilistic beliefs she believes herself to possess, then I misconstrue the reasons underlying her choice of action. Whenever I say that preferences, beliefs and rationality principles explain actions, I actually mean to refer to beliefs about preferences, beliefs about rationality and beliefs about a number of other aspects of the game-playing situation (available actions, outcomes, and so on).[15]

Of course, players have to have numerous beliefs about their own role in the game. The Epistemic Programme has helped to uncover the fact that players must have beliefs about each other. Games in which players do not have a clue about the preferences of their opponents are reducible, for them, to games against nature; all entries in the game matrix—except their own—can be removed, as far as they are concerned. Acting on principles from decision theory, they would choose on the basis of exogenous information. In the game shown in Figure 1.2 player 1 may form the belief that 2 will not play his second strategy on the basis of the information 1 has about 2's preferences, because player 1 observed that 2_2 is strictly dominated by 2_3. Without beliefs about 2's preference ordering and rationality, player 1 would not have been in the position to make the argument.[16]

In summary, for the Epistemic Programme in game theory to make sense, a player of a normal form game must have beliefs about her and her opponents' possible actions, about her and her opponents' preference orderings, about her and her opponents' rationality principles, about her and her opponents' actions to be performed, and, with recursion to generate the necessary hierarchy of beliefs, about her and her opponents' beliefs. This five-tiered structure reoccurs in game-playing situations of extensive games—but with crucial complications.

---

[14] They state that 'Knowledge of one's own payoff function may be considered tautologous', 'Epistemic Conditions for Nash Equilibrium', *Econometrica*, 63 (1995), 1162. This is not to exclude Bayesian games and other models of incomplete or imperfect information. The aim of modelling such games is to capture situations where individuals are less than fully informed about the preference relations of their opponents, or even about certain exogenous features of the actual world. This is compatible with the claim I defend, because—to stay within the framework of the Epistemic Programme—reasoning about solution concepts of such games involves considering larger games in which these pieces of information have been 'endogenised'. For a textbook treatment, see Martin Osborne and Ariel Rubinstein, *A Course in Game Theory* (Cambridge: MIT Press, 1994), 24–27.

[15] I use the concept of *game-playing situation* rather loosely. Cf., e.g., Adam Brandenburger, 'The Power of Paradox: Some Recent Developments in Interactive Epistemology', *International Journal of Game Theory*, 35 (2007), 465–492.

[16] Again, these assumptions can be somewhat relaxed. Sensible conclusions can be derived about the likelihood of opponents performing certain strategies even when a player only has approximate, not entirely accurate beliefs about opponent utility functions. For further discussion see Section 2.4.

1.1.3 Extensive Games: The One-Shot Interpretation

Extensive games model temporally extended, sequential strategic interaction. Normal form games, by contrast, model one-shot events. That, at least, is the common view. For von Neumann and Morgenstern, however, normal form and extensive form games are different possible models of one and the same sort of thing. 'Imagine,' they write,

> that each player... instead of making each decision as the necessity for it arises, makes up his mind in advance for all possible contingencies.[17]

Then this is no restriction of his freedom of action

> because the strategy is supposed to specify every particular decision only as a function of just that amount of actual information which would be available for this purpose in an actual play.[18]

One thing can be modelled in two different ways. The *normalised* form is more appropriate for proving general results, the *extensive* form is better when one wishes to analyse particular cases, but for the founding fathers of game theory the two forms are 'strictly equivalent'.[19] To analyse the characterisation results that the Epistemic Programme has obtained about extensive form games, it pays to set out these two interpretations more clearly, and I do this by distinguishing between a one-shot interpretation and a many-moment interpretation of extensive form games. The latter interpretation is most sensitive to sequentiality, but the one-shot interpretation, too, is still different in subtle respects from pure normal form games.

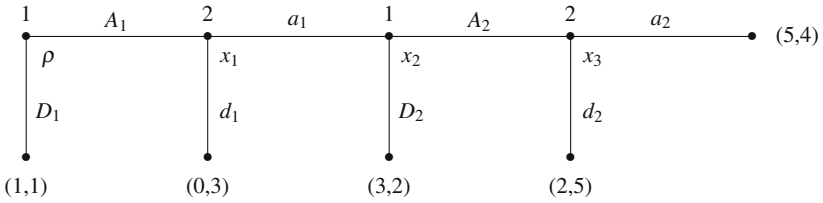

Fig. 1.3

According to the one-shot interpretation, players of an extensive game act simultaneously, and choose between a set of strategies fixing the actions at any of their

[17] Op. cit. 79

[18] Ibid.

[19] Ibid., 85. It is not terribly clear what 'strict equivalence' means here. For further discussion see Boudewijn de Bruin, 'Game Transformations and Game Equivalence', ILLC Technical Note X-1999-01 (University of Amsterdam, 1999), and Susan Elmes and Philip Reny, 'On the Strategic Equivalence of Extensive Form Games', *Journal of Economic Theory*, 62 (1994), 1–23.

decision nodes in the game. In the Centipede game shown in Figure 1.3, for instance, player 1 can choose between the strategies D_1D_2, D_1A_2, A_1D_2 and A_1A_2. In order to find the assumptions on one-shot extensive game-playing situations one could expect, first, that copying the conditions on the five elements of normal form game-playing situations suffices. Preference orderings surely satisfy the von Neumann–Morgenstern axioms, players act on principles of rationality, and they ought to have beliefs about possible strategies, performed strategies, preferences, rationality principles, and, with the usual recursion, about beliefs. Moreover, similar to playing normal form games, players decide on precisely one *full strategy* (a function, here, mapping any decision nodes of hers to an immediate successor) and choose simultaneously.

Under the one-shot interpretation, players of an extensive game can choose between a set of possible full strategies prescribing a unique action at any decision node. At every decision node more than one action is possible, and on that account the set of possible full strategies contains at least two elements. From these possible strategies, players pick precisely one, without the game structure allowing them to have information about the choices that the others make. In short, they choose simultaneously.

Preference orderings in normal form game-playing situations range over strategy profiles combining the strategic choices of all the players. Different strategy profiles give rise to different outcomes with different—or identical—utility for one or more players. Preference orderings in one-shot extensive game-playing situations, by contrast, are defined over the terminal nodes of the game tree, the genuine outcomes of the game, and the same terminal node can often be reached by more than one full strategy. In the Centipede , for instance, strategy profiles (A_1D_2, d_1d_2), (A_1A_2, d_1d_2), (A_1D_2, d_1a_2) and (A_1A_2, d_1a_2) determine the same outcome.

More subtle issues arise when we consider rationality in the one-shot interpretation, as we have to decide whether one-shot principles of rationality differ from normal form principles. Using normal form rationality as one-shot rationality, a full strategy in an extensive game is rational for a player given her beliefs and utility function whenever the full strategy is rational in the normal form version of the game. Different notions of rationality for normal form games (weak domination, strict domination, perfect rationality, and so on) will give rise to different notions of rationality for extensive games, without, however, essentially referring to their sequential character. An alternative way to define one-shot extensive game rationality does exploit the extra structural properties of extensive games by considering a full strategy not only in the whole game but also in all of its subgames. Imagine, for instance, that player 2 in the Centipede wishes to quit the game at the first decision node, because this is presumably the best thing to do given the player's beliefs and preferences. Looking at the normal form of the game, it does not matter whether she plays d_1d_2 or d_1a_2, because the outcome will be the same with either of these two full strategies. The latter prescribes a bad action in the subgame generated by x_3, though, for player 2 would lose one unit of utility in comparison to playing d_2. That being the case, a notion of rationality sensitive to subgames excludes player 2 from playing d_1a_2 without, it is important to note, transcending the

one-shot framework—it only involves player 2's one-time event of choosing a full strategy.

If rationality principles come in two versions, it might be expected that beliefs come in a subgame-insensitive and a subgame-sensitive version also. The subgame-insensitive conception takes the full strategies of players' opponents as objects of their beliefs. In order to get some impression of the ways a subgame-sensitive conception can be defined, consider the ways that player 1's belief that 2 will quit the Centipede game as soon as possible can be phrased. Since the one-shot interpretation holds to the view that the only objects of choice are full strategies, player 1's beliefs should be about 2 playing d_1d_2 or d_1a_2. She could believe that player 2 plays the former, she could believe he plays the latter, and she could believe that either he plays the former or he plays the latter, but in all three cases she believes that at x_1 player 2 quits the game. For the rational evaluation of player 1's choice of a one-shot full strategy makes no difference which of the three beliefs she uses, because all that matters is what she does at the root of the game. In order to evaluate a full strategy in, say, the subgame generated by x_2, it matters what exactly she expects of her opponent in that subgame.

There are two ways out. First, players may be supposed to have beliefs for all subgames. Player 1 may believe that player 2 will play d_1a_2 in the entire game, but that in the subgame generated by x_2 he will play d_2, for instance. Alternatively, we could presuppose that a player's beliefs about the *entire* game are serious. If a belief entails a certain action at a certain node which, according to that very belief, will not be reached, the belief about this action still expresses serious reasoning about the opponent's strategic situation. Player 1 should, following this latter logic, believe that 2 will play d_1d_2 in the entire game, and that belief would be reusable in the subgame generated by x_2 stating that 1 believes 2 to play d_2 at x_3.

I do not have a strong opinion about which of the two is the best, because, I believe, they are both antithetical to the one-shot interpretation of extensive game-playing. In a one-shot game-playing situation, players make up their minds about full plans of action, without envisaging any future moments of decision. Players who do not envisage future moments at which they can decide, do not envisage future moments at which they have beliefs either, and if I am right here, special subgame-sensitive beliefs, under the one-shot interpretation, do not make sense. As a consequence, rationality principles which pay attention to subgames do not make sense according to the one-shot interpretation for the same reason. Beliefs about subgames and rationality, in other words, are about what would happen if subgames were reached at some later point in time. But under the one-shot interpretation it does not make sense to consider such possibilities. Subgames are not reached at all.

Ultimately, the one-shot interpretation comes in only one version—a radical one. Playing an extensive game in the one-shot interpretation is playing its normal form version without rationality principles that go beyond the normal form, and Chapter 3 discusses important consequences for the epistemic characterisation results adopting the one-shot view.

1.1.4 Extensive Games: The Many-Moment Interpretation

The one-shot interpretation seems to stay closest to von Neumann and Morgenstern's dictum that normal form and extensive form games are equivalent. To motivate considering a many-moment interpretation of extensive game-playing, a possible counterargument against the equivalence would hold that a player who thinks of the game as consisting of various subsequent decision moments will just not develop a full plan of action for the game—and hand it to the umpire overseeing it—but will rather think only about the choice of action to be made at the decision node she thinks she is at. The many-moment interpretation indeed conceives of a game-playing situation as a sequence of decision moments. Game-play, that is, entails a run through the game tree terminating in some outcome, but it entails more, because every decision moment ought to contain the necessary elements to explain the action performed at any decision moment in the run—preferences, principles and beliefs. A game-playing situation is accordingly a run through the game tree plus some extra information.[20]

The essential difference between the one-shot and the many-moment interpretations is that the former has players choose a full strategy at one single point in time, while the latter sees players as making up their minds at various points in time. In the many-moment interpretation, players may exert influence on the outcome more than once, respond to their opponents' moves, and block entire subgames by their choices of action. Since a many-moment game-playing situation is a sequence of decision moments, the logical form of the various components of decision moments is different from the corresponding element of the one-shot interpretation. At every decision moment a player has to act rationally on the basis of some preferences and beliefs. A player's beliefs may, first, be thought of as relating all of her possible actions at some decision node to the terminal nodes she expects to ensue if she were to choose that action, and however minimal this conception of beliefs may be, it is enough to make rationality principles work: choose the action you expect to lead to the best possible terminal node. Yet as it stands this conception of beliefs is in almost all cases unserviceable to the Epistemic Programme's purposes, because only very occasionally will the utility structure of the game be sufficient for the players to base their expectations on. Outcomes that are worst for everyone may be excluded, and those that are best for everyone expected, but such cases are rare, and statistical data and the like are clearly ruled out by the ban on exogenous information.

As long as the players' beliefs concern terminal nodes, they may be said to possess beliefs about a restricted number of possible eventualities only, and I therefore propose to add more structure to beliefs to make them fit for the many-moment

[20] Another interpretation comes into sight when we consider players deciding on the performance of entire strategies, but not necessarily once and for all. Such an interpretation is mentioned by Wlodek Rabinowicz, 'Grappling with the Centipede: Defence of Backward Induction for BI-terminating Games', *Economics and Philosophy*, 14 (1999), 99. See also ibid., 'To Have One's Cake and Eat It, Too: Sequential Choice and Expected-Utility Violations', *Journal of Philosophy*, 92 (1995), 586–620.

interpretation. In this interpretation, players have beliefs about all possible future decision nodes, giving rise to an expected path through the game and its subgames.

Apart from beliefs about choices of action, players have beliefs about possible actions, preferences, principles, and, with recursion, beliefs, plus—new for the many-moment interpretation—beliefs about the actions played prior to the current decision moment. Beliefs about possible future actions and beliefs about previous actual actions are beliefs about the tree structure of the game and the position therein. Beliefs about preferences are beliefs about the ordering over terminal nodes.

Formal renderings of those beliefs, as well as the even more obvious beliefs about beliefs, may disagree on the details of the logical form, but these issues need not detain me here. What should be discussed, however, are the different ways in which the many-moment interpretation can deal with the history of game-play. The one-shot interpretation has players form beliefs about their opponents' choice of full strategies. The many-moment interpretation has players form beliefs about their opponents' choices of action at each and every possible decision node of theirs.[21] Now, imagine a player at some decision moment who wishes to form a belief about what is going to happen at some future decision node. If it is her own decision node, she could simply commit herself to performing some action there, but that would be strikingly incoherent with the many-moment interpretation, because she would in that case adopt a full strategy as in the one-shot case. To be a genuine many-moment agent, she should imagine herself in the particular future decision node, find out what preferences, principles and beliefs she will act upon then and there, and deduce from that a belief about her own choice of action. This is no different from the way players form beliefs about each others' prospective choice of action at future decision nodes. Beliefs about future preferences, principles and beliefs, that is, are used to deduce predictions of future actions.

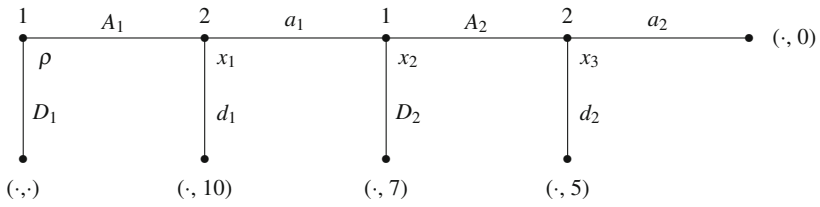

Fig. 1.4

[21] In the one-shot interpretation, beliefs may also involve subgames and in the many-moment interpretation, beliefs may also be about terminal nodes. As I have shown, however, these are either insufficient to run the Epistemic Programme, or incoherent given the further details of the interpretation.

How can players obtain any information about these elements? As with normal form games and the one-shot interpretation, general conditions on game-playing situations such as common true belief about rationality and utility may be used. A new source of information springs, however, from the decision nodes and their positions in the game tree. Consider, for instance, a variant of the Centipede, shown in Figure 1.4, of which only the relevant parts of the utility functions have been drawn. Player 1 at the root of the game wants to predict what is going to happen at x_3, and tries—in a first attempt—to find out what 2's preferences, principles and beliefs will be at x_3. She may reason that player 2 is rational and that going down at x_3 is the only rational move whatever player 2 believes, concluding that at x_3 player 2 will play d_2. She could also—a second attempt—argue that imagining player 2 at x_3 entails imagining player 2 having chosen a_1 at x_1. Playing a_1 at x_1 is, player 1 believes, not rational for player 2, whatever his beliefs are, and therefore imagining player 2 at x_3 entails imagining player 2 with a past of irrational play. Player 1 may go on to note that there are various ways to make sense of x_3 being reached. Player 2 may have made a mistake at x_1, he may be genuinely irrational, and so on, and depending how player 1 continues the image of player 2 at x_3 different expectations about his action at x_3 follow. If that is right, without extra assumptions it is not clear what beliefs player 1 forms.

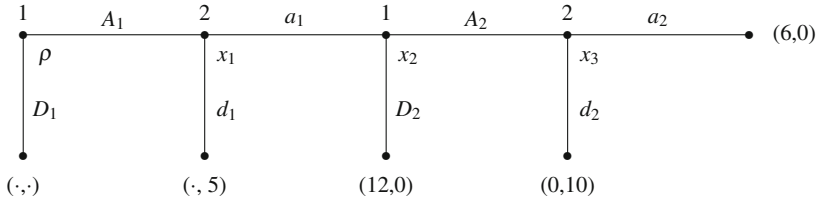

Fig. 1.5

The example suggests that players who take a *history-sensitive* look at decision nodes are in a worse position than those who adopt a *history-insensitive* view, because history often involves more than one, possibly conflicting explanation. To reinforce this point, consider yet another Centipede, shown in Figure 1.5. Player 1 at the root of the game again wants to know what is going to happen at x_3. If she takes the history of x_3 seriously, she has to consider 2's choice of a_1 at x_1. As a_1 is rational only if 2 expects 1 to play A_2 at x_2, picturing x_3 as being reached means picturing not only 2's action and principle at x_1 (as in the previous example) but also 2's beliefs at x_1. In order to form a belief about what is going to happen at x_3, it matters what 1 believes about 2's beliefs at x_1; it matters, for instance, to the question whether 1 ought to consider x_3 the result of mistakes, irrational play, or otherwise.

In summary, the most far-reaching consequences of the many-moment interpretation of extensive game-play pertain to beliefs, partly because of their content (terminal nodes, or actions at all decision nodes of the subgame) and partly because of their formation (history-sensitivity, or not). I will briefly examine assumptions that the many-moment interpretation places on further elements of game-playing situations.

A weak total ordering of the terminal nodes representable by a von Neumann–Morgenstern utility function captures the preferences of the players, and the distinction between strictly and weakly dominated actions can be used here, too, to formalise different concepts of rationality. Equally clearly, under the many-moment interpretation players of an extensive game choose from a set of possible actions at any decision node on the path followed through the game tree. Yet while the agent who moves at some decision moment of a game-playing situation chooses exactly one action among the possible actions, the many-moment interpretation has players play in turn, not simultaneously; exactly one player plays at a decision moment.

At a decision moment, the player who moves has to know where she is in the game tree. This general condition entails that she knows what has happened up to that moment, which actions have been chosen and which ones have been disregarded. She needs to know the entire past and every possible future of the entire past. This condition further entails that the player knows which actions she can choose. She needs to know every possible future of every possible action and by the same token she also has to know what her opponents can do at later decision moments. If these conditions do not hold, no genuinely game-theoretic account can be given of her actions, because particular features of the game theorist's model (for instance, particular preference orderings over terminal nodes of some subgame) would not play a role in the agent's own motivation for her actions.[22] In addition, a player should know her own preferences and those of her opponents, and she should know on which principle she acts at the decision node she is at.[23]

It is clear that without any beliefs about the opponents' rationality principles at possible subsequent decision moments the player cannot form any argument about the expected outcome—unless she reasons on the basis of statistical data and other forms of information that were excluded by the ban on exogenous information. However, how one sees such belief formation depends on how one sees decision nodes—with or without appreciation of their histories. The obvious requirement for the history-insensitive view is that players ought to believe that their opponents are rational at every possible decision node. There does not seem to be an obvious requirement for the history-sensitive conception.

Players should know what they play at a decision node and they should know what they have played (a consequence of earlier conditions, called *perfect recall*)

[22] It may be that these conditions will often fail to hold. What concerns me here, however, is not their conceptual plausibility or empirical adequacy. Rather, I indicate here what makes an explanation a game-theoretic one.

[23] This does not mean that she has to know the principles to be applied at any possible (or even subjectively probable) future decision node. For a discussion of weakening knowledge about opponents' utility functions, see Sections 2.3 and 2.4.

but they should not be required to know what they will play at possible future decision moments, because that would approximate the one-shot interpretation. Past actions of the opponent, however, ought to be known (a consequence of earlier requirements, called *perfect information*).[24] Yet players do of course generally have beliefs about future actions and about beliefs about future actions, for without beliefs about future beliefs a player would not be able to form any beliefs about what actions she would choose at that future decision node. Since she cannot, under the many-moment interpretation, truly commit herself to some future action already, she must form a belief about her future actions on the basis of beliefs about her future preferences, principles and beliefs. Similarly, beliefs about opponents' future beliefs are needed. Beliefs about past decision nodes, however, seem only required under the history-sensitive view. [25]

1.1.4.1 Identity Over Time

This admittedly rather rote discussion shows how conditions required for normal form game-playing situations as well as for the one-shot interpretation of extensive games can be adapted to the specific characteristics of many-moment game-playing situations. It is not all old wine in new bottles, though. I will now turn to conditions without earlier analogues, and ask how the five components of decision moments (possible actions, preferences, principles, beliefs and choices of action) evolve over time. Is it necessary to assume conditions relating those ingredients at different decision moments? Such a question does not of course make sense for possible actions, preferences and choices of actions, because if these ingredients changed the game-theoretic model proposed by the theorist would not be the game really played. But for principles and beliefs the question is more than pertinent.

In epistemic characterisation theorems game-theoretic explanations follow decision-theoretic explanations in that actions are taken to result from rational choice, and as there is more than one such principle, players may change their rationality principles over time. A player may start playing only strictly undominated actions and gradually come to play weakly undominated ones too, for instance. Such changes may be relatively uncommon, but similar changes in beliefs are, of course,

[24] This does not mean that games with imperfect information or imperfect recall are neglected here. The way I treat such games in Chapter 5, however, follows the one-shot interpretation and the reason is that only if players can foresee future informational asymmetries can a genuinely game-theoretic explanation be given. Overstating it slightly, games with imperfect information and games with imperfect recall do not make sense under the many-moment interpretation.

[25] Players have beliefs about the future, which may be justified and even true, but in the explanation of the players' actions the theorist cannot go beyond belief and make essential use of the fact that the beliefs, in fact, constitute knowledge. This also applies to normal form and one-shot situations. This does not mean that knowledge (as opposed to mere belief) never adds to the explanation of an action. Timothy Williamson, *Knowledge and its Limits* (Oxford: Oxford University Press, 2000), 60–64 gives an argument in favour of the 'causal efficacy' knowledge may have for human agency. That aspect of a knowledge constituting belief that makes it causally efficacious (something like its justification) has not been dealt with in the Epistemic Programme in game theory.

entirely natural. In fact, (theoretic) rationality entails that several beliefs change, most notably beliefs about the position in the game. It is useful to distinguish between the beliefs a player has about herself and beliefs she has about her opponents, because players have introspective access to aspects of the (possible) development of their agential identity that they do not have about their opponents. Beliefs about the players' own pasts will not generally change, as they know on which preferences, principles and beliefs they acted, but beliefs about their possible futures may change quite radically. Suppose that in some extensive game at x_k you believe that at x_m you will play on a principle of rationality. At some intermediate decision moment x_l of yours, however, you may feel differently about that situation. Introspection may tell you that your attitude to the game has changed—which does not necessarily mean that your preference ordering over terminal nodes has changed—and this may make it less likely that you will act on a principle of rationality at x_l. This does not strike me as a very common phenomenon of special use to game-theoretic modelling, but it does not seem incompatible with the view of action explanation underlying the Epistemic Programme. So, beliefs about your possible future principles may change, and beliefs about your beliefs about possible principles may change also.

Similarly, a player's beliefs about her opponent's past principles and beliefs may change. At x_l you may believe that your opponent's action at some preceding x_k was the result of a mistake. At some later x_m you may have seen more of his decisions, and this new information may force you to reconsider your earlier beliefs, but these beliefs are irrelevant in terms of future decision making. However, beliefs about the future principles and beliefs of a player's opponents ought not to change. If we adopt a history-insensitive view of decision nodes, picturing future decision nodes is independent of any information about previous play; and if we adopt a history-sensitive view of decision nodes, the picture of future decision nodes already takes care of all possible information about past play—as counterintuitive as it may sound.

1.2 A Logic for Game Theory

As I have shown, the research presented in this book studies game theory from two perspectives. It contributes internally to the Epistemic Programme in game theory itself by developing an epistemic logic for game theory, and it criticises externally the applicability of game theory as a descriptive and normative endeavour from the point of view of epistemology.[26]

The conceptual study of normal form and extensive game-playing situations as the Epistemic Programme views it cannot be used to derive stable results as long as no appropriate formalism is available to capture what would otherwise remain tacit and imprecise. I will now present the bare bones of the formalism. In the next two chapters, I will use the formalism to represent a number of existing results from the

[26] I have articulated my views on the interrelations between epistemic logic and epistemology in Boudewijn de Bruin, 'Epistemic Logic and Epistemology', in V. Hendricks and D. Pritchard (eds.), *New Waves in Epistemology* (Basingstoke: Palgrave Macmillan, 2008), 106–136.

Epistemic Programme, both to enable better and more precise comparisons, and to obtain new epistemic characterisation results. The last two chapters will then refer back to the formal results and use them in their critical evaluation of descriptive and normative game theory as well as in a comparison of the Nash Equilibrium Refinement Programme and the Epistemic Programme.[27]

1.2.1 A Logic for Normal form Games

Given an N-person normal form game with multi-matrix $(p_{i,k_1,\ldots,k_N})_{i,k_1,\ldots,k_N}$ representing the utility structure, I will define a formal language to describe all aspects of game-playing that are relevant to the study of the epistemic and rationality assumptions underlying game-theoretic solution concepts.[28]

The logical symbols used are $\neg$ (*not*, negation), $\wedge$ (*and*, conjunction), $\vee$ (*or*, disjunction), $\rightarrow$ ('if…, then…', implication), and $\leftrightarrow$ ('…if and only if…', equivalence). No quantifiers $\forall$ ('for all') or $\exists$ ('there exists') are needed. The conjunction (disjunction) of all sentences from a finite set Σ is abbreviated by $\bigwedge\Sigma$ ($\bigvee\Sigma$), assuming commutativity. If the φ_i enumerate Σ we may also write $\bigwedge_i \varphi_i$ ($\bigvee_i \varphi_i$).

In order to obtain a genuine modal system, the set of logical symbols is enlarged with three operators. Depending on the precise proof system, the $\Box$-operator (historically with interpretation 'it is necessary that…') has a *doxastic* reading ('it is believed that…') or an *epistemic* reading ('it is known that…'). Since we are dealing with more than one player it makes sense to index the operators $\Box_i$ for each player i. The *dual* of the $\Box_i$ is written $\Diamond_i$; that is, $\Diamond_i$ abbreviates $\neg\Box_i\neg$. The $\mathbf{E}_I$-operator stands for 'every player $i \in I$ believes (knows) that…', and the $\mathbf{C}_I$-operator is used to speak about *common* belief (knowledge)—all players believe (know)…, and all players believe (know) that all players believe (know)…, and all players believe (know) that all players believe (know) that all players believe

[27] The Epistemic Programme has benefited from an inspiring number of studies in logic and games, and the present framework, while original in its formalisation of rationality in characterisation of iterated dominance solution concepts, is indebted to the work of various authors including Johan van Benthem, 'Games in Dynamic-Epistemic Logic', *Bulletin of Economic Research*, 53 (2001), 219–248, ibid., 'Extensive Games as Process Models', *Journal of Logic, Language and Information*, 11 (2002), 289–313, Oliver Board, 'Dynamic Interactive Epistemology', *Games and Economic Behavior*, 49 (2004), 49–80, Thorsten Clausing, 'Doxastic Conditions for Backward Induction', *Theory and Decision*, 54 (2003), 315–336, ibid., 'Belief Revision in Games of Perfect Information', *Economics and Philosophy*, 20 (2004), 89–115, Aviad Heifetz and Philippe Mongin, 'Probability Logic for Type Spaces', *Games and Economic Behavior*, 25 (2001), 31–53, Graham Priest, 'The Logic of Backwards Inductions', *Economics and Philosophy*, 16 (2000), 267–285, Robert Stalnaker, 'On the Evaluation of Solution Concepts', *Theory and Decision*, 37 (1994), 49–73, ibid., 'Knowledge, Belief and Counterfactual Reasoning in Games', *Economics and Philosophy*, 12 (1996), 133–163 (repr. with proofs in C. Bicchieri, R. Jeffrey and B. Skyrms (eds.), *The Logic of Strategy* (New York: Oxford University Press, 1999), 3–38), ibid., 'Belief Revision in Games: Forward and Backward Induction', *Mathematical Social Sciences*, 36 (1998), 31–56, and ibid., 'Extensive and Strategic Forms: Games and Models for Games', *Research in Economics*, 53 (1999), 293–319.

[28] For further discussion of notations, definitions and some theorems, see the Appendix A.

(know)..., and so on *ad inf.* An abbreviation for $\mathbf{E}_I \ldots \mathbf{E}_I \varphi$ with n occurrences of $\mathbf{E}_I$ is $\mathbf{E}_I^n \varphi$. Furthermore, $\mathbf{E}_I \varphi \wedge \mathbf{E}_I^2 \varphi \wedge \cdots \wedge \mathbf{E}_I^n \varphi$ is written $\mathbf{E}_I^{\leq n} \varphi$. This is referred to as common belief (knowledge) *up to level n*. Probabilistic expressions $\mathbf{P}_i(\cdot) = \cdot$ represent i's probabilistic beliefs and arbitrary finite sums of such expressions $\mathbf{P}_i(\varphi_1) \cdot \mathbf{q}_1 + \cdots + \mathbf{P}_i(\varphi_n) \cdot \mathbf{q}_n \geq \mathbf{q}$ are allowed as long as they are not mixed over players (as $\mathbf{P}_i(\varphi_1) \cdot \mathbf{q}_1 + \mathbf{P}_j(\varphi_2) \cdot \mathbf{q}_2 \geq \mathbf{q}$ would be for $i \neq j$), and obvious abbreviations use $\sum$.

The non-logical symbols include proposition letters to speak about games. Proposition letters $\mathbf{i}_m$ stand for the statement 'i plays her mth strategy i_m'. The formal analogue of the statement that $u_i(1_{k_1}, \ldots, N_{k_N}) = r$ for some real number r is $\mathbf{u}_i(k_1, \ldots, k_N) = \mathbf{r}$. In order to be able to make all relevant statements involving utility, we need countably many symbols to refer to the real numbers (never all, sometimes finitely many). Rationality conceptions, finally, correspond to proposition letters of the form $\mathbf{meu}_i$, $\mathbf{rat}_i$, $\mathbf{prat}_i$ and $\mathbf{mrat}_i$, which I will explain later.

For modal logics, a Hilbert-style proof system is common and convenient. A proof in such a system of a sentence φ from a set of sentences Σ is roughly a finite sequence of sentences that are either taken from Σ, or axioms (typical for the particular Hilbert system), or statements derived (by rules typical for the particular system) from sentences occurring earlier in the sequence, such that the last sentence is φ. If such a sequence exist—if φ is derivable from Σ—one writes $\Sigma \vdash \varphi$. Generally, the axioms contain all modal or non-modal instances of tautologies from (classical) propositional logic. No extra creativity is needed to derive statements of the form $\varphi \vee \neg\varphi$ or $\varphi \rightarrow (\psi \rightarrow \varphi)$.

For the part of the language concerned with modality the following axioms are needed.[29] They are routinely stated for completeness and future reference.

Prop All classical propositional tautologies.
Dual $\Diamond_i \varphi \leftrightarrow \neg\Box_i\neg\varphi$.
K $\Box_i(\varphi \rightarrow \psi) \rightarrow (\Box_i\varphi \rightarrow \Box_i\psi)$.
T $\Box_i\varphi \rightarrow \varphi$.
D $\Box_i\varphi \rightarrow \Diamond_i\varphi$.
4 $\Box_i\varphi \rightarrow \Box_i\Box_i\varphi$.
5 $\Diamond_i\varphi \rightarrow \Box_i\Diamond_i\varphi$.
E $\mathbf{E}_I\varphi \leftrightarrow \bigwedge_i \Box_i\varphi$.
C $\mathbf{C}_I\varphi \leftrightarrow \mathbf{E}_I(\varphi \wedge \mathbf{C}_I\varphi)$.

In proof systems including the E-axiom every axiom for the $\Box_i$ is provable for the $\mathbf{E}_I$.[30] For instance, if the T-axiom and the E-axiom are available, all instances of $\mathbf{E}_I\varphi \rightarrow \varphi$ can also be appealed to. In proof systems including the C-axiom as well as the rule of induction from below, every axiom for the $\Box_i$ is provable for the $\mathbf{C}_I$.

[29] Not all axioms are always needed, and this feature, highlighted in the discussion of the epistemic characterisation results in the next two chapters, is conceptually as well as technically quite interesting.

[30] The Dual-axiom and the E-axiom are, in some way, not genuine axioms, but definitions of operators. This does not apply to the C-axiom, because the **C**-operator cannot be defined in the finitary language applied here.

Taking $\Box_i$ as the epistemic modality, the K-axiom expresses that what is known to be a logical consequence of something known is known; the epistemic subject i is *logically omniscient*. The Dual-axiom fixes the meaning of the $\Diamond_i$-operator. The T-axiom captures *veridicality*; what is believed is true. The D-axiom states the *consistency* requirement that what you believe is not believed not to hold. The 4-axiom formalises *positive introspection* of doxastic or epistemic states; you know that you know what you know. The 5-axiom, in turn, formalises that you know that you do not know something if you know not to know it—*negative introspection*. The E-axiom determines that the $\mathbf{E}_I$-operator expresses the beliefs everyone has. The C-axiom, finally, captures rather cryptically what it means that something is commonly known: everyone knows it, everyone knows that everyone knows it, and so on *ad inf*.

To discuss linear (in)equalities (to do the calculation necessary to solve maximisation problems) the following axioms are needed.[31]

0-term $\sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k \geq \mathbf{r} \leftrightarrow \sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k + \mathbf{P}_i(\varphi_{k+1})\cdot 0 \geq \mathbf{r}$.
Per $\sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k \geq \mathbf{r} \leftrightarrow \sum_k \mathbf{P}_i(\varphi_{l(k)})\cdot\mathbf{q}_{l(k)} \geq \mathbf{r}$ for l any permutation.
AddCoef $(\sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k \geq \mathbf{r} \wedge \sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}'_k \geq \mathbf{r}') \rightarrow \sum_k \mathbf{P}_i(\varphi_k)\cdot(\mathbf{q}_k + \mathbf{q}'_k) \geq (\mathbf{r} + \mathbf{r}')$.
MultCoef $\mathbf{c} \geq 0 \rightarrow (\sum_k \mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k \geq \mathbf{r} \leftrightarrow \sum_k \mathbf{c}\cdot\mathbf{P}_i(\varphi_k)\cdot\mathbf{q}_k \geq \mathbf{c}\cdot\mathbf{r})$.
Dich $t \geq \mathbf{r} \vee t \leq \mathbf{r}$ for t any term.
Mon $\mathbf{q} > \mathbf{r} \rightarrow (t \geq \mathbf{q} \rightarrow t > \mathbf{r})$ for t any term.

To allow for probabilistic reasoning the Kolmogorov axioms are essential.

NonNeg $\mathbf{P}_i(\varphi) \geq 0$.
True $\mathbf{P}_i(\top) = 1$.
False $\mathbf{P}_i(\bot) = 0$.
Add $\mathbf{P}_i(\varphi) = \mathbf{P}_i(\varphi \wedge \psi) + \mathbf{P}_i(\varphi \wedge \neg\psi)$.
Dist $\mathbf{P}_i(\varphi) = \mathbf{P}_i(\psi)$ whenever $\varphi \leftrightarrow \psi$ is a propositional tautology.

In order to ensure that probabilistic and non-probabilistic beliefs are related in the right way, two additional axioms are useful.

Cons $\Box_i\varphi \leftrightarrow \mathbf{P}_i(\varphi) = 1$.
KnProb $\varphi \rightarrow \Box_i(\varphi)$ for φ an i-probability sentence.

The first of these interrelation axioms is a consistency requirement on the relation between non-probabilistic and probabilistic beliefs. In a sense it guarantees the maximum possible, showing that as the $\Box_i$ become redundant, the investigations can in principle be carried out in a stricter language. A weaker interrelation axiom can be defended by demanding an implication in the direction from left to right only. But since it is not necessary here either to allow for cases in which a player holds some proposition φ possible without assigning positive probability to it (or the converse of this case) or to distinguish between beliefs with probability one and non-probabilistic beliefs, I will ignore this subtlety.

[31] For what follows, see Ronald Fagin and Joseph Halpern, 'Reasoning About Knowledge and Probability', *Journal of the Association for Computing Machinery*, 41 (1994), 340–367.

The second axiom (where i-probability sentences are sentences starting with $\mathbf{P}_i$ or Boolean combinations thereof) yields the players quite some measure of introspective power. It is easily seen, for instance, that the KnProb-axiom together with Cons-axiom and the necessitation rule entail positive as well as negative introspection.

The proof rules are thus:

MP If $\Sigma \vdash \varphi \to \psi$ and $\Sigma \vdash \varphi$, then $\Sigma \vdash \psi$.
Nec If $\vdash \varphi$, then $\vdash \Box_i \varphi$.
Ind If $\vdash \varphi \to \mathbf{E}_I(\varphi \wedge \psi)$, then $\vdash \varphi \to \mathbf{C}_I \psi$.

The first two rules of *modus ponens* and *necessitation* appear in the proof system of every epistemic logic. The last rule of *induction* is specific for proof systems that contain the E- and C-axiom.

To capture normal form game-playing situations we need four additional axioms.

$\text{Strat}_{\geq 1}$ $\bigvee_m \mathbf{i}_m$.
$\text{Strat}_{\leq 1}$ $\bigwedge_{m \neq n} \neg(\mathbf{i}_m \wedge \mathbf{1}_n)$.
KnStrat $\bigwedge_m (\Box_i \mathbf{i}_m \leftrightarrow \mathbf{i}_m)$.
KnUt $\mathbf{u}_i(k,l) = \mathbf{r} \to \Box_i \mathbf{u}_i(k,l) = \mathbf{r}$.

These axioms determine what players do, and what they know, when they play normal form games. The first axiom stipulates that every player plays at least one strategy, while the second axiom forbids any player to play more than one strategy. The KnStrat-axiom requires a player to have correct beliefs about what strategy he chooses. The KnUt-axiom requires players to have correct beliefs about their own utility functions.[32]

The precise formalisations of the game-theoretic solution concepts follow in the next two chapters. For the Nash equilibrium, no extra formal material is needed. For the iterated dominance solution concepts, I will develop a new way of axiomatisation.

1.2.2 A Logic for Extensive Games

Negation, connectives, and abbreviations are as before, as are, for the one-shot interpretation, the modal operators. For the many-moment interpretation doxastic or epistemic modalities $\Box_i^x$ are used to represent player i's beliefs or knowledge at the decision moment at which decision node x is reached, and super-scripted versions of $\mathbf{E}_I^x$, $\mathbf{C}_I^x$ and $\mathbf{P}_i^x(\cdot) = \cdot$ are defined similarly.

[32] Since any proof system contains the necessitation rule, players also believe (or know) these axioms to be true, believe them to believe them to be true, and so forth. This yields common beliefs about the possible actions, and about the fact that players know their utility. In the epistemic characterisation of mixed iterated strict weak dominance the last axiom takes a different form. See Section 2.4.

The roles decision nodes and decision moments play in extensive game-playing situations require me to restructure the original normal form language a little. On the basis of an arbitrary enumeration of all full strategies of some extensive form game, proposition letters $\mathbf{i}_k$ denote the kth such strategy; more precisely—the statement 'i plays her kth full strategy'. For decision node x (which is not necessarily a decision node where i has to move), proposition letter $\mathbf{i}_k^x$ states that player i chooses according to the kth strategy at all decision nodes in the subgame generated by x, and incidentally $\mathbf{i}_k(x)$ is used for the statement that, at her decision node x, player i chooses the action prescribed by her kth full strategy.[33] Utility statements need to be relativised as well. As before, the statement $\mathbf{u}_i(k,l) = \mathbf{r}$ captures the fact that 'the utility for player i, when full strategies k and l are being played, is r', and $\mathbf{u}_i^x(k,l) = \mathbf{r}$ restricts this statement to the subgame generated by x, meaning that 'the utility for player i, when the restrictions of full strategies k and l to the subgame generated by x are being played, is r'. Finally, a number of proposition letters are needed for the various principles of rationality such as $\mathbf{anrat}_i$, $\mathbf{nrat}_i$ and $\mathbf{rrat}$, to be explained later. Super-scripted, they express the respective relativised statements.

We need most of the earlier modal axioms in order to study extensive game-playing situations: the doxastic and epistemic axioms, the axioms for linear (in)equalities, the axioms for probability theory and the interrelation axioms, as well as the three proof rules. For one-shot analysis, we can adopt the earlier normal form version, but for the many-moment analysis we need to employ axioms and rules for such modalities as $\Box_i^x$, $\mathbf{P}_i^x(\cdot) = \cdot$, and so forth. It is plain how this is done consistently, though.

1.2.2.1 The One-Shot Interpretation

Two proof systems formalise the two interpretations of extensive game-playing situations. The one-shot interpretation has the following axioms.

$\text{Strat}_{\geq 1}$ $\quad \bigvee_m \mathbf{i}_m$.

$\text{Strat}_{\leq 1}$ $\quad \bigwedge_{m \neq n} \neg(\mathbf{i}_m \wedge \mathbf{i}_n)$.

KnStrat $\quad \bigwedge_m (\Box_i \mathbf{i}_m \leftrightarrow \mathbf{i}_m)$.

KnUt $\quad \mathbf{u}_i(k,l) = \mathbf{r} \rightarrow \Box_i \mathbf{u}_i(k,l) = \mathbf{r}$.

Sub_1 $\quad \mathbf{i}_k^x \leftrightarrow \bigvee_{l \in D} \mathbf{i}_l$ where D contains those strategies coinciding with k on the subgame generated by x.

Sub_2 $\quad \mathbf{i}_k(x) \leftrightarrow \bigvee_{l \in D} \mathbf{i}_l$ where D contains those strategies coinciding with k on decision node x.

UtSub $\quad \mathbf{u}_i^x(k,m) = \mathbf{u}_i^x(l,n)$ whenever i's kth and lth, and j's mth and nth strategies coincide on the subgame generated by x.

KnUtSub $\quad \mathbf{u}_i^x(k,l) = \mathbf{r} \rightarrow \Box_i \mathbf{u}_i^x(k,l) = \mathbf{r}$ for all decision nodes x.

[33] It is immaterial whether $\mathbf{i}_k^x$ and $\mathbf{i}_k(x)$ are really new proposition letters, or only abbreviations of the disjunction of the proposition letters of those strategies that coincide with the kth strategy on the subgame generated by x.

The first four axioms are copies of the conditions on normal form game-playing situations. Players pick exactly one full strategy, they know what they do, and they know what their utility functions are. Sub_1 states that the use of super-script is to talk about the restriction of some full strategy to the relevant subgame. Sub_2 ensures that function notation is present to report the action taken at some decision node. UtSub guarantees that the super-script works well when applied to utility functions. The KnUtSub-axiom, finally, is there to endow players with knowledge about their utility function in subgames.

1.2.2.2 The Many-Moment Interpretation

The following axioms fix the many-moment interpretation.

$\text{Strat}_{\geq 1}$ $\bigvee_m \mathbf{i}_m$.
$\text{Strat}_{\leq 1}$ $\bigwedge_{m \neq n} \neg(\mathbf{i}_m \wedge \mathbf{i}_n)$.
Sub_1 $\mathbf{i}_k^x \leftrightarrow \bigvee_{l \in D} \mathbf{i}_l$ where D contains those strategies coinciding with k on the subgame generated by x.
Sub_2 $\mathbf{i}_k(x) \leftrightarrow \bigvee_{l \in D} \mathbf{i}_l$ where D contains those strategies coinciding with k on decision node x.
UtSub $\mathbf{u}_i^x(k,m) = \mathbf{u}_i^x(l,n)$ whenever i's kth and lth, and j's mth and nth strategies coincide on the subgame generated by x.
KnStratM $\mathbf{i}_k \leftrightarrow \bigwedge_{\rho \preceq x} \Box_i^x \mathbf{i}_k(x)$.
KnUtM1 $\mathbf{u}_i(k,l) = \mathbf{r} \rightarrow \Box_i^x \mathbf{u}_i(k,l) = \mathbf{r}$ for all decision nodes x.
KnUtM2 $\mathbf{u}_i^y(k,l) = \mathbf{r} \rightarrow \Box_i^x \mathbf{u}_i^y(k,l) = \mathbf{r}$ for all decision nodes x and y.
KnWhere $\Box_i^x \bigwedge_j \bigvee_{y \prec x, j_k \in D} \mathbf{j}_k^y$ where D contains the proposition letters for those full strategies that are consistent with reaching x.

The first five axioms are those axioms from the one-shot interpretation that do not contain a $\Box_i$, and their motivation is similar. Of the last four axioms, KnStratM ensures that at every moment of a game-playing situation players know what they choose then and there.[34] The next two axioms ensure that players know their utilities in the entire game as well as in all subgames. The last axiom is there to guarantee that players know, at some decision moment, which decision node has been reached.

Finally, the solution concept of backward induction is defined in Chapter 3.

[34] This means that I adopt an *at choice*, rather than a *pre choice* conception of game-playing situations where the beliefs and preferences are considered just before the moment of choice, not at the moment of choice. See Wlodek Rabinowicz, 'Grappling with the Centipede: Defence of Backward Induction for BI-Terminating Games', *Economics and Philosophy*, 14 (1998), 115–119.