Preferences, Counterfactuals and Maximisation: Reasoning in Game Theory

Philipp U. Beckmann

London School of Economics
& Political Science

Thesis submitted to the University of London
for the completion of the degree of a
Doctor of Philosophy

June 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis explores two kinds of foundational issues in game theory. The first is concerned with the interpretation of the basic structure of a game, especially the definitions of outcomes and payoffs. This discussion leads to the second issue; namely the nature of solution concepts and their relation to both explicit and implicit assumptions in game theory concerning hypothetical reasoning.

Interpreting utility functions in game theory, I argue that the notion of revealed preferences is ill-suited for counterfactual reasoning and for taking account of the implicit normativity of instrumental rationality. An alternative interpretation is outlined that treats preferences as determinants of choice. Accordingly, outcomes have to be individuated so as to capture everything that matters to an agent. I consider whether this is problematic when properties of outcomes depend on choice processes themselves. Turning to a decision theoretic problem, I question Verbeek's (2001) claim that modal outcome individuation conflicts with axioms of consequentialism. Next, I critically assess Rabin's (1993) model of fairness equilibria. Hypothesising about unilateral deviation is shown to be incompatible with belief-dependent utility definitions.

Counterfactuals in games are then analysed more generally. It proves to be crucial for solution concepts whether our formal framework allows us to differentiate between indicative and subjunctive conditionals. Stalnaker's (1996) prima facie counterexample to Aumann's (1995) theorem that common knowledge of rationality implies a subgame perfect equilibrium is questioned on the grounds of a plausibility criterion. Again drawing on what has been established about the structure of a game and the meaning of its elements, Gauthier's (1986) notion of constrained maximisation, an attempt to overcome the non-cooperative equilibrium of the finitely iterated prisoner's dilemma, is shown to be incompatible with orthodox game theoretical methodology. The approach of treating the unit of agency as endogenous is addressed.

To my parents

Contents

Acknowledgements

First and foremost, I would like to thank Richard Bradley for his great support and dedication as a supervisor. His comments have been inspiring and eye-opening.

I am indebted to Ned McClennen, Shepley Orr and Arndt von Schemde for useful comments on chapter drafts of this thesis. Furthermore, I am grateful to Peter Ainsworth, Eleanor Chiari, Damien Fennell and Tamsin Scurfield for their help with the final draft.

I am also grateful to the Department of Philosophy Logic and Scientific Method, the Centre for Philosophy of Natural and Social Science as well as the German National Merit Foundation for their generous support.

I thank Stefanie Hagenburger for her faith and patience. Moreover, I am obliged to my friends in London and Munich. In particular, I thank Momoko Ashizawa, Philipp Dorstewitz, Florian Englmaier, Vincent Guillin, Kathy King, Valérie Lafitte, Antti Saaristo, Marco Schönborn and Bettina Woll. I am grateful to Marco Rupp for shelter in Brussels.

Last but not least, I would like to express my deep gratitude to my parents, my sisters Eva and Anna, and my uncle Ludger.

Introduction

This thesis is a contribution to the question of what game theory as a method in social sciences, normative and descriptive, can and cannot deliver. My approach, though not exhaustive, is to consider how games represent reasoning of interacting agents and how to justify certain solution concepts. Among other components of games, I discuss the concepts of utility, consequentialism and outcomes, epistemic assumptions and hypothetical reasoning, as well as agency and rational choice as maximisation. I will show that the limits of game theory as a tool of the social sciences exceed the practical problems of its application. As will be seen, it is not clear whether game theory, in its current form, can capture all human interaction of a strategic nature even in principle.

I will not make a prior distinction between the normative and the positive interpretations of game theory. As will be argued, both are crucial and constitutive sides of the theory. Furthermore, normative intuitions about how we, both as agents and as game theorists, should reason about situations that are represented as games will be an important factor in appreciating conceptual components and solution concepts of games. This is particularly crucial for some of the 'paradoxical,' or rather counterintuitive, results, which have been a driving force in mathematical game theory as well as debates in social, political, and even moral theory.

The thesis can be read as falling into two main parts that are interwoven in a number of ways. The first part deals with the conceptual interpretation of some of the 'building blocks' of the theory. More specifically, I will look into the specification of outcomes in games and the corresponding utility figures. First, I will try to understand what such utility figures represent. I consider possible answers such as 'preferences' or 'choices.' Then I will turn to the objects of measurement of utility figures, namely possible outcomes of games. The critical question is whether everything that matters to players, everything that makes a difference with respect to their preferences and/or choices, can be specified as part of the outcomes. Next, I discuss attempts to capture process-dependent properties, namely beliefs formed throughout a game, directly in the utility functions of a game. The resulting so-called 'psychological games' require an amended equilibrium concept that will be assessed.

The second part of the thesis deals with solutions to games. Two main themes will be addressed. The first is captured in the word 'counterfactuals' in the title of the

thesis. It concerns the question of how to support solution concepts by reasoning about off-equilibrium situations in games, or rather by *reasoning about* players' reasoning in such situation. The second theme concerns attempts by theorists to avoid intuitively implausible solutions, such as non-cooperation in finitely iterated prisoner's dilemmas, by altering the concept of rationality as maximisation.

Throughout the different chapters, I will point out a number of interesting connections between the topics discussed. To name some examples, hypothetical reasoning is not only crucial for justifying equilibria; it also decisive in individuating outcomes in games. The individuation of outcomes in turn is crucial for our understanding the meaning of utility functions. Both the interpretations of utilities and the way we capture hypothetical reasoning in games, however, impact on the discussion of rationality as utility-maximisation.

The thesis is mainly conceptual and less formal. It is not meant to be a contribution to mathematical game theory. The value added lies rather in making sense of existing formal approaches (or identifying the absence of such sense). Accordingly, I will not introduce more formalism than is necessary to convey a conceptual point.

A similar minimalist principle applies to some of the many grand philosophical topics that lurk in the background of my discussion. Among them are major problems of modal logic, theory of probability, theory of action, and the normativity of instrumental rationality. As each of these topics fills shelves in philosophical libraries, I develop my viewpoint on them just as thoroughly as is expedient for the question at stake.

The plan of thesis is as follows. There are five chapters. The first three deal with preferences and outcome specification. Chapter 1 is concerned with the question of how to interpret utilities in games. I present three kinds of arguments for why reading preference as choice fails as a constitutive concept of game theory. I subsequently offer an alternative definition of utility, 'effective preference,' which has a number of methodological and conceptual virtues but also seems to limit the scope of situations that can be represented as games.

Chapter 2 addresses an argument by Verbeek (2001), which doubts that the explicit and implicit axioms of a certain notion of consequentialism are consistent. This would undermine Hammond's expected utility theory. The argument builds on a counterexample in which the fairness of a distribution (here understood as random selection with equal chances) is a relevant aspect of the outcome-specification. I show how this argument goes wrong and draw conclusions about the adequate moment of

outcome-specification in the modelling process. But I move on to discuss how far process-dependent properties do constitute a problem for certain expected-utility theories.

Chapter 3 deals with psychological games. In these games, payoffs, as opposed to outcome descriptions, are endogenous with respect to beliefs about how players will actually play in a game. I introduce Rabin's (1993) model of 'fairness equilibria' as a special case of psychological game analysis and raise immanent concerns about the concept of psychological equilibria by scrutinising the hypothetical reasoning supporting such equilibria. Next I discuss whether fairness equilibria capture the intuitions that first motivate process-dependent outcome definitions in games. I consider amendments to the concept of fairness equilibria.

While Chapter 3 already touches on the foundation of solution concepts, Chapters 4 and 5 focus on this issue. They do so in two distinct ways. Chapter 4 discusses how solution concepts derive from epistemic assumptions in games, most prominently common knowledge of rationality. Starting with a debate between Aumann (1995) and Stalnaker (1994, 1996), I argue that only Stalnaker's representation of epistemic assumptions is sufficiently rich to justify the actual (equilibrium) play by what players are modelled to take everyone's beliefs and belief revision policies to be in counterfactual scenarios. Moving to the vexed question whether common knowledge of rationality implies backward induction, a counterexample by Stalnaker (1998) is discussed and put in relation with game theoretical orthodoxy. However, Stalnaker's list of sufficient assumptions for supporting backward induction seems very strong. This motivates my own attempt to present a weaker set of assumptions in conjunction with a plausibility criterion regarding the permissibility of belief revision policies.

Chapter 5 finally turns to the debate about whether not cooperating in a finitely iterated prisoner's dilemma is or is not rational. Gauthier's (1986) argument is presented and put into the context of the Hobbes' debate of two kinds of rationality. I then take another look at the conceptual implications of game theory as a representational form focussing on inter- and intrapersonal separability of choices as well as the conceptual space between choice and preferences. Trying to make sense of Gauthier's insufficiently specified notion of constrained maximisation, I consider alternative 'routes' out of the prisoner's dilemma, showing that they either collapse into game theoretical orthodoxy or conflict with the conceptual limits of state-of-the-art game theory. I touch on the question of whether McClennen's approach, which is in many respects related to the

idea of constrained maximisation, redeems the outlined problems. I also consider the approach of treating the unit of agency as endogenous.

Chapter 1

Utility as effective preference

1. Introduction

Among other things, game theory should enable us to represent and analyse strategic human behaviour while remaining a formal discipline with well-defined components. This poses a number of problems addressed in the diverse chapters of this thesis. One challenge is to interpret the utility figures in games. The problem is vexed because utility payoffs fulfil a formal role in solution concepts and at the same time ought to express something meaningful about potential outcomes of the situation to be represented. In particular, they need to be mathematical determinants of the outcome of the game while having to allow for meaningful interpretations of counterfactual outcomes of the game. So an interpretation of utility in games needs to be equipped for making sense of some accepted solution concepts.

Another complication lies in the ambivalent status of game theory as both a positive and a normative discipline. By virtue of the normative status of instrumental rationality, it seeks to explain and predict behaviour as well as to advise players on how to pick strategies.

Traditionally, game theorists derive a notion of utility from another economic theory, namely revealed preference theory. They often simply take utility in games to stand for consistent choices. Others hold that preferences, although a separate concept from choice, could at least always be derived from observed choice. Neither of these interpretations lives up to the challenge of defining utility payoffs in games as just outlined. Hence they have rightly been challenged by Sen (1973). This chapter will support Sen's critique, but go beyond it.

There are critical parts and constructive parts to the chapter. In Sections 2-4, I bring forward three kinds of arguments for why reading preference as choice fails as a constitutive concept of game theory. Section 2 picks up arguments by Sen (1973) and Hausman (2000) on how the role of beliefs in games contradicts this interpretation. Section 3 lays out in how far preference-as-choice in combination with other standard assumptions excludes the possibility of crucial counterfactual reasoning in these games.

Section 4 explains how far equilibrium strategies in games have an implicit normative or prescriptive interpretation. Such interpretations, however are impossible in conjunction with preference-as-choice. In Section 5, I offer an alternative definition of utility by going through the requirements it needs to fulfil. The new notion, "effective preference," is designed to preserve some key virtues of preference-as-choice. However, I will also hint that this new interpretation of utility payoffs implies some limit of scope to the kind of situations which can be represented as games. Section 6, explores related problems of expected-utility theory in games.

2. Preferences, beliefs and choices

Paul Samuelson (1938) initiated revealed preference theory. He proved that given certain choice behaviour, such that when an agent chooses x (and not y) in the presence of x and y he will not also choose y in the presence of x and y, a transitive and complete (preference) relation can be inferred from this agent's choices.[1] This theorem (henceforth simply RPT for revealed preference theorem) has fuelled a behaviourist agenda to rid economics of psychological notions. Authors like Little (1949) reckoned that the theorem allowed economists to dispense with the notion of preference altogether. All that would be necessary was talk of consistent choices. It is this behaviourist interpretation of "preference" (and utility notions derived from it through further representation theorems) as a synonym for choice which I will refer to as the preference-as-choice interpretation, henceforth PCI.

It is to be distinguished from other revealed preference theories which endorse different interpretations of Samuelson's theorem. Sen (1973) spelled this point out clearly:

> "The rationale of the revealed preference approach lies in the assumption of revelation and not in doing away with the notion of underlying preferences, despite occasional noises to the contrary." (p.57)

Hausman (2000) went a step further, arguing that, if anything, the correspondence between choice and a preference relation *legitimises* the talk of subjective preference.

[1] The axiom in the antecedent of this theorem has become known as the Weak Axiom of Revealed Preference (WARP). Houthakker (1950) introduced a stronger axiom (SARP) for a more general theory. Sen (1971) generalised the approaches, proving SARP largely redundant. The interpretative issues which I am after in this chapter, however, do not require a review of this technical debate.

This reading of Samuelson (1938) is as an epistemological theory of how preferences are (always) potentially deducible from observed choice, given certain assumptions. I therefore label it epistemological revealed preference theory, henceforth ERPT. Both Sen and Hausman are not only critical about PCI, but also about ERPT.

Sen's critique

Sen (1973) formulated his critique of ERPT in what can be described as four categories:

(1) The problem of indifference
(2) The problem of incompleteness
(3) The problem of non-preferential motives for action
(4) The problem of strategic interaction

The first two problems concern all sorts of economic and non-economic decisions; the last two problems concern social interaction and are thus relevant for conceptual debates in game theory. (1) is the concern that an indifferent chooser, like Buridan's ass[2] confronting two identical haystacks, is better off choosing *some* items than not choosing at all for want of a decisive choice criterion. But there are versions of RPT which include indifference relations.[3] So Sen is more concerned with (2), the case in which there is no underlying preference at all. Equipped only with RPT, no choice pattern can reveal the absence of any underlying preference between two or more options. This problem can only be avoided by assuming completeness along with the other assumptions.

Even more worrying for Sen, however, is (3) that in interaction not all choice is driven by what we usually consider to be preference. Motives such as sympathy or fear of social stigma, might still be incorporated into preference theory by adequate redefinition of choice sets. Sen, however, worries that the motive of acting in a socially responsible manner, if in conflict with personal preferences, could pose a more fundamental problem for ERPT. The conflict that Sen describes springs from the "dual link-up between choice and preference on the one hand and preference and welfare on the other." (1973, p. 69) As far as I understand his point correctly, he seems to doubt

[2] A paradox of medieval by the French medieval philosopher Jean Buridan concerning the behaviour of an ass who is confronted with two piles of food of equal size and quality. It has no reason to prefer one pile to the other and therefore starves.
[3] An overview is given in Sen (1971).

that any definition of preference that includes aspects beyond personal welfare (like social responsibility) can at the same time be identified with choice.

While this might be a deep worry for the project of unifying branches of economics, I will ignore the issue for the moment. Sen (1973) does not deliver a detailed treatment of this concern anyway. However, I provide some arguments for the limits of incorporating motives into outcome definitions in decision and game theory in Chapters 2 and 3 as well as towards the end of this chapter.

The deepest worry for ERPT with respect to interaction is point (4). Sen argues with the example of the prisoner's dilemma, a version of which is represented in Figure 1.1.

		Column	
		Co-operate	Defect
Row	Co-operate	2, 2	0, 3
	Defect	3, 0	1, 1

Figure 1.1

The argument is that rational players' choices, which result in mutual defection, do not reflect the fact that both players prefer the outcome of mutual cooperation. The fact that they are not able to attain this result due to non-constrainable individual choice, keeps them from revealing their preference for one equilibrium outcome over the other.[4] This seemingly straightforward argument has a dubious status however. It seems to be driven by a concern for the Pareto-suboptimality of the defection equilibrium. But the fact that both players could be better off than they are in equilibrium in itself does not spell out an inconsistency of either PCI or ERPT. After all, players in the defection equilibrium reveal what they would individually prefer given any move by the other player. Hausman provides better arguments on the same issue.

Hausman's critique

Hausman (2000) picks up on Sen's critique. For him, the conflict of strategic choice and ERPT is part of a more general concern with ERPT. Hausman argues that the use of

[4] More on rational choice in the prisoner's dilemma will be said in Chapter 5.

(rational choice theory in) folk psychology is such that it cannot dispense with beliefs. Preferences only help to make predictions about choices *in conjunction* with assumptions about beliefs. Inversely, beliefs are needed to induce information about preferences from observed choice behaviour. According to Hausman, it is this failure to account for beliefs which renders ERPT inapplicable to game theory. What exactly that means shall be explored towards the end of this section. Let us first look into Hausman's argument more closely.

Hausman argues that the only way to dispense with beliefs in folk psychology is to embed them in preferences such as to form 'final preferences.' These final preferences are then adequate to determine the specific choice to be explained; no mediation through beliefs is required as these are already part of final preferences.[5] So could utilities be read as representing final preferences? Hausman's answer is negative. He has a number of arguments to support this answer, which he could distinguish more clearly than he does. Let me explain them in turn.

Hausman at one point (pp. 107-108) argues that if utilities in games were to represent final preferences, they would include beliefs about the action of the other player(s) and therefore bypass the predictive task of game theory. The challenge of predicting outcomes in games is to break through the interdependency of players' beliefs about each other's play to arrive at an equilibrium. Plugging actual (equilibrium) beliefs into the definition of preferences is to presume the equilibrium beliefs in defining the game.[6] Hausman at one point says that this is to trivialise the task of prediction. But the concern should go deeper: Presuming equilibrium beliefs for each outcome is to obscure what it is to solve these games at all. I will come back to this point.

In the same breath as addressing this concern about prediction in games, Hausman also expresses the thought that final preferences would trivialise prescription in games, i.e. the reading of games which takes rational equilibrium strategies as advice for players. I will dedicate a section of this chapter to this point.

---

[5] Hausman points out that "Preferences for actions might be like this in Savage's decision theory." (2000, p. 105) He does not develop this claim. I will return to Savage's decision theory and its notion of outcomes in Section 6 of this chapter.

[6] This is a central concern in Chapter 3 as well.

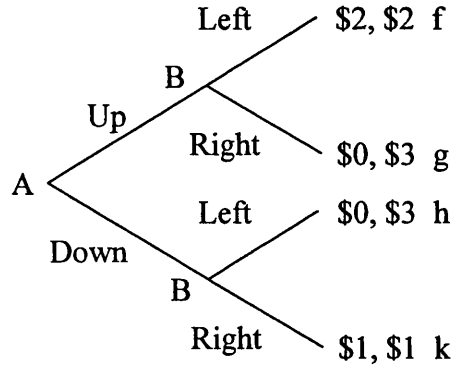


Figure 1.2

At another point of his argument, it sounds as if Hausman is worried that, given the final preference reading, certain outcomes of a game cannot – in the sense of a logical impossibility – be compared. For instance, in the game in Figure 1.2, player *B* could not possibly compare outcomes *f* and *h* such as to derive final preferences.[7] This is because, even hypothetically, the two outcomes could not be available at the same time since they are defined in terms of two mutually exclusive actual plays of the game.[8] This certainly causes a categorical problem for PCI. But does it also spell trouble for ERPT? This depends on whether it can be argued that some choice situation, which is sufficiently similar to the ones over which preferences are to be formed, can be constructed. This will be discussed in Section 5 of this chapter.

The same holds for the next point that Hausman makes. He claims the following about Figure 1.3:

		Column	
		Left	Right
Row	Up	1, 1	3, 0
	Down	0, 2	2, 3

Figure 1.3

[7] Compare Hausman (2000), p. 107-109.
[8] This point about hypothetical choice is developed in Hausman (2005b); a paper that I will return to at the end of this chapter.

> "The numbers could not possibly represent [final preference] rather than preference. If they did then Row would know that Column would play *left* if Row plays *up* and that Column would play *right* if Row plays *down*, and so Row should play *down*! This is a bizarre and perhaps even incoherent way to read the normal form [...]" (p. 111, italics in original)

Hausman probably tries to convey that, since final preferences encapsulate beliefs about the actual play, they no longer allow for causally separate reasoning of players about their play. While this point remains undeveloped, it seems to hold more water than Sen's point about the prisoner's dilemma. Hausman's criticism is about what kind of hypothetical reasoning in games is possible in conjunction with PCI. This deserves even more attention and will be treated in the next section. However, I would like to remark that the argument presented, even if sound, does not undermine ERPT.

3. Counterfactuals

Any discussion of solution concepts in games, whether represented in normal or extensive form, must be based on an analysis of the counterfactual situations in a game.[9] That is to say equilibria are to be justified on the basis of subjunctive conditionals of the type "If move x had occurred, move y would have occurred (or: be the best strategy)..." or so. A formal tool of analysis must take this into account. It must be able to accommodate an account of hypothetical reasoning. Chapter 4 will focus on the representation of such reasoning and the underlying representation of epistemic states of players in games. This chapter will only touch on some of these issues. The main aim of this section is to demonstrate that hypothetical reasoning in games is not possible in conjunction with PCI. This idea is explicitly addressed by Skyrms (1998). So his argument will be presented first. However, I will argue that Skyrms' solution to the problem is misguided, as he holds on to PCI for the wrong reasons.

Skyrms' critique

Skyrms argues that the whole idea of a hypothetical reasoning in games, which underlies all intuitive justifications of equilibrium play, becomes redundant – and even meaningless – when a certain revealed preference theory is applied to games. More

[9] For pioneering works see, e.g., Bacharach (1987), Bicchieri (1988), Binmore (1987). A more systematic literature overview is given in Chapter 4.

precisely, he claims that the following three criteria imply backward induction, and *a fortiori* Nash equilibrium, by *reductio ad absurdum*:

(1) Complete information concerning payoff and structure of a game
(2) PCI
(3) Players' probability spaces are governed by (1) and (2).

Skyrms' original version of condition (2) is "Revealed Preference: Payoffs are in terms of utilities (interpreted as dispositions to choose)." But the exact meaning of "disposition" is not elaborated at all by Skyrms. PCI, however, seems implicit to how Skyrms exploits condition (2). It also seems faithful to Skyrms' initial statement: "In this paper we investigate subjunctive reasoning under a strict revealed preference interpretation of utility. This interpretation has the virtue of making the logic of utility maximization crystal clear." (p. 546)

The *reductio* argument can be stated most concisely for a normal form game like the prisoner's dilemma, as in Figure 1.1. Given (1)-(3) (and the usual assumption of common knowledge of rationality), all outcomes except for mutual defection are "impossible" by assumption. No matter what each player thinks about the other's move, he cannot be rational unless he defects. Skyrms point is that supposing that one of the players co-operates immediately leads to a contradiction:

> "In a world in which Row plays [Co-operate] Row is either playing a different game or perhaps is not coherent enough to be said to be playing any game at all. This means that certain counterfactual suppositions which, on the face of them, are sensible, are not satisfied in any world. There is no answer to what Column would do if Row were to play [Co-operate] in [the prisoner's dilemma]." (p. 554)

If this point is accepted, a proof by induction shows that the backward induction path is realised in games of perfect information with no payoff ties (Skyrms 1998, pp. 556–557). Graphically, this can be represented by crossing out backwards those branches of game trees which will not be played in any possible world leaving a single branch springing from each decision node.

Take the game in Figure 1.4a, as used by Skyrms (1998) but first introduced by Myerson (1991, p. 192). It allows for the following natural forward induction argument: 'If rational player *A* plays Across it can only be to obtain a payoff of 9. *B* knows this and plays Up at his first node.' According to backward induction, however, *B* would

play Across at his first node. Assumptions (1)-(3) render this backward induction solution trivially true. The forward induction solution can not even be coherently grasped under these assumptions. A version of the game which is (partly) reduced in accordance with (1)-(3) is depicted in Figure 1.4b. It mirrors the fact that no sense can be made of any hypothetical reasoning about the non-backward induction play. In fact, a fully reduced game would just consist of a single line leading from *A*'s first node to the outcome (2,0).

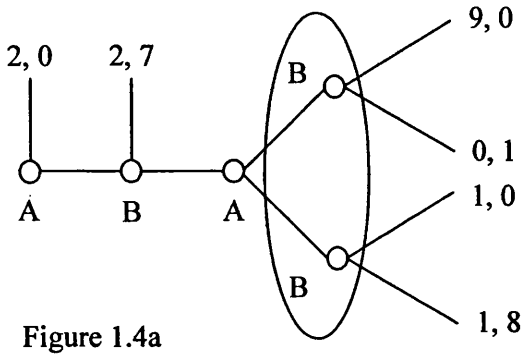


Figure 1.4a

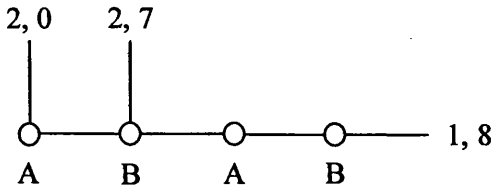


Figure 1.4b

Note a background assumption of this argument is the "usual partitional account of knowledge and common knowledge," going back to Aumann (1976, 1995) so that "rationality of the players is common knowledge at all times in all possible worlds." As I will discuss in much detail in Chapter 4, counterfactual reasoning is impossible to make sense of in this formal framework. I will come back to the implication of this observation later in this section.

The wrong remedy

For now, we need to ask what to make of Skyrms result. Clearly the 'good news' of having found a simple way to solve a large class of games, when accepting (1)-(3), is outweighed by the 'bad news' that the solution is deprived of all intuitive support in the form of hypothetical reasoning.

One obvious remedy is to relax one of the assumptions (1)-(3). Skyrms seems to want to hold on to assumption (3) keeping "rationality of all players a tautology in each of the players' own probability spaces" (555) at any cost. As will be discussed in Chapter 4, this assumption about how the players' beliefs and their revision is to be thought of is also dependent on the representation of epistemic assumptions. For the

moment, let me simply accept Skyrms' view on this. Skyrms also wants to hold on to assumption (2), as he wants to preserve the "crystal clear" interpretation of utility maximisation. As I will argue, of course, this reason to hold on to (2) is insufficient.

Nonetheless, Skyrms decides to relax assumption (1), and assumption (3) accordingly, introducing a version of Harsanyi's incomplete information approach, according to which there is always infinitesimal uncertainty about players' types. To be more exact, types are defined over agents, where a player consists of a number of agents, one for each information set. A type is specified by a utility function and a belief set over other agents' types. Skyrms labels this approach "limits of reality" because while doubt is presumed to be realistic and certainty unrealistic, infinitesimal doubt is thought to constitute the limit of reality. Under this assumption, players update beliefs according to Bayes' rule over the course of a game. This re-introduces sufficient complexity to the game to lend subjunctive conditionals about off–equilibrium situations a meaning. In addition, backward induction only holds under a number of considerably strong conditions.[10] It is not an analytically trivial way to reduce games any more.

Giving up condition (1), however, comes at a price; namely giving up on a central and practical idealisation of probability-one beliefs or knowledge as applied in much of classic game theory. At the same time, it leaves a bad aftertaste because even in an idealising framework, where the probability for some off-equilibrium situation is just zero, the subjunctive conditionals with antecedents that refer to these off–equilibrium situations in the antecedent should still be meaningful. Even when *certainly* not realised, an off-equilibrium situation should be modelled as logically possible.

Also, when some play is causally excluded, so that player *A* knows, e.g., that player *B* will not make a certain move, it should still be treated as "epistemically open" to player *A* that he does. In fact, player *A* should be endowed with a belief revision policy which is conditional on the (counterfactual) information that player *B* did not make the expected move.

In Chapter 4, I will defend a framework developed by Stalnaker (1994, 1996, 1998) which allows for representing off-equilibrium plays of a game as well as for

[10] The conditions are: (1) "Truth," i.e. agents' near-certain beliefs are actually true; (2) "Belief," i.e. in each player's priors there is common belief in (1); (3) "Richness," i.e. each player's beliefs must give every strategy profile non-zero probability; and (4) "Independence," i.e. in each player's priors all agents are treated as epistemically independent. (pp. 564-568) Very similar conditions will re-appear in Stalnaker (1998) as discussed in Chapter 4.

representing belief revisions by the players. This framework is semantic and inspired by Kripke (1963). It is based on models in relation to which conditional statements about a given game, and thereby solution concepts, are deemed true or false.

In contrast, Skyrms introduces a framework designed to deal with the pragmatics of subjunctive conditional in games. He generally believes (following Ernest Adams) that "conditionals do not function as factual claims but rather as bearers of conditional probability." (Skyrms 1994, p. 13) Vis-à-vis game theory, he believes that statements of the sort 'If player x were to do act A, then consequence C would ensue' are to be seen as carrying a probabilistic 'assertability value' which helps the agent to calculate his subjective expected value of an act. In this sense, subjunctive conditionals are "decision making conditionals" (Skyrms 1998, p. 549).

While the pragmatic view of subjunctive conditionals might have its justification, it does not suffice for supporting solution concepts in game theory. In games, the *meaning* of a counterfactual statement about off-equilibrium situations should be clear. The truth-value of such statements should determine actual play. Skyrms' pragmatic reading of conditionals in conjunction with the auxiliary assumption that every play might be the actual play with some small probability[11], however, does not provide the framework to make sense of counterfactuals.

This might also explain why Skyrms seems to endorse (or refuse to relax) PCI. If all subjunctive statements about games and all hypothetical beliefs by players are just probabilistic statements about the actual world, there is no conflict in principle with interpreting preferences as *actual* choices. Utilities can be seen as representing preferences which in turn represent consistent choice behaviour, some of which is simply infinitely unlikely to occur. For the same reason, Skyrms sees no need to divert from a partitional representation of epistemic states which does not allow for representing hypothetical reasoning and counterfactual play in games either, as mentioned above.

To sum up, Skyrms is correct in pointing out that PCI, along with other assumptions in game theory, makes it impossible to make sense of counterfactual play in games, and hence of solution concepts as derived from hypothetical reasoning. However, he is misguided in his solution to the problem. Instead of holding on to PCI, he should realise that PCI itself only allows for conceptualising (more or less probable)

[11] This is at least according to the "realist interpretation of infinitesimals." Compare Skyrms (1998), p. 564.

actual plays of a game, never counterfactual play. He should endorse a conceptual semantic framework for subjunctive conditionals in game theory all of whose components allow for attributing meaning to subjunctive conditionals. Preferences cannot be actual choices but must be some notion which is conceptually separate from actual choice, in order to make sense of discussing hypothetical choice. Furthermore, the representation of epistemic assumptions must be such that knowledge of rationality does not imply certain play as a matter of necessity. It must enable coherent and structured reasoning (by the players and the game theorist) about non-actual play.

Whether ERPT is endorsed or not is a separate question. If it is, one will have to present a version which is in harmony with subjunctive reasoning about games. I will argue in Section 5 that this is problematic as well.

4. Prediction and prescription

As mentioned before, Hausman implies that game theory has normative appeal. He finds that (what I labelled) ERPT is not compatible with this task: "Revealed-preference theory [...] trivialises the tasks of advising players on what strategy to employ or predicting what strategies will be played." (2000, p. 107) This section will defend this criticism. It does so in three steps. First I will explain how far game theory, as a rational choice theory, has a normative dimension. I will argue that this argument is not question-begging. Second I will argue that this is not just a reading of game theory which can be separated from a positive reading. Third I will look into what this implies for the definition of utility in games.

The normativity of instrumental rationality

There is some controversy about the use and purpose of rational choice and game theory. Some authors see both as (at least partly) normative exercises, meaning that equilibrium strategies in games based on rationality assumptions can be read as prescriptions for action. Sugden (1991), for instance, rejects a strictly positivist stance, taking the normative appeal of instrumental rationality as the basis for its predictive or descriptive applicability:

> "I shall be concerned with a different view of rational choice theory, in which the theory is seen as having a genuinely normative content: it tells us how, as rational agents we *ought* to choose. If the theory also has predictive

power, this is because, in a non-tautological sense, human beings have some tendency to act rationally." (p.752, italics in original)

Other authors see game theory as a purely positive discipline. They find that game theory should be able to predict or explain human behaviour in strategic contexts.[12] Luce and Raiffa (1957) went so far as to address the normative reading of games under the heading 'Some Common Fallacies.'[13] Their point also reminds us of the interpretation of utility which is at stake in this chapter:

> "[...] we may think of the [utility] theory as a guide to consistent action. Here, again, certain (simple) preferences come first and certain rules of consistency are accepted in order to reach decisions between more complicated choices. [...] The point is that there is no need to assume, or to philosophize about, the existence of an underlying subjective utility function, for we are not attempting to account for the preferences or the rules of consistency. We only wish to [...] represent them." (pp. 31-32)

The controversy about the normativity of game theory has its roots in the standard assumption of rationality of players in a game and, usually, common knowledge of it or belief in it.

In a number of philosophical traditions, instrumental rationality has been taken as the source of some sort of 'ought.' Kant (1785), for instance, refers to the so-called hypothetical imperative. For him it is an analytic truth that, given a desired end, a person should will the means which are known to bring this end about.[14] Kant labelled this imperative 'hypothetical' because of its (contingent) dependence on ends. But he did not cease to attribute normative authority to it, although not universal authority, of course. Critics of Kant's moral philosophy, such as Humean philosophers, have traditionally raised concerns with Kant's conclusions about the categorical imperative, but rarely questioned the validity of the *hypothetical* imperative. Korsgaard (1997, p. 218) writes that "both empiricists and rationalists have supposed that the instrumental principle itself needs either no justification or has an essentially trivial one."

But although there might be broad agreement that some sort of 'ought' comes with instrumental rationality, it is not obvious what moral and logical status this 'ought' has. Broome (1999, 2003) argues that the instrumental principle has only the status of a

[12] For a recent advocate of this view compare Binmore (1987/1988), for example.
[13] This comment is made in relation to von Neumann and Morgenstern's (1944) expected utility theory.
[14] He adds that identifying the causal means-ends relation might require synthetic reasoning as well.

'normative requirement,' which can never be reason-giving. This is because their logical structure is that of "wide-scope oughts."[15] Take, for example, the sentence '*p* requires *q*', where *p* is a desired end and *q* is the necessary means or action to attain it. Then this sentence can be captured as $R(p \rightarrow q)$ or $O(p \rightarrow q)$. In these expressions, *R* could be read as 'there is reason to,' *O* could be read as 'one ought to.' The arrow represents an implication and parentheses mark the scope. According to Broome, it cannot be inferred from $R(p \rightarrow q)$ that $p \rightarrow Rq$. Equally,we cannot infer $p \rightarrow Oq$ from $O(p \rightarrow q)$. That is, it cannot be inferred from normative requirements relating *p* and *q* that one ought to *q* given *p*. Normative requirements do not entail free-standing 'oughts' which instruct us what to do. They only regiment the relation of certain beliefs (or desires) and actions.

Broome's way to think about the logic of instrumental 'oughts' is perfectly compatible with the premise of this section. According to Broome's stance, rational choice theory can be read normatively without saying that it generates reasons for action which stand alone.[16] If we take rational agents in game seriously as intentional agents, rationality *requires* them to take action consistent with their preferences. And game theorists advise them to do so.

There are drawbacks to Broome's theory, however. For instance, it is hard to spell out a wide scope 'ought' as a sentence in ordinary language. It seems that, when formulating the imperative of rational choice (in a game) in English, the scope of the 'ought' only includes the consequent: 'Given the preferences, a player *ought* to choose that strategy which is calculated to yield the greatest expected payoff.' This lack of an equivalent of a wide scope 'ought' in ordinary English is a serious problem for Broome's theory, which I cannot address here. For the time being, the logical form of the normativity of instrumental rationality must therefore remain an open issue.

Begging the question?

The objection comes to mind, however, that this argument presupposes that preferences in games represent some given subjective attitude – the claim to be argued for in the first place. To this I would like to react by first pointing out that it equally applies to a predictive use of game theory, as Hausman's quote from the beginning of this section correctly implied. Secondly, I would like to argue that the intuitive appeal of the

[15] Broome's points are much more thoroughly and carefully developed. The reader is asked to forgive this rough representation of those points which seemed relevant for my ongoing argument.

[16] For a critique of neo-Humean projects, which attempt such a deduction of practical reasons from instrumental rationality, compare Hubin (2001).

rationality axioms, which underlie the representation theorems necessary for reading games as mere representation of choices, seem to spring from normative intuitions about agency. These intuitions, however, always presume some sort of subjective ends or preference. To deny these presumptions in interpreting game theory while at the same time referring to them in the context of representation theorems seems conceptually incoherent to me.

Separating prediction and prescription?

Kadane and Larkey (1983) are worried about the fact that in most analyses the normative and the descriptive dimensions of game theory are intertwined. In order to avoid this "confusion," they suggest an analytical separation of the positive and the normative theory. The positive bit is to be designed to be somehow testable and is to be empirically validated. The normative bit is to be validated in its success in prescribing decision procedures that lead to the highest average payoff.

The problem with this method is that the normative analysis of a game can never stand alone, because a player or his advisor need both a predictive theory with respect to the other player *and* a prescriptive theory to guide the own choice of strategy. This interrelation is complex because the prediction of the other player's behaviour will depend on the first player's action, which in turn follows the prescription etc. This problem holds, at least, as long as I make symmetric rationality assumptions about the players. Most solution concepts, however, do assume symmetry of players.

So far I have argued that prescription is inherent to the notion of rationality which in term is the key to solving games, that prescription in games needs a descriptive basis, and that the descriptive and the prescriptive parts of the analysis of games are so interwoven that they cannot be treated separately. Let me now spell out what that implies for the notion of utility in games.

Rationality as utility maximisation

As mentioned before, PCI is incompatible with the prescriptive reading of games. If what is represented by the payoff is the outcome of a choice itself, I arrive at statements such as: 'A player should choose a strategy such as to maximise the (expected value of the) measure which represents his choices.' This seems tantamount to the odd imperative: 'Do what you are doing anyway.' Accordingly, it is futile to think of rationality as a normative requirement at all. Starting with PCI, the game theorist might

see a task in identifying the rational equilibrium play. But then, by assumption of PCI, the players have been assumed to play rationally all along. The game theorist has been involved in a redundant exercise. His theory cannot add any predictive or prescriptive value.

To re-gain such usefulness, the notion of subjective preference in games must be conceptually divorced from the concept of choice. Preference has to be the backdrop against which instrumental rationality can be motivated in the first place. It has to be the factor with respect to which rationality has normative authority, if only a relative one as laid out by Broome.

Having said that, utility and preferences must also have causal efficacy. That is, conjointly with the beliefs assumptions, the assumption of rationality as utility maximisation must be a determinant of action in a predictive sense. Under the usual assumptions, there should not be a gap between payoff maximisation and what the individual does. This is necessary in order to use any of the rationality-based solution concepts in games at all. As I have pointed out, even a normative reading of some player's rational strategies often presupposes a predictive reading of some other player's action. I will come back to the requirement of causal efficacy in the next section.

5. Effective preferences

In this section, I outline an interpretation of preference in game theory which, I believe, lives up to its conceptual role. I will call this new interpretation of utility in games 'effective preferences.' The definition I will attempt here is not exhaustive. The aim is simply to highlight characteristics which the notion of effective preference has to combine if it is to evade the problems discussed above. It will be pointed out what the requirement that preference has to serve to determine outcomes in games implies.[17] (This requirement of causal efficacy, of course, is what lends its name to the label 'effective preferences.') Then I will address the idea that payoffs in games represent some sort of 'goodness.' Finally, I will come back to the question of what use RPT is for empirically substantiating preferences in games.

[17] Note that designing the utility notion so that it serves as the basis for defining rationality is not the same as trivialising the solution of games. Solving games can still be a complex, if determinate, matter.

All things considered?

As said earlier, Sen (1973) argues that there might be "non-preferential" motives for choice. Such motives, he argues, cannot just be defined as one of the constituents for a technical term of preference, because we will have to use the (same) notion of preference in welfare economics. Sen (1976) also argues against incorporating all motives into a single notion of preference.[18] For him the rational agent which is projected by assuming a single preference ordering must be "a bit of a fool" and "close to being a social moron." Sen is convinced that economic theory's preoccupation with such fools is misguided. Sen (1997) even introduces conventional rule following as a motive for agents. He argues that these are triggered in particular choice contexts; they *immediately* influence the *act* of choice.[19]

Hausman (2005b) does not agree with Sen's conceptual multiplicity with respect to modelling preferences. He wants to regiment the economist's use of "preference." Concretely, he suggests:

> "An agent's preferences consist of his or her overall evaluation of the objects over which preferences are defined. This evaluation implies a ranking of these objects with respect to everything that matters to the agent: desirability, social norms, moral principles, habits – everything relevant to evaluation. Preferences thus imply all-things-considered-rankings. [...] It should be the single correct usage of the term "preference" in economics and decision theory." (p. 37)

Among the stronger arguments[20] for the all-things-considered-rakings interpretation (henceforth ATCRI) provided by Hausman are that it comes close to the usage of the notion in rational choice theory, game theory and expected utility theory, as we know them. Conversely, he debunks one of Sen's arguments against that standard usage as a non-sequitur: While a player who might reason in terms of a single preference ranking

---

[18] In particular, he thinks that "commitment" which helps one to overcome the temptation of free-riding etc., should not be subsumed under the same notion as egoistic motives or sympathy (which simply reflects externalities of other people's outcomes on one's welfare). The solution offered by Sen (1976) is introducing (moral) meta-rankings over first-order preferences. To my knowledge, this notion remains formally undeveloped in Sen's work.

[19] By way of an example, Sen (1997) demonstrates how one such convention ("strategic nobility") might impact on choice behaviour in a game. He effectively argues for a "solution" to that game which, to me, seems irreconcilable with anything orthodox game theorists call a solution concept.

[20] There are two quite weak, if not irrelevant, arguments as well. The first is that ATCRI in tendency conforms to the ordinary meaning of "preference" (Compare Hausman (2005b), p. 45). In another article, Hausman himself points out that this cannot be a criterion for the economist (Compare Hausman (2000), p. 105). The second argument is that ATCRI allows to distinguish the question of what things are taken into consideration in specific people's preferences from the question what preferences are.

might be a 'rational fool,' the theorist who *models* the player this way need not be a fool at all.

The key argument, however, remains that any interpretation of preferences short of ATCRI will lead to a picture of game and decision theory where "rational choice need no longer be determined jointly by beliefs, knowledge of the game form, and preferences over outcomes" (p. 47). So ATCRI fulfils the requirement that I was trying to express as 'causal efficacy.' At the same time it does not collapse into PCI. So preferences (mediated by belief) fully determine choice, but they are not identical with the choice. This seems to be what game theorists should be after. It is a good first draft for what effective preferences should be.

There are two potential kinds of problems with ATCRI, however. The first kind consists of practical questions: Who takes care of investigating the process of preference formation, if not the economist? How do we ever recognise a game (in the technical sense) when preference formation is a black box? Hausman (2005b) is aware of these problems. He recognises that the task of identifying preferences resides in "a sort of limbo" (p. 47), but does not offer a remedy. In another paper (Hausman 2005a), however, he seems to suggest that game theorists should not only be concerned with how to solve games but also with identifying what games are really being played and when. I shall not be concerned too much with addressing these issues here, as that might threaten the applicability of game theory, but not its coherence.[21]

The second kind of problem is a concern about the coherency of ATCRI. It goes back to the strong implicit presumption that "all things considered" by agents can be coherently defined as part of the *outcomes* over which the all-things-considered ranking is defined. I believe that there are (at least) two potential problems with the presumption.

One problem comes with questioning the implicit assumption of 'proto-consequentialism' as defined by Bacharach and Hurley (1991). This assumption addresses the general concern that when context of choice is re-defined as part of the outcome of choice, a new (second-order) context might be intuitively relevant with respect to the new decision situation. It claims "that there must be some level of description of decision problems at which all acceptable distinctions of context relative

[21] In Chapters 3 and 4, however, I will address the importance of (and problems with) identifying games correctly.

to naïve characterizations of goods have been captured, and such that no context dependence of evaluation relative to it can be rational" (p. 13). Bacharach and Hurley, however, argue that nothing guarantees that this holds in principle. Logically, there need not be a natural ending to the chain of context-of-outcome and context-of-context-of-outcome etc. But, as the authors mention, empirically there might be limits to what order of context the game theorist or the players might assign any kind of intuitive significance to. I believe this to be true, at least for folks whose intuitions have not been spoiled by moral theory. But I also suspect that these limits vary from situation to situation, game to game. But while identifying those limits implies a lot of work to be done by the theorist, it does not constitute an impossible task.

A further concern with the presumption that all things considered can be coherently defined as part of the outcomes points towards such impossibilities. Hausman (2005b) writes:

> "Moreover, all-things-considered preferences depend on *beliefs* as well as on motivating factors. [...] Belief and preferences can be input to choices, as they are when my desire for water and my belief that the liquid in front of me is water lead me to drink. Beliefs and preferences can also influence preferences, as, for example, they are when my preferences among flavors and textures and my aversion to early death coupled with my beliefs about the consequences of alternative diets determine my preferences among alternative foods." (p. 38)

Hausman correctly observes that beliefs in decision situations, strategic or not, are often part of the preference-formation, not only in calculating the utility-maximising act. But there might be problems with belief-dependence of outcomes in decision and game theory if the beliefs are beliefs about the process of choice itself.

I will argue in Chapter 3 that problems with the conceptual justification of Nash equilibrium arise if preferences are endogenous with respect to beliefs over the strategies played in a game, as first suggested by Geanakoplos, Pearce and Stacchetti (1989). So even if games could be defined with belief-dependent preferences in that sense, the resulting game could not be solved with the established solution concepts. This point is independent of the expected-utility theory applied and therefore poses a serious problem to ATCRI.

To conclude, there are a number of unresolved problems with defining outcomes as context-dependent. ATCRI simply presumes such problems to be surmountable or

irrelevant. This seems unreasonable. So my alternative notion of effective preference must contain a clause to the effect that only so much context is to be accounted for in the outcome-description that no incoherency conflicts arise. This, however, might come at the cost of reducing the scope of game theory significantly.

Preference and goodness

Broome has long been concerned with pointing out the difference between reading 'utility' as usefulness, as (a representation of) preference, or as goodness.[22] He has also pointed out that if one reads decision theory as theory of practical reasoning, utility is best understood as some sort of goodness. His argument is that he cannot "envisage any process of reasoning that could take you from a preference to an intention." He continues, "A preference can cause an intention, but I do not see how this causal process could be one of reasoning." (1999, p. 102) The notion of goodness that Broome has in mind is not specified. But it can be read as fully subjective, diverting from what might be judged as in some sense objectively good for the agent.

Without going into the details of Broome's theory of practical reasoning, I would like to dismiss any notion of goodness as an interpretation of utility in game theory. The reason has been given before. Even if the prescriptive reading of an equilibrium play of a game might seem compatible with such a goodness-interpretation of utility the predictive element to this reading would not be. Utility-as-goodness, be it subjective or objective, conceptually allows for intervening factors such as weakness of the will or impulse to come between utility and action. So prediction could never be deterministic, as is needed for solution concepts to 'get off the ground.' Of course, goodness could be re-defined such as to exclude intervening factors. But this would simply render Broome's re-interpretation void.

Hypothetical choice and testability

Hausman argues that preferences in games cannot represent actual choice. He also correctly argues that it could not even represent hypothetical choice (2005b, p. 36). Consider his supporting counterexample, the marriage game. It consists of his choice between Proposal and Non-proposal and her (subsequent) choice between Acceptance and Rejection. He has to have a ranking over, e.g., (Proposal, Acceptance) and

[22] Compare, for instance, Broome (1991c).

(Proposal, Rejection). But, as Hausman observes, this preference could not even be deduced from a hypothetical choice situation, because he cannot choose Acceptance or Rejection on her behalf.

If we cannot envision a state in which certain hypothetical choices would be possible, then we cannot artificially set up such choice situations in the real world either. This implies that there could not be effective experiments, for example, to test for the underlying preferences. That is, of course, unless we can exclude the possibility that certain contextual aspects matter to an individual. But if we could have such information we would probably have some direct access to an individual's preferences in the first place.

If such experiments are not possible, however, even ERPT, the epistemological interpretation of RPT game theory, loses much of its bite. This is because there are many games containing outcomes which could not possibly be objects of (hypothetical) choices. So in many games, utilities could not even stand for choices that we could observe *in principle*. But then it is quite unclear what significance RPT could have for the understanding of utility in game theory at all. The empirical content and testability of utility in games seems doubtful. Not only is it methodologically illegitimate to dispense with the substantive notion of subjective preference, as suggested by PCI, but it also seems prima facie unjustified to suppose that any such notion could be tested by observing choice, as suggested by ERPT.

6. Probabilities and expected utility

So far I have used 'utility' and 'preferences' as almost synonymous. This is because I do believe that utility should express preferences. But in order to weigh utilities with probabilities, as done when defining games of imperfect information or mixed equilibria, a notion of expected utility is required. That is, we need cardinal utility, in the technical sense of a 'measure,' which is robust with respect to linear transformation. But how do we get from preferences to expected utility?

Neoclassical economists, such as Jevons and Marshall, thought of cardinal utility as 'measurable' in the sense of being derivable from marginal utilities. Marginal utility in turn was thought of as derivable from observed exchange of goods.[23] One of the many concerns with this method, epistemological and conceptual, is that it presupposes

[23] Compare the critical review of historical notions of 'measurable' utility in Ellsberg (1954), especially pp. 533-538.

certainty of prospects. This problem, however, can prima facie be avoided by relating utility to choices under uncertainty in the form of representation theorems.

Von Neumann and Morgenstern expected utility

Most famously, von Neumann and Morgenstern (1944) (henceforth vNM) introduced such a representation theorem. According to the Luce and Raiffa (1957) reading, the theorem states: Given that a set of consistency assumptions[24] is fulfilled, people act *as if* they are maximising expected utility. Some remarks on the applicability of von Neumann and Morgenstern's representation theorem for the notion of effective preference are due.

Firstly, this representation theorem, just like RPT, ultimately relates utility to acts. This does not itself imply an interpretation of utility. But, as Luce and Raiffa's reading exemplifies, it tempts game theorists to jump to the conclusion that cardinal utility *means* consistent choice. This is analogous to inferring PCI from RPT. And the two inferences are equally spurious. In fact, reading utilities as consistent acts leads to many of the problems explored in Sections 2-4.

In addition, the axioms of consistent choice in the vNM theorem are stronger then the weak axiom of revealed preference, WARP. Whether these rules are applicable or not in most situations has been the subject of an extensive literature.[25] Much of this literature has rightfully questioned whether these axioms hold (even in tendency) descriptively. Even highly educated agents have been observed to have problems with applying some basic rules of probability theory. In addition, some authors even questioned the axioms' normative appeal.[26] If the descriptive doubts hold, game theory which is based on vNM-type expected utility would at least be severely limited in scope. If the normative doubts prevail, vNM expected-utility would not even be a justified concept for idealised agents.

Finally, the vNM axioms presuppose *objective* probabilities in games. Probabilities are explicitly conceptualised as "frequency in the long run" (vNM 1944, p. 19). As will be explained out soon, this assumption is vital for game theory's main

[24] These are usually organised as the axioms of transitivity, completeness, the sure-thing principle and continuity (in probabilistic terms). Compare, e.g., Friedman and Savage (1952). But there are alternative axiomatic systems.
[25] Compare, most prominently, Kahneman & Tversky (1979).
[26] See, e.g., McClennen (1991).

solution concepts. But it remains a demanding assumption about the world modelled in games and agents' epistemic access to it.[27]

Savage expected utility

There are alternative representation theorems, however, which are based on subjective probabilities, such as in Savage (1954). The crucial question is whether they fit the game theoretic framework. Sugden (1991) finds that "deep problems are raised" (p. 764) when applying Savage's framework to game theory. Kadane and Larkey (1982), in contrast, defend such a transfer of Savage's framework to game theory. In a prompt reply, however, Harsanyi vehemently argues that Kadane and Larkey's proposal would mean ridding game theory of any interesting solution concepts as they are dispensing with the powerful implications of the usual (common knowledge of) rationality assumptions:

> "What do Kadane and Larkey put in place of von Neumann's theory? Only the highly uninformative statement that in two-person zero-sum game, just as in any other game, each player should try to maximize his expected payoff in terms of his subjective probability. [...] this amounts to no more than saying that he should do whatever he thinks is best." (1982, p. 122)

Harsanyi is right in arguing that Kadane and Larkey are too optimistic about reforming the interpretation of probability while holding on to the rest of game theory. On the other hand, it seems hasty to consider an interpretation of probability justified solely on the basis of its usefulness in 'solving' games. But then I have no alternative to offer myself.

Apart from this deep conflict of Savage's approach with traditional solution concepts, however, there is another problem. Savage's decision theory presumes a space of consequences which is distinct from the space of states. Acts, according to his theory, are functions from states to consequences. Furthermore, expected-utility can only be deduced from this framework on the assumption that players can form preferences over all prospects, i.e. all (formally) possible state-outcome combinations. This is the so-called 'rectangular-field assumption'[28] which Valentyne (1993) formulates concisely:

---

[27] Furthermore, I suspect that this assumption might be incompatible with Aumann's (1987) reading of mixed equilibrium strategies as an equilibrium of subjective expectations of players about the others' play. This is not the place to develop this point though.

[28] See Broome (1991a, pp. 80-81 and 115-117).

> "For all outcomes o_1-o_n, if o_1-o_n are possible outcomes under states s_1-s_n respectively, then $<s_1,o_1;...;s_n,o_n>$ is a prospect." (p. 27)

This assumption, however, becomes quite implausible once modal properties are attributed to outcomes. The assumption and an example of its conflict with process-dependent outcome-specifications will be explored in the next chapter. The example referred to in Chapter 2 will be taken from decision theory, not game theory. But by introducing additional agents, the argument for modal properties of outcomes can only gain more support. The significance of process-dependent outcomes will also be addressed in Chapter 3.

Furthermore, the rectangular field assumption raises problems with Savage's state-outcome distinction in general. In games, one player's actions could be read as the states faced by the other agent. Now consider the marriage game from above again. According to Savage's rectangular field assumption, at least one prospect includes a state of the world in which she rejects but the outcome or consequence of his and her interaction would still be marriage. Without unorthodox assumptions about the nature of the situation represented by this game, however, it seems hard to grasp the meaning of (the strategy option) 'rejection,' if it is compatible with the consequence denoted 'marriage.' This has to do with the fact that 'rejection' as the description of the state makes reference to a certain set of outcomes already. Vice versa, the outcome description 'marriage' refers to certain preceding actions. I shall leave the discussion at that here.[29]

[29] For a discussion of Savage's notions of 'state,' 'act' and 'consequence,' as well as their distinction, compare also the correspondence between Aumann and Savage (1971). Aumann raises critical points about the fuzziness and conceptual distinguishability of the concepts mentioned. These problems are also at the heart of the problems with RFA: In order to capture consequences fully they have to be described state-dependently, but these state-dependent outcome descriptions lead to absurd act descriptions. Savage concedes these problems, but holds on to the idea "that decision situations can be usefully structured in terms of consequences, states, and acts in such a way that the postulates of [Foundations of Statistics] are satisfied." But his idea remains vague: "Just how to do that seems to be an art for which I can give no prescription and for which it is perhaps unreasonable to expect one – as we know from other postulate systems for application."(p. 79)

7. Conclusion

I have argued that the interpretation of utility as consistent choice based on revealed preference theory (PCI) is conceptually inapt for game theory. The same holds for an epistemological interpretation of revealed preference theory, according to which utility does not mean *consistent choice* but could be derived from it even in strategic contexts (ERPT). My critique takes its origin in Sen (1973) and Hausman (2000). Their critiques contain a lot of interesting arguments and ideas some of which I have tried to follow through in more detail. One major theme is that ERPT (and PCI) do not allow for a distinguishing preference and beliefs in games. This undermines the motivation of equilibrium play.

A related criticism is that hypothetical reasoning in games, the driving force behind rationalistic solution concepts to games, is impossible to grasp in conjunction with PCI. If preferences just represent actual choice then nothing meaningful can be said about what rational players would do in off-equilibrium situations. I argued that Skyrms (1998), who endorses this argument, suggests the wrong remedy. I hinted at an alternative framework to his.

I have also argued that there is an inherent normativity to rational equilibrium play. I explored the specific status of this normativity, which derives from the notion of instrumental rationality, and its entanglement with descriptive aspects of rational play in games. Prescriptive readings of equilibrium play, however, are also incompatible with PCI.

Pointing at a remedy, I have introduced the alternative reading of utility in games as effective preference. This first outline of a definition draws a line between choice and preference, so as to allow for counterfactual as well as normative reasoning, while trying to preserve as the causal efficacy of the preference concept. The new notion must therefore comprise all that motivates the agent. On the other hand it must avoid inconsistencies when referring to process-dependent outcome specifications. This limits the extension of effective preferences and thereby the scope of game theory which builds on this notion.

Finally, I have raised concerns about deriving expected utility from preferences. I pointed at problems which arise from representation theorems, which bridge utility and preference, distinguishing those that build on subjective probabilities and those that build on objective probabilities.

Chapter 2

Consequentialism and process-dependent properties of outcomes

1. Introduction

Consequentialist reasoning lies at the heart of decision and game theory. If it is refuted so are the theories that build on it. It must be kept in mind, however, that consequentialism is a vague notion unless we specify precisely what counts as a consequence. In decision and game theory this role is assigned to utility, as developed out of preference by representation theorems. In Chapter 1, I discussed the difficulties that are raised by interpreting these representation theorems. In this chapter, I shall be concerned mostly with a related problem: Normatively speaking, is consequentialism binding given that we are sufficiently tolerant regarding the question of what counts as a relevant outcome description? One reason to doubt this is that in many cases choices and outcomes do not fall apart as neatly as the theorist might wish. In fact, there are decision situations in which intuitively relevant properties of the outcomes depend on how the outcome is reached.

Verbeek (2001), whose arguments will be at the centre of this chapter, presents an example first brought up by Diamond (1967). It addresses decision theory, but its discussion concerns the interactive decision situations of game theory as well. In the example, a random move by nature – the toss of a coin – adds an aspect of fairness to outcomes. Verbeek argues that trying to take this aspect of fairness into account is incompatible with other axioms of decision theory, which can be defended on independent grounds. He concludes that consequentialism is incompatible with the aforementioned notion of fairness and other values which are dependent on choice processes as well as outcomes (naïvely construed). Verbeek points out that, if he is right, Hammond's (1988) project of justifying expected utility theory by deriving it from certain principles of consequentialism would also fail.

I will argue that Verbeek is wrong by debunking his case about Diamond's example as attacking a straw man. I will argue with Hammond and Broome (1991a) that aspects such as fairness need to be modelled as process-dependent (or modal) properties

of outcomes in decision theory. This does not lead to the problems that Verbeek identifies, as long as the individuation of outcomes is done in the right phase of modelling the decision situation.

I will then argue, however, that some worries about process-dependent properties *are* to be taken seriously. One concerns the pragmatic foundations of transitivity. I argue that it could be resolved by introducing a (rather paternalistic) substantial restriction on the re-individuation of outcomes. Another problem concerns the plausibility of an axiom underlying some expected utility representation theorems when allowing for context-dependent outcome re-individuation. While this affects Savage's decision theory, other decision theories might be immune to this problem. Game theory based on objective probabilities may not be affected either.

The plan of the paper is as follows. Section 2 outlines one project that is at stake in Verbeek's argument, i.e. Hammond's reduction of expected utility theory to consequentialism. Section 3 presents Verbeek's argument. Section 4 shows how it goes wrong and draws conclusions about when in the modelling process outcome-refinement should be completed. Section 5 discusses whether and how a maximin regret chooser can be represented coherently in a consequentialist framework. Section 6 explores how far process-dependent properties constitute a problem for certain expected-utility theories. Section 7 concludes the chapter.

2. Hammond's project

Hammond (1988) tries to derive the existence of expected utility functions from consequentialism. What makes his approach interesting is that it avoids using (arguably) strong assumptions like Samuelson's (1952) independence axiom or Savage's (1954) sure-thing principle. Instead Hammond deduces expected utilities from (what he takes to be) more basic principles of consequentialism.

His notion of consequentialism is particular, in so far as it rests on a special feature of von Neumann and Morgenstern's (1944) game theory. Hammond basically demands that behaviour norms, which regiment an agent's behaviour in some decision tree, should be those which are applied in any other decision tree with the same normal form. This is consequentialist insofar as, once outcomes are fully specified, the order of the choice process should not play a role; in particular choices in sub-trees ('continuation trees') should be coherent with the choices in the whole tree. This notion

might be in conflict with other technical notions of consequentialism[30] as well as definitions of the term in moral philosophy.[31] But let me accept it as a starting point.

Hammond's approach starts with extensive form representations of decision problems. However, he cannot assign payoffs or utilities to terminal nodes in decision trees, since those are to be derived. Instead he has to assign outcome descriptions to those terminal nodes. Hammond appreciates that this is not trivial. In fact he makes interesting introductory comments about the connection of consequentialism (which he reads as a normative principle) and outcome specifications. Let me present a quote, which takes us to core issue of this chapter:

> "As a normative principle [...] consequentialism requires everything which should be allowed to affect decisions to count as a relevant consequence – behaviour is evaluated by its consequences and nothing else. If regrets, sunk costs, even the structure of the tree itself, are relevant to normative behaviour, they are therefore already in the consequence domain." (1988, p. 26)

Interestingly, Hammond identifies a number of outcome properties here which, on a naïve account, manifest themselves in the *process* of deciding or choosing. On a more substantial normative account of consequentialism, they might even be excluded as outcome properties to be regarded at all. And, as we shall see, it is not unproblematic to include them.

But while Hammond identifies an area of potential problems, he also brackets it as irrelevant for his task. He thinks that identifying relevant outcomes is a task to be addressed by "practical normative theories," while his consequentialist analysis is about another task, which concerns the structure of normative behaviour. He writes:

> "Since the content of the consequence domain is really a subject for practical normative theory, I shall avoid it by taking as fixed the state contingent consequence domains [...] for a fixed finite set [...] of possible states of the world." (1988, p. 27)

As I will explain, Verbeek's critique implies that such a bracketing of the outcome-individuation issue is not possible. In fact, he attempts to develop a knockout argument against consequentialism by considering how tolerance with respect to process-

[30] Compare Levi (1991).
[31] Compare footnote 1 in Verbeek (2001).

dependent outcome-specification conflicts with further axioms of rational decision making that Hammond endorses.

If Verbeek's critique passed, Hammond's project would indeed be threatened. This is reason enough to consider it. However, a flaw in Verbeek's point does not by itself vindicate Hammond's project. Showing that Verbeek's point does not hold water, as I will attempt soon, only exonerates Hammond's project of one major charge. And, in fact, I do not wish to defend Hammond's notion of expected utility in this thesis as it contains a number of conceptual elements and proofs which call for closer examination.

3. Verbeek's argument

Verbeek builds his argument against consequentialism around the following example. Mom can give an indivisible treat to either Jane or Peter. She has three options: (*A*) letting a coin decide, (*B*) giving the treat to Peter; or (*C*) giving the treat to Jane. She is indifferent between the last two options but prefers the first to the last two. Verbeek claims that "Mom's preferences pose a dilemma from a consequentialist perspective" (2001, p. 184).

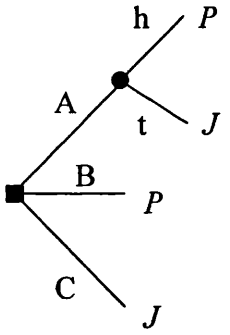


Figure 2.1a

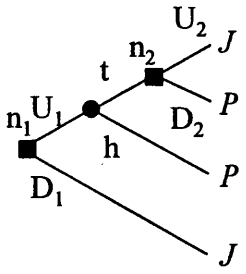


Figure 2.1b

The argument for this claim starts with the prima facie observation that Mom's decision situation can be represented by two decision trees, Figures 2.1a and 2.1b.[32] As in Verbeek (2001), boxes represent Mom's choice and circles the coin flip, resulting in heads (*h*) or tails (*t*) with fifty-fifty odds. *J*, of course, represents 'Jane gets the treat' while *P* represents 'Peter gets the treat.' In Figure 2.1b, Mom gets to choose two times; she can play up or down at the first node (i.e. U_1 or D_1) and up or down at the second node (i.e. U_2 or D_2). But strategically, it could be argued, it does not matter whether

[32] There are slight notational deviations from Verbeek's figures.

Mom's decisions are spread over two nodes. In fact, Figures 2.1a and 2.1b share the same normal form representation, so long as Mom's strategies *A*, *B*, and *C* in Figure 2.1a are identified with Mom's strategies $\{U_1, U_2\}$, $\{U_1, D_2\}$, and $\{D_1\}$, respectively, in Figure 2.1b.

The crucial question is, of course, whether saying that both trees represent Mom's problem equally, is to ignore Mom's preference for fairness by contingency. But let me follow Verbeek's line of reasoning.

The dilemma

The alleged dilemma consists of the impossibility for Mom of meeting two core axioms of decision theory (and of consequentialism, according to Verbeek), NEC and SEP. Here are Verbeek's definitions:[33]

> "*Normal-form/extensive-form coincidence* (NEC): Let *T* be any decision tree with associated set of plans *S* and let T^n be the normal-form representation of *T*, then for any $s \in S$, *s* is acceptable in *T* if and only if it is acceptable in T^n." (p. 186)

> "*Separability* (SEP): Let *T* be a tree and $T(n_i)$ be a separate tree, identical to the tree continuation of *T* from n_i onward. Let $s(n_i)$ stand for the plan continuation in *T* from n_i on. Every plan available in *T* from n_i on is also available in $T(n_i)$. Consequentialism requires that $s(n_i)$ is an acceptable plan continuation if and only if its corresponding plan in $T(n_i)$ is also acceptable." (p. 187)

NEC captures the intuition that time *per se* does not matter in forming rational strategies, only information. So under the assumption of perfect information, the normal and the extensive form (tree) representations should be strategically equivalent. As a biconditional, the NEC warrants that the same holds for two trees which share the same normal form, as do the trees in Figures 2.1a and 2.1b. SEP requires that plans are acceptable at a given node (for the rational chooser) if they remain rational at every decision node onward. It preserves the credibility of plans.

---

[33] The following quotes are Verbeek's reconstructions of McClennen's (1990) definitions which in turn root in interpretations of Hammond (1988). The concept of acceptability is defined very broadly by McClennen: "Let the acceptable set *D(X)* of any set *X* be that subset of *X* consisting of just those alternatives in *X* that are judged acceptable by whatever criteria are being employed." (1988, p. 39). Hammond specifies these criteria as "consistent behaviour norms" (e.g. 1988, pp. 27, 34). Note that whatever constitutes acceptability in Hammond, it cannot make reference to principles first to be derived by consequentialism, like expected utility.

Now assume Mom satisfies SEP in Figure 2.1b. If she does, she should consider both U_2 and D_2 at n_2 given her indifference between J and P. But this would violate NEC, because there would no longer be a binding plan U_1U_2 which is equivalent to strategy A (according to which there is no option of Peter getting the treat once the coin lands tails-up). There is a difference between what counts as acceptable plans in Figures 2.1a and 2.1b.

Conversely, assume that Mom satisfies NEC. In Figure 2.1a, the only acceptable plan for her is A, i.e. to go for the lottery, because she has a preference for fairness. But to secure that option in Figure 2.1b Mom has to somehow commit herself to U_2 at n_2. However she has no reason to do so according to SEP. For Verbeek, this constitutes a dilemma.

There is no point in contemplating this alleged proof at length, however, because it rests on an ambivalent assumption about Mom's preferences. Verbeek makes confused assumptions. On the one hand, Mom is implicitly assumed to be indifferent between all the final outcomes in Figures 2.1a and 2.1b, when assuming strategic equivalence between 1a and 1b. On the other hand, she is supposed to strictly prefer either J or P as lottery results over any non-randomly determined outcome when it comes to assessing acceptable outcomes. In other words, the 'dilemma' disappears once we take into account that the outcomes in Figure 2.1a are not all sufficiently described by 'Jane obtains the treat' or 'Peter obtains the treat.'

The trilemma

But to be fair, Verbeek himself argues that the best response to the alleged dilemma is to introduce a notion taken from Broome.[34]

> *"Principle of individuation by [rational] justifiers* (IRJ): Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them."

Its motivation is the following. What is needed to grasp what is really going on in Mom's decision problem is some representation of her preference for fairness *qua* lottery. This can be done by re-individuating the outcomes, so as to indicate in the label of an outcome whether or not it was reached by a fair method. As wanting a lottery to

[34] Compare Broome (1991a), p. 103. This notion was adopted by Verbeek (2001), p. 191.

decide over the distribution is a reason that Mom can bring forward to motivate a finer individuation of outcomes, this re-individuation is in accordance with Broome's principle.[35]

But Verbeek believes that introducing this principle of individuation (IRJ) raises a new problem. This time it is an alleged *tri*lemma among NEC, SEP and IRJ. Verbeek claims that Mom cannot satisfy all three principles at once. I should point out again that, if any of the parts of the above argument sound confusing, this might be because they are confused. I will try to argue so in Section 4.

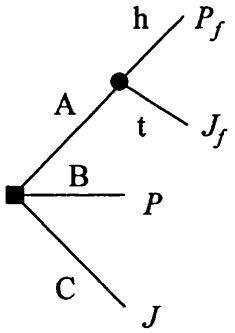


Figure 2.2a

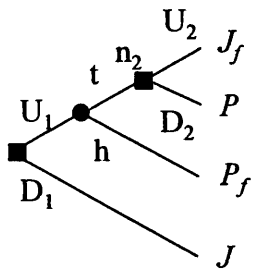


Figure 2.2b

	h	t
A	P_f	J_f
B	P	P
C	J	J

Figure 2.2c

	h	t
U_1U_2	P_f	J_f
U_1D_2	P_f	P
D_1	J	J

Figure 2.2d

Before the alleged trilemma is laid out, consider a new set of figures, 2.2a and 2.2b. The subscript '*f*' indicates the re-individuation 'by a fair procedure.' Note that Figure 2.2b appears in Verbeek (2001).

Assume Mom satisfies NEC and SEP. Next, consider the following argument. P_f and J_f in Figure 2.2b are false or even meaningless expressions, since they falsely indicate that they are outcomes which have been reached by a randomisation over two options. This, however, is not the case unless Mom commits herself to choosing U_2 at n_2, which she cannot do by assumption of SEP. So while Figure 2.2a is meaningful in that sense, we have to give up 2.2b. This leaves us with Figure 2.1b to be compared

[35] Broome originally motivated the introduction of rational justifiers in the context of defending transitivity as a rationality axiom. This issue is independent of matters of fairness, decision under uncertainty etc. Compare Broome (1991b) or Broome (1991a), pp. 100-107.

with 2.2a. But in Figure 2.1b, telling from the outcome descriptions, Mom is indifferent between $\{U_1, U_2\}$, $\{U_1, D_2\}$, and $\{D_1\}$, whereas in Figure 2.2a only plan A is acceptable. So in order to recover NEC while holding SEP fixed, we have to relax IRJ such as to constrain the legitimate representations of the situation to Figures 2.1a and 2.1b. But this is to reject Mom's, and indeed most people's, intuition of fairness.

The next step of Verbeek's 'proof' is invalid: Assume that NEC and IRJ are satisfied. It must then be the case that the lottery outcomes are equally feasible in Figure 2.2b as they are in Figure 2.2a. But for that, again, a commitment to U_2 would be necessary in Figure 2.2b. But that would violate SEP since the agent is indifferent between the outcomes of U_2 and D_2 at n_2.

Assume finally that Mom meets IRJ and SEP. Now we have to give up the belief that the two forms of representation have the same normal form. Verbeek writes: "Mom cannot acknowledge the value of fairness in the dynamic case [as in Figure 2.2b] whereas she can do so in the normal form reduction of that tree [as in Figure 2.1a]." (p. 196). So, Verbeek argues, NEC would be violated.

4. Verbeek's straw man

Verbeek's trilemma only arises because Verbeek's analysis proceeds without a clear order of modelling Mom's choice. The most economic way for me to argue this point is to describe (what I will argue is) the correct order of modelling and show that the trilemma vanishes as a consequence.

The natural figure to start with is Figure 2.2a. (It is revealing that this figure never appears in Verbeek's paper.) Figure 2.2a spells no trouble in terms of the meaningfulness of the outcome definitions relative to the preceding decisions because Mom's intuition is captured by representing fairness with subscripts to the outcome symbols. In that sense, IRJ is assumed to be yielded. This is crucial because Mom's intuitions should be considered. After all, the whole discussion started with Diamond's intuition of fairness.

Having drawn Figure 2.2a, it is legitimate to represent the same decision problem in the normal form. This leads us to Figure 2.2c. Note that the result P_f can be reached only by plan A in conjunction with the coin landing heads-up. Now consider Figure 2.2b. What has it got to do with Figure 2.2a? As it is, not a lot. It certainly does not have the same normal form. Its normal form is Figure 2.2d, where the outcome of

the choice-lottery combination $\{U_1D_2; H\}$ is P_f. So we have no justification to identify plan B with $\{U_1, D_2\}$. NEC simply does not conflict with IRJ (in conjunction with SEP) in the example Verbeek presents. Figure 2.2b no longer poses a problem vis-à-vis Figure 2.2a (qua NEC and SEP) since it has a different normal form.

This observation comes as a relief because I have no intention to contest NEC and SEP, at least not within the current debate. SEP seems to apply simply by construction of decision trees. It just states part of what it *means* to model a decision to happen at a certain node. There seems to be no point to contest it unless we want to introduce a different representation of choice altogether.[36] Furthermore, there is no need to challenge NEC, which I find intuitively appealing. Independent of that, I will later argue that Verbeek's (2001) defence of NEC is flawed for reasons similar to those brought up against his formulations of the trilemma between SEP, NEC and IRJ.

One can now ask, whether Figure 2.2b represents anything meaningful at all, as Verbeek did in the first step of his 'proof' of the trilemma. Is there a proper lottery in Figure 2.2b, i.e. a lottery like that represented in Figure 2.2a, given that n_2 represents a genuine possibility for Mom to give the treat to Peter? And, accordingly, can there be fair outcomes in Figure 2.2b in the same sense as in Figure 2.2a? I will argue in Section 6 that fairness in Diamond's sense is a modal property of outcomes which relies on counterfactual propositions about the outcome. An outcome can only be fair, in that sense, if a random process (which generates events with equal probability etc.) could have determined an alternative, and hence equally fair, outcome. If this is how we approach the semantics of 'fairness,' then Figure 2.2b has to be considered meaningless.

The order of modelling

This leads to another question. How can we avoid producing such meaningless representations of situations involving fairness considerations? This, I believe, can once again be answered by yielding a natural order of modelling the situation. A natural order here means that the first tree to be drawn is that which yields the real (temporal) order of decisions by agents and nature, like Figure 2.2a. In this original representation qualifying outcomes as fair is guaranteed to make sense. But will this hold for any normal-form-equivalent representation? Or does NEC imply that we will have to accept

[36] I agree with Hammond (1988) and Verbeek (2001) in that respect. In contrast, McClennen (1990) tries to resolve intuitions about rational choice in paradoxes of decision theory and game theory by relaxing SEP. I find that this results in an ambiguity of extensive representations: choice nodes do and do not represent choices. I will say more on this in Chapter 5.

some meaningless trees? I believe that the answer is 'no', at least for the fairness property discussed here.

Consider Figure 2.2c. Here there is a row of outcomes which contains only fair outcomes. Also, there is no row which contains both fair outcomes and non-fair outcomes. But then there can be no move by Mom which ultimately yields a fair outcome when the coin lands on one side and a non-fair outcome when it lands on the other side. So in any extensive form representation of Figure 2.2c in which Mom has a choice subsequent to the coin-toss two things hold. Firstly, at least one of her moves must have fair outcomes. Otherwise there would be no row in the normal form that does.

Secondly, if a move of Mom's yields a fair outcome when the coin lands one way, it also yields a fair outcome when the coin lands the other way. Otherwise there would be a row in the normal form that contains both fair outcomes and non-fair outcomes.

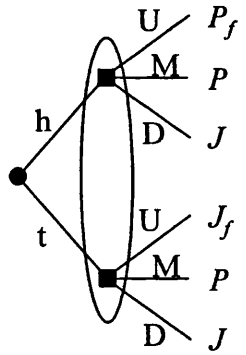


Figure 2.3

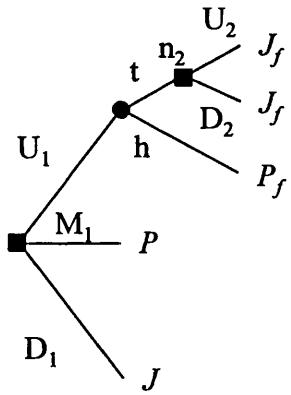


Figure 2.4a

	h	t
U_1U_2	P_f	J_f
U_1D_2	P_f	J_f
M_1	P	P
D_1	J	J

Figure 2.4b

So meaningful trees containing fair lotteries, and hence outcomes that are meaningfully labelled 'fair,' will only yield normal-form-equivalent trees that also contain fair lotteries and outcomes which are meaningfully labelled 'fair.' Figure 2.2a, for instance, could yield Figure 2.3, which has the normal of Figure 2.2c except for the names of the strategies. Since Mom cannot distinguish between the events 'coin lands heads' and 'coin lands tails,' playing up for her is like committing to the lottery outcome. These outcomes can therefore meaningfully be labelled 'fair.' If one reads

normal-form-equivalence in terms of reduced normal forms,[37] Figure 2.4a also counts as normal-form-equivalent to Figure 2.2a, because Figure 2.4b, Figure 2.4a's normal-form, can be reduced to Figure 2.2c. In Figure 2.4a, Mom has a pseudo-choice when the coin lands heads. Again, Diamond's fairness-intuition is preserved and the outcome descriptions of Figure 2.4a are meaningful.

5. Lessons from the maximin regret chooser?

Verbeek's argument for the alleged trilemma builds on a defence of its three underlying assumptions. As a defence of NEC, Verbeek presents the example of a so-called maximin chooser.[38] I want to argue here that, independent of the critique of the alleged trilemma, this particular defence is incompatible with Verbeek's overall argument, as it violates IRJ. Verbeek's argument regarding the maximin chooser could, however, be transformed into a defence of a substantive version of IRJ. Let me explain Verbeek's argument first.

	E_1	E_2	E_3
A	\$10	\$5	\$4
B	\$2	\$9	\$5
C	\$8	\$0	\$10

Figure 2.5

Regret	E_1	E_2	E_3	Max
A	0	4	6	6
B	8	0	5	8
C	2	9	0	9

Figure 2.6

Regret	E_1	E_2	E_3	Max
A	0	4	1	4
B	8	0	0	8

Figure 2.7a

Regret	E_1	E_2	E_3	Max
B	6	0	5	6
C	0	9	0	9

Figure 2.7b

Regret	E_1	E_2	E_3	Max
A	0	0	6	6
C	2	5	0	5

Figure 2.7c

Suppose there are three lotteries, *A*, *B* and *C*, with undefined probabilities attached to three states of the world, called E_1, E_2 and E_3. These lotteries are depicted in Figure 2.5. The maximin chooser (henceforth 'Maxi') always chooses as follows. He

[37] The reduced normal form of a game represents payoff-equivalent strategies as one strategy. Compare, e.g., Myerson (1991), pp.54 ff.

[38] Prior to the defence of NEC, Verbeek explains an argument by Levi (1991) as an example for an argument in favour of the relaxation of NEC. I will not enter the debate, since I have argued that there is no trilemma and hence, without further argument, no pressure on the NEC principle. Also, my argument in the current section can be made without referring to Levi.

considers only the lotteries which are presented to him at the given point of time. Maxi calculates his regret as the difference of the payoffs *within* each given state of the world between all lotteries presented to him. Next he identifies the maximum regret that might occur for each lottery. He picks the lottery for which this number is the smallest.

When choosing between *A*, *B* and *C* at once, Maxi's decides in favour of *A* as can be read in Figure 2.6. But when the decisions are between *A* and *B* first, *B* and *C* second, and *C* and *A* last, Maxi can no longer identify an optimal choice for this sequence of decisions (Figures 2.7a-c). In fact, Maxi's choices would be circular. Verbeek argues that such impossibility of maximisation is unacceptable. He wants to change the ways of the maximin chooser by instructing him with NEC. This allegedly makes NEC indispensable. But how could NEC put Maxi on the right path in the first place?

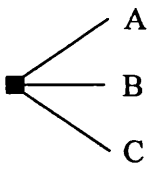


Figure 2.8a

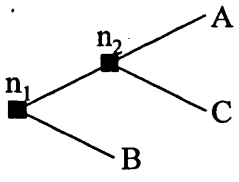


Figure 2.8b

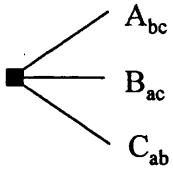


Figure 2.9a

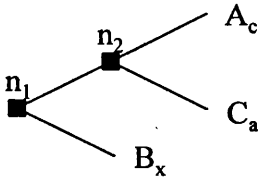


Figure 2.9b

Verbeek argues with two trees, Figures 2.8a and 2.8b. They represent the simultaneous occurrence of all three lottery-options and a choice sequence between the options respectively. The point is that, if one could convince Maxi to accept NEC, he could no longer be a maximin chooser, since he would have to come up with the same optimal choice in normal-form-equivalent games such as Figures 2.8a and 2.8b. But, as far as backward induction applies, Figure 2.8a has the result *A* while Figure 2.8b leads to *B*. In other words (to stay closer to Verbeek's wording), if we relax NEC, Maxi would not be a maximiser in the sense just outlined.

The maximisation requirement is not further motivated in Verbeek (2001). But it could be driven by a pragmatist intuition. Possibly, Verbeek wants to save Maxi from falling pray to a money-pumping lottery organiser who offers Maxi the chance to enter ever new lotteries for small fees, while always just offering a choice of two at a time.[39] Under certain assumptions, such an organiser could ruin Maxi. I will present a pragmatic argument by referring to Broome (1991b) in a short while.

[39] The order could be: *A* and *B* first, *B* and *C* second, and *C* and *A* third, *A* and *B* fourth etc.

But even when accepting the maximisation requirement, we need to be careful with Verbeek's conclusion that Maxi's example validates NEC. In fact, consider Figures 2.9a and 2.9b. In these representations, the outcomes have been re-individuated such that the immediate alternatives to an option are captured as relevant properties of that same option. Hence A becomes A_{bc} when juxtaposed to alternatives B and C in Figure 2.9a, etc.; and A becomes A_c when juxtaposed to alternative C in Figure 2.9b etc.[40] In contrast to Figures 2.8a and 2.8b, the normal forms of Figures 2.9a and 2.9b are not the same, of course. Hence, even if Maxi does accept NEC, he is not forced to give up on his selection principle as long as he is able to re-individuate his expected outcomes suitably. And it seems that Maxi can do so, according to IRJ. After all, regret is the key emotional factor to Maxi's assessment of monetary payoffs. In that sense, Maxi is not an example of what can 'go wrong,' if NEC is relaxed.

Instead, if the example of Maxi is to be expounded in some way, it could be used to support a substantive extension of the principle regimenting outcome-individuation. Such an extension would forbid the transformations like those of Figures 2.8a/b into Figures 2.9a/b. It is derived from Verbeek's normative requirement of maximisation. In fact, if Verbeek wishes to preserve his intuition about maximisation throughout decision sequences, i.e. situations which are spread out in time, but can be perfectly anticipated, he should do so by making an explicit assumption of what should count as an outcome. The resulting substantive version of IRJ, which will be labelled SIRJ, would have to be similar to the following principle:

> *Substantive principle of individuation by rational justifiers* (SIRJ): Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them. Furthermore, candidates for justifiers that can lead to intransitive choice when presenting the agent with subsets of the full choice set (whose elements are not re-individuated outcomes) sequentially, do not count as rational and are to be excluded.

I shall not be concerned with the question of whether this postulate is watertight in that it does the job it is supposed to. Moreover, the principle might not be a plausible

[40] The subscript x to B_x is a placeholder for whatever alternative is expected as a turnout of n_2. This is not an unproblematic notational feature however. In fact, the modal outcome description B_x is endogenous to an expectation over the outcome of a subsequent decision. This is not very intuitive. Furthermore, Hammond's derivation of expected utility from consequentialism requires outcomes to be given, not derived. This weakens my critique of Verbeek's defence of NEC. It also points at a potential conflict between IRJ and SEP, which is not addressed by Verbeek.

rationality requirement by some standards. It is certainly paternalistic to some extent, as it potentially ignores people's intuitions about what matters in outcomes. It actually comes close to what Hammond calls a "practical normative theory." Finally, I should point out that SIRJ, as opposed to IRJ, is impotent as a means to *check* on transitivity, one of IRJ's original purposes. SIRJ cannot fulfil this function because it simply *demands* transitivity.

All in all, SIRJ as such will be very hard to defend. It was simply introduced to substantiate the argument that Verbeek's maximin argument, while failing as a defence of NEC, can be exploited as a defence of a substantive strengthening of IRJ.

Stealing by offering

Another way to look at Verbeek's concern with Maxi's selection mechanism is to reformulate it as a concern about transitivity of preferences (as a rationality requirement). This would be analogous to a point made by Broome (1991b), who argues that the standpoint of moderate Humeanism is inconsistent. This standpoint is defined as taking preferences of a rational agent as purely exogenous while expecting the rational agent to exhibit transitivity in choices. That argument proceeds by considering a (putatively) rational agent's excuses in terms of outcome re-individuation when confronting him with prima facie intransitive choices of his.

Broome introduces the example of Maurice, who makes the following choices. He chooses staying home (H) when offered a trip to Rome (R) as an alternative; he chooses R when offered a hike in the mountains (M) as an alternative; but he chooses M when the alternative is H. Maurice justifies his action by re-individuating staying home relative to the alternatives at stake. So there are the alternatives H_r, i.e. staying home when the alternative is the trip to Rome, and H_m, i.e. staying home when the alternative is the hike. Alternatives matter here, because if the alternative is mountaineering, Maurice feels like a coward when staying home. This makes it impossible to assess Maurice's transitivity (and hence rationality) on the basis of the described observations only.

Next, Broome's argues that the preference order between H_r and H_m can never be determined by observing choice. Since the order between these two options is crucial for testing transitivity, however, another way to determine it becomes indispensable. Here some degree of deliberation *about* preferences (e.g. in terms of rational justifiers) will be necessary. This, however, violates Humeanism as defined by Broome. Nothing

more shall be said about Broome's main point, because I am mainly interested in how Broome handles the money-pump argument as a challenge to his argument. This is where the analogy with Maxi appears.

The money-pump argument, if it applies to Maurice's case at all, would work as follows: Maurice is irrational in differentiating between H_r and H_m, because he could be exploited when being offered (the outcomes prima facie individuated as) H, M and R – one new option at a time – in circles, demanding a sufficiently small payment from Maurice in return for each swap. So here is Broome's counterargument:

> "Suppose Maurice is right [about cowardice, when offered the alternatives of staying home and the hike]. And suppose he is now planning to stay home, having turned down a trip to Rome. Then, if you come and offer him a mountaineering trip, you are by that very action making him worse off. You are, in effect, moving him from H_r to H_m, which is justifiably lower in his preference ordering. Maurice is willing to buy his way out of this position. It is as though you stole his shirt and then sold it back to him. Rationality cannot protect Maurice from that sharp practice. So the fact that he is susceptible to it is no evidence of irrationality. The money-pump argument fails, therefore." (1991, p. 58)

Projecting this argument onto the example of Maxi, it should first be noted that, given Maxi's selection mechanism which evaluates options as dependent on alternatives, he is equally susceptible to the kind of 'stealing by offering' as is Maurice. That means that what Maxi possesses in terms of regret-value at one stage can be taken away from him or, at least, diminished in value simply by offering him additional options.

But while Broome's argument concerning the money-pump can be extended to the case of Maxi, I am not sure whether this kind of reasoning defeats the application of the money-pump argument in either Maurice's case or Maxi's case. *If* we accept the authority of the money-pump argument with respect to more straightforward cases of intransitivity, i.e. cases for which context-dependence is harder to argue for, we might just as well have to accept it in the cases of Maxi and Maurice. After all, the money-pump is a *pragmatic* argument. And I believe that there is a substantial implicit premise in the pragmatic perspective, according to which emotions such as regret should be disregarded if they cause potentially unlimited material loss.[41] So far as we endorse this

[41] Strictly speaking, the loss need not be material. It just needs to be loss of something of which more is always better.

hidden premise, Maxi and Maurice might just have to bite the bullet that the way they form preferences makes them susceptible to the money-pump argument. Whether that is enough to dub them irrational, I will not try to answer here.

Instead, I would like to return to the case of Mom. In what follows I would like to explore why it will be hard to capture her decision problem in terms of expected utility.

6. Expected utility theory

In Section 4, I asked how far we could talk of a proper lottery in Figure 2.2b (or 2.2d for that matter) given that n_2 represents a possibility for Mom to determine with certainty that Peter gets the treat. It seems questionable that calling the outcomes of $\{U_1U_2; H\}$ (or $\{U_1D_2; H\}$) and $\{U_1U_2; T\}$ fair is meaningful, given what Mom means by 'fair.' This is so because fairness in Mom's sense can only be attributed to outcomes whose alternative is also fair. In addition, both alternatives must be directly subject to chance. Let me rephrase this point once again.

Fairness as a modal property

Fairness is a modal property. It depends on what would have been the case, had the coin landed the opposite way to how it *actually* did. It depends on what is true for certain other possible worlds; more precisely, it depends on what would be true for possible worlds, in which there was also a preceding toss of a coin, but in which the coin landed on the other side. If the allocation of the treat in those worlds subsequent to the coin-toss is not the opposite of the allocation in the actual world subsequent to the coin-toss, the word 'fair' seems misplaced.[42] But Mom's choice in n_2 of Figure 2.2b (between Jane and Peter obtaining the treat), is not the opposite of Peter obtaining the treat in the modal sense that was just outlined.

Earlier I argued that this problem could not contribute to the alleged trilemma, because there is no such trilemma. But the problem should cause Verbeek worry elsewhere, namely when it comes to assigning expected utility. In fact, the modal character of some outcome descriptions leads to conceptual conflicts with an assumption which is made in certain expected-utility theories.

[42] I endorse Broome's (1991a) notion here.

The rectangular field assumption

Savage's (1954) decision theory builds on the so-called the "rectangular field assumption," henceforth RFA.[43] The RFA is a necessary condition for a certain separability theorem[44] to hold. This theorem in turn is a presumption of the indispensable axiom (P1) of Savage's expected utility theory, which postulates that there is a simple ordering between all so-called acts. Remember that the core of this theory is a representation theorem. This means that, unless RFA is fulfilled, we are back to a situation in which we cannot assign utility figures to decision problems under uncertainty.

This is true, of course, only as long as we assume that Savage's expected-utility theory is the only one we have. This is not the case – not even for decision theory, as I will point out later. In the case of game theory, the commonly assumed expected-utility is another one altogether, anyway; namely that of von Neumann and Morgenstern (1944). It uses objective probabilities, which, as I discussed in Chapter 1, is crucial for standard solution concepts. However, the objective-probability assumption is often considered to be hard to defend. Hammond's expected-utility theory, as discussed earlier, also incorporates the RFA, despite the author's explicit recognition of its weaknesses.[45] I shall not attempt to judge between different expected-utility theories here. The fact of the matter is that we will need *one* such theory if probabilities are to play a role in our decision theory or game theory. In that sense RFA remains a *potential* problem.

The RFA only makes sense in terms of Savage's distinction between states, acts and consequences, as discussed in Chapter 1, Section 6. Broome (1991a) explains the assumption:

> "Take the set of all outcomes: the set of all the possible outcomes that any prospect may lead to. Now let us go through the states of nature one by one, and to each assign, quite arbitrarily, some outcome of this set. This operation will define an arbitrary prospect: the prospect that delivers, in each state, the outcome we have assigned to that state. The assumption says

[43] The term was introduced by Broome (1991a).

[44] The theorem states that an ordering is strongly separable if and only if it is additively separable. A proof is given by Debreu (1959). Strong separability means that all subsets of the arguments of the ordering are separable in the utility representation. Additive separability means that the utility of all arguments of the utility function is a sum of utilities over each separate argument. Compare, e.g., Broome (1991a), pp. 69-70.

[45] Compare Hammond (1988), pp. 30 and 62-66.

that *any* arbitrary prospect constructed this way has a place in the preference ordering." (p.115, italics in original)

Again, being a prospect implies having a position in the preference ordering of the agent whose preferences are eventually to be represented in form of utilities. I will soon point out what the assumption means for Mom's problem.

So does RFA hold for Mom's problem? I have already addressed the intuition to answer that question above. Mom would have to be able, for instance, to:

(a) imagine or grasp the meaning of the prospect that Peter receives the treat in a fair manner if the coin lands heads-up and that he obtains it in a non-fair way if the coin lands tails-up, as well as

(b) form a relative preference for this prospect.[46]

Step (a) will be difficult, to say the least, given that it would mean that Mom would have to imagine that the lottery happened and did not happen at the same time. (Note the similarity of this point with Hausman's (2005b) argument that certain choices could not even occur hypothetically, as discussed in Section 5 of Chapter 1.) But even if Mom could somehow invoke the mental picture, it is still not clear that it would be the kind of thing that she could rank relative to other prospects.

So what can we do to resolve this problem? Should we simply come up with a substantial principle of outcome-individuation once again, so as to exclude all re-individuations which jeopardise RFA?[47] The answer must be 'no'. While the aforementioned SIRJ appears decidedly paternalistic, it can at least make reference to the pragmatic money-pump argument to support it as a normative principle. This time there is no such normative backdrop at all. Excluding problems with the RFA in the form of a restriction to the outcome individuation would have no intuitive justification, not even a contentious one.

Instead, there are different ways to react to the problem of the RFA vis-à-vis modal outcome properties as just described. Firstly, we could simply bracket the problem of justifying expected utility theory and move on under the presumption that

[46] Compare Broome (1991a), p. 116, and Valentyne (1993), pp. 28-29, for a similar discussion of Diamond's example.

[47] Sugden (1991) seems to suggests that Savage might be read as restricting the realm of what "counts as a consequence" (p. 763). However, he goes on to argue that since that would exclude plausible intuitions about outcomes, we face a dilemma between accepting Savage's axioms and accepting transitivity as a rationality requirement.

some expected-utility theory, other than Savage's, can be vindicated; von Neumann and Morgenstern's approach might be a candidate. Secondly, we could restrict our formal analysis to problems of certainty while treating entire lotteries as outcomes, as was done in the example of the minimax regret chooser. Thirdly, we could make an effort to found an expected utility theory on a set of axioms which does not include RFA.

The first way seems openly presumptuous, yet not a priori unacceptable. The second way would limit the scope of consequentialist analysis enormously. While it would not *exclude* the analysis of situations involving risk, consequentialism would no longer *permeate* the decision under risk. It would restrain the analysis to that part of the situation which includes lotteries as certain outcomes. This also implies that we would have to presume that there is some way to determine some ordinal utility measure for such complex lotteries-outcomes. This, in turn, seems a questionable starting point given that Maxi's example showed that intransitive preferences over lotteries are intuitively explicable. The third option poses a new challenge, which I will address next.

Dispensing with the assumption

Some authors have taken on this challenge, if in a different context. In addressing Harsanyi's attempt of founding utilitarianism axiomatically, Broome (1990) and Bradley (2005) have adopted an axiomatic approach to expected utility theory which works without the RFA. This approach, as introduced by Bolker (1966) and Jeffrey (1983), has some interesting features. For instance, it does not identify propositions that (exclusively) play the role of outcomes or consequences. Instead, probabilities and preferences are defined on a single set of propositions to which a complete, *atomless* Boolean algebra without the necessarily false proposition is applied. The utility of any prospect, which is a set of possible worlds, can therefore be grasped as the disjunction of sub-prospects (consisting of complementary subsets of the set of possible worlds which described the original prospect). Therefore the utility of any prospect can be represented as the sum of the utilities of its sub-prospects weighed by their conditional probabilities. The Bolker-Jeffrey approach thereby avoids setting up causally impossible gambling scenarios, over which preferences must be formed in Savage's approach.

Sketching this possible escape route from Savage's axiomatic system, serves to put the problem with the RFA in relation. However, the Bolker-Jeffrey approach, might

not take us very far. Causal decision theorists have raised caveats.[48] In the so-called 'twin prisoner's dilemma' (where 'twins' are perfectly symmetrical agents), e.g., the probability updating method by Bolker-Jeffrey supports the strictly dominated (cooperative) equilibrium. This is because the first agent's cooperation, e.g., might be taken as a *cause* for the second agent to cooperate. This is a mistaken interpretation of the observation that the probability of the second agent's cooperative response would have to be updated after the first agent cooperates.[49] While such an updating might seem adequate from the perspective of a third person, who collects *evidence* for what the first player is likely to do after observing the second, the agents cannot plausibly read this probabilistic relation as information about their causal powers. The probabilities cannot inform what they have reason to rationally do in the dilemma.

The fact that the Bolker-Jeffrey framework invites such confusion might not be a problem for the "valuational interpretation" (Broome 1990, p. 106) in discussing Harsanyi's social choice theory/political philosophy. But it seems to pose a serious problem for decision theory, hence, for the kind of decision problems raised by Verbeek. I cut off the discussion here, concluding only that an alternative to Savage's framework for motivating utility figures in the cases discussed might not be easily found.[50]

7. Conclusion

The narrow conclusion of this chapter is that Verbeek's adaptation of Diamond's example of fairness-qua-lottery failed to prove that consequentialism is an inconsistent notion. It is not the case that three of its basic principles form a trilemma. In fact, if the principle of rational justifiers is taken seriously all the way through in Verbeek, the alleged conflict between the normal-form/extensive-form coincidence condition (NEC) and the separability condition (SEP) does not even arise. This teaches us a lesson about the order of modelling decisions. Outcomes need to be fully specified – including their process-dependent properties – before conclusions are being drawn from normal-form equivalence of extensive forms.

---

[48] For a good overview compare the section entitled 'Decision versus valuation' in Broome (1990), pp. 486 ff.

[49] Compare 'The symmetry fallacy and counterfactual reasoning' in Chapter 5, Section 5.

[50] Prima facie, von Neumann and Morgenstern (1944) offer an axiomatic approach to expected utility that works without RFA. On second sight, however, they simply avoid the problem by assuming (the plausibility of) a space of lotteries over outcomes with objective probabilities over outcomes. They also simply assume that agents have preferences over these lotteries.

Since Verbeek's attack on consequentialism fails, Hammond's project is not subject to the kind of doubt that Verbeek raises. Verbeek has not shown that consequentialism could *not* serve as "an unproblematic and minimal requirement of rationality." (2001, p. 182) But, of course, this does not prove that it could. In fact, some authors challenge the plausibility of one of the basic axioms directly. McClennen (1990), e.g., attempts to undermine the authority of SEP, Levi (1991) criticises the status of NEC in consequentialism. I did not need to look into these critiques in greater depth here.

However, I did address two recalcitrant concerns regarding process-dependent outcomes as such. The first is the pragmatic concern with transitivity. While agents with process-dependent preferences might theoretically fulfil transitivity as long as their outcomes are re-individuated adequately, they might still be susceptible to practises which I subsumed under 'stealing by offering.' They might be made worse off simply by being given additional choices. This can only be avoided by introducing a substantial restriction to outcome-individuation.

The second concern with process-dependent properties of outcomes is with the rectangular field assumption. It might spell trouble for attempts to support expected-utility theories building on a distinction between states of the world, acts and consequences, like Savage's (1954) theory. Doubts were raised also whether the Bolker-Jeffrey approach provides a remedy.

Chapter 3

Psychological games and fairness equilibria

1. Introduction

Applications of game theory usually start with the specification of a game, be it in normal or extensive form. Abstracting from the potential problems of indeterminacy in the equilibrium analysis, payoffs are understood to comprise all information necessary to determine the rational moves for a given player. Payoffs are the sole motives for his actions in that sense. This was discussed in Chapter 1. Accordingly, game theorists usually point out that, say, non-material, other-regarding or process-regarding aspects of preferences are all taken to be encapsulated in the payoffs.

Chapter 2 addressed problems that arise when specifying outcomes (prior to deriving utilities) that are dependent on the process of play. The current chapter considers game theoretical approaches that 'skip' the conceptual backdrop of outcomes *descriptions* and incorporate process-dependence immediately into the utility functions. More specifically, these approaches try to model utilities as dependent on the beliefs that players form over the course of the game.[51]

Geanakoplos, Pearce and Stacchetti (1989), henceforth GPS, were the first to derive such "psychological games" with belief-dependent payoffs from "material games." Rabin (1993) carried GPS's idea of psychological games and the corresponding solution concept of psychological Nash equilibrium forward by introducing the notion of "fairness equilibrium." It specifically takes into account other-regarding beliefs. According to the underlying model, each agent's beliefs about what the other agents will choose, and what he believes the others expect him to choose etc., will impact on the fully defined payoffs. A great part of my discussion will be a critique of these approaches.

My discussion does not provide a final assessment of the threat for game theory posed by problems with psychological games and equilibria. This is because my critique falls into two great parts and challenges both the immanent conceptual viability of equilibrium in psychological games as well as the intuitive support of belief-dependent

[51] This need not - and sometimes cannot - be interpreted temporally, as will be discussed.

utility (and hence the necessity to model it). The immanent conceptual critique concerns mainly the circular relation between beliefs and choices in psychological games. The intuitive critique starts by questioning whether the intuitions listed in support of modelling utilities as belief-dependent are compatible with the conceptual baggage of modern utility and rational choice theory which the theory of psychological games builds upon.

The plan of the chapter is as follows. Section 2 relates the debate to follow with a general discussion about face-value forms of games by contrasting it to the final forms of games and corresponding notions of consequentialism. Section 3 introduces the general approach of psychological games as introduced by GPS. Section 4 introduces Rabin's model of fairness equilibria as a special case of psychological game analysis. Section 5 raises immanent concerns about the concept of psychological equilibria. I scrutinise the hypothetical reasoning supporting such equilibria. Section 6 focuses on the question of whether psychological equilibrium and fairness equilibrium are concepts that capture the intuitions which first motivated process-dependent outcome definitions in games. Section 7 considers an amendment to Rabin's approach as suggested by Hargreaves Heap and Varoufakis (2004), henceforth HHV, which aims to introduce a notion of intention to psychological game analysis.

2. Payoffs and consequentialism

The applicability of game theory as an analytical tool – be it predictive, descriptive or prescriptive in purpose – depends on whether the analyst can somehow pin down the game which was or will be played. One of the most delicate parts of this definitional process is to identify the effective payoffs at hand. In Chapter 1 it was argued that such utility-payoffs need to represent all motives that matter in a game. This leaves the, arguably, most difficult task of applying game theory at the modelling stage at which the theorist has to identify the game that is *really* played. The task is so difficult, because, among other things, the theorist has to incorporate in the payoffs all aspects concerning the process of play which matter to the players. That is unless we either assume that processes, or actions for that matter, do not matter or should not matter to agents. But before I return to this point concerning the underlying notion of consequentialism, let me explain a useful distinction between two kinds of outcomes and payoffs.

Culmination and comprehensive payoffs

In the introduction to his discussion of the relation between of acts of choice and utility maximisation, Sen (1997) offers a helpful terminology. He distinguishes between "preferences over *comprehensive* outcomes (including the choice process)" and "conditional preferences over *culmination* outcomes *given* acts of choice" (p. 745, italics in original). Culmination outcomes, in this sense, blend out aspects of the preceding choice process. For a given level of description, they capture all significant results of an action as the outcome, but no aspect of the action itself. Comprehensive outcomes, in contrast, also capture all relevant aspects of the action itself as part of the outcome.

This distinction is helpful for my purposes. It can be transferred to payoffs, distinguishing between culmination payoffs and comprehensive payoffs. This distinction is important because the works on psychological games start from the premise that only comprehensive payoffs capture all intuitively relevant motives for agents. Other authors, like Hausman (2005a), hold that there is (or, rather, that there has to be) a tacit premise in game theory that only culmination payoffs are a suitable interpretation of utilities in games. I will come to this point next.

However, for the practical reason of making my arguments recognisable for the ongoing discussion, I adopt the rather misleading distinction between 'material payoffs' and 'psychological payoffs,' as e.g. in Rabin (1993). Strictly speaking, however, I thereby continue an equivocation. Calling culmination payoffs 'material payoffs' and comprehensive payoffs 'psychological payoffs' does not do justice to Sen's distinction. The crucial distinction is between measures for preference which do or do not take process-dependent aspects into account, not between material and psychological measures.

Formal and substantial consequentialism

Game theory is clearly a consequentialist theory, as its solution concepts make reference to the utility (the consequence of players' choices) as the only source of motivation. But what this trivial statement really implies depends on how we grasp consequentialism. Hausman (2005a) offers the following definition:

> "[By 'consequentialism'] I mean [...] that an agent's choices and preferences among actions depend exclusively on (a) constraints, (b) the

> agent's beliefs about the consequences of the alternative actions the agent is aware of, and (c) the agent's evaluation of these consequences. Consequentialism denies that actions have some other intrinsic value or that they are governed by principles that are not concerned with the consequences of the actions." (p. 2)

Hausman distinguishes his definition from Peter Hammond's (1988) consequentialism as a rationality criterion. The difference is that for him consequentialism presupposes that choices depend (predictively, explanatorily) on an evaluation of outcomes whereas in Hammond utilities are *assigned* to consistent choice. This need not imply a deeper conflict, because Hammond is after a representation theorem for expected utility theory, while Hausman is after a conceptual foundation for applying game theory.

But Hausman's definition can be characterised in another important way. It is a definition of a *substantive* consequentialism that denies "that actions have some other intrinsic value." Hausman believes that for game theory to be applicable, a "default principle" has to hold according to which preferences exclusively depend on what has been characterised as culmination outcomes above.[52] He explains that this does not exclude the possibility of other-regarding consequentialism, but that it does exclude, e.g., the possibility of reciprocal altruism as captured in process-dependent outcome properties. Hausman continues that if the default assumption fails game theory is "consequentialist only in form" (p. 9). It would still work as a formal construction, but supposedly fail to be applicable for explanation or prediction.

So, crucially, we can identify another form of consequentialism, namely *formal* consequentialism. This version does not make any normative claim about what should and what should not matter to players. It settles for the much weaker requirement that, once all relevant motivational aspects of players choices have been captured in the payoffs, the payoffs are all that is to be regarded.[53] But the interesting question is, of course, whether all motivational aspects can be so captured.

Hausman seems to have doubts in this matter, as he argues that a failure of the default principle (which leaves the theorist only with formal consequentialism) entails problems for the game theorist. But he is not sufficiently clear about what the problems

[52] To be sure, Hausman does not deny that grasping strategic situations in terms of comprehensive games might resolve some intuitive problems we have with 'paradoxical' equilibria in games like the prisoner's dilemma or the centipede game as defined in monetary payoffs, e.g. He states that players might construct the real games (whose equilibria determine the actual play under the usual assumptions) by envisioning comprehensive outcomes. But then he expresses doubts whether rationality any longer binds players in such constructed games. Compare Hausman (2005a), p. 16.

[53] Note that an inverse version of this is held by Hammond (1988), p. 26. Compare also Chapter 2.

are. Hausman identifies three problems. The first is epistemic. The game theorist cannot know what matters to players, unless it is an observable outcome. The second answer refers to complexity. It is supposed to be particularly pressing for games with incomplete information: "Sometimes preferences over outcomes are so heavily dependent on features of the strategic interaction and in such a complicated way, that the players can scarcely be said to be playing a *game* at all." (p.14, italics in original) So Hausman implies that sometimes the players cannot derive comprehensive utilities from the situation defined in terms of culmination outcomes due to cognitive limitations. Both problems concern the applicability of game theory, which shall not be my concern here.

The third problem, however, seems to hint at a problem with formal consequentialism *in principle*. The point seems to be that there is some kind of problematic regress, in that "game theory may often derive players' rankings of strategies from preferences over outcomes that derive from their rankings of strategies" (p. 10). Unfortunately, Hausman does not elaborate on this idea. However, I will develop a critique against psychological games along these lines later.

For the moment, let me just note that one important question at stake in the debate about the viability of psychological games and equilibria is whether formal consequentialism without a substantive normative "default principle" is sustainable.

3. Psychological games and equilibria

GPS (1989) introduce the concept of "psychological games." In a psychological game, payoffs depend on belief, beliefs about beliefs, and possibly even beliefs of a higher order. Payoffs are endogenous with respect to beliefs in that sense. As GPS explain, they are not always unique for a given "material game".[54] The idea behind this partial alteration and extension of game theoretical analysis is to capture agents' motives which are dependent on expectations (like "surprise, confidence, gratitude, disappointment, embarrassment, and so on", p. 61) and on what is learnt over the course of a game. In that sense, GPS do not leave the process of payoff specification as a black box

[54] The term "material game" is actually only introduced by Rabin (1993). But it is helpful to use it here already in order to denote that structure which, under the consideration of beliefs, gives rise to psychological games. As said earlier though, the material-vs-psychological contrast in the terminology misses the point. For instance, there might be belief-independent payoffs which are non-material, etc.

presuming that all that matters to agents is somehow given in terms of exogenous payoffs. GPS write:

> "A player's emotional reactions cannot in general be independent of his expectations and of his interpretation of what he learns in a play of a game. Hence, we argue that in many cases the psychological payoffs associated with a terminal node are endogenous, in the same sense as equilibrium strategies are. Indeed, in some examples, no single set of payoffs adequately summarizes the strategic situation." (p. 61)

This section is mainly concerned with the construction of GPS's approach. It will not be further discussed here whether modelling belief-dependence is, even on a merely intuitive analysis, a good way or the only way to capture sentiments like surprise, confidence, gratitude, disappointment, or embarrassment. But we will come back to these questions in the discussion of fairness equilibrium which builds on the concept of psychological equilibrium.

Construction and results

GPS mention learning in their motivation for psychological games. Interestingly enough, GPS provide an analysis in which learning plays no role. In fact, their analysis of solutions to psychological games offers re-definitions of established solution concepts for (extensive form) psychological games in which only beliefs at the beginning of a game matter. The authors merely express optimism concerning the possibility of modelling psychological games as dependent on (dynamically) changing beliefs.[55] In that sense, extensive form representations effectively play a shallow role in the analysis of solution concepts in GPS (1989).

But before GPS even come to the analysis of extensive games, they reduce complexity in the analysis of normal form psychological games in three very general ways to allow for Nash equilibrium to be applicable (at least) on the formal level. These formal requirements need to be introduced in order to make sense of any example of a "psychological Nash equilibrium" (p. 65). But we must also keep in mind that the formal viability of a concept does not entail its conceptual viability.

First of all, GPS exclude the possibility of inconsistencies between beliefs of different orders by assuming coherency over belief hierarchies. Secondly, they assume that in equilibrium all beliefs conform to "some commonly held view of reality." (p. 65)

[55] Compare GPS (1989), pp. 70 and 71.

This way they avoid the potential problem that the beliefs which define the payoffs are not consistent with the beliefs that derive from those payoffs in the equilibrium. Thirdly, they drastically simplify the payoff-belief relation in equilibrium and certain structural assumptions about belief (hierarchy) sets. These assumptions are used to define psychological Nash equilibria and to prove their existence for normal form games.

Under these assumptions and some additional ones, the results for strategic form equilibria also apply to the extensive form. In addition, the refinements of subgame perfection (and trembling hand equilibrium respectively) as well as sequential equilibrium can be defined.

As for positive results, GPS prove the existence of subgame perfect and sequential equilibria for psychological games in extensive form. As for negative results they make the case that the analogue to (trembling hand) perfect equilibria need not exist for psychological games. Also, they explain that backward induction cannot be applied even under the usual assumptions. I will return to the significance of these findings vis-à-vis the conceptual status of psychological Nash equilibria later.

Some examples

Let me now consider two examples of psychological games and equilibria by GPS (1989) to add some 'flesh' to the discussion. First, let me consider the atomic example of a psychological game in Figure 3.1, to illustrate GPS's idea. Here, only one player acts, but a second player (or group of players) forms beliefs about player 1's choice which in turn impacts on 1's outcomes. In that sense the situation has some interactive dimension. The question is whether 1 acts boldly or timidly. The probabilities over these choices are p and $1\text{-}p$, respectively. Player 2's estimate of p is p', which appears in 2's payoffs. Player 1's estimate of p' is p''. So player 1 cares about what the other expects him to do.

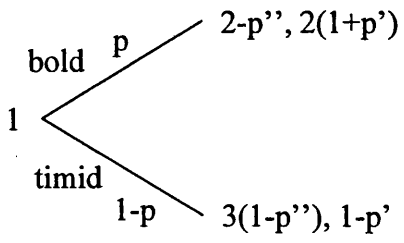


Figure 3.1

The possible psychological Nash equilibria of the game, in which all beliefs of all orders conform to reality, are $p=p'=p''=1$ yielding payoffs (1, 4), $p=p'=p''=0$ yielding payoffs (3, 1), and $p=p'=p''=1/2$ yielding payoffs (3/2, 7/4).[56] GPS suggest the following interpretation of these outcomes:

> "Player 1 is best off when his friends expect little, but if their expectations are high he is trapped into meeting them, and then in a vicious cycle (for player 1) they are justified in holding such a lofty opinion of him." (p. 67)

While this might sound plausible at first, I would like to call attention to some peculiar aspects of this example. The first is that multiple Nash equilibria arise, although only one player moves. This new kind of embarrassment of riches will be addressed later. Furthermore, it seems very hard to see why player 1 should be caught in a "vicious circle" in the bold-play equilibrium. After all, there is no significant dynamic to the belief formation in this game.

More generally, it remains unclear how any of the equilibria suggested can be defended in terms of player 1's reasoning. Suppose 1 'finds himself' in the bold-play equilibrium (ignoring for the moment that I cannot spell out what that means). How is he to reason about player 2's beliefs and payoffs in the hypothetical state in which he chooses to act timidly instead of boldly? I do not know. In fact, I believe that the concept of psychological Nash equilibrium is conceptually ill-equipped to give any answer to such questions at all.

To raise a related concern, we can ask who takes action in this game and why. If we assume that player 1 takes the initiative we will have to explain how he could do so rationally. Remember that what is rational for 1 depends on what 2 expects him to do, which in turn coincides with what 1 does in equilibrium. So it seems that 1 will have a hard time deciding what to do. On the other, it seems implausible to assume that player 2 takes the initiative by simply forming an expectation in this game, to which player 1 then tries to conform.[57] I will expand on this criticism later.

---

[56] The equilibria in pure 'strategies' are found by assuming $p=0$ and $p=1$, respectively, and testing for 1's rationality. The equilibrium in mixed 'strategies' are found by assuming that 1 is indifferent between playing timidly and boldly and solving $2\text{-}p''=3(1\text{-}p'')$ for p'' etc.

[57] It is also unclear what the probabilities in the mixed equilibrium represent. This will be discussed more in the next example.

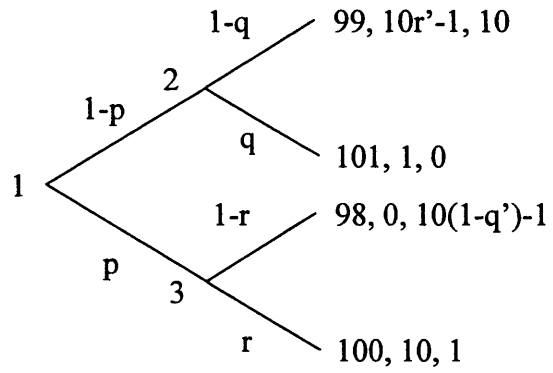


Figure 3.2

Let me first present another example, as depicted in Figure 3.2. In this game, players make their choices by deciding over the probabilities *p*, *q* and *r*, respectively. In addition, *r*' represents player 2's expectation about (3's choice of) *r* and *q*' represents player 3's expectation about *q*.

According to GPS, players 2 and 3 in this game are modelled to develop feelings towards 1's expected 'intention' to lower their payoff. The way in which this is captured is that they both would experience increasing payoffs when playing a strategy which diminishes player 1's payoff, the more they believe that they will end up in a relatively bad outcome, if they do not get to make a move themselves. This, strangely, makes player 2's feelings about wanting to punish player 1 dependent on player 3's move and player 3's payoffs of punishing player 1 dependent on player 2's move. But I do not want to argue how well GPS manage to capture our intuitions about the feeling of revenge.

Instead, I would like to comment on two other aspects of this example. The first is that there is only one equilibrium in this game, namely one in mixed strategies. It consists of $q'=q=4/5$ as well as $r'=r=1/5$ and $p=0$. I am personally completely overtaxed in trying to understand what these probabilities represent; choices by randomisation or subjective expectations or something else altogether? More will be said on this in my critique to follow, concluding that the worry about mixed psychological equilibria is real but not necessarily more pressing than the concern with pure psychological equilibria.

Secondly, I would like to highlight the fact that the two examples, as just discussed, are peculiar in the sense that the only 'psychological' belief dependence

modelled is between players who do not interact strategically in the traditional sense.[58] The players 2 and 3 in Figure 3.2, for instance, have no influence on the other's feasibility of material outcomes. Only player 1 immediately interacts with 2 and 3 from a non-psychological perspective. In Figure 3.1, there is no interaction at all in the material sense. My suspicion is that GPS chose these examples because they allow for a 'neat' calculation of psychological equilibria. In fact, in games in which players interact traditionally, by excluding the feasibility of certain outcomes for each other, *and* via altering psychological outcomes through beliefs, even the algorithm for calculating psychological equilibria can be problematic. This is exemplified by a game introduced by GPS that will be discussed in Section 5 (Figures 3.5a and 3.5b).

Requirements for psychological Nash equilibrium

Let me return to the simplifications which underlie GPS's concept of psychological Nash equilibria because they are the source of some of the conceptual deficiencies of GPS's and Rabin's models.[59] A closer look at the assumptions is revealing.

Collective coherency of belief hierarchies

For GPS, a player's first-order beliefs over actions are probability measures over sets of mixed strategies of all other players. Second-order beliefs are probability measures over cross products of the other players' first-order beliefs and their mixed strategies. This allows for (intuitively plausible) correlation between beliefs about others' strategies and beliefs about others' first-order beliefs. For all orders of belief of a player, this definition is continued inductively over the order of the respective player's beliefs. The complete belief set (hierarchy) of a player is the product of beliefs of all orders.

"Coherency" is reached if, for a given player, the different levels of beliefs are not contradictory. That is, a player's belief about the other players' strategies of order k must be coherent with, in the sense of 'derivable from,' the belief set of order k+1.[60] For instance, player 1's beliefs about what player 2 believes will be played must be coherent with what player 1 believes will be played. "Collective coherency" is reached if it is commonly known that all players' beliefs hierarchies are coherent. More specifically, it

[58] Similar observations hold for a third example used by GPS (1989). Compare p. 67.

[59] My explanations of GPS's ideas will be technical, but not formal. For formal treatments see GPS (1989), pp. 64-69. Note that the underlying formal framework here partly goes back to earlier works like Brandenburger and Dekel (1985).

[60] This amounts to the restriction that belief of order k has to be the marginal probability measure of beliefs of order k+1 about strategies of order (k-1 and lower). For the formalism see GPS (1989), p. 64.

is defined as the product of all individual belief hierarchies according to which all beliefs are certainly coherent. GPS reason that collective coherency is implied by the common knowledge of rationality. This presupposes some stronger underlying sense of rationality, however; namely one that demands rational beliefs. But whether it has to be assumed separately and explicitly or not, psychological Nash equilibria require collectively coherent beliefs.

Conformity of belief sets and psychological Nash equilibrium

A psychological Nash equilibrium is a tuple consisting of a belief profile, which is an element of the collectively coherent belief set, and an actually played strategy profile such that

(a) the belief profile is actually true and
(b) no player could be better off by unilaterally playing another strategy.

These two requirements seem to pose a puzzle. Roughly, the question is what comes first in the construction of the equilibrium, the equilibrium beliefs about the play (and the payoff which depends on them) or the equilibrium play (which should be rational *given* the payoffs)? This puzzle reappears in GPS's explanation of the relation between payoffs and expectations, which refers to the example of a gift exchange between a couple:

> "As in the standard theory of games, the payoffs for players are defined first on the outcomes (given any belief profile [...]) and only afterward extended by taking expectations to all of the mixtures included in [the product of all sets of mixed strategies]. The reason is that player i is presumed not to observe the mixture used by any player $i \neq j$. On the other hand, there is no requirement that the payoffs be linear in the beliefs. In particular, player i may get the same payoff if he expects j to deliver chocolates, and she does, as he would if he expected her to deliver flowers and she did. Yet his payoff might be very different if he expected her to randomize and she did (perhaps because he might think she is indecisive)." (p. 65)

Whether this definition can be a basis for a formalisation or not is one question. But on the conceptual level we are left with an awkward double role of expectations. They seem to be both source and outcome of payoffs (and vice versa). As I will explain later on, this conflicts with the intuitive-conceptual support of Nash equilibrium.

Summary form and continuity

Nash's original equilibrium idea was celebrated because it cut right through the regress of reasoning about beliefs of increasingly high order. This simplicity made it feasible to investigate the existence of Nash equilibria and to compute them for given games. By making payoffs dependent on belief hierarchies, however, it appears as if GPS re-introduce the problem of regress into higher orders of beliefs. Realising this potential pitfall, GPS introduce the "summary form" (p. 68) of a game, which reduces the complexity of the utility expression.

In the summary form, payoffs by definition only depend on such beliefs that are themselves a function of a strategy profile, such that beliefs are actually true in equilibrium.[61] In addition it is assumed that beliefs are collectively coherent and that beliefs have certain formal properties,[62] which render the utility function of the summary form continuous. This in turn enables the existence-proof for a psychological Nash equilibrium. So how strong a restriction is the summary form reduction?

Continuity in itself is already quite a strong restriction. GPS deliver an example in which continuity is violated. In this example, utility is a function over the entire hierarchy of beliefs such that beliefs of arbitrary order can lead to 'disappointment.'[63]

What is more worrying, however, is the strength of the summary form assumption itself. More precisely, the worry is about what *pre*sumption the summary form makes. At first sight, it might seem as if the summary form is simply a definition of a state which is as problematic (or unproblematic) as the orthodox Nash equilibrium definition.

But the implications of the definition of best reply (needed to define Nash equilibrium) which is defined on the basis of the summary form go even further than that. The definition presupposes that in psychological Nash equilibrium, so far as it exists, the beliefs *which feed into the payoffs* are stable even when a player unilaterally hypothesises about some other play.[64] They are implicitly held fixed not only for the

---

[61] A candidate for equilibrium is defined as (b, σ), where b is the set of all players' belief hierarchies "reflecting common knowledge of σ." For a particular player i, utility is a function $u_i(\beta_i(\sigma), \tau)$ of his beliefs $\beta_i(\sigma)$ (which are, by assumption, actually true in equilibrium) about the strategy profile and the strategy profile itself, τ. The summary utility function $w_i(\sigma, \tau) := u_i(\beta_i(\sigma), \tau)$ is directly defined over σ and τ.

[62] Compare GPS (1989), p. 68.

[63] Compare GPS (1989), p. 69.

[64] The definition is: "Let $BR_i : \Sigma \rightarrow \Sigma_i$ be player i's 'best response' correspondence, defined for each $\sigma \in \Sigma$ by $BR_i := \{\tau^*_i \in \Sigma_i \mid w_i(\sigma, (\tau^*_i, \sigma_{-i})) \geq w_i(\sigma, (\tau_i, \sigma_{-i}))$ for all $\tau_i \in \Sigma_i\}$." Here, Σ_i is the set of all of player i's mixed strategies and Σ is the cross product of all Σ_i, i.e. the set of all mixed strategy profiles. See GPS (1989), p. 68.

equilibrium itself, but also for the hypothetical scenarios in players' reasoning, which support the equilibrium. In this sense, the summary form grants the existence of psychological Nash equilibrium, but at the same time undermines its conceptual plausibility.

Again, the implicit assumptions of the summary form of the game 'cut through' the circularity concerning the determination of payoffs and beliefs, raised earlier: The definition of the expected play itself is a determinant in the payoff specification, which in turn feed into the expectations about the play. This mutual endogeneity between payoffs and beliefs, however, is a much more dubious assumption still than the usual assumption of circularity of beliefs in equilibrium, as will be discussed later.

GPS extend their approach to extensive form games. To do so, they have to introduce a richer formal structure. My critique however makes no reference to this additional structure.[65]

4. Fairness equilibria

In identifying the key feature of psychological games, it is crucial to note that the question whether beliefs have *some* influence on the outcome definition of these games can largely be conceptually separated from the *specification* of this influence in terms of a belief-dependent outcome function. Rabin's (1993) analysis of fairness equilibria, as special cases of psychological Nash equilibria, builds on this insight.

Rabin bases his fairness equilibria approach fully on GPS's analysis of psychological games. However, he develops a specific view on how psychological games arise. He hypothesises about how beliefs enter utility as a result of fairness intuitions.[66] For this purpose, Rabin distinguishes between material games and psychological games. As discussed above, material games should better be labelled 'culmination games,' following Sen's (1997) terminology as discussed above. They represent the outcomes of strategy profiles with respect to everything except for the *process* of choice. Psychological games can be derived from material games by

[65] Compare GPS (1989), pp. 70-71. This structure is used to define "behavior strategies" for players as subjective probability distributions over their respective possible actions for each information set, as well as utilities depending on the terminal vertex reached and initial beliefs. The dependence on initial beliefs only (as opposed to dependence on beliefs at subsequent choice nodes) is another simplifying assumption, which grants equivalence with the normal-form analysis of psychological games.

[66] Note the difference between Diamond's (1967) notion of fairness, as referred to in Chapter 2, and Rabin's notion of fairness. Diamond's notion has got to do with providing fair chances to a number of agents while Rabin's notion rather aims at capturing an intuition about reciprocity.

calculating the comprehensive utilities from the material payoffs as well as first and second order beliefs of the players about each other's choices.

More specifically, Rabin defines psychological games by utility functions for each player which positively depend on expected[67] material payoff, the player's own "kindness" as well as the opponent's beliefs about his kindness. A player's kindness is defined as a function of his choice and his beliefs about the other player's choice. To be more precise, it depends on what the player expects to be the deviation of the other's payoff from his 'entitlement' which is the arithmetic mean of the lowest and highest payoff possible given the choice the other player is believed to take. This is normalised by the range between the lowest and highest payoff possible given the choice the other player is believed to take.[68]

One's belief about someone else's kindness can then be defined congruently as a function of one's belief about the other's choice and one's belief about the other's beliefs about one's own choices.

---

[67] Expectations are defined to be correct in psychological equilibrium.

[68] The expression for player ***i*'s kindness** towards *j* is: $f_i(a_i,b_j) \equiv \frac{\pi_j(b_j,a_i)-\pi_j^e(b_j)}{\pi_j^h(b_j)-\pi_j^{\min}(b_j)}$. Here, a_i is player *i*'s choice; b_j is player *i*'s belief about player *j*'s choice; $\pi_j(\cdot)$ are *j*'s payoffs: $\pi_j^h(b_j)$ is his highest payoff given b_j and $\pi_j^l(b_j)$ his lowest payoffs of all Pareto-efficient outcomes given b_j, respectively. $\pi_j^{\min}(b_j)$ is the worst possible outcome for *j* given b_j. The expression $\pi_j^e(b_j)$ captures the average of $\pi_j^h(b_j)$ and $\pi_j^l(b_j)$.

Player ***j*'s belief about *i*'s kindness** is defined accordingly as $\tilde{f}_i(b_i,c_j)$, where c_j is *j*'s belief about *i*'s belief about *j*'s choice. (In equilibrium it holds by assumption that $\tilde{f}_i(b_i,c_j) = f_i(a_i,b_j)$.)

Player ***i*'s psychological payoff** is $\Psi_i = \tilde{f}_j(b_j,c_i)\cdot\left[1+f_i(a_i,b_j)\right]$, which is equal to $f_j(a_j,b_i)\cdot\left[1+f_i(a_i,b_j)\right]$ in equilibrium. In **psychological games**, payoffs are added to material payoffs for a given (hypothetical equilibrium) outcome. Compare Rabin (1993), pp. 1286-1287. This aggregation can be done with or without weights. Compare HHV (2004).

The 'material' game:

		Player 2 C	Player 2 D
Player 1	C	3, 3	0, 4
	D	4, 0	1, 1

Maximum/minimum payoffs and entitlements:

		π_1^h	π_1^l	π_1^e
Player 1	C	3	0	1.5
	D	4	1	2.5

	Player 2 C	Player 2 D
π_2^h	3	4
π_2^l	0	1
π_2^e	1.5	2.5

Kindness:

		Player 2 C	Player 2 D
Player 1	C	f_1=0.5	f_1=0.5
	D	f_1=-0.5	f_1=-0.5

		Player 2 C	Player 2 D
Player 1	C	f_2=0.5	f_2=-0.5
	D	f_2=0.5	f_2=-0.5

Psychological payoffs:

		Player 2 C	Player 2 D
Player 1	C	Ψ_1=0.75	Ψ_1=-0.5
	D	Ψ_1=0.25	Ψ_1=-0.25

		Player 2 C	Player 2 D
Player 1	C	Ψ_2=0.75	Ψ_2=0.25
	D	Ψ_2=-0.5	Ψ_2=-0.25

Psychological games with different weights (1 and 4, respectively) for psychological payoffs added to the material payoffs:

		Player 2 C	Player 2 D
Player 1	C	3.75, 3.75	-0.5, 4.25
	D	4.25, -0.5	0.75, 0.75

		Player 2 C	Player 2 D
Player 1	C	6, 6	-2, 5
	D	5, -2	0, 0

Figure 3.3

Spelling out a special case of psychological Nash equilibrium, the so-called fairness equilibrium is defined by two requirements. (1) Actions, as well as first-order beliefs and second-order beliefs about these actions coincide. (2) Each player maximises psychological payoffs. These requirements correspond to GPS's simplifying assumptions of belief-conformity and the summary form, as well as the traditional Nash requirements of individual rationality.

An example

As opposed to GPS's games there are no variables in Rabin's psychological games, because the calculation of Rabin's psychological games already incorporates the conformity of beliefs assumptions and the summary form assumption. When the psychological game is already derived, the analysis is completed by simply calculating the Nash equilibrium in the psychological game. In that sense, each utility figure in a psychological game captures the payoffs that would be yielded were the outcome, to which these payoffs are assigned, the actual equilibrium. This calls for concrete explanation.

Consider Figure 3.3, which is borrowed from HHV (2004), who discuss Rabin's model in some detail. It is the transformation of a material prisoner's dilemma into two psychological games. Building on the definitions of a psychological game and fairness equilibrium, as listed in Footnote 68, minimum and maximum (Pareto efficient) payoffs are identified first and entitlements are derived. Next, the kindness of players 1 and 2 is derived, respectively. This is done for each outcome of the material game under the (unmotivated!) assumption that this outcome was the equilibrium. Then the psychological payoffs are calculated for each hypothetical outcome in a similar fashion. This can be done by exclusively using kindness figures – instead of using both kindness figures and belief of kindness figures, as in the definition – because in the psychological equilibrium to be calculated beliefs are assumed be true. Finally, the weighted psychological payoffs are added to the material payoffs, yielding psychological games. Given a weight of 1, e.g., the derived psychological game is a prisoner's dilemma; given a weight of 4 it changes to a game that has an alternative equilibrium in (what, in the material game, was referred to as) cooperation.

Again, questions immediately arise. What could possibly justify the assumption that some outcome is an equilibrium *before* the full psychological payoff is even justified? Or, inversely, what is the authority of the psychological equilibrium in a

certain outcome, when it can only be grasped on the assumption that this outcome is already an equilibrium? It seems that any psychological equilibrium would be an equilibrium in two senses then, by assumption and as the result of calculating psychologically optimal responses; it would be overdetermined in that sense. Finally one wonders: How exactly would player 1 in the defective equilibrium in one of the psychological games, for instance, reason about what payoffs would await him should he cooperate (and, thereby, reason whether such a unilateral deviation would be rational)?

I will explain in Section 5 that, similar to the case of the GPS model, no satisfactory answers can be given to these questions. But before I go through this critique a good look should be taken at the methodological motivation for modelling psychological games in the first place. Are psychological games methodologically indispensable?

Why model psychological games?

The focus of Rabin's theoretical investigation is on the nature and relation of Nash equilibria and fairness equilibria in certain types of games. Among other things, the propositions presented hold that Nash equilibria coincide with fairness equilibria in those types of games where players mutually minimise or mutually maximise each other's outcomes and in which material payoff differences are small.[69] But none of these results in particular are at stake here. Rather, I am interested in what motivates Rabin to build on GPS's framework.

Rabin presents another argument for belief-dependent outcomes referring to empirical evidence found by authors like Greenberg and Frisch (1972), for example:

> "Psychological evidence indicates that people determine fairness of others according to their motives, not solely according to actions taken. In game theoretic terms 'motives' can be inferred from a player's choice of strategy from among those choices he has, so what strategy a player could have chosen (but did not) can be as important as what strategy he actually chooses." (Rabin 1993, p. 1289)

This statement, however, presents no immediate argument for belief-dependent outcome definitions and fairness equilibria. Rather, it seems to be an argument for the importance of counterfactuals in game theoretic analysis. But the importance of counterfactuals or

[69] Compare the propositions and definitions in Rabin (1993), pp. 1290-1292.

hypothetical reasoning of players need not be captured in actual beliefs which are part of outcome-definitions. But Rabin provides a better argument.

In fact, consider Figure 3.4a, the 'Battle of the Sexes' game. Rabin argues that in the psychological transformation of the material 'Battle of the Sexes' game there could be an asymmetric fairness equilibrium {Opera; Boxing}. This requires that additional motives come into play to the effect that, e.g., players wish to 'hurt' each other because they mutually belief that the other will knowingly – and even at the price of sacrificing their own material payoff – play such a hurtful strategy, which is payoff-diminishing to the other. He continues to argue that this can only be consistently captured in a belief-dependent payoff model, not in an orthodox (purely material) model where payoffs only depend on action. This is because given the possibility of the psychological game description above there could be an equilibrium in {Boxing; Boxing} if the beliefs are not mutually distrustful in equilibrium or {Opera; Boxing} if they are. Rabin argues that "these statements would be contradictory if payoffs depended solely on the actions taken."[70] But is he right in this justification for the introduction of psychological games?

He is right, of course, in the sense that ceteris paribus two mutually exclusive strategies cannot both be strictly optimal for a player in a material game. But why not just assume that the (cumulative) game played by players 1 and 2 which incorporates all significant motives of action is represented by Figure 3.4b, allowing us to simply proceed with orthodox game theoretical solution concepts? Rabin's answer to this question might be that in this game the psychological effect is mistakenly projected onto outcomes. This is problematic because if, e.g., the players arrive at {Opera; Boxing} by false beliefs about each other's moves, e.g., the additional psychological payoffs of (the feeling of) revenge would no longer be justified.

		Player 2	
		Opera	Boxing
Player 1	Opera	2, 1	0, 0
	Boxing	0, 0	1, 2

Figure 3.4a

		Player 2	
		Opera	Boxing
Player 1	Opera	2, 1	3, 3
	Boxing	3, 3	1, 2

Figure 3.4b

[70] See Rabin (1993), p. 1286.

So let me consider another putative alternative for capturing the effect of a revenge motive on the equilibrium in the Battle of the Sexes. We could say that when players 1 and 2 are confronted with a material Battle of the Sexes, it is not a priori clear what psychological versions (each with belief-independent outcomes) of the material game is played, Figure 3.4a or Figure 3.4b. This would be to capture the belief-dependence of payoffs on the level of the game selection. But this would be a highly unsatisfactory approach, of course, because it would turn what is traditionally a task for the game theorist into a task for the players. This is troublesome as we cannot just model some meta-game over Figures 3.4a and 3.4b. This in turn is because which one of these games is played is supposed to be neither the player's decision (in isolation) nor a move by 'nature.'

So far then, it seems as if Rabin has a good point in advocating a solution in which payoffs are endogenous to beliefs. However, we should not yet conclude that the analysis of rational equilibrium psychological games is methodologically justified. This is because, apart from the question of whether sensible solution concepts for psychological games can be presented, it is not plausible to assume that people attach emotions to any choice by another player if the belief about his choice is *derived* from the underlying common belief in (or knowledge of) rationality. This will be addressed in more detail in Section 6.

5. Immanent problems with psychological equilibria

For psychological games to count as a useful extension of game theory rather than a new kind of theory altogether, it has to be possible to apply a plausible solution concept to these games. This section investigates how Nash equilibrium in general, some standard refinements, and the concept of mixed equilibria fare in the psychological game framework. The question here is whether these concepts can be applied in principle.

Psychological Nash equilibrium

The mutual dependence of beliefs and payoffs in psychological Nash equilibrium is not at all an unproblematic extension of the usual circularity of the Nash equilibrium. Nash equilibrium in regular games is conceptually supported by a clear intuition. It concerns

the hypothetical reasoning behind assessing unilateral deviation from an equilibrium candidate. This intuition does not extend to Nash equilibria in psychological games.

In a Nash equilibrium of a regular game, it remains an unresolved question how the players got to the exact beliefs which support the equilibrium. As I will explain in more detail in Chapter 4, Sugden (1991) argues that we might run into difficulties when mixed strategies are interpreted as subjective beliefs, as long as a correlating device is excluded.[71] The problem is that one needs to assume that, for there to be such an equilibrium, players have common priors with respect to each strategy played. But this assumption cannot be founded on rationality assumptions alone.

In certain psychological games the situation is potentially even worse. It is unclear why it is justified for players to *maintain* the beliefs they have in equilibrium. This is because when a player considers an alternative move to the one in the alleged equilibrium, he cannot rely (without further assumptions) on the fact that the payoffs structure would be stable. In order to assess the potential gain of a deviation, he would have to determine the payoffs and therefore the other player's beliefs for the hypothetical case of a deviation. More specifically, he has to find out what the other would believe in such a case given that all know that everyone is rational. But then he is thrown back to the original question of what would be rational to do. To answer this question, however, he needs to assess unilateral deviation.

The infinite regress of players' reasoning in the problematic class of psychological games therefore cannot be stopped by contemplating unilateral deviation, because the content of beliefs in these hypothetical scenarios is indeterminate. Beliefs about other players' choices and beliefs cannot be derived from the structure of the game and the assumption of common knowledge of rationality, because rationality is not defined without payoffs. Payoffs in turn depend on beliefs. Beliefs, however, need not conform to reality outside of equilibrium.

Again, this is more problematic than the circularity of beliefs in a regular Nash equilibrium, because beliefs in a regular Nash equilibrium can at least be justified by unilateral deviation as long as we start by assuming the equilibrium state. Here at least

[71] Compare Sugden (1991), pp. 765-768.

there can be a definite answer to the question whether it would be *rational* to deviate unilaterally. [72]

Remember the example of Figure 3.1. Even in this game, which is not strategic in the traditional sense, it is unclear how to defend a psychological Nash equilibrium like $p=q=q'=1$. Without further ad hoc assumptions, it seems hard for player 1 to hypothesise, whether it would be rational to 'deviate' and play $p=0$. What would be the payoffs? Would it still be the case that $q'=1$? On what basis should q' be specified?

Alternatively, take the prisoner's dilemma and its psychological game versions in Figure 3.3, as derived according to Rabin's fairness equilibrium approach. It is not possible to defend the cooperative Nash equilibrium in the weight-4 case by arguing that it can be seen in the matrix that none of the players would be individually rational to deviate from cooperation. The payoffs (5, -2), e.g., describe the payoffs for the case that $\{D; C\}$ is the *actual* equilibrium strategy profile and all beliefs are *actually* aligned accordingly. It is not evident, however, that the payoffs would be (5, -2) if player 1 were to deviate from $\{C; C\}$ by playing D, because (5, -2) are only the payoffs if both players believe that $\{D; C\}$ is the equilibrium. So, *implicitly*, Rabin's structure makes untenable assumptions about how player 1 should hypothesise about his payoff outside the actual equilibrium. Furthermore, Rabin's structure is ill-equipped to make any more sensible *explicit* assumptions about how player 1 should hypothesise about his payoff outside the actual equilibrium.

This problem is formally bypassed by GPS who prove that there are always equilibrium choices in psychological games which are compatible with coherent beliefs. But, as in Rabin's case, these beliefs could hardly be considered to *support* the equilibrium, because they in turn depend on what counts as a rational action.

Furthermore, I should point out that there is a difference between what Rabin (1993) and GPS (1989) imply about best response hypothesising of players. GPS imply that, starting from a psychological equilibrium, the beliefs that determine the payoffs which are referred to when hypothesising unilateral deviation are fixed.[73] Rabin, on the other hand, implies that beliefs and payoffs adapt to choices in the unilateral deviation

---

[72] Note that Stalnaker complains about the indeterminacy (and implicit assumptions) of counterfactual reasoning even within the orthodox framework with common knowledge of rationality assumptions. Compare Chapter 4.

[73] The strategy profile σ, which determines the beliefs which in turn determine the payoffs, is the first argument in the summary-form utility function $w_i(\cdot)$ of the best response expression $BR_i:=\{\tau^*_i \in \Sigma_i \mid w_i(\sigma, (\tau^*_i, \sigma_{-i})) \geq w_i(\sigma, (\tau_i, \sigma_{-i}))$ for all $\tau_i \in \Sigma_i\}$, as mentioned earlier. It remains constant in this expression while i's choices $\tau_i \in \Sigma_i$ are varied.

scenario. But since none of these authors offer the framework to support their implicit assumption in an explicit manner, I shall not pay attention to their differences.

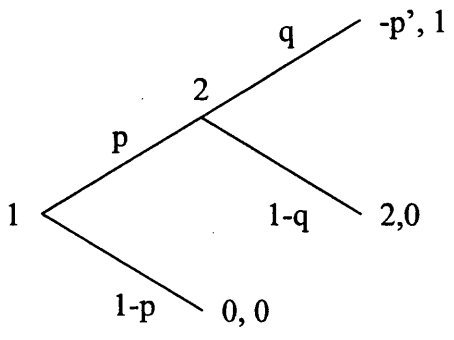
The mutual dependence of beliefs and payoffs also means that any causal interpretation of the belief-action relation is doomed. HHV (2004) address this problem:

> "To get to a unique utility assessment of an outcome, we must assume an equilibrium of beliefs. This turns what used to be a simple unidirectional system of causation, running from utilities to rational beliefs to equilibrium, into a form of circularity. [...] To use the apparatus of game theory to predict what rational people will do, we need to know what beliefs *actually* obtain. But if one knows this then the apparatus of instrumental rationality is no longer really needed to explain how people act. It is true, of course, that action can be instrumentally rationally constructed once the beliefs are known, but knowing the equilibrium beliefs about action is enough to predict what actions will be taken and it seems almost simpler to say that people's actions have been guided by the prevailing norm." (pp. 284-285)

This quote also touches on the overdetermination of equilibria, as mentioned in Section 4. Common belief in (or knowledge of) rationality has become redundant in psychological equilibrium, because equilibrium choices have to be assumed in order to deduce equilibrium beliefs. But then there is no need to apply Nash equilibrium at all.

Refinements

We can assess the failure of perfect Nash equilibrium as a universally applicable refinement for psychological games in a similar fashion. Consider the counterexample in Figure 3.5a taken from GPS (p. 73). Player 1's belief about 2's belief about p is given by p'.



Figures 3.5a

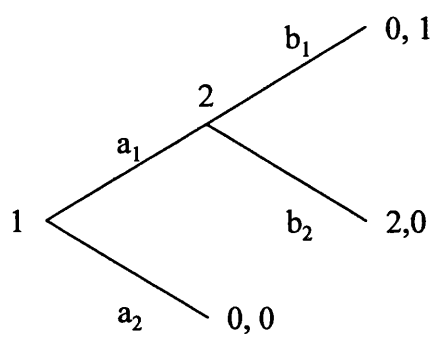


Figure 3.5b

GPS argue that it is certain that player 2 plays Up according to standard assumptions. This leaves us with p=0 in the psychological Nash equilibrium. But since by assumption p=p' in equilibrium, we arrive at the "induced game" depicted in Figure 3.5b, for which $\{a_2; b_1\}$ is not a perfect Nash equilibrium. So there can be no trembling hand perfect Nash equilibrium in the original psychological game.

The trembling hand refinement tests a Nash equilibrium's robustness against minimal perturbation. This perturbation can be thought of as enabling hypothetical reasoning because all strategies are modelled as being played with some probability and hence can be grasped as possible turns of the game. But we have seen earlier that the psychological Nash equilibrium is in conflict with hypothetical reasoning.

Remember also that trembling hand perfection (henceforth perfection) is a refinement for the static form. So the trembles allow making comparisons between different possible worlds in static reasoning.

GPS's argument against the existence of a perfect equilibrium in Figure 3.5a is based on the unspecified value of p'. The tremble thought experiment forces us to think about what value it would be known to take on at the second node. The answer is p'=1, in which case player 1 would have been irrational in choosing to play Up. So that scenario is excluded by assumption. But then if 1 is known to play Down in virtue of being rational, p' would be zero, which would once again make playing Up the weakly dominant strategy for player 1 etc. This problem arises because outcomes cannot be conceptualised as robust in non-equilibrium situations in psychological games. Hence the rationality of moves has to be assessed on the basis of moving criteria, namely belief-dependent utilities. But, of course, the underlying beliefs in the different counterfactual states depend on what would be rational moves in these states etc. This is an immediate consequence of the mutual endogeneity of beliefs and payoffs discussed above.

GPS make it sound as if the failure of trembling hand perfection is not all that worrying because for extensive form psychological games a number of refined equilibria exist, e.g. sequential and subgame perfect equilibria, which are usually perceived of as at least equally strong refinements as perfection only in a different form of representation.[74] But this would be a hasty conclusion because for extensive form psychological games the actual beliefs for all information sets can simply be

[74] Compare GPS (1989), pp. 73-74.

proclaimed, just as it was done for unrefined psychological Nash equilibria in the strategic form. The onus to prove that extensive form equilibrium refinements have a sensible conceptual interpretation still lies with GPS. So despite the formal existence proofs for sequential and subgame perfect equilibria, the struggle with the trembling hand perfect psychological Nash equilibria should be taken seriously.

Furthermore, GPS report that backward induction does not apply either as solution algorithm to psychological games because "when a node is reached, it does not capture adequately the state of the game: the node identifies a history of play but not the players' beliefs." (p. 63) The failure of backward induction should not come as a surprise, though, after what HHV (2004) said about the order of causal reasoning in psychological Nash equilibrium. Backward induction commits one to a specific direction of reasoning, deriving beliefs from payoffs. But the mutual dependence of the beliefs and payoffs in psychological games does not allow for this commitment.

Mixed equilibria

As for another problem with psychological games, HHV (2004) argue that psychological Nash equilibria in mixed strategies are impossible because there can be no such thing as a mixed strategy in a psychological game (p.280). This contradicts GPS who even present alleged examples of such equilibria (p. 73). HHV's argument can be recapitulated as follows.

(1) Mixed equilibria presuppose the possibility of randomisation.
(2) Randomised strategies depend on objective probabilities.
(3) The notion of objective probabilities conflicts with the fact that beliefs in psychological games must be endogenous (as part of the infinite circle of mutual dependence with the utilities).

---

(4) Therefore mixed equilibria do not exist in psychological games.

Premise (1) goes back to HHV's disbelief in Aumann's reading of mixed strategies as based on players' subjective beliefs about each other's strategies. HHV's argument is close to that of Sugden's (1991) point, mentioned a few paragraphs earlier, that an the implicit statement about subjective beliefs in such a mixed strategy equilibrium is extremely strong and yet is supposed to be grounded on no psychological premise whatsoever.

I tend to agree with Sugden on this point. However, I would like to point out that the beliefs in pure psychological Nash equilibria 'are built on air' in a very similar fashion. As was pointed out above, they are also circularly endogenous and in that sense derived from themselves. So the standard of intuitive plausibility which undermines psychological mixed equilibria as based on subjective probabilities, also renders obscure psychological Nash equilibria in pure strategies. So if premise (1) is taken to be true, there seems to be a problem with what is suggested in premise (3). The reasons supporting premise (1) seem to render the concept of beliefs as part of an endless circle of mutual dependence with payoffs meaningless.

Having said that, I am not defending GPS's account of mixed psychological equilibria. Rather I would conclude that the concept of mixed psychological equilibrium is even more obscure than that of regular mixed Nash equilibrium.

In addition, GPS's approach remains obscure when the formalisation of expectations over outcomes is introduced. Remember that GPS write that "the payoffs for players are defined first on the outcomes (given any belief profile [...]) and only afterward extended by taking expectations to all of the mixtures included in [the product of all sets of mixed strategies]." At the same time they explain that "there is no requirement that the payoffs be linear in the beliefs." As with pure strategies, it is not clear what it means, in terms of agents' *reasoning*, to 'first' define outcomes and then define expectations, if there is a circular dependence between the two.

6. Intuitive problems with belief-dependent payoffs

Leaving behind the immanent problems that arise with GPS's approach, I would now like to return to the question of whether their approach corresponds to the intuitions which made these authors look beyond the orthodox approach in the first place. The quick answer to this question is a qualified 'no.' That is, even if belief-dependent outcomes were coherent notions, it is not clear whether they could capture what we want them to capture.

Many of the intuitions that GPS set out to represent in psychological games cannot be represented in the structure that GPS provide. GPS mention the number of "psychological considerations" which they think orthodox game theory cannot capture, namely "surprise, confidence, gratitude, disappointment, embarrassment, and so on" (p. 61). Rabin adds "fairness," which, as remarked in a footnote earlier, stands for a version

of reciprocity. Considering how each of these notions can be taken account of in psychological games reveals shortages in GPS's framework. Let me present some of these deficiencies in turn.

Dynamics

GPS's framework works for representations of games in which time plays no substantial role for the 'psychological' aspect of outcomes. Sure, he presents extensive form representations of psychological games. But even in these only initial (conforming) beliefs determine the effective outcomes. What matters for these comprehensive outcomes in GPS's framework are static beliefs in unrepeated games. There is no historical information and consequently no learning about, say, the type of the other player flowing into the psychological perception of outcomes in these games.

But without this substantive role of the temporal dimension it seems hard to model surprise or disappointment, for instance. These notions clearly refer to a contrast between the expected and the experienced and hence (if they are about belief at all) demand at least two temporal stages of belief. In the kind of rational equilibrium of correct expectations that the GPS adopt from orthodoxy, there are by definition no surprises, be they positive or negative deviations from the expected, at all.

More arguably, this critique can be extended to other emotions, like gratitude and embarrassment. Admittedly, one could argue alternatively that these emotions could be understood as being based on contrasts between beliefs in two possible states of a game. Accordingly, gratitude, e.g., could be a feeling based on the contrast between what another player does in the actual state and what he would do in some possible state. But this cannot serve as a defence of GPS's approach either, because GPS put all the weight of representing process-related emotions on actual beliefs, not on hypothetical reasoning in games, as mentioned earlier.

Finally, some might argue that even intuitions about fairness as reciprocity have an irreducibly temporal aspect. The supporting argument could be that feelings towards the action of another player can only develop on the basis of historical experience of interaction (with possible updating mechanisms, etc.). However, the empirical evidence that players in experimental setups rationalise their action with arguments about reciprocity even in one-shot games, for instance, might be taken to indicate that

fairness-intuitions do not presuppose history.[75] But there is a deeper concern still about the representational value of psychological games.

Rational choice and rules

Notwithstanding the header, this subsection is not a survey of the theory of action. Rather it is to make the point that (some of) our feelings about the action of other agents in strategic interaction are directed towards what type of choice (and chooser) we are confronted with. For instance, we seem to appreciate people who cooperate in certain types of situations, e.g. those described as material prisoner's dilemmas, *no matter* whether this is rational in some sense or not. In fact, my impression is that many of us particularly appreciate people who cooperate even if it is (materially) irrational in some sense or at least before the rationality of a cooperative choice has been assessed. Conversely, we appreciate those cooperative actions less which have been chosen as the most rational option after a process of deliberation about payoff maximisation of some sort.

To put this point in yet another way, we seem to particularly like people who cooperate as a matter of principle or rule. Some might insist that such a rule must have been chosen intentionally or consciously at some stage in order to evoke these feelings in us.[76] But still the point remains that it often seems to be a necessary condition for us to feel that someone acts in some sort of praiseworthy manner, that his action was not calculated as being the expedient choice for himself.

As I will argue in a short while, this intuition might abandon us once we try to understand what it means to say that other-regarding motives are already captured as part of the outcome definitions of a game. This is because we might not feel that it is in any way less praiseworthy when someone acts altruistically as the outcome of some altruism-maximising calculus. But if other-regarding motives are captured in the other agents' utility functions, there is no need to depart from game theoretical orthodoxy in order to capture motives of reciprocity.

Coming back to the original point, it seems that beliefs about what an agent is going to do in a game are in themselves not what our feelings concerning fairness, gratitude etc. are about. This is because such beliefs are (in GPS's model, at least)

---

[75] Compare, e.g., Güth et al. (1982).

[76] So then there would be another necessary condition for our appreciation of someone else's actions. I will return to this point in a few paragraphs.

determined by further beliefs about the other's rationality and further epistemic assumptions. But if rationality can be identified as the driving factor behind the agent's choice the kinds of feelings that GPS and Rabin try to capture are (often) not at stake at all.

Some caveats are due at this point. I mentioned that players might plausibly develop feelings towards rules, which guide other agents' actions. However, I am not suggesting that a notion of rules can easily or at least in principle be introduced into the game theoretic framework without reinventing the discipline. Being rule-driven to pursue one kind of action could be represented as the rule-abiding action being strictly dominant. But if we decide to represent the rule of 'Always cooperate!' in what was originally a prisoner's dilemma, that would change the game into something other than a prisoner's dilemma and hence render the rule 'Always cooperate!' meaningless in the resulting game. Apart from the question of formalisation, there is in fact a semantic problem which casts serious doubts on this possibility. The question is how to understand a rule, given that utilities are supposed to be all-motives-encompassing numbers. It is hard to grasp what it means to follow the rule 'Always cooperate!' e.g., when cooperation itself is a very abstract notion depending on the structure of the game and on utilities encapsulating other-regarding motives. I shall not be concerned with solving these issues here.

Belief-dependent outcomes and other-regarding utility functions

I would like to briefly return to the question of what is left of our intuitions about strategic decision-processes, which led GPS and Rabin to extend on orthodox game theory, once we take into consideration that the utility definitions in material games[77] might already contain more than self-regarding motives. The point is that defining utility as all-motives-encompassing lends greater complexity to the intuitive understanding of strategies in games. To repeat my previous example, it is not easy to intuitively comprehend what it means for someone to 'cooperate' in a prisoner's dilemma when (at least) one player's 'defecting' is driven by an altruistic utility function.

[77] Here, again, the term "material games" is misleading. The more descriptive term 'culmination outcome games' would make it harder to forget that – even before process-dependent aspects are captured in outcomes – utilities capture more than material or selfish motives.

In fact, if players develop process-dependent feelings towards (what they believe will be) each other's choice of action in such a game, these feelings will have to be understood as higher-order motives or feelings. These conceptually come posterior to all the motives captured in the utilities of material games.

The conceptual discussions in both GPS (1989) and Rabin (1993) do not properly take this into account. In fact, the examples mentioned in these works seem to implicitly appeal only to intuitions about process-dependent preferences in games which are material games in the literal sense. They seem to make reference only to games in which the utilities exclusively represent selfish assessment of material payoffs. But, of course, utility theory has moved a long way from such presumptions about what counts as the objects of preferences. So the onus lies with the authors of these theories to illustrate that we have process-dependent (or belief-dependent, for that matter) preferences over, say, altruistic utilities.

GPS and Rabin have to go a longer way in interpreting their examples to convince us of what exactly these higher-order motives refer to intuitively. If they fail to do so, this would severely diminish the scope of their theory. But it would not necessarily render the scope void.

Rational choice and intentions

We often say that it is the idea or intention behind an act that matters.[78] This might, among other things, be a way to express the fact that we are aware of the possibility that some actions are accidental or do not lead to an anticipated effect even though the agent wanted to do something praiseworthy or blameworthy. In games with common knowledge of the game structure and rationality, it might be argued, there can be no such accidents. But intuitions could be relevant in another sense. It could matter, for example, to an agent whether another agent intends to cooperate in a situation. So it might be tempting to define outcomes as process-dependent in the sense of intention-dependent.

But since game theory uses a rational choice framework the follow-up question is whether and how intentions in games can be accounted for. My immediate reply would be pessimistic: As it is, the rational choice framework bypasses any account of intentions. Assumptions about utility, rationality and further epistemic relations replace

[78] A minor linguistic observation might be of interest in this context: The German equivalent of the English proverb 'It's the thought that counts' literally translates as 'The intention counts.'

talk of intentions. That does not mean that intentions are refuted; they just cannot be represented adequately in the rational choice framework as it is. Very coarsely speaking, in game theory, actions are (apart from struggles with solution concepts and equilibrium selection) determined by belief (expressed in the epistemic assumptions) and desires (expressed as utilities).

While authors like Bratman (1987) have explored the nature of intention and its relation to the belief-desire model in more depth, I have no intention myself to explore this issue here. In principle, it might well be possible to formalise a belief-desire-intention model of rational choice and adopt it in game theory. But so far it has not been done; and I have the feeling that it will not be easy.

Having said that, HHV (2004) present an amended version of Rabin's fairness equilibrium approach, which HHV explicitly set up to take account of intentions. Let us consider their suggestion.

7. Introducing Intentions

HHV (2004) criticise Rabin's approach for similar conceptual reasons which I laid out. In particular, they are concerned about the conceptual status of psychological equilibria, as mentioned earlier. They also formulate concerns about the intuitive value added of Rabin's model. More specifically, they argue that Rabin's definitions of fairness as reciprocation presuppose some sort of 'eye-for-an-eye' ethos which stands in contrast with other, more forgiveness-oriented moral intuitions. Also they find that the underlying notion of kindness is just one value among many (e.g. justice or honour) and that it is arbitrary to take account of it alone.

But despite the severity of their concerns, HHV consider a modified version of Rabin's existing model. They take it to be an improvement in the sense that their amended definition of entitlement (and the derivative definition of fairness equilibria) better match our intuition of fairness. Their re-definition is designed to be not only sensitive towards beliefs over choices, but also intentions. I will consider how far HHV's model is an improvement vis-à-vis Rabin's model and whether it is correctly described as being intention-sensitive.

Let me first comment on HHV's new definition of entitlement. Remember that the original definition of entitlement is the arithmetic mean of the lowest and highest payoff possible given the choice the other player is believed to take. HHV impose two

requirements on their re-definition of entitlement; for agent *j* to have any non-zero entitlement, given his (correct) anticipation of *i*'s move,

(1) *j*'s actual strategy must sacrifice utility as compared to his Nash reply, and
(2) for *j*'s entitlement to be positive, *i* must gain utility from *j*'s actual strategy as compared to the scenario in which he plays his Nash reply; for *j*'s entitlement to be negative *i* must lose utility.[79]

One crucial feature of these requirements is that they define Nash equilibrium play as a reference point of moral neutrality. Also, HHV suggest a specific entitlement formula which fulfils (1) and (2).[80] According to this formula, *j*'s deviation from Nash play counts towards his moral desert if *i* profits from it and vice versa. This is a proportional relation: The more *j* 'sacrifices' for *i*'s good, the more *j* deserves; the more *j* 'sacrifices' for *i*'s disadvantage, the less *j* deserves.

The motivation behind requirements (1) and (2) is to make a player's entitlement, as manifest in the perception of the other player, dependent on how well he 'intends' the other player to fare in the outcome, materially speaking. HHV argue that the term 'intention' is suitable here, because it can be applied to cases where *i* determines *j*'s entitlement on the basis of what *i* believes *j* will do and from what *i* believes *j* believes *i* will do. In such cases, *i* could, for instance, detect when *j* makes (or '*intends*' to make) a sacrifice to 'hurt' *i*. I will argue, however, that this intentionality-interpretation of HHV's amended model is not easy to follow when the solution concept of psychological Nash equilibrium is applied to calculate equilibrium play in the psychological game and, hence, the underlying material game.

HHV's definition of fairness (kindness or nastiness) is derived from their notion of entitlement. One of its additional features, however, is that *i*'s kindness or nastiness, e.g., is set to zero, if *i* is 'irrational.' This is the case when there would have been a strategy for *i* yielding him at least the same (material) payoff while yielding *j* greater or lesser payoff, respectively.[81]

---

[79] Compare HHV (2004), pp. 285-286. The formal definitions are (1) $\pi_j(s_i, s_j) < \pi_j^n(s_i)$, where $\pi_j^n(s_i)$ represents *j*'s Nash reply payoff given s_i, and (2) $\pi_i(s_i, s_j) > \pi_i^n(s_i)$ and $\pi_i(s_i, s_j) < \pi_i^n(s_i)$, respectively. The strategies s_i and s_j are equilibrium strategies which, by assumption, conform to everyone's beliefs of all orders. HHV's fairness definition thus need not refer to beliefs directly. Given the conformity assumption, this is merely a notational matter.

[80] Compare HHV (2004), p. 286.

[81] Compare HHV (2004), p. 285.

In isolation, HHV's re-definitions of entitlement and fairness offer improvements with respect to our intuitions of these notions. It seems plausible to treat Nash play as morally neutral, as it is what might be considered the (rational) default scenario. It also prima facie appeals to my intuition that someone's entitlement is relative to how this person is expected to trade off his own payoff against that of others.[82]

At the same time, my intuitive agreement ends when HHV imply that their approach, in conjunction with psychological Nash equilibrium, captures concerns about the kindness of players' intentions. This is because in psychological equilibrium (i.e. the re-defined fairness equilibrium) players are still assumed to act rationally. They maximise *psychological* payoffs, of course. But the crucial point is that they are driven by a maximisation calculus, not a concern about anyone else's payoffs, material or psychological. The picture of psychological payoff-maximising players leaves no space for picturing players as being (directly) driven by a concern for other players' payoffs.

In other words, in HHV's fairness equilibrium, players are modelled as developing emotions towards kind or nasty 'intentions' on the level of the material game, while they also commonly know that these alleged intentions only come about as rational choice on the level of the corresponding psychological game. This seems like an awkward model of interaction, because choices are systematically thought to be overdetermined, namely by character traits like nasty intentions *and* rationality. Players think of each other as rationality-driven, while they also judge each other's behaviour as if it was not.

In this light, it even seems awkward to refer to the material Nash equilibrium as a morally neutral state. After all, once the psychological game is introduced as the frame of reference for rational choice, the material Nash equilibrium can no longer be treated as the scenario of what is to be rationally expected. That is unless two potentially conflicting levels of rationality are deemed acceptable in the model.

A related problem arises in HHV's new fairness definition itself. Remember that an agent's kindness or nastiness is set to zero, if he is 'irrational.' But the underlying definition of rationality refers to the material game. Again, it is not obvious what the

[82] HHV's model also has the advantage of rendering $\{C; C\}$ the only psychological Nash equilibrium if sufficient weight is given to psychological payoffs. (Prima facie, this might correspond to some people's intuition that cooperation in the prisoner's dilemma is rational in some sense. Compare Chapter 5.) Unfortunately, this result presupposes applicability of psychological Nash equilibrium as a solution concept.

status of this 'rationality' is in the presence of a competing notion of rationality, based on the psychological game.

All in all, my critique of HHV's intention-based amendment of fairness equilibrium exemplifies a point that I tried to argue earlier. It is hard to integrate intentions or rule following into a rational choice framework, especially into game theory with common knowledge of (or belief in) rationality assumptions. Furthermore, my critique is an extension of the immanent problems with any psychological GPS-type equilibrium approach, as discussed by HHV themselves. That is because the circular definition of equilibrium beliefs reappears in HHV's definitions of intention, entitlement and fairness. This in turn obscures the intuitive interpretation of these notions.

8. Summary and concluding remarks

Trying to define payoffs in games as belief-dependent challenges the game theoretical habit of presuming that we can simply define utilities as all-encompassing motives. This is important, as we need to understand more about whether and how a feedback mechanism between motives of play and the play itself could work. Attempts at modelling such feedback processes, as in GPS (1989) and Rabin (1993), are important because they shed light on the foundations of game theory.

However, I tried to argue for greater scepticism towards the idea of psychological game than that expressed by these authors. In particular, belief-dependent outcome specifications seem to generate more problems than solutions. In this sense, we can specify Hausman's (2005a) insufficiently specified concern about the possibility of process-dependent consequences in rational choice theories further. This is not, however, a final verdict about the possibility of modelling process-dependent payoffs in games.

The problems with psychological games, i.e. games with belief-dependent outcomes, do not occur on a purely formal level. They arise only when we try to fill in the 'conceptual flesh' around the 'skeleton' of the new assumptions and solution concepts. I addressed problems with the concept of psychological Nash equilibrium in pure and mixed strategies, as well as the trembling hand refinement. The core problem here is a mutual endogeneity of payoffs and beliefs in psychological Nash equilibrium. This mutual dependence is harmful as it obscures the intuition behind Nash's condition

that there be no incentive for unilateral deviation. This is because the content of this condition, which depends on counterfactual reasoning, cannot be assessed when it is not clear what the rational choice would be in counterfactual situations in which beliefs – and thus payoffs – do not conform to the actual beliefs. This problem is not solved by assuming one or the other equilibrium belief profile will prevail, because such an assumption begs the question. It does not address the extent to which the assumed equilibrium is supported by hypothetical reasoning. I argued that this problem reoccurs in the GPS's refinement discussion regarding psychological Nash equilibria.

Apart from these immanent problems with solving psychological games, the second concern raised in this chapter is whether the introduction of psychological games helps at all in representing intuitions about gratitude, disappointment and fairness. This is an important concern, as it challenges one methodological motivation behind dealing with psychological games. This in turn raises doubts about how much is at stake in the theoretical debate about psychological games.

On the other hand, I do not claim that there are no intuitively plausible cases for which psychological games are the only adequate form of game theoretical representation. Also, the fact that the concept of psychological Nash equilibrium makes limited sense does not entail that psychological games as such make no sense. They just cannot be solved.

I raised three specific concerns with regards to our intuitions. The first is that temporally extended learning plays no role in psychological game models. This is worrying insofar as surprise-based feelings, which motivated GPS to model psychological games, require at least two temporally distinct events; namely that of the expectation-formation and the occasion on which the expectation is contradicted.

The second concern is that game theory as rational choice theory might be incompatible with modelling feelings that we have about each other's mode of choosing. If by assumption all moves are (known to be) rational, then other players' choices hardly seem to be the kind of things we would categorise as kind, nasty or fair. They would just be recognised as rational. But rational choice does not seem to be the kind of decision mode which we differentiate emotionally. Rather emotions seem to be applicable to action which is driven by non-endogenous, stable intentions, like intentional rule following. This point is strongly related to my criticism against HHV's amended version of Rabin's model. HHV introduce a notion of intentions on the level of material games which competes with the rational choice based equilibrium concept

on the level of psychological games. The intuitive and conceptual status of that model remains unclear.

The third concern is that I doubt whether our intuitions are equipped to grasp what it means to face process-dependence of utilities on top of an already complex utility definition which captures, e.g., other-regarding motives. The term 'material games' has done some damage by giving the impression that non-process-dependent utility functions are necessarily selfish or equivalent to monetary payoffs. Part of the challenge of justifying the notion of fairness equilibria as a broadly applicable notion, in fact, lies in arguing that process-dependence of fairness intuitions still makes sense when applied to a broader range of utility functions, including non-selfish ones.

Let me conclude by remarking on the game theorist's options, should he find that some motives in strategic interaction, if to be represented in game theoretic terms at all, can only be captured in psychological games. Since we have seen that psychological Nash equilibrium cannot be supported as a solution concept in these games without begging the question, two ways of escaping the task of solving psychological games come to mind. The first is to define the scope of game theory such as to exclude the problematic type of strategic situation. The other, more controversial, solution would be to impose a substantive normative principle on players, in so far as possible. Such a principle could be, for instance, substantive consequentialism, according to which intrinsic aspects of the process of choosing must never influence choice.

Chapter 4

Epistemic assumptions and counterfactuals in games

1. Introduction

The applicability of game theory to real social interaction depends on whether we can give plausible interpretations of the foundational concepts which are needed to first define meaningful games. Chapters 1 through 3 dealt with the interpretation and conceptual limits of some of these concepts, such as consequentialism and utility. But the usefulness of game theory, both as a normative and positive theory, also depends on whether we can conceptually support solution concepts to games. More specifically, the credibility of games depends on how solution concepts derive from epistemic assumptions in games and on how conceptually credible these assumptions are.

After Nash's (1951) breakthrough, the debate about solution concepts developed as a discussion about how to refine Nash equilibrium, which often seems to leave game theory with too many solutions for some games, with too few or implausible solutions for others. But a number of authors came to the conclusion that a systematic assessment of solution concepts requires an assessment of epistemic concepts and counterfactual reasoning in games. For example, it might be necessary to define what exactly is meant by assuming that there is common knowledge of rationality in a perfect information game and which solution concepts are supported by this assumption. The answer to this will depend on how knowledge is formally represented. It is also inseparable from how counterfactual reasoning, of the sort 'What would player *j* believe/know if player *i* were to deviate from the equilibrium?' is conceptualised.

The literature about these questions has become increasingly systematic. We will focus on a controversy between Aumann (mainly 1995) and Stalnaker (mainly 1994, 1996, 1998). These authors develop and discuss their positions explicitly. They are both concerned with the question of what assumptions support backward induction in perfect information games.

There are two ways to consider the differences between Aumann and Stalnaker. There is a level of formal representation and a level of conceptual interpretation. These levels are interdependent. The formalism, once set, limits the way a solution to a game

can be read in terms of the players' reasoning. The interpretation of what it means, e.g., that there is common knowledge of rationality might in turn be seen as a criterion for the design of the formal representation. According to Aumann, knowledge is necessarily true belief. Knowledge of an event implies the event; even knowledge about knowledge (etc.) about an event implies the event. Stalnaker suggests a definition of knowledge which avoids this implication. This allows him to grasp counterfactual reasoning about non-equilibrium play in games, which in turn is crucial to define equilibrium play. In particular, Stalnaker's framework allows him to differentiate between the semantics of the claim, e.g., 'If player *i* were to reach node *v*, he would choose rationally' and 'If player *i* reaches node *v*, he will choose rationally.'

I will lay out the two frameworks and discuss their respective conceptual background and implications. In doing so, I will defend Stalnaker's viewpoint. But I will offer an explanation for the source of the deficiency in Aumann's framework.

I will then address the question of which assumptions suffice to imply backward induction. According to Stalnaker, common knowledge of rationality does not suffice. Offering a remedy, he postulates an epistemic independence assumption which in combination with common belief in perfect rationality yields backward induction for perfect information games in which for each player different outcomes yield different payoffs. I point out the strength of the additional assumption and hint at what might work as a weaker alternative assumption to attain backward induction, all other assumptions being equal. This alternative assumption is an intuitive criterion about the permissibility of the conjunction of certain beliefs and belief revision policies.

The plan of this chapter is as follows. Section 2 introduces trends in the relevant literature. Section 3 explains Stalnaker's and Aumann's frameworks and some results. Section 4 relates these frameworks to the authors' respective conceptual positions on (common) knowledge of rationality in games. Halpern's (2001) and Samet's (1996) attempts to accommodate counterfactual reasoning in a partition structure representation of knowledge are discussed. Section 5 explains a crucial distinction between two kinds of conditionals in language and in game theory. Furthermore, it lays out two interpretations of what a game in game theory represents, which might explain the source of the weakness in Aumann's framework. Section 6 explores which assumptions are sufficient for backward induction to hold according to Stalnaker and Aumann, respectively. The strength of Stalnaker's sufficient assumptions motivates Section 7,

where a plausibility criterion regarding the permissibility of belief revision policies in conjunction with other assumptions is discussed.

2. Some literature

I will now point out some landmarks of the discourse on epistemic assumptions and counterfactuals in games. Doing so, I will confine myself to pre-1995 literature, so as to outline the setting of the discussion in the following sections.[83]

Problems with solution concepts are as old as game theory itself. However, solution concepts (most famously the Nash equilibrium) predate the discussion of epistemic assumptions and counterfactuals in games. This is why the more recent history of game theory is a history of successive refinements of solution concepts. Arguments regarding the epistemic assumptions and counterfactuals in games were first offered to support particular refinements to Nash (1951). Selten (1975), e.g., introduces "subgame perfectness" and "perfectness" to make use of his famous assumption of errors (in choosing) occurring with infinitesimally small probability, the "trembling hand."[84] It is used, of course, for making "unreached parts of the game" (p. 25) subject to assessment in terms of player's rationality. Counterfactual choice situations are simply replaced by unlikely events in this framework.

It is all the more surprising then that Selten and Leopold (1982) were among the first to proclaim that "[i]n order to do justice to the spirit of decision and game theory one must bravely face the issue of counterfactuals." (p. 191). They even briefly assess different theories of counterfactuals before proposing the (partially specified) "parameter theory" of counterfactuals themselves, an essentially probabilistic theory.

Parallel to those developments in game theory, modal and epistemic logic are developed. Prominently, Kripke (1970) develops a semantic approach to modal logic. Both syntactic and semantic approaches to modal logic will later play roles in the game theoretical literature. The treatment of subjunctive conditionals in modal logic is also related to the theory of counterfactuals by, for instance, Stalnaker (1968) and Lewis (1973).

[83] There is, of course, important post-1995 literature which I do not address, e.g. Battigalli (1996), Ben-Porath (1997), and Clausing (2004). The following account is non-exhaustive and loosely chronological.
[84] He also introduces the agent normal form.

Despite these developments, Aumann (1976) introduces the partition representation of knowledge, which will be crucial to his later work on counterfactuals in games, and which constitutes a limiting case of Kripke semantics.

The refinement contributions of the 1980's continue to raise questions about epistemic assumptions and counterfactuals in games. In their proposal of sequential equilibria Kreps and Wilson (1982), for instance, not only try to tackle the question of what equilibria are credible in games which have just one subgame (e.g. in 'Selten's horse'), but also which beliefs are consistent in situations which are expected to happen with probability zero (thereby going beyond the requirements of perfect Bayesian equilibrium, which does not demand beliefs to adhere to Bayes' rule in probability-zero scenarios).[85] As an alternative to the 'trembling hand' metaphor, Kreps and Wilson (1982) introduce the idea of consistent beliefs over pure strategies as beliefs over completely mixed strategies which converge to pure strategies. But while these authors make belief formation more explicit than Selten, they still resort to a method of dealing with counterfactuals which is similar in spirit, namely dealing with them in terms of low probability events.

Two further refinements underpin the view that Nash equilibria are too restrictive (in addition to being too weak a requirement in other games), namely correlated equilibrium (Aumann 1974) and rationalisability (Bernheim 1984, Pearce 1984). Later works by Aumann (1987) and Brandenburger and Dekel (1987) show that both of these notions can be captured as results of common knowledge of rationality and Bayesian updating (with common or differing priors respectively). But these contributions still lack an explicit discussion of how to model reasoning about off-equilibrium situations.

Bacharach (1987) finally introduces an (incomplete) theory of games in terms of a logic, composed of a first order predicate logic and operators from Hintikka's (1962) epistemic logic. He aims to capture the description of games axiomatically, including the players' knowledge, preferences, available choices and actions. He then proposes some general results on the definition and existence of solutions to games based on a 'class of theories' about games. Bacharach (1994) discusses the advantages of this approach, which he calls 'deductive,' as compared to semantic/Kripkean approaches as,

[85] The authors refer to the extensive form. We will not engage in the discussion of the relevance of normal-form and extensive-form representation to solution concepts. For a discussion compare Kohlberg and Mertens (1986).

for example, in Stalnaker (1994). He points out that the deductive approach allows for a better understanding of how agents come to knowledge, as this is mirrored by theorems about the game. Furthermore, he argues that semantic approaches are more prone to imply intuitively problematic notions of logical omniscience and "cloisteredness" (the limitation of knowledge to logical implications).

Binmore's (1987/1988) versatile discussions of how to model rational players touch on some of the same problem areas as Bacharach, but with a completely different focus and different ideas of ways to fix those problems. Among other things, Binmore is concerned with the lack of reality in assumptions of hyper-rational players; the notion of rationality that does not assess procedural aspects of reasoning; and the fact that counterfactual reasoning is needed to support equilibria and yet leads to logical inconsistencies on some accounts of common knowledge. Binmore proposes to abandon purely axiomatic approaches to modelling players in favour of a constructive algorithmic approach considering how Turing machines could optimally cope with games. Importantly, counterfactuals are read as possibilities of errors, which are (unlike in Selten) to be interpreted as correlated. This general approach was taken forward by Binmore and Shin (1992), who develop, among other things, a definition of knowledge for algorithmic players. Note at this point that this chapter is not concerned with questions of computational attainability and realism of rationality assumptions.

Bicchieri (1988) is also concerned with justifying rational equilibria based on a theory of belief formation in extensive form games. Starting from the refinement discussion, she suggests capturing intuitions from both the trembling hand and the sequential equilibrium in a theory which explicitly models how players would revise beliefs in counterfactual situations, building on contributions by Gärdenfors and Levi. Bicchieri (1989) refers to Binmore's argument that a certain conception of common knowledge of rationality (in conjunction with further parts of a theory of a game) leads to an inconsistency of that theory. She discusses sufficient beliefs to support backward induction and discusses the role of communication as a way to manipulate beliefs.

Pettit and Sugden (1989) make a related point with specific emphasis on the example of the finitely repeated prisoner's dilemma. They highlight that it is hard to picture players' consideration to leave the backward induction path when staying on it is a logical necessity by definition of common knowledge of rationality.

Bonanno (1991) makes an interesting contribution to the application of logic to extensive form perfect information games. He treats all conditionals as material.

Starting with formulae much weaker than common knowledge he defines rational solution concepts which neither imply backward induction solutions nor Nash equilibria. While the simplicity of this approach marks a limiting case of possible formalisations, it is vital to consider how faithful this representation is to intuitions about epistemic states of rational agents.

Expounding Aumann's (1976) representation of knowledge, Brandenburger (1992) develops a system of sufficient conditions for solution concepts in terms of knowledge of the structure of a game, rationality, choices, and belief conjectures. In the same year, Reny (1992a, 1992b) publishes a number of results on the logical impossibility of common knowledge (in Aumann's sense) in many perfect information extensive form games, threatening those solution concepts requiring this common knowledge assumption (e.g. backward induction, sequential equilibrium, rationalisability) for these games. Reny (1992c) tries to salvage sequential equilibrium from such inconsistency by weakening it to a requirement which comes closer to Selten's (1975) notion of perfectness.

Taking Lewis' (1973) notion of counterfactuals seriously, Shin (1992) introduces a metric of possible worlds in the possibility space, which helps him to deal with counterfactuals in decision theory and normal form games. He generalises his case to support a solution concept which amounts to that of Aumann (1987).

Lismont and Mongin (1994) discuss the distinction of syntactical and semantic approaches of dealing with modal logic in with games. They defend a syntactical approach whose expressions they consider as better "regimented" and understood. Lismont and Mongin establish soundness and completeness relations between Kripkean semantics and sets of axioms.

Note that my own discussion avoids systematic discussions of different modal logics. It is acknowledged though, that such debates might hold answers for some questions that are raised in this chapter. For instance, the relation of partition structure representations[86] and other Kripke semantics, like the one advocated by Stalnaker, may be more accessible when discussed syntactically.[87] But this lies outside the scope of this chapter. Let me instead explore how Stalnaker (1994) and Aumann (1995) address the debate about counterfactuals in games.

[86] It is the equivalent to the axiomatic system often referred to as S5.
[87] For a more recent overview compare, e.g., the introduction of Bacharach et al. (1997).

3. Two ways to think about counterfactuals in games

In this section, and indeed most of the remaining chapter, I will focus on the controversy between Aumann and Stalnaker about the role of counterfactuals in game theory. Both authors developed a formal and conceptual framework to address the topic. Furthermore, they address and contest each other's viewpoints, often referring to the same examples of games. This controversy climaxes in two prima facie contradictory theorems about the validity of backward induction as a solution concept under certain assumptions. While subsequent sections will scrutinise the sources of the controversy, this section sets up the problem.

Stalnaker's framework

The objective of Stalnaker's approach as developed in his 1994/1996/1998 papers is to provide a way to evaluate solution concepts in games by assessing their validity against so-called models of games. Stalnaker intends to move beyond simply postulating a number of solution concepts and subsequently assessing their plausibility (or the plausibility of the solutions generated) against a number of (sometimes conflicting) criteria, as it has been done throughout the history of refinements to the Nash equilibrium.[88] His idea is to develop a semantic of solution concepts as inspired by Kripke's (1963) systematic evaluation of alternative modal logics.

According to Kripke, the validity of a modal sentence in a formal language is to be defined as the sentence's truth relative to a class of models. Roughly a model specifies features of the actual and other possible worlds. Theorems are to be accepted only if a corresponding class of models, in virtue of which the theorems are valid, is intuitively supported or considered plausible. Stalnaker states that the validity of a solution concept should be determined analogously. A solution concept is to be accepted if and only if a class of models of a game is accepted, such that in all models of the class one of the strategy profiles generated by the solution concept is realised and that each strategy profile generated by the solution is realised in at least one of these models. He argues that "A solution concept defines a set of strategy profiles, as a logic defines a set of theorems." (1994, p. 49).

A model of a game is to be thought of as a representation of a particular situation, in which the game is set, and of factors determining how the game is played.

[88] An earlier attempt to a systematic assessment of solution concepts goes back to Kohlberg and Mertens (1986).

Accordingly, model classes can be thought of as types of such situations, which are equal in specified ways. As said above, a model may specify several possible worlds in addition to the actual world. The play in the actual world can thus be justified in a sense to be specified on the basis of what would happen in other possible worlds. This does not seem so awkward once we consider, e.g., that a rationalisation (used in a non-technical sense here) of our own action in a situation usually specifies what would have happened had we acted otherwise.

At this stage, I should outline which parts of Stalnaker's approach to game theory are orthodox in game theory's history of thought and which parts dissent from this orthodoxy. Stalnaker is in line with orthodoxy when it comes to the structure of games, notions of payoffs[89] etc. Methodological individualism and consequentialism are endorsed as starting points, or the scope for that matter, of game theory. Furthermore, the notion of rationality is orthodox in the sense that it is equivalent to expected utility maximisation, independently of whether such rationality leads to socially suboptimal outcomes etc. So Stalnaker has quite a different methodological agenda from, e.g., Gauthier (1986) and McClennen (1990). He analyses game theoretic situations as compounds of multiple decision theoretic problems:

> "[...] I assume that all of the decision problems in the game are problems of individual decision making. There is no special concept of rationality for decision making in situations where the outcomes depend on the actions of more than one agent."[90]

Still, the notion of expected utility maximisation is quite complex in Stalnaker because the players in his models determine their expectations in a different way than is the case in game theoretical orthodoxy. This is done because Stalnaker distinguishes between two kinds of counterfactuals, as will be explained later. One of the upshots of this distinction is that even when a proposition that renders a certain situation counterfactual is *known* (in some sense), this proposition can potentially be revised once information is attained to the effect that the supposedly counterfactual situation actually occurred. More will be said later. What matters at this stage is that to accommodate the possibility of belief revision in the presence of knowledge, Stalnaker also needs a notion of knowledge (in addition to an account of probabilistic belief) which leaves enough

[89] See qualifications of the notion of preference in Chapter 1.
[90] Stalnaker (1996), p. 136.

'conceptual space' for representing the coexistence of knowledge and belief revision policies. Although game theory textbooks are usually not explicit about their underlying notion of knowledge, Stalnaker's account seems to go against what might be *implicitly* assumed in these sources. To become more specific about Stalnaker's framework, and some of his basic results, we need to introduce some of his formal notation.

The structure of a game is captured by the usual expression $\langle N, \langle C_i, u_i \rangle_{i \in N} \rangle$.[91] N is the set of players, C_i is player i's set of available strategies and u_i is his utility function. The structure of a model of a game is represented as $\langle W, \boldsymbol{a}, \langle R_i, P_i, S_i \rangle_{i \in N} \rangle$, where W is the set of all possible worlds w, and $\boldsymbol{a}$ is the actual world within W. S_i is a strategy function which assigns a strategy from the set C_i to each possible world in W, so that $S_i(w)$ represents player i's strategy in the possible world w. The expressions P_i and R_i represent player i's beliefs; R_i are the full (subjective, probability-1) beliefs and P_i the partial (subjective, probability lower than one) beliefs.

R_i is a binary relation between two possible worlds x and y, such that player i's beliefs in world x are subjectively compatible with another world y. It can be thought of as an epistemic accessibility relation. So the set of i's beliefs in world x can be represented as a set of possible worlds y, i.e. $B_i(x)=\{y:xR_iy\}$.[92] Here xR_iy expresses that world y is subjectively compatible with world x for i. But if to worlds are subjectively compatible, the two worlds imply the same strategies for the player, so if xR_iy then $S_i(x)=S_i(y)$. Also, the necessary and sufficient conditions for two worlds to be subjectively indistinguishable for player i, expressed as $x \approx_i y$, can be formulated as $(z)(xR_iz \leftrightarrow yR_iz)$. In words, subjective indistinguishability between two worlds x and y is represented by those worlds which are subjectively compatible with both x and y.[93] Furthermore, belief in some proposition ϕ can be represented as the set $\{x:\{y:xR_iy\} \subseteq \phi\}$, i.e. that set of possible worlds for which the set of epistemically compatible worlds lies within ϕ.[94] R_i is limited by the following four assumptions, whose significance I will shortly touch upon. Any R_i is

[91] Throughout this chapter the original notations are adjusted only so much as to obtain some minimal conformity. The account refers mainly to Stalnaker (1994), but continues to include amendments from (1996) and (1998). In the following $(x)\ldots$ stands for $(\forall x)\ldots$ etc. Also the personal indices i and j are neglected for some of the expressions (such as Q, B and R) for reasons of simplicity.

[92] Compare Stalnaker (1996), p. 146.

[93] Compare Stalnaker (1998), p. 34.

[94] Compare Stalnaker (1996), p. 156.

- *not necessarily reflexive*, i.e. *not* $((x)xR_ix)$ Otherwise it would imply the impossibility of false beliefs which would have unwanted consequences for Stalnaker's attempt to represent counterfactual reasoning.
- *serial*, i.e. $(x)(\exists y)xR_iy$. So being in world x, there exists at least one possible world y for player i which is compatible with his beliefs.
- *euclidean*, i.e. $(x)(y)(z)((xR_iy \wedge xR_iz) \rightarrow yR_iz)$. This implies the positive introspection assumption that if something is believed in a possible world then it is believed to be believed in any accessible (i.e. compatible) world.
- *transitive*, i.e. $(x)(y)(z)((xR_iy \wedge yR_iz) \rightarrow xR_iz)$. This amounts to negative introspection. If something is not believed in a world then it is not believed in worlds accessible from this world either.

The expression P_i is a probabilistic measure of a player's belief in each possible world, such that this probability can be added up over possible worlds. This fact is exploited when expressing player i's partial beliefs in proposition ϕ in world x as the ratio $P_{i,x}(\phi) = P_i(\phi \cap \{y: xR_iy\})/P_i(\{y: xR_iy\})$. This notion is of limited significance in Stalnaker's analysis here. However, it should be noted that while the expression $P_{i,x}(\phi)$ follows the definition of P_i, P_i itself is an exogenously defined additive measure which gives some positive value to each subset of W. How the P_i's themselves come about is not discussed in Stalnaker's framework.

The above notion of belief allows Stalnaker to formalise common belief elegantly. Common belief in ϕ is defined as the set $\{x:\{y:xR^*y\} \subseteq \phi\}$, where R^* is the transitive closure of the set of all and everyone's R-relations.[95] One intuitive implication of this definition is that within ϕ no chain of beliefs (about someone's beliefs, etc.) can ever 'reach' a compatible possible world outside the event ϕ. So if we define A to be the event that everyone is rational[96], e.g., the proposition that there is common belief that everyone is rational is: $Z=\{x \in W:\{y \in W:xR^*y\} \subseteq A\}$.

Stalnaker does not settle for this structure of models, but enriches it by yet another relation, Q_i, making R_i redundant. Hence the more comprehensive structure of a model of a game is $\langle W, \boldsymbol{a}, \langle Q_i, P_i, S_i \rangle_{i \in N} \rangle$. The Q-relation is imported from AGM belief revision theory into game theory.[97] When given some intermediate information ϕ, Q regulates the relation between prior and posterior beliefs by ordering all possible

[95] It is the superset of the R-relations such that whenever (x, y) and (y, z) are in the superset, (x, z) is also in the superset.
[96] See Stalnaker 1996, p. 142, for a standard definition of expected utility maximisation.
[97] 'AGM' stands for Alchourron, Gärdenfors and Makinson. Compare Gärdenfors (1988).

worlds. In this respect, it is the basis for a belief revision function. It is the job of Q to define the belief revision function[98] $B(\phi)$ such that four conditions are met.[99]

(1) New information is believed. So, $B(\phi)$ must be a subset of ϕ; i.e. $B(\phi) \subseteq \phi$.

(2) Prior beliefs should be preserved as far as possible, so that if at least one world lies in both B and ϕ (i.e. $B \cap \phi$ is nonempty) then the posterior beliefs are just the intersection $B(\phi) = B \cap \phi$.

(3) New information always leads to some consistent posterior belief state, i.e. if ϕ is nonempty so is $B(\phi)$.

(4) The order of occurrences of new information does not matter to a posterior belief state so that $B(\psi \cap \phi) = B(\psi) \cap \phi = B(\phi) \cap \psi$.

Q is defined in the following way: xQy if and only if $y \in B(\{x, y\})$. In words, world y is epistemically weakly more plausible to world x exactly in those cases in which y is a member of the set which is itself the outcome of the belief revision based on that information containing both x and y. It follows from conditions (1)-(4) that Q is a reflexive, transitive and connected ordering of all possible worlds.[100] Inversely, we can define $B(\phi)$ in terms of such an ordering Q and show that (1)-(4) hold. That is for any ϕ, $B(\phi) = \{x \in \phi$: yQx *for all* $y \in \phi\}$.

As said above, Q, the epistemic prioritisation relation, makes R, the relation defining prior beliefs, redundant. It does so in virtue of two biconditionals,

(5) $x \approx_i y$ if and only if ($xQ_i y$ or $yQ_i x$) and

(6) $xR_i y$ if and only if $wQ_i y$ for all w such that $w \approx_i x$.

(5) redefines subjectively indistinguishable worlds as those worlds for which an epistemic plausibility function exists (and needs to exist). (6) expresses that prior beliefs are exactly those beliefs which Q ranks equally as the highest in all indistinguishable

[98] Set B is supposed to be the set of possible worlds representing prior beliefs, defined as the worlds with the highest epistemic priority. Now if B' is the (super)set (containing all elements of B) of all possible worlds which are compatible with any potentially occurring new information, then the set $B(\phi)$ of posterior beliefs given some specific new information ϕ is a subset of B'.

[99] Compare Stalnaker (1996), p. 144. I adopt Stalnaker's notation. $B(\cdot)$ represents the belief revision function while B, not followed by an argument, represents the prior belief set.

[100] The proof is not given. Compare Stalnaker (1998), p. 38.

worlds. So Q only makes a difference among beliefs incompatible with the prior beliefs.[101] This is why prior beliefs can be expressed in terms of epistemic priorisation.

The richer framework allows for expressions of more complex propositions which turn out to be relevant in Stalnaker's take on game theory. For one thing, "perfect rationality" (1996, p. 149) can be defined. A perfectly rational agent makes the utility maximising choice even on occasions which only arise if the agent was in error about something. A first-degree error in some possible world is not to hold everything that is true in that world as a prior belief, so it is the set $E_i^1=\{x \in W$: *not* $xR_ix\}$. In these new terms, Stalnaker can define an (at least) first-degree error as $E_i^1=\{x \in W$: *for some* $y \in W$ *such that* $x \approx y$: *not* $yQ_ix\}$; it is the set of worlds for which it is not the case that all subjectively indistinguishable worlds can be found at the top of the Q-ranking. Higher degree errors are defined recursively, so an (at least) k+1 degree error is $E_i^{k+1}=\{x \in E_i^k$: *for some* $y \in E_i^k$ *such that* $x \approx y$: *not* $yQ_ix\}$.

Stalnaker exploits this expression to define perfect rationality recursively as well. Rationality in a state without error is conventionally defined as expected utility maximisation $r_{i,x}^0 = r_{i,x} = \{s \in C_i$: $eu_{i,x}(s) \geq eu_{i,x}(s')$ *for all* $s' \in C_i\}$. Rationality in a state with at least a $(k+1)$-degree error is $r_{i,x}^{k+1} = \{s \in r_{i,x}^k$: $eu_{i,x}(s|\ E_i^{k+1}) \geq eu_{i,x}(s'|\ E_i^{k+1})$ *for all* $s' \in r_{i,x}^k\}$. Player i's perfectly rational strategies can thus be expressed as $r_{i,x}^+ = \{\cap r_{i,x}^k$ *for all* k *such that* $E^k \cap \{y$: $x \approx y\}$ *is non-empty*$\}$, meaning that i's strategy choice would be expected utility maximising for any imaginable degree of error. It is of interest for sections to come, that perfect rationality is Stalnaker's approximation of what Aumann's notion of "substantive rationality" (1995, p. 16) attempts to capture.

Furthermore, the Q-relation allows Stalnaker to define his "defeasibility conception of knowledge" concisely. As the notion is to be discussed at a later point, only so much shall be said here: The defeasibility conception of knowledge (henceforth knowledge^D) describes a coincidence of belief and truth that continues to hold throughout any possible appearance of evidence in that world. It is a true belief that cannot be defeated by any true information. So 'i knows^D ϕ' is the set $\{x:\{y:xQ_iy\} \subseteq \phi\}$, which is structurally parallel to 'i believes ϕ' defined as $\{x:\{y:xR_iy\} \subseteq \phi\}$. Accordingly, 'it is common knowledge^D that ϕ' is the set $\{x:\{y:xQ^*y\} \subseteq \phi\}$, where Q^* is the transitive closure.

[101] Compare Stalnaker (1996), footnote 11.

Characterisations

With the help of the structure introduced so far, Stalnaker is also able to define some ways to single out relevant classes of models, which are then shown to be equivalent to the application of certain solution concepts. The most obvious maybe is (simple) "rationalisability"[102], i.e. the class of models in which there is common belief in rationality. If on top of common belief of rationality, there is common belief in no error (in terms of beliefs about each other) and there actually is no error, then the class of models is characterised by "strong rationalisability."

More precisely, Stalnaker's formal expression for strong rationalisability requires that all R_i relations in the transitive closure $\{w:aR^*w\}$ be reflexive (in addition to the usual assumptions), which makes them equivalence relations. He interprets this as the assumption that "it is common belief that no players have any false beliefs" (1994, p. 61). But the formal expression is stronger: Everything that is commonly believed in the actual world is also true.[103] Note also that the above requirements do not imply that *all* beliefs are common belief; there might be ignorance about the others' beliefs, for example.

Both model classes can be narrowed down further by replacing 'rationality' with 'perfect rationality.' This defines "perfect rationalisability" and "strong perfect rationalisability" (1996, p. 155).

In his 1994 paper, Stalnaker proves a number of characterisations of solution concepts (i.e. certain sets of strategy profiles) in terms of model classes. As sketched earlier, a characterisation is the following two-way relation: In the actual world of each model of the class, the strategy profiles played are members of the characterised strategy profile set; and for every characterised strategy profile in the set there is model in the class such that in this model's actual world the strategy profile is realised. [104]

Firstly, (the class of models representing) rationalisability characterises the solution algorithm of iterated elimination of strictly dominated strategies for games in strategic form (*Theorem 1* in Stalnaker (1994)). Further interesting results concern the relation of Nash equilibria and model classes. Iterated elimination of inferior[105] strategy profiles is characterised by strong rationalisability (*Theorem 3*). A basic result is that

[102] Compare a similar notion of 'rationalisability' introduced by Pearce (1984) and Bernheim (1984).
[103] Compare (1994), p. 61, and (1996), p. 155.
[104] Compare Stalnaker (1994), p. 55.
[105] A strategy profile c is inferior relative to some subset of the set of strategy profiles of the game if and only if for some player i there exist a (possibly mixed) strategy which generates equal or more utility than c_i given the strategies of the other players. Compare Stalnaker (1994), p. 62.

Nash equilibria are characterised by models in which each player knows the other's beliefs about his strategies and knows that the other is rational (*Theorem 2*)[106]. This is because the model class constrains the space of possible worlds to those in which players maximise utility, while the probability measures P_i are defined to be the same in all worlds.

The assumption of knowledge which is implied by reducing the set of possible worlds in this way is stronger than the belief assumptions implied by strong rationalisability (although, on a different account complete knowledge does not imply common belief in rationality)[107]. But for a specific type of game, perfect information games, Stalnaker (1996) declares that "all and only Nash equilibrium strategies are strongly rationalisable" (p. 160).[108] The statement had to be weakened in order to hold true. In fact, only for perfect information games in which every subgame has a unique Nash equilibrium (or only equivalent and interchangeable equilibria), every strongly rationalisable strategy profile is a Nash equilibrium. This condition is met for zero-sum games.[109] The other direction of the original claim, i.e. that all Nash equilibria are strongly rationalisable, can, however, be proven by stipulating a model containing a single possible world for each equilibrium.[110]

A final important result of Stalnaker's framework is the interrelation of model classes with respect to the solution concepts they characterise. Generally speaking, the fact that one class is a strict subset of another does not imply that the narrower class characterises a stronger solution concept.

The class of (models representing) common belief in rationality includes the class of common knowledge^D of rationality, which in turn includes the class of strong rationalisability. Accordingly, the class of common belief in perfect rationality includes the class of common knowledge^D of perfect rationality, which in turn includes the class of strong perfect rationalisability. However, the strategy profiles realised under common belief in rationality are the same as those characterised under common knowledge^D. In fact, it can be shown that any strategy profile realised in a model of the first class is also

[106] This is proved for 2-person games, but can be generalised for *n*-person games. See Stalnaker (1994), p. 60.
[107] See the counterexample in remark (1) in Stalnaker (1994).
[108] A defective proof can be found in Stalnaker (1994).
[109] This remark is taken from a personal exchange with Robert Stalnaker. The mistake in the 1994 paper was pointed out in a revised version of the paper in Bacharach et al. (1997).
[110] Remarkably, Stalnaker does not comment (in the context of this proof) on the fact that single-world models do not allow one to represent hypothetical reasoning. However, he does not advocate strong rationalisability as a model class in particular either.

realised in a model of the latter class; so the latter is not more restrictive regarding the solutions of a game.[111] Let me offer an intuitive way to grasp this: Knowledge^D only constitutes a stronger constraint than common belief in combination with *perfect* rationality, because only perfect rationality guarantees that in unexpected worlds (which will not *actually* occur according to knowledge^D) expected utility will be maximised.

Accordingly, the strategy profiles realised under common knowledge^D of perfect rationality are the same as those characterised under strong perfect rationality. This is because knowledge^D of *perfect* rationality by definition implies that belief in rationality needs no revision even if new information comes to light.

Let me graphically represent this by the following figures depicting sets of strategy profiles (not classes of models). They illustrate which characterisations have the same 'extensions' in terms of realised strategy profile for all games, and which characterisations are more restrictive as others.[112]

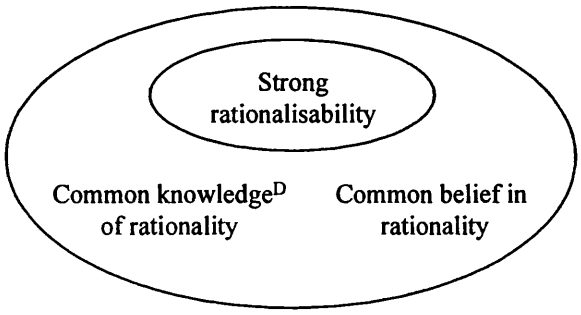


Figure 4.1

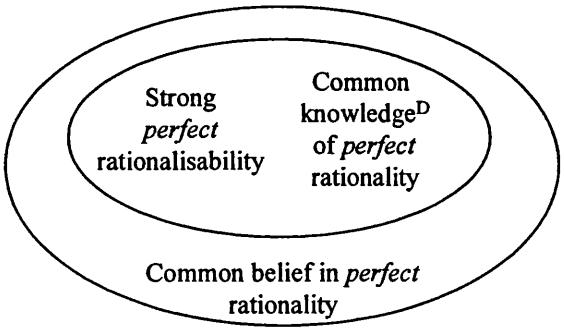


Figure 4.2

Aumann's framework

Aumann's (1995) framework is technically simpler and closer to the game theoretical mainstream.[113] Let us start with his formal notion of knowledge. There is a set of possible worlds[114], W, which is structured by a partition $\mathcal{K}_i$ for each person i. So, for i, subjective distinguishability of two possible worlds is equivalent to these two worlds being in different atoms of the partition. An *event* E is a set of possible worlds, while i's

[111] Stalnaker (1996), footnote 19, outlines the idea of the proof for this.
[112] The strict inclusions indicated in this figure are actually weak inclusions for some games.
[113] As mentioned earlier, it can be shown to be syntactically equivalent to a system (S5), which makes more restrictive assumptions about the epistemic accessibility relations than Stalnaker's system.
[114] Aumann uses the expression "possible states of the world" (1995, p. 8).

knowledge of an event is represented by the union, labelled K_iE, of those atoms of $\mathcal{K}_i$ which are in E. Such knowledge is an event itself. The event that everyone (all persons $i=1, 2, ..., n$) knows E, labelled KE, is thus the intersection of events $\cap_i K_iE$. So common knowledge of E, CKE, is the intersection $KE \cap KKE \cap KKKE \cap ...$ etc. *ad infinitum*. Note that one implication of this framework is that knowledge of an event logically implies the event.

One such event E could be that player i is rational. Everyone being rational is the intersection of all these events. As usual, rationality is defined as utility maximisation.

Note that a complete *knowledge system*, Aumann's counterpart to Stalnaker's notion of a model, also is completed by a function which maps the states of the world, on the players' strategy profiles. Strategies are distinguishable for i, which means that they are the same if they fall into the same atom of $\mathcal{K}_i$.

Furthermore, Aumann explains that his framework can be expanded such as to capture probabilistic knowledge. To do so one would have to specify a probability distribution over each atom of $\mathcal{K}_i$, specifying the probabilities that i assigns to each possible world at each atom. The corresponding notion of rationality would be the usual expected utility maximisation.[115]

Refining his notion of rationality, Aumann points out that the proposition that a player is rational, as often applied in assumptions for solution concepts, is meant as a "substantive"[116] claim. Like Stalnaker, Aumann sees a conceptual challenge in the fact that the material conditional makes it impossible to justify an equilibrium by reference to counterfactual turns of a game. Any such materially conditional claim is true, since its antecedent is false. Aumann thinks that we need to make a statement of "subjunctive mood" (p. 13) about counterfactual decision nodes in order to express, e.g., that a player is rational in game. However, Aumann does not provide a formal representation of subjunctive conditionals in his framework. He merely refers to Samet (1996), whose

[115] On the other hand, Aumann mentions at a later point in his paper (p. 18) that there are problems involved in replacing knowledge as certainty with probability-1 belief. If certainty is replaced by probability-1 belief, ex post rationality of a move (assessed at the time of a choice) need no longer imply its ex ante rationality (assessed before the game). Aumann writes: "If you knew something at the beginning of play, then you certainly know it later; but if you ascribe probability 1 to it at the beginning of play, you need not do so at a later vertex, if that vertex initially had probability 0. One can try to fix this, but then something else unravels. [...]" (1995, footnote 11). But it is precisely this kind of potential for belief revision at unexpected choice nodes which takes us to the heart of the conflict between Stalnaker and Aumann, as will be addressed below.

[116] Note that Binmore (1987) uses the word differently.

account we will address soon.[117] So Aumann verbally defines substantive rationality as habitual utility maximisation. He postulates that the substantively rational player will maximise utility wherever the course of the game could hypothetically take him, i.e. at every node of a game tree. It remains to be seen whether knowledge of substantive rationality can be expressed meaningfully in Aumann's formal account at all.

Common knowledge of rationality and backward induction

On the basis of these definitions, Aumann proves that common knowledge of rationality implies the backward induction solution.[118] The form of the proof is reminiscent of the verbal argument for backward induction that is presented to novices in game theory. The proof proceeds inductively against the direction of the game showing that the backward induction path is taken at all nodes, qua common knowledge of rationality. The base case is the backward induction choice at the last nodes; the inductive step shows that if the backward induction choice was made at all subsequent nodes it is made at the current node etc.

Stalnaker (1998), however, seems to prove that Aumann's theorem is wrong. He gives an example of a non-subgame-perfect Nash equilibrium in which there is "common knowledge" that players are "substantively rational," i.e. rational even in choices which are counterfactual. As this example could not be represented in Aumann's formal framework, Stalnaker refers to it as a "prima facie counterexample" (p. 47). (Note also that Stalnaker (1996) constructs an alternative version of the counterexample with a different model of the same game, namely by assuming perfect rationalisability.[119] The main difference between the two versions of the counterexample is that the definition of perfect rationality presupposes the formal representation of possible worlds other than the actual world.)

Consider Figure 4.3. There are three Nash equilibria in pure strategies in the unreduced normal form of this game, namely $\{A_1, A_2; a\}$, $\{D_1, A_2; d\}$ and $\{D_1, D_2; d\}$. Only one of these is subgame perfect, namely $\{A_1, A_2; a\}$. But there is a way to defend

[117] Compare Aumann (1995), footnote 9.

[118] Compare Aumann (1995), pp. 10-12. Aumann also proves (see *Theorem B*) that, for all perfect information games, there is a knowledge system such that common knowledge of rationality exists. It is, to stipulate one example for all such games, the system in which there is just one possible world, namely that in which players choose according to backward induction. Possibly, Aumann considers this result important because there are arguments that the notion of common knowledge of rationality is inconsistent in some games. Compare, e.g., Bicchieri (1989). The problem with Aumann's *Theorem B* is that it stipulates a single possible world, which forbids any counterfactual argument.

[119] Compare Stalnaker (1996), p. 151.

another Nash equilibrium as compatible with common knowledge of substantive rationality, namely $\{D_1, A_2; d\}$. For the purpose of his argument, Stalnaker stipulates this equilibrium to be the only possible state. Furthermore, Stalnaker stipulates the following belief revision: Bob stops believing that Ann is rational once the second node is reached. [120] Actually he now believes that Ann is irrational in the sense that she will choose the option with lesser expected utility, once faced with a choice.[121] The justification for his belief revision is that Ann knows that Bob will play *d* once at node 2, which would leave her a maximum of 1. She knows this because she knows his belief revision policy. (More will be said about this circular use of the belief revision policy later.) A payoff of 1, however, is less than the payoff of 2 she could secure by choosing D_1. Hence Ann would be irrational when choosing A_1.

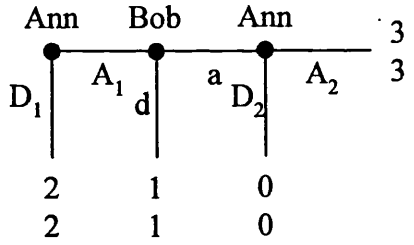


Figure 4.3

So let us check for substantive rationality in Stalnaker's fashion. Given that Ann believes that Bob would play *a* at node 2, Ann is substantively rational because she chooses D_1 at node 1. She is also substantively rational at node 3, playing the expected utility-maximising option A_2. Bob is substantively rational because at node 2 he believes that Ann will play D_2, leaving him with a payoff of 0, rather than the payoff of 2 he could secure by choosing *d*.

So what do we make of the prima facie contradiction between Aumann's and Stalnaker's proofs? The crux of the 'contradiction' lies in the formal representation of common knowledge and rationality. Before I move on to assess these representational

[120] Formally, there are no possible worlds to represent this reasoning. But Stalnaker's point is exactly that without sufficient logical possibilities in the formal representation, a lot can be stipulated verbally. I will come back to this point. Again, for a version of Stalnaker's argument that uses several possible worlds, see Stalnaker (1996), p. 151.

[121] It is not a priori clear what one should think upon learning that someone did not act rationally. One might infer that this person is an expected utility *minimiser*, or that there is a risk (with defined probabilities) of an irrational choice or that there is uncertainty.

differences between Stalnaker and Aumann, let me investigate how deep the conceptual differences between the two authors go.

4. Two ways to think about knowledge and rationality

Stalnaker (1996) writes about his stance on knowledge and belief in games extensively. He argues that the partitional representation of knowledge violates our intuitions regarding the possibility of assessing the truth of a proposition by introspection:

> "If we conflate knowledge and belief, assuming in general that *i* knows ϕ if and only if *i*'s degree of belief for ϕ is one, then we get a concept that combines the introspective properties appropriate only to the internal, subjective concept of belief with the success properties appropriate only to an external concept that makes claims about the objective world. The result is a concept of knowledge that rests on equivocation." (1996, p. 153)

According to Stalnaker, necessarily true belief ceases to be belief in a plausible sense at all, since it implies that introspection informs us about the truth or falsehood of our beliefs. Stalnaker deems it implausible to assume that someone can infer the truth of a proposition ϕ from the fact that he knows ϕ, or even from the fact that he knows that someone knows that ϕ. Crucially, Stalnaker criticises that with such a notion of belief "there can be no disagreement[122], no surprises, and no coherent counterfactual reasoning" (p. 154), because all true beliefs would be non-revisable and the agent would know this.[123]

As a conceptual alternative to Aumann's notion of knowledge, Stalnaker offers the actual (instead of necessary) coincidence of belief and truth, i.e. the coincidence in the actual rather than all possible worlds. All that is actually commonly believed is true and commonly believed to be true. This is the definition of strong (perfect) rationalisability. For Stalnaker, this is a way to make belief and knowledge coincide without equivocating them. While for the theorist false beliefs are still conceivable for non-actual worlds in this framework, it is assumed that *actually* all beliefs are true. That

[122] This expression makes reference to Aumann (1976). He is unimpressed by Aumann's result that agents with the same priors cannot agree to disagree (on posteriors), because he believes that the result is stipulated more or less trivially by the choice of the framework. Compare Stalnaker (1996), footnote 15.

[123] It could be misunderstanding to say that there could be *no* coherent counterfactual reasoning under such assumptions. As I will discuss later in this subsection, Halpern delivers an extension of Aumann's framework, which even makes it possible to defend Stalnaker's counterexample based on a partition structure notion of knowledge.

is, in the actual world, all that is commonly believed is true. On the basis of these assumptions Stalnaker can argue against Aumann's backward induction theorem.

However Stalnaker defends an even weaker notion of knowledge that allows for coexistence of false beliefs and knowledge in the actual world. This is knowledge^D. Remember that according to its definition, knowledge^D is the set of possible worlds such that no information that is compatible with the actual world could alter the prior belief (of probability one). This notion is weaker than strong rationalisability because it makes it possible to define situations in which some other ϕ is known^D while some φ is falsely believed.

Also, introspection does not have the power to inform an individual about the truth of his beliefs. Having said that, no actual experience (in the sense of new information) could prove knowledge^D wrong.

But how does the notion of substantive rationality fit into this debate? Aumann simply defines substantive rationality as utility-maximisation at each (hypothetically reached) vertex. Aumann interprets his definition in terms of players programming automata for each decision node of the game before the game starts:

> "The idea is that when programming his automaton at *r*, the player does so as if *r* will be reached – even when he knows that it will not be! Each choice must be rational 'in its own right' – a player may not rely on what happened previously to get him 'off the hook.'" (1995, p. 12)

While Stalnaker agrees that the meaning of substantive rationality must depend on subjunctive conditionals that start with the antecedent 'if player so- and-so where at (counterfactual) node such-and-such,' he also thinks that Aumann oversimplifies the matter. He believes that one should even revise 'knowledge' about some future event once one finds out that the more fundamental 'knowledge' (e.g. about someone's rationality) from which the prediction was derived is in conflict with observations made. In terms of the example of Figure 4.3, Stalnaker's point is this:

> "Assume that since Bob knows that [Ann] chooses D_1A_2, he knows even if she doesn't choose D_1, she will still choose A_2. The idea seems to be that Bob should reason as follows: 'I know that [Ann] will choose D_1A_2. The only part of her strategy that is relevant to my choice is A_2. So I will assume that if my node is reached, she will choose A_2.' But why should Bob assume this if he knows that if his node is reached, than [Ann] did *not* choose D_1A_2." (1998, p. 48, italics in original)

The formulation of this critique is mind-boggling. But Stalnaker would reply that this is so only because Aumann does not even provide an adequate formal framework to accommodate the concept of subjunctive reasoning and substantive rationality that he informally advocates. Without amendment, the partition structure of knowledge makes it hard to represent reasoning about states which are excluded by what is known. There could not be any kind of learning, not even counterfactually.

Stalnaker's remedy comes in two steps. First, he opts for a weaker notion of knowledge. Secondly, he suggests distinguishing (i.e. not conflating) the questions of what one thinks about a counterfactual situation and what one would think (if one were) in that situation. This will be further discussed later.

Stalnaker represents his intuition about substantive rationality in the form of perfect rationality. As explained earlier, a substantively rational player chooses rationally even if it turns out that he was in error about something. This holds for all orders of error.

While this is a quite demanding assumption, it is still much less restrictive than substantive rationality in Aumann's account. According to Stalnaker, hypothesising about an error means to imagine oneself in a situation, e.g. a decision node, which was not believed to occur. This might trigger a belief revision which changes what qualifies subsequent choices as rational. Aumann does not reason like this. Translating his framework into Stalnaker's terms, it is as if Aumann always holds what counts as rational fixed even when hypothesising about states which contradict the initial knowledge, relative to which the choices were deemed rational.

Aumann also touches on this issue of time-dependent assessment of rational choices. He distinguishes between *ex ante* and *ex post* rationality. The first refers to the knowledge at the beginning of the game, when strategies are programmed; the latter is relative to knowledge at a given subsequent point of the game when the player has to choose. Importantly, Aumann claims that ex post rationality implies ex ante rationality. This would make the latter the weaker assumption and in this sense the more desirable assumption for a theorem. His argument is that "when the time comes for a player to move, he certainly knows at least as much as he did when play started." (p. 17) But this is true only in the sense that (presuming perfect memory) knowledge can only increase. As far as the points in the game at which ex post rationality is assessed, however, are counterfactual, i.e. contradict what is *known* to be the case, one cannot assume such monotonic increase of knowledge. Some things which were 'known' would have to be

accepted as false. But that means that ex post rationality does not imply ex ante rationality without further assumptions.

Halpern's synthesis

Halpern (2001) has tried to capture the essence of the difference between Aumann and Stalnaker in formal terms. His formulation of the difference might seem surprising as Halpern claims that "Stalnaker's result can be obtained using exactly the same definition of (common) knowledge and rationality as the one Aumann (1995) used." (p. 426) In fact, Halpern proceeds to reconstruct Stalnaker's 'counterexample' with a partition structure representation of knowledge and Aumann's concept of (substantive) rationality as utility maximisation at each node of a tree.

The crucial supplement to Aumann's framework that allows him to generate Stalnaker's result, however, is a selection function. In fact, this selection function is the only difference between Aumann's model and Stalnaker's extended model. A selection function selects the closest possible world for a given node and a given possible world. It is a function from the set of possible worlds and the set of nodes in a game to the set of possible worlds. For instance, it could express that at some node υ in world w, the possible world w' is closest, i.e. $f(w, \upsilon) = w'$. The so-selected world is to be taken as the scenario relative to which rationality is assessed. For instance, a player in world w at node υ is rational if he chooses so as to maximise utility in world $f(w, \upsilon) = w'$. The selection function in Halpern (2001) is not player-specific, but there could be a number of such functions indexed for the players.[124]

The example states five strategy profiles.[125] In terms of Figure 4.3, they are $s_1 = \{D_1, A_2; d\}$; $s_2 = \{A_1, A_2; d\}$; $s_3 = \{A_1, D_2; d\}$; $s_4 = \{A_1, A_2; a\}$; $s_5 = \{A_1, D_2; a\}$. There are also five possible worlds, such that in w_1 strategy s_1 is played, in w_2 strategy s_2 etc. Ann always knows in what world she is, so $K_{Ann}(w_i) = \{w_i\}$ for all worlds i. Bob usually does too, but he cannot distinguish worlds 2 and 3, so $K_{Bob}(w_2) = K_{Bob}(w_3) = \{w_2, w_3\}$. The selection function is defined by $f(w_i, \upsilon_1) = w_i$ for all i; $f(w_1, \upsilon_2) = w_2$; $f(w_i, \upsilon_2) = w_i$ for $i = 2,...,5$; $f(w_1, \upsilon_3) = f(w_2, \upsilon_3) = w_4$; $f(w_3, \upsilon_3) = w_5$ and $f(w_i, \upsilon_3) = w_i$ for $i = 4,5$. Crucially, all players are substantively rational in w_1. Once υ_2 is reached, players would be in w_2, which Bob cannot distinguish from w_3. So he considers it possible that Ann will play

[124] See footnote 3 in Halpern (2001).
[125] Compare Halpern (2001), pp. 429-430.

D_2. Hence Bob chooses substantively rationally. The selection function $f(w_1, \upsilon_3) = w_4$, however, informs us that Ann would not do so. Hence Ann chooses substantively rationally.

How can we make sense of selection functions intuitively? Are they (inter)subjective or objective? Are they about what rationality *means* or about how beliefs are updated? Before answering these questions, it is helpful to recall that Stalnaker (1968) himself introduces selection functions in addition to Kripke's models in modal logic in order to select "for each [counterfactual] antecedent A, a particular possible world in which A is true." (p. 45) With their help, subjunctive conditionals can then be semantically evaluated. If its consequent is true in the selected possible world, the whole counterfactual conditional is true.

Halpern's selection function fulfils all requirements that Stalnaker postulates.[126] But although Halpern feels that his interpretation of Stalnaker's 'counterexample' is in the same spirit as the original, it is noteworthy that Stalnaker suggests a more complex framework for the analysis of counterfactual conditionals in games.[127] He does so to account for counterfactual conditionals as players' mind states which again can be the contents of further mind states etc. In contrast, Halpern's use of selection function(s) makes it hard to give an intuitive interpretation of players' reasoning throughout the hypothetical courses of the game. The selection functions represent the players' belief revision only in the sense that the theorist can formulate what players would know in other possible worlds. At the same time, players know with certainty that they are in the actual world. This is irrespective of whether selection functions are indexed for players or whether there is only one.

So the combination of partition knowledge and selection function(s) yields a hybrid between Aumann's and Stalnaker's account which fails to capture Stalnaker's intuitions about players' reasoning in games. Hence, Halpern's emphasis that

---

[126] The requirements are

- For all antecedents A and base worlds α, A must be in $f(A, \alpha)$. Compare (*F1*) in Halpern (2001).
- For all antecedents A and base worlds α, $f(A, \alpha)=\lambda$ (where λ is the absurd world in which all contradictions are true) only if there is no world possible with respect to α in which A is true.
- For all antecedents A and base worlds α, if A is true in α, then $f(A, \alpha)=\alpha$. Compare (*F2*) in Halpern (2001)
- For all base worlds α and all antecedents B and B', if B is true in $f(B', \alpha)$ and B' is true in $f(B, \alpha)$, then $f(B', \alpha) = f(B, \alpha)$.

Halpern also has a condition (*F3*) that the strategies chosen in $s(f(\omega,\upsilon))$ and $s(\omega)$ are the same in the subgame starting with υ.

[127] Halpern (1999, footnote 2) supposes that his interpretation is "essentially the model that Stalnaker had in mind" at a roundtable where Aumann, Stalnaker and Halpern met in 1998.

Stalnaker's 'counterexample' can be defended within a partition representation of knowledge does not vindicate this representation of knowledge for Stalnaker's account in general. Indeed, Halpern mistakenly implies that he has shown that Stalnaker's concern about the partition representation of knowledge is futile.[128]

Nonetheless, Halpern's account helps us to understand the formal differences between Aumann and Stalnaker better. He demonstrates that a 'counterexample' to Aumann's (1995) theorem can be given without altering the formalisation of knowledge as a partition structure. But then again Halpern does not provide a basis for deciding between Aumann's and Stalnaker's framework.

Having criticised Halpern's account for not capturing Stalnaker's intuitions, it should be remembered that Stalnaker's 'counterexample' also fails to provide a formal framework to capture hypothetical reasoning, as there is just one possible world in his model:

> "There is just one state x. The strategy pair realized there is $(D_1, A_2; d)$, The knowledge partition, of course, is $\{x\}$ for both players, so in this model there is common knowledge of everything that is true." (Stalnaker (1998), p. 47)

So this model formally excludes the counterfactual reasoning, belief revisions and substantive rationality which Stalnaker's counterexample refers to verbally. However, Stalnaker should not be faulted because the whole point of the counterexample was to demonstrate the poverty of Aumann's framework. Stalnaker shows that Aumann's account is unable to deal with (non-formal) counterfactual hypotheses supporting the coexistence of common knowledge of rationality and subgame imperfect equilibrium.

Samet's approach

Samet (1996) offers another way to extend the partition structure representation of knowledge, such as to accommodate hypothetical reasoning in solving games. There are lessons to be learnt from studying his approach because his critique of Aumann starts out in similar manner to Stalnaker's, while his constructive points differ.

Samet's critique of Aumann's notions of knowledge comes in two forms. He criticises Aumann for violating the basic idea that we can only call an action rational or

[128] Halpern (1999) writes: "The definition [in his account of Stalnaker's 'counterexample'] of knowledge is the standard one, given in terms of partitions. I stress this point because Stalnaker (1996) has argued that probability-1 belief is more appropriate than knowledge when considering games." (p. 429)

irrational if we can hypothesise that there was an opportunity to take that action. Secondly, Samet argues that the hypothetical reasoning, as in "Player *i* (who chooses action a) thinks that had she chosen action b, player *j* would have known she, player *i*, was irrational" (p. 235), cannot simply be represented as a composition of the antecedent and consequent involved.

Instead, Samet suggests introducing an additional operator for hypothetical knowledge. This binary operator assigns values to tuples of events and hypotheses according to an individual selection function. Importantly, the underlying selection function concerns epistemic states. It maps the actual world and the hypothesis on some possible world(s). Hence a certain event is hypothetically known in worlds which together with the relevant hypothesis are mapped into a world inside the set of worlds which constitute the event. Samet then introduces some formalism for models of games, which contain the so-extended information structure and a function mapping worlds on temporally extended courses of the game. The definition of someone being rational at some node is the intersection of the events that the respective node is reached, that up to the node the player did not know a utility-enhancing move, and that the player hypothetically (imagining all possible actions at the current node) knows that the current action is optimal.

One result that Samet derives from this setup bears a resemblance to one of Stalnaker's conclusions, namely that common knowledge of rationality does not imply backward induction in perfect information games. But Samet's counterexample is built up differently: Player *i*'s knowledge of player *j*'s rationality at each vertex υ is equivalent to player *i* knowing that player *j* hypothesises that he, *j*, would be rational at υ; but that does not imply that *i* would hypothesise about the events at υ in the same way. Therefore common knowledge of rationality does not imply that everyone would always act rationally:

> "Now the common knowledge assumption tells us that *i* knows that the player who plays at [vertex] *a*, say *j*, hypothesises he is rational at *a*. [...] But there is nothing that bounds *i* to hypothesise that if a is reached, *j* would behave rationally." (p. 243)

So, somewhat similar to Stalnaker's 'counterexample,' there is a difference between knowing that someone would be rational at some point and knowing that, if that point is reached, this person will act rationally. However, as will be discussed in the

next section, Stalnaker makes a distinction between two *kinds* of conditionals, not only between how different persons hypothesise.

So Stalnaker might be convinced by Samet's critical analysis of Aumann's framework and by Samet's intention to depict knowledge (of rationality) in games as hypothetical. But he would be critical about the way hypothetical knowledge is defined. This critique would point out that Samet's axioms postulate a single hypothetical knowledge function (one for each player) for all possible worlds. This makes it impossible to model revision of what a player knows in the actual world. A player could not change a hypothetical belief, should he find the antecedent of this belief to be true.

In Samet's approach, the additional assumption which is needed to support backward induction is a common hypothesis of node rationality. According to this assumption, all players hypothesise equally about all nodes as far as players' rationality is concerned. I suppose that Stalnaker would not be impressed by Samet's theorem on the sufficient conditions for the application of backward induction. After all, the common hypothesis assumption in conjunction with the way rationality is defined amounts to assuming that it is common knowledge that no player ever changes the belief about common knowledge of rationality. But this implication is everything but trivial, as will be discussed in the two next sections.

5. Two kinds of *if*s

Some of Stalnaker's arguments make reference to a distinction between two kinds of conditionals, containing two kinds of "*if*s".[129] The first *if* is causal, the second epistemic. An example, used by Stalnaker (1996 and 1998), might help to illustrate the distinction. Consider the statements:

(1) If Shakespeare had not written Hamlet, someone else would have.

(2) If Shakespeare did not write Hamlet, someone else did.

According to Stalnaker, (1) contains a causal *if*, because a statement is made to the effect that Shakespeare was not the only person who had the causal power to write this drama. It makes a statement about what would be the case in a world in which Shakespeare did not write Hamlet. (2) contains an epistemic *if* in the sense that it

[129] This expression is borrowed from Stalnaker (1996). The distinction appears under different labels in earlier theories as will be discussed soon.

reflects the speaker's belief revision policy. It makes a statement about what would happen if the speaker found himself confronted with reliable information that Shakespeare did not write Hamlet. One such information could be, for example, that Shakespeare did not exist while the drama Hamlet does exist.

The independence of (1) and (2), or of causal and epistemic *ifs* more generally, is exemplified by the fact that it cannot be faulted in principle to jointly hold (2) and the negation of (1), namely

(1a) If Shakespeare had not written Hamlet, no one else would have.

One might believe that only Shakespeare had the causal power to write this drama, while one might still have the same belief revision policy as before. In Stalnaker's words "[o]ne believes the counterfactual [(2)], but one would give it up if one learned that the antecedent were true." (1996, p. 46)

Stalnaker argues that there can be a divergence between the path one takes to revise one's belief when confronted with the information that is described in the antecedent (Shakespeare did not write Hamlet) and the path that one's imagination takes in the thought experiment picturing the (nearest) possible world in which Shakespeare did not write Hamlet. Is this legitimate?

To support the possibility of such a divergence, consider another example, borrowed from Lewis (1973). Like Adams (1970), he distinguishes between subjunctive and indicative conditionals, where the earlier is roughly equivalent to Stalnaker's causal *if*, the latter to the epistemic *if*.

(3) If Oswald had not killed Kennedy, then someone else would have.
(4) If Oswald did not kill Kennedy, then someone else did.

Again it seems to be perfectly reasonable to hold (4) in conjunction with

(3a) If Oswald had not killed Kennedy, then no one else would have.

For instance, the following situation would render the joint belief of (3a) and (4) reasonable: One strongly believes that it would take a unique madman like Oswald to shoot Kennedy, i.e. that no other living person at that time would have been in the

mental and physical position to commit that crime. So when imagining a world where Oswald did not exist, one does not picture anyone at all murdering Kennedy. On the other hand, one does not believe that Kennedy lived beyond 22 November 1963. So when one learns (reliably) that Oswald did *not* shoot Kennedy, one retains the belief that Kennedy is dead.

What has this to do with game theory? According to Stalnaker the distinction and possible divergence of causal and epistemic counterfactuals has been implicitly excluded in game theory as pursued by Aumann – and harmfully so. As Stalnaker argues in his 'counterexample' to Aumann's theorem, it might well be the case that one player believes (or knows in some sense) the other player to be rational, but would give up this belief if confronted with information to the opposite effect. He believes that it could be possible to jointly hold the following two beliefs as a player or a game theorist:

(5) If player A were to reach node υ, she would choose the utility-maximising move.
(6) If player A does reach υ, she will not (deliberately[130]) choose the utility-maximising move because she proved to be irrational qua reaching υ.

Rendering this set of beliefs reasonable, Stalnaker (1998) presents an argument regarding his counterexample (as in Figure 4.3):

> "Bob has the following initial belief: [Ann] would choose A_2 on her second move if she had a second move. This is a causal 'if' – an 'if' used to express Bob's opinion about [Ann]'s disposition to act in a situation that they both know will not arise. Bob knows that since [Ann] is rational, if she somehow found herself at the second node, she would choose A_2. But to ask what Bob would believe about [Ann] *if* he learned that he was wrong about her first choice is to ask a completely different question – this 'if' is epistemic; it concerns Bob's belief revision policies, and not [Ann]'s disposition to be rational." (p. 48)

Remember that, in this example, the belief that both players act rationally is correct while the rationality of Ann's choice is supported by her (equally correct) belief that Bob would cease to believe in Ann's rationality if she departed from the equilibrium path at the first node.

[130] See the remark on the opposite of rationality in Footnote 121.

As Halpern (2001) illustrates, using Aumann's notions of (common) knowledge and substantive rationality, there is no space to entertain such reasoning. This is because Aumann's framework does not allow one to differentiate between two kinds of *if*s. In fact, when Aumann defines substantive rationality of a player as his utility-maximisation at each hypothetically reached node of a game, this is assessed *relative to beliefs* about players' rationality which are implicitly assumed to stay *fixed.* Thereby the common knowledge of rationality assumption implicitly forbids any belief revision.

Stalnaker thinks that Aumann should be explicit about assuming that beliefs about rationality are never revised on the basis of observations made during a game, instead of making this assumption implicitly by assuming causal independence. In fact, Stalnaker accuses Aumann of committing the fallacy of inferring that, for any player, belief (5) entails the belief revision policy (or rather belief *preservation* policy)

(6a) If player *A* does reach υ, she will choose the utility-maximising move.

I believe that Stalnaker's analysis of Aumann's flaws is correct.

A possible source of Aumann's fallacy

Not trying to *justify* Aumann's approach, I would like to point at a possible source of its weakness. In fact, Aumann and Stalnaker seem to think differently about what games in game theory represent. Aumann pictures games as 'proper' games like chess. [131] All that seems to matter when representing a proper game are moves within a game relative to a set of well-defined rules and the players' epistemic states, which in turn are only about moves and other players' epistemic states. Stalnaker's way to think about games in game theory, in contrast, might be to understand them as representations of complex (real world) social events.[132] This distinction might seem to matter because the real-world interpretation usually evokes a much richer picture of the set of possible worlds.

Aumann points out that a statement like 'If he had pushed his pawn, Black's Queen would have been trapped' is a counterfactual whose truth value is more easily assessed than a statement about the *real* world history, such as 'If Hitler had crossed the channel after Dunkirk, he would have won the war.' He reckons that this is so because

[131] Compare Aumann (1995, p.15).

[132] Stalnaker is not explicit about the issue. But the Shakespeare example indicates that reading games as representations of complex social events is playing a role in Stalnaker's take on counterfactuals in games.

in order to imagine the latter we would have to picture a course of history which differs from the actual history in "a myriad of ways." (p.15) Imagining the antecedent of the claim about the chess example, on the other hand, would only require limited imagination. In the chess example, all that would be needed is to picture the pawn in one spot rather than another in the given state and rules of the game.[133]

But even if we accept the interpretation of games as proper games, it is misguided to infer that knowledge of rationality excludes the possibility of revising beliefs about rationality. In a proper game, it might be *easier* to spell out what it means that someone acts rationally at a given point of the game. But this does not mean that we can dispense with representing what would be believed were the player to choose irrationally.

In contrast, if we take games to represent real social scenarios it seems that there is a broader range of counterfactual antecedents. It is often necessary to reason about counterfactual states using rather unspecific antecedents like 'Let's just suppose that we got to such-and-such a situation *somehow*...' So, since identifying the closest possible world in which the antecedent holds would be a rather fuzzy task here, we might be intuitively more willing to treat it as independent of belief revision policies.

But regardless of that impression, even the representation of a proper game needs to be complex enough to allow for picturing rational moves at all hypothetical decision points and, at the same time, picturing what players would think should irrational moves occur. Accordingly, if it is supposed to be the case that beliefs in rationality are robust in a proper game, this should not just be treated as the default assumption but instead be modelled explicitly within a sufficiently rich framework.

6. Epistemic independence and backward induction

Remember that if one endorses Stalnaker's take on counterfactual conditionals in games, backward induction does not follow from common knowledge of (substantive) rationality. A fortiori, it does not follow from common belief of rationality. If backward induction is to be applicable further assumptions have to be made, for instance a constraint to the effect that no event in the game changes anyone's beliefs. A version of this assumption is Stalnaker's (1998) 'epistemic independence' assumption in

[133] Compare Aumann (1995), pp. 14-15. Note that not all counterfactuals about chess are easily assessed. But picturing the crucial aspects of a counterfactual state in a chess game is still much simpler than picturing a counterfactual state of affairs in most real life interactions.

conjunction with the so-called agent form representation of games. Both notions need explanation.

For Stalnaker, ϕ is an information *about* a player if and only if for any worlds x and y which are subjectively indistinguishable for that player either both x and y are elements of ϕ, or none. So, roughly speaking, a player is characterised by what worlds he cannot distinguish. Furthermore, two propositions ϕ and ψ are *epistemically independent* for player i in world x if and only if $P_{i,x}(\phi \mid \psi) = P_{i,x}(\phi \mid \sim\psi)$ and $P_{i,x}(\psi \mid \phi) = P_{i,x}(\psi \mid \sim \phi)$.[134] That is to say, the probabilistic belief in one proposition must be perfectly uncorrelated to another proposition and vice versa. Joining these two definitions allows Stalnaker to rigorously formulate the assumption that *information about different players is epistemically independent.*

The agent form of a game, as introduced by Selten (1975), treats a single player as different agents at different information sets (e.g. nodes) of an extensive-form game, as long as all these agents share the same preferences and original information (but not necessarily the same beliefs).

With these definitions, one can comprehend a theorem proved by Stalnaker (1998):[135]

> "Let Γ be a perfect information game in agent form in which for each player different outcomes have different payoffs. Let M be a model for Γ in which it is common belief that all agents are perfectly rational, and that all agents adopt belief revision policies that treat information about different *agents* as epistemically independent. Then in M, the subgame perfect equilibrium strategy profile is realised." (p. 43, my italics)

The proof of the theorem is quite straightforward.[136] It formalises the intuition that no choice in the game ever changes the beliefs about the subsequent subgame. Each possible sequence of nodes reached contains as many different agents as there are nodes, because even if a player gets to move more than once in game his choices are represented as choices by different *agents*. So no move ever leads to a belief change. Also, the notion of perfect rationality is such that it retains its meaning throughout the game even in cases in which there was an error of some degree.

---

[134] Compare Stalnaker (1998), p. 42.
[135] Compare Skyrms (1998), p. 567, for a similar set of requirements for backward induction.
[136] Compare Stalnaker (1998), Appendix A.

What is astonishing about this proof is not the result, but the fact that Stalnaker introduces the agent form as a premise without much ado. The agent form can be interpreted in two ways. Either it is read as a constraint to the effect that only games are taken into consideration in which all players only have one possible choice point in a game. Or it is read as implying that it makes no analytical difference if a player is actually represented as a sum of agents for each node as long as these agents all have the same information and preferences. Stalnaker explicitly entertains the first interpretation.[137] This is worth pointing out, given that it seriously limits the scope of the theorem. In fact, it does not even include Stalnaker's 'counterexample,' in which Ann gets to choose twice. On the other hand, it would be even more awkward if Stalnaker were to claim that it does not matter whether the *identity* of an agent is preserved throughout a game. Effectively, to assume epistemic independence then would be to imply that nothing could ever be learned about a player (and his rationality, e.g.) from observing his action.[138]

Stalnaker makes this assumption explicitly, but blames Aumann for *implicitly* making it:

> "What Aumann's definition of substantive rationality does is implicitly to build an epistemic independence assumption into the belief revision policies of all players, an assumption that is *considerably stronger* than the epistemic independence assumption discussed above. Aumann's assumption is that not only beliefs about different players, but even beliefs about different parts of a single player's strategy are epistemically independent." (1998, p. 48, my italics)

So Stalnaker argues that by representing knowledge as a partition structure (and, as Halpern would add, not providing any selection functions[139]), Aumann's definition of

---

[137] Compare Stalnaker (1998), p.43.

[138] This assumption exhibits parallels to a set of assumptions ("entrenched common belief in rationality") in support of backward induction discussed by Sugden (1991), p. 772. These amount to assuming that perfect information and rationality remain commonly believed throughout a game, and that it is common belief that this is so.

[139] In Halpern's framework, once selection functions have been introduced, subgame perfection is implied by common knowledge of substantive rationality if another constraint on selection functions is fulfilled. This condition states: "For all players i and vertices υ, if $\omega' \in K_i(f(\omega, \upsilon))$ then there exists a state $\omega'' \in K_i(\omega)$ such that [the strategies actually played in ω' and ω'] $s(\omega')$ and $s(\omega'')$ agree on the subtree of Γ below υ." (Compare condition (*F4*), p. 431.) It ensures that there is no surprise that could lead to a change of strategies; starting from some possible world, the selection function could only lead to a reduction of the strategies considered possible. To capture the even stronger epistemic independence assumption, which Aumann implicitly makes, it has to be granted that the set of strategy profiles considered possible is not altered *at all* by the selection function. Not only would there be no surprises then, but there would not be any increase in knowledge about any one of any sort. This is guaranteed by

(common) knowledge of rationality at each node is tantamount to claiming that no matter what (possibly irrational) move(s) are observed throughout the game, the certain common belief in rationality would not change.

But making a similarly strong assumption explicitly does not render the assumption itself acceptable. Epistemic independence in conjunction with agent form representation amounts to a very strong and implausible constraint. So if this were the weakest condition sufficient for the applicability of backward induction, its scope would be negligible. In the next section, I will say more about this problem and argue for a potential remedy. It comes in form of a plausibility argument and it will force us to look at interpretations of belief revisions that undermine backward induction.

7. A plausibility argument for backward induction

Stalnaker shows that epistemic independence *cum* agent form representation is sufficient to support backward induction. I argued, that this assumption is quite strong. Let me elaborate on this.

One could easily argue that epistemic independence between different players seems implausible. For instance, it is not unreasonable for Ann to regard her two opponents Bob and Chris as part of a population whose members tend to have traits in common. So she could (reasonably) update her beliefs about Chris after having been surprised by Bob's behaviour. It seems even more implausible to assume that beliefs about a person are never revised, even if one observes something surprising about that person. In fact, such recalcitrant beliefs call for an explanation. Such an explanation could be that the player with such robust beliefs fails to be able to learn at all or that this player fails to see other players as subjects whose actions exhibit some continuity over time.

So if we want to support the solution concept of backward induction in certain perfect information games, it would be desirable to do so on a weaker premise than epistemic independence *cum* agent form representation. I want to argue that this can be done by taking a closer look at what kinds of belief revisions seem impermissible, given the other assumptions made.

---

replacing the 'if' in (*F4*) by 'iff'. (Compare Halpern (2001), footnote 6.) But again, Halpern's framework only captures the intuition of epistemic independence in an indirect way. It seems hard to decide over the plausibility of the constraint (*F4*) without looking at examples, or indeed retranslating it into Stalnaker's terms.

The 'counterexample' revisited

Let us take another look at Bob's beliefs and his belief revision policies in Stalnaker's 'counterexample.' On the one hand, Bob believes (at the first node) the counterfactual claim that, if Ann got to choose at node 3, she would choose rationally, i.e. A_2. On the other hand, he would cease to believe that Ann will choose rationally if he found out that she had the chance to choose. That is, he believes that if node 2 were reached, this would be a signal that Ann is not rational. So Bob believes that if subsequently node 3 were reached through his own choice, Ann would play D_2. Something is awkward about this. Bob's belief revision policy seems self-referential in a problematic sense.

Bob's belief revision policy is self-referential because Bob revises his beliefs about Ann's rationality upon observing an action which counts as irrational *if* (among other assumptions) it is assumed that common knowledge of this very belief revision policy is robust. This is because Ann believed that Bob would choose d only because she knew of his belief revision policy; but Ann's so-derived belief about Bob's choice in turn motivates his belief revision. So Bob's belief revision policy makes sense of action in the light of itself. The intuitive awkwardness of this might be most apparent in the following reformulation of the rationalisation behind the alleged equilibrium: Bob would be surprised by Ann's choice, viz. A_1, because he knows that Ann knows that he would be surprised in this way at this point. So the justification of Bob's belief revision needs to make reference to the fact that Bob will revise his beliefs according to this belief revision policy. This amounts to a kind of bootstrapping. It is a belief revision policy that calls itself into existence.

One might reply to this that it is not a relevant question how an equilibrium state comes into existence, only why it is stable once it does occur somehow. But this attitude can lead to conceptual problems. Consider an argument by Sugden (1991) which has some similarity to my argument.[140] Sugden claims that in some games (like Matching Pennies) the unique Nash equilibrium need not be a solution to the game. That is for cases in which mixed strategies are interpreted as subjective beliefs and in which a correlating device (as discussed in Aumann (1987)) is excluded. The argument is that, in the absence of such a device, one needs to assume that players have common priors with respect to each strategy played in order to support an equilibrium in mixed strategies. For example, players A and B would have to attach the same prior probability

[140] For the complete argument compare Sugden (1991), pp. 765-768.

to *A* playing a certain (pure) strategy. But the interpretation of this assumption of common priors is obscure. Sugden suggests that it could only be thought of as some psychological impulses occurring with probabilities known to *A* and *B*. This, however, seems be a psychological premise which cannot be defended on the grounds of rationality alone.

The parallel between Sugden's argument and mine is that we both challenge an equilibrium which is supported by a belief which in turn is hard to support without reference to some psychological notion in addition to the rationality assumptions. My argument, however, goes further in that it finds the beliefs necessary to support the 'counterexamples' to backward induction to be *in conflict* with our intuitions about rational agents. Not only is it unclear how the common knowledge of (or common belief in) a bootstrapping belief revision policy could be derived from the rationality assumptions, because a crucial bit of the justification for such a belief revision policy is itself. It is also dubious how rational agents can commonly believe in such belief revision policies while also commonly believing in rationality. This needs to be explained.

It is common knowledge in the 'counterexample' that Ann knows (at node 1) that Bob has the belief revision policy that he has. This, of course, implies that Bob knows that Ann knows that he has this belief revision policy. He continues to believe this at node 2. In fact, it is only on the basis of this belief that he is moved to change his belief in Ann's rationality at node 2. Let us spell out Bob's beliefs at node 2: On the one hand, he believes that Ann understands exactly when he will change his believes about her rationality. On the other hand, he no longer believes that Ann is rational. This is odd, because Bob's first belief only seems to be a sensible belief if he also assumes that Ann knows what it *means* to be rational. But if Ann knows what it means to be rational, it seems strange for Bob to suppose that she would not *act* rationally. Bob would have a bizarre picture of Ann's mind state. And it is even stranger to assume that such a belief revision policy is common knowledge of rational agents at the root of the game.

Generally, it seems conceivable that someone understands a normative standard of action, but does not act accordingly, because of a weakness of will, for instance. But when we spell this out for rationality in game theory, it leaves us with a bad aftertaste. Remember that utility is supposed to capture everything that could motivate the agent and that rationality means expected utility maximisation. If we assume that a player knows all he wants (which he does in a perfect information game) and how to best

achieve it (in virtue of understanding the concept of rationality), the picture of a player not rationally pursuing his wants seems conceptually obscure.[141]

Again, the argument presented here is a plausibility argument to exclude certain belief revision policies in connection with the assumption of common belief of perfect rationality.

The general argument

In what follows, I present a more general verbal argument that raises the bar for finding *any* counterexample to the backward induction result in certain games given the assumption of common belief of perfect rationality (and the assumption that players are perfectly rational). The idea of the argument is to ask how to defend a deviation from the backward induction path under the assumption of common belief in perfect rationality. In doing so, it is assumed that belief revision policies are concerned only with belief in rationality or belief in belief in rationality etc., not any other aspect of a game. The scope of the argument is perfect information extensive form games, in which for each player different outcomes have different payoffs. For the purpose of this argument I will use the abbreviations 'BI' for backward induction and 'CBPR' for common belief of perfect rationality. I will call the path through the entire game tree that is predicted by backward induction the 'BI path' and the play predicted by backward induction for any subtree of the entire tree the 'BI play.'

I now argue why the scope of the plausibility criterion presented earlier is not restricted to one example. The idea of the argument is to show that under the assumption of CBPR players would not be perfectly rational when being the first to deviate from the BI path. This is because, on any node of the BI path including the root of the game, deviating from the BI path by definition yields less utility than staying on it, given that CBPR would not break down at any later node. Now by assumption, CBPR does not break down on the actual path through the game tree chosen by the players. So even when assuming an actual one-off deviation from the BI play, everyone's projected actual play from that point on would still have to be the BI play – by assumption. Hence a deviation from the BI path could only be rational for a player if he expects that someone else would deviate from the BI path later should he himself

[141] This is related to the question whether the hypothetical imperative is analytic. I briefly touched on this broader question in Chapter1.

stay on it. Otherwise, staying on the BI path would have been the rational strategy. But how could this expectation be supported given CBPR?

Some agent[142] *A* will only rationally deviate from the BI path (as the first agent to do so in a game) if he supposes that some agent *B* would do so later in the game if he, *A*, stayed on the BI path.[143] But *A* would think so only if he thought that *B* expects some breakdown of CBPR *because* all agents up to his move stayed on the BI path. *B* could then think that some later agent *C* would deviate from the BI path, unless *B* deviates first. More precisely, a number of beliefs, if held by *A*, could lead to *A*'s conclusion that it is rational for him to leave the BI path:[144]

- the belief that someone proved irrational by staying on the BI path.
- the belief that *B* (an agent choosing after *A*) would believe that someone would be irrational because everyone up to *B*'s node stayed on the BI path; or
- the belief that *B* would believe that some agent *C* (an agent moving later than *B*) would lose the belief in someone's rationality should everyone up to *C*'s node stay on the BI path;
- the belief that *B* would believe that *C* would believe that some agent *D* (an agent moving later than *C*) would lose the belief in someone's rationality should everyone up to *D*'s node stay on the BI path;
- etc.

Let us consider whether any of the candidates on this list are plausible justifications for *A*'s deviation. Agent *A*'s belief revision policy implied in the first candidate would only be in harmony with CBPR if *A* believed that everyone's staying on the backward induction path so far was non-utility-maximising. But if *A* does not make reference to some future agent's belief revision policy (as in the other candidates in the list above), her revision only 'makes sense' in the light of assuming itself. It seems that *A* is only justified to revise her beliefs if it is (commonly) believed that *A* believes that someone's backward induction move would trigger her belief revision policy. No other reason comes to mind for questioning CBPR in that situation, given

[142] I use 'agent' instead of 'player' here to account for the fact that a single player might get to choose several times in a game. The beliefs of agents constituting a player, however, are assumed to be interdependent here.

[143] The point here is that deviation from the backward induction path can only be explained by the expectation of future players' deviation up to some point. Some later player must be supposed to deviate for another reason.

[144] I believe that the list (if continued accordingly) is exhaustive as a list of beliefs that could potentially justify *A*'s deviation as rational.

that everyone has stayed on the backward induction path so far. However, the reason just given exhibits the critical self-referentiality.

All other kinds of belief revisions on the list (concerning beliefs in beliefs, etc., in rationality) make a reference to further belief revision policies by agents moving later in the game. Crucially, there always has to be reference to a belief revision policy that infers someone's irrationality from observed backward induction play at the end of this chain of references (of belief revisions about belief revisions...etc.). As by assumption everyone has stayed on the backward induction path so far, nothing could initiate anyone's revision of CBPR but some (indirect) reference to another belief revision policy. So there needs to be at least one person who believes that someone believes ... that some (rational) agent has a belief revision policy of the problematically self-referential type.

In other words, some agent, in rationalising his own choices, has to ultimately make reference to someone's bootstrapping belief revision policy, while still believing everyone to be actually perfectly rational. But then, on the assumption that the above list of possible beliefs (that could lead some agent *A* to leave the backward induction path) is exhaustive, I have shown that, given common belief in rationality, no departure from the backward induction path can be argued for without reference to some implausible belief revision policy.

More examples

Let me look at some more examples to see how a bootstrapping belief revision policy could be supposed to support a subgame-imperfect equilibrium. These examples do not contribute substantially to the argument. They might help one to grasp it better however. The game in Figure 4.4 can be treated somewhat similarly to Stalnaker's 'counterexample.'

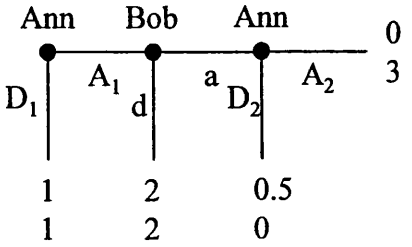


Figure 4.4

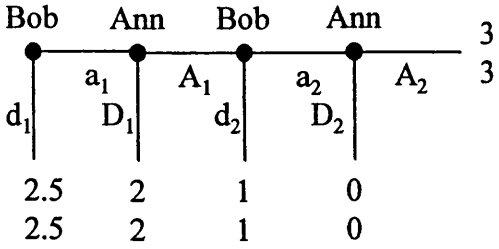


Figure 4.5

There are also three Nash equilibria in pure strategies in the unreduced normal form of this game. Only one of these is subgame perfect, namely $\{A_1, D_2; d\}$. But let me try to support $\{D_1, D_2; a\}$ on the assumption of common knowledge of substantive[145] rationality. Let me stipulate that Bob stops believing that Ann is rational once node 2 is reached. He now believes that Ann is irrational in the sense that she would choose less expected utility if node 3 was reached. Bob would revise his beliefs this way because Ann knows that Bob will play a once at node 2, which would leave her a maximum feasible payoff of 0.5 at node 3. This is less than what she could secure by choosing D_1.

Let me check for substantive rationality in Stalnaker's fashion. Ann is substantively rational because she chooses D_1 at node 1 given that she believes that Bob would play a at node 2, which in turn is due to her knowledge of his belief revision policy etc. Bob is materially rational because he never gets to choose. He is furthermore substantively rational because he would believe, at node 2, that Ann will play A_2. So move a would leave him with a payoff of 3, rather than the 2 he could have secured. Ann, however, would play the expected utility maximising option D_2 at node 3. Note that in this solution the structure of underlying (beliefs in) belief revision is the same as in Stalnaker's 'counterexample.' It is also implausible for the same reasons.

Next, consider the game of Stalnaker's 'counterexample' preceded by another vertex (Figure 4.5). Here the subgame perfect equilibrium is $\{A_1, A_2; a_1, a_2\}$. But consider the Nash equilibrium $\{D_1, A_2; d_1, d_2\}$ as a potential counterexample. This can be supported with reference to two belief revision policies (rather than one, as in the other examples). The first belief revision policy (temporally speaking) at work is as follows. Should Ann ever arrive at node 2, she would cease to believe that Bob believes that she is rational. This revision is made because it is common knowledge that Bob will play d_2 at node 3. Anticipating this, Ann would rationally play D_1 at node 2. But this would leave Bob with less than he could have secured by choosing d_1 at node 1. A *second* belief revision policy, implying (common knowledge of) move d_2, is that Bob would change his beliefs about Ann's rationality at node 3. This revision is made (as in the original 'counterexample') because Bob believes that Ann knows this second belief revision policy held by Bob, leaving us with the same kind of self-reference as before.

Let me check for perfect rationality. At node 1, Bob is rational to pick d_1 because he believes that Ann will play D_1 at node 2. At node three, he is rational to pick

[145] This notion could be specified in terms of perfect rationality as in Stalnaker's (1996) formulation of his 'counterexample.'

d_2 because he now believes that Ann chooses irrationally. Ann is rational to pick D_1 at node 2 because she believes that Bob would play d_2 next. At node 4, she would maximise utility by playing A_2.

Again, the point of going through this longer game is to exemplify how, at *some* stage in the argument, the defence of a subgame-imperfect equilibrium under the assumption of common knowledge of substantive rationality requires justifying an expected deviation from the backward induction path by reference to some self-referential belief revision.

8. Summary and concluding remarks

This chapter touched on some aspects of a vast topic in the foundations of game theory, namely the relations between epistemic assumptions, hypothetical reasoning and solution concepts. Mainstream game theory often bypasses the aspect of hypothetical reasoning, i.e. the reasoning about counterfactual situations in games. It attempts to support solutions concepts directly on the basis of epistemic assumptions like common knowledge of rationality.

But this form of bypassing means to make counterfactual claims about games implicit, where an explicit discussion about non-equilibrium situations in games like, e.g., nodes outside the backward induction path on the centipede game, is urgently required to intuitively support the equilibrium in the first place. Furthermore, the plausibility of an epistemic assumption as such depends on what it implies about players' hypothetical reasoning. Methodologically speaking, it should therefore count as a virtue if a formal framework of epistemic assumptions can represent hypothetical reasoning in games explicitly.

Aumann (1976, 1995) delivers a framework that, more or less ex post, provides a formal representation of common knowledge or rationality supporting the results that game theory has traditionally taken this assumption to yield. For extensive games, for example, this partitional representation of knowledge validates the method of backward induction. This result, however, stands and falls with the credibility of implications of Aumann's framework regarding hypothetical reasoning in games.

I argued that this credibility in turn is rightly questioned by Stalnaker (1994, 1996, 1998). The partitional representation of common knowledge of rationality does not allow the game theorist to consider the possibility that players commonly and

correctly believe that everyone will actually choose rationally while ceasing to believe so under counterfactual circumstances. Explicitly, partitional representation excludes this possibility; implicitly, it treats common belief in rationality as robust against all counter evidence. The first of these two facts allows Stalnaker to verbally argue for an example where common knowledge of rationality is compatible with a non-backward induction equilibrium to a game.

Stalnaker's alternative notions of knowledge and common knowledge of rationality, which facilitate his 'counterexample,' are weaker, of course. But I support Stalnaker's stance that these definitions are more justifiable than Aumann's epistemic notions, because they are merely representations of subjective mind states which do not allow for logical inference about the state of the world.

Among other side discussions, this chapter addressed two attempts, by Samet (1996) and Halpern (2001), to bridge the divide between Stalnaker's and Aumann's respective results on backward induction by a unifying formal framework. I challenged the intuitive value added of these approaches on the grounds that neither of them sufficiently appreciates Stalnaker's urge to make the formal framework rich enough to represent crucial features of hypothetical reasoning.

One of the implications Stalnaker's approach is that it demands stronger assumptions sufficient for supporting backward induction. But how strong do these assumptions have to be? Stalnaker makes the case for an epistemic independence assumption, which excludes any learning throughout a game. By the way of a constructive addition to my critical discussion, I outlined an argument according to which this rather implausible assumption can be avoided by restricting what kinds of belief revision policies are permissible. I argued that belief revision policies which support a breakdown of common belief of perfect rationality by reference to themselves are to be excluded. I conjectured that this restriction in conjunction with common belief of perfect rationality supports backward induction. It is subject to further research, whether this conjecture can be proved within a formal framework that is yet to be developed.

More generally, more research should be conducted focusing on the translation of Stalnaker's framework to applied game theory. Some kind of a 'manual' indicating permissible solution concepts for different types of situations (considering especially the epistemic relations between players) seems vital if Stalnaker's rather demanding technical discussion is to have repercussions in main stream game theory.

Finally, I would like to emphasise key aspects concerning the relevance of this chapter vis-à-vis the foregoing chapters. First, counterfactual reasoning, and its appropriate formal representation, are of paramount importance in solution concepts. The fundamental concepts in games, like outcomes and preference, therefore have to be suited for this kind of reasoning as well. Furthermore, I have presented a plausibility criterion for equilibria that exhibits parallels to my critique of the concept of psychological Nash equilibrium in Chapter 3. Both of these arguments draw on intuitions about whether or not we can spell out the players' hypothetical reasoning in order to support the equilibrium in question.

As will be seen in the next chapter, counterfactual reasoning even has implications on what is a reasonable conception of rationality as maximisation.

Chapter 5

What is "constrained maximization"?

1. Introduction

Like the previous chapter, this final chapter also considers how solutions to games are derived from the (common knowledge of) rationality assumption. This chapter, however, will also complement the previous discussion by questioning the orthodox notion of rationality itself. This questioning is motivated by one of the classic 'paradoxes' of game theory.

In ethics and political philosophy there are various schools of thought linking rationality and morality. One of these traditions sets out from what could be labelled instrumental rationality. Two contemporary advocates are David Gauthier and Edward McClennen. One focal point of their theories is to point a way out of the (finitely iterated) prisoner's dilemma, which is often thought to epitomise situations of conflict. I shall be mainly concerned with Gauthier's (1986) *Morals by Agreement* (MbA). Importantly, his theory belongs to the family of moral theories which do not settle for a genealogy of morals or, more specifically, of cooperation. Gauthier does not presume that the question of what one *should* do in the prisoner's dilemma can be reduced to some naturalistic account of how we came to our moral intuitions.

Regarding the scope of his work, Gauthier realises that the question of whether one should cooperate in the prisoner's dilemma, does not exhaust all matters of morality. But he believes that stronger, more complete accounts of morality can be given on the basis of studying cooperation.[146] Whether this last claim is true shall be bracketed from this discussion. Furthermore, within Gauthier's (1986) argument, I shall focus on the core matter of cooperation as "constrained maximization" (CM), which Gauthier himself considers to be the "most fruitful idea" (1993, p. 185) of MbA.

I argue for two main claims. Firstly, Gauthier is not always clear about what conceptual departure from classical game theory is doing the work in his argument for CM. Secondly, once spelled out more clearly, possible interpretations of his approach either (i) collapse into established solutions to, or dissolutions of, the dilemma in

[146] Compare Gauthier (1990), pp. 145 ff, and Gauthier (1986), pp. 13 ff, respectively.

mainstream game theory or (ii) seem to become incompatible with the implicit conceptual baggage of game theory.

Let me say a word on the spirit on my critique. I approach Gauthier's ideas by asking how far it is compatible with the inherent limits of (state-of-the-art) game theory, as a form of representation. This should not be read as a categorical pledge for methodological conservatism with respect to game theory. A useful version of game theory should be designed so as to fit our (normative and descriptive) intuitions, not vice versa. Having said that, I find that the onus is with the 'reformers' of game theory to show that their revised version of the theory merits the standard of game theory's orthodoxy as defined by von Neumann and Morgenstern (1944, p. 7): "The theory finally obtained must be mathematically rigorous and conceptually general."

The plan of the paper is the following. Section 2 recapitulates Gauthier's core argument. Section 3 traces the roots of the alleged conflict of instrumental rationality and cooperation, which lies at the heart of Gauthier's discussion of the prisoner's dilemma, back to Hobbes. It is shown that prisoner's dilemma-like social constellations invited a confusion of different notions of rationality long before the rise of game theory. Section 4 takes a closer look at the conceptual implications of game theory and their relation to some concepts suggested by Gauthier. The focus is on the concepts of inter- and intrapersonal separability of choices as well as the conceptual space between choice and preferences. Section 5 considers alternative routes out of the prisoner's dilemma, concluding that they are either 'old news' in game theory or incompatible with its current conceptual framework. Section 6 investigates whether McClennen's (re)formulation of Gauthier's idea redeems the outlined problem. Section 7 presents some concluding remarks.

2. Gauthier's argument in a nutshell

In the beginning of MbA, Gauthier foreshadows the conclusion of his argument: "To choose rationally, one must choose morally" (p. 4). In fact, Gauthier's project is to derive a normative principle, i.e. cooperation in the finitely iterated prisoner's dilemma, from allegedly "non-moral premisses" (p. 5). These premises are those taken from rational choice theory. They include a maximising notion of instrumental rationality, and a subjectivist, preference-based notion of value. Gauthier explicitly contrasts his premises to a "universalistic" (p. 7), e.g. Kantian, notion of rationality and to any

objective premises about values or meta-preferences. What remains is Gauthier's explicit endorsement of subjective causal expected utility theory in its contemporary axiomatic form.[147] Interpersonal comparison is not presumed to be possible. Accordingly, Gauthier assigns reason the role of the "slave of passions" (p. 12) in Hume's famous dictum. Rationality is defined as utility maximisation, where the notion of maximisation has yet to be specified.

Gauthier endorses a contractarian approach in a Hobbesian tradition, appealing to self-interested but (ex ante) mutually disinterested subjects. He builds on four core conceptions. The first is the market as a hypothetical backdrop of a state in which morals are neither needed nor present. It is "a foil against which morality appears more clearly" (p. 13), namely as a means to overcome market failures. The second notion is a claim concerning an allegedly rational outcome of a bargain over some amount of goods; it is dubbed the "principle of minimax relative concession".[148] Thirdly, there is CM. Like the second conception, the fourth and fifth conceptions are analytically separable from CM. The fourth conception is the (Lockean) proviso "that prohibits bettering one's position through interaction worsening the position of one another" (MbA, p. 16). The fifth conception is the Archimedean point which plays conceptually a somewhat analogous role to Rawls' (1971) 'original position.'

My focus is on CM. Several components can be identified which jointly constitute something of an abstract version of the 'State of Nature' on the basis of which Gauthier motivates his notion of morals. These are self-bias (or self-interest exceeding the interest in others), awareness of variable scarcity and awareness of positive and negative externalities. What emerges is the mutual dependence of agents, both in the sense of opportunities from a potential division of labour as well as competition over resources. Dealing with (idealised) rational agents, the upshot of this situation is supposedly a strategic conflict as represented by the prisoner's dilemma. Some standard game theory follows: On the assumption of common knowledge of rationality and a finite repetition of games, backward induction yields equilibrium in mutual defection. It is Pareto-inferior to the co-operative outcome because at least one person could be made better off while making no one else worse off. This divergence of rational

[147] The problems arising from these premises deserve a discussion of their own. Compare Broome (1993) and Mandler (2001).

[148] As Gauthier later concedes, replying to critics like Binmore (1993), this concept is "simply undercut by the non-cooperative approach" (1993b, p. 177). Since it also the case that the truth or falsity of the 'principle of minimax relative concession' has no immediate implications for Gauthier's general claims about cooperation, I shall leave aside any discussion about the principle.

equilibrium and optimality poses the core problem, which Gauthier sets out to resolve by showing that on the second look "co-operation is a rational mode of interaction." (MbA, p. 118) He follows Hobbes in calling the disposition to cooperate in such situations "justice" (MbA, pp. 113, 150).

Gauthier departs from Hobbes' when it comes to the alleged redemption via introducing the idea of a sovereign. Calling this move a political rather than a moral solution (MbA, p. 163), Gauthier effectively accuses Hobbes of escaping the dilemma rather than resolving it – an allegation that, I will argue later, applies to Gauthier as well, if in a more subtle form. Gauthier himself seeks a moral solution by exploiting the double role of "reason" (MbA, p. 163)

He contrasts two kinds of agents, both rational in some sense. The first is the "straightforward maximizer," who behaves like the rational agent in game theory textbooks. That is he will defect in the prisoner's dilemma and his pledges for mutual cooperation lack credibility. The second kind is the constrained maximiser who has a "disposition" to cooperate should this be advantageous taking into account the best available information. That is, in exactly these cases this agent will act on his disposition. Gauthier also calls this engaging in a "joint strategy" (MbA, pp. 166/167).

It is assumed that agents can have dispositions to choose one way or another. These dispositions are (to some degree) externally recognisable. This allows the agents to commit (in some sense) to some particular future choice, in this case cooperation. Furthermore, it allows agents to know more, or rather know something *different*, about their opponents than that they are rational and partake in common knowledge of rationality. In fact, agents actually have probabilistic knowledge about the dispositions of their opponents. This assumption is called "translucency" (p. 174) of dispositions, a gradually weakened version of transparency.

Given these additional assumptions (whose consistency remains to be scrutinised), calculating the optimal strategy allegedly comes down to weighing the expected utilities of cooperating and defecting respectively against each other. The result will depend on the relative size of the potential payoffs and the (presumably known) probabilities of meeting straightforward or constrained maximisers. Roughly, it follows that the better the constrained maximisers become in telling apart friend from foe, the more likely cooperation becomes.[149]

[149] Compare MbA, pp. 171-181, for a formal treatment.

All this is methodologically dubious, but at the same time it is crucial for the vindication of Gauthier's account. It seems that Gauthier cannot vindicate cooperation as a "rational" strategy in the circumstances described, without establishing some possibility of agents disposing themselves to certain modes of interaction (even if this disposition is conditional). So I will focus on this issue.

Note finally that there is an evolutionary aspect to the discussion of CM and its potential for success. In fact, Gauthier acknowledges an "interesting parallel" (MbA, p. 187) between his conclusion about CM and Trivers' result of evolutionary stability of reciprocal altruism. While this parallel remains largely undeveloped, I will argue later why any such parallel is highly problematic.

3. The Foole's challenge: Historical roots of the debate

> "For the question is not of promises mutuall, where there is no security of performance on either side; as when there is no Civill Power erected over the Parties promising; for such promises are no Covenants: But either where one of the parties has performed already, or where there is a power to make him performe; there is the question whether it be against reason, that is against the benefit of the other to performe, or not. And I say it is not against reason." (Leviathan, p. 204)

This quote is Hobbes' reply to what he calls the "Foole's" challenge to the hypothetical social contract, the "Covenant." Following Gauthier's interpretation, this contract essentially consists in establishing cooperation in the prisoner's dilemmas posed by Hobbes' infamous 'State of Nature.' Accordingly, the challenge posed by the Foole involves spelling out the rationale of the straightforward maximiser, namely never to cooperate in the absence of credible commitment or effective sanctions.[150]

What is peculiar about the quote is the use of the word "reason." Earlier in his argument, Hobbes clearly says that seeking one's advantage is the reasonable thing to do; it is what reason demands. Now a choice is dubbed reasonable which, prima facie, is

[150] Of course, the situation Hobbes describes is better captured as a sequential variant of the prisoner's dilemma in which the first chooser has already (rationally or not) made the cooperative move. This was observed by the second player. Furthermore, I must point out that on one reading of the following passage Hobbes might actually have in mind a certain potential for sanctions namely the exclusion of the Foole from (cooperative) society. But this rationale of reputation via observation of behaviour is not doing the work in Gauthier's version of Hobbes' contractarianism, as we will see.

in conflict with this non-cooperative policy.[151] As a Hobbes scholar, Gauthier is aware of this conflict. In his *Logic of the Leviathan* (1969), he discusses a crucial distinction, which underlies the seeming contradiction. It is the well-known distinction between the 'Right of Nature' and the 'Law of Nature,' both of which are normative concepts. In rough terms, the first essentially claims that one should seek one's advantage (in the State of Nature), the second that one should lay down this right, if doing so is advantageous. Of course, Hobbes goes through much greater pain in listing several particular Laws of Nature.[152] But what I am concerned with is how these Laws, as a whole, relate to the Right of Nature. In a reading of Hobbes, which explicitly stands in contrast with some of Hobbes' own statements, Gauthier describes the relation as follows.

> "Furthermore, the laws of nature would not themselves impose limitations on the right of nature. In telling us what is rationally required, they would serve the useful role of enabling us more effectively to act in accordance with reason, or to exercise the right of nature, but they would not affect that right in principle." (1969, p. 39)

Thus Gauthier claims that what is easily misunderstood as a conflict between two normative claims, is really a complementary set of claims.[153] He tries to overcome the apparent contradiction, by extending the notion of rationality to comprise both notions of 'Right of Nature' and of 'Laws of Nature.' But this will remain a harmful equivocation unless the relative authority of both normative notions is clearly defined in the unifying notion. This, I believe, is the main challenge for Gauthier's reading of Hobbes and, more importantly, for his *own* theory.

But before I move on to Gauthier's theory, let me spell out in different terms what the equivocation of "reason" at its worst might consists of. On the one hand, reason is defined as instrumental, narrowly understood. This notion of reason is often dubbed Humean.[154] In the State of Nature, given that I cannot change others and given

---

[151] Note, by the way, that Hobbes evidently does not accredit this reasoning motivational efficacy. Otherwise he would not introduce the sovereign as the ad hoc remedy.

[152] Compare Hobbes (1651a), Part I, Chapter 15.

[153] Compare Gauthier (1969), p. 62, and Gauthier (1993a), p. 30.

[154] One note of caution concerning historical roots seems appropriate at this stage. When Gauthier's approach is labelled "Humean" this is always just related to a fraction of Hume's theory, viz. his definition of the role of reason. It would thus be flawed to appeal to the Humean concept of justice, an artificial virtue, as a way out of the prisoner's dilemma. For one thing, Gauthier is not interested in a genealogy of morals; and that's what the discussion of natural and artificial virtues is all about. Furthermore, if Gauthier intended to plug in already established (artificial) virtues as pre-existent

that our end is to survive, we shall practise our Right of Nature. On the other hand, we recognise – from an equally consequentialist perspective – that cooperation would be advantageous. So rationality also seems to command cooperation. But this is not the same instrumental rationality. After all, we are still aware that, all other things equal, it would be even more advantageous if all cooperated while we defected in the State of Nature. So as long as nothing changes about the decision context, it would take a different notion of rationality to command us to cooperate, e.g. Kantian rationality.

It is not my intention to accuse Hobbes of this conflation. My point is rather that Hobbes might have inspired Gauthier to commit this mistake. Symptomatically, Gauthier directs our attention (1969, p. 46) to a passage in Hobbes' *De Corpore Politico* which has much of a Kantian air: "And therefore he that violateth a covenant, willeth the doing and the not doing of the same thing, at the same time, which is a plain contradiction". [155] The way that Hobbes uses the concept of volition in this quote almost suggests a kind of rational willing which must stand the universality test. In other words, while it would be no violation of Humean rationality at all to want all others to cooperate while we want to defect, it would be irrational (in a Kantian sense) to want to defect and thereby want all others to defect equally. [156]

Gauthier himself explicitly denies any commitment to (the universalistic aspect of) Kantian rationality.[157] But he does so most clearly in his paper *The Incomplete Egoist* in which he refutes the following alternative line of attack against the straightforward maximiser: "The egoist's commitment to maximization is thus rejected as insufficient for the universality inherent in true rationality" (1990, p. 268). Having said that, there are passages in which exactly such a universality of reasoning seems to do the work in defence of cooperation as a rational decision:

> "[...] as Hobbes insists it is rational for her to make an agreement only on terms that recognize her as equally rational with her fellows. Each must then be accorded the rights that the others would demand, were they in her position." (1993a, p. 38)

---

preferences his argument would fail to dispense with moral premises. Finally, Hume's genealogy of artificial virtues is based on a rather complex interplay of passions and reason, which Gauthier's framework, as it is, cannot accommodate.

[155] See Hobbes (1651b), p. 96.

[156] Compare Sugden (1991), p. 756. Sugden suggests that Kant might hold that 'Defect in the prisoner's dilemma' could not be *desirable*, while it might not be inconsistent to *will* that such a maxim become a universal law.

[157] Compare MbA, pp. 6, 183 and 236.

I do not claim that a latently universalistic supplement to an otherwise Humean conception of rationality is doing the work of establishing CM as a rational concept throughout MbA. As will be argued in Section 5, Gauthier seems to take different routes to support cooperation at different points of his argument. From a historical perspective, however, it might be interesting that one of these routes leaves a distinctly Kantian aftertaste.

4. The conceptual baggage of game theory

Game theory consists of a set of formal rules; that is why it works as a branch of mathematics. On top of these rules, we can give the entities and relations of game theory natural interpretations; that is why social (or indeed natural) scientists can use it. The most fundamental of these interpretations are the basic conceptual features like agents[158], choice, and rationality. Discussing these (normatively or positively) can, to some degree, feed back into the formal treatment. For instance, epistemic assumptions can be altered, solution concepts can be refined. But the reformulation of basic conceptions is limited by the formal representation applied. In fact, there is something like a conceptual baggage implied by the (current) game theoretical representation of social phenomena.

In this section, I try to outline some basic features of this baggage and show why they are in conflict with some of the notions that Gauthier employs in defence of CM. Of course, the formal framework of game theory is not immutable. A modified framework might not imply the same conceptual limitations, as said earlier. But let me start from the status quo.

Intrapersonal separability and dispositions

One conceptual feature of game theory is intrapersonal separability of choice. The easiest way to think of it is in terms of the extensive form of a game. Once the vertices (or decision nodes) are defined, there is nothing in the formal setup which demands that the rational agent perceive his choices as connected other than in their compound effect on the payoff. In fact, some authors have pointed out that there is an implicit separability assumption in game trees. This condition demands that any player reaching a decision node should play as if this node were the root of a new game (i.e. the

[158] I use 'agent' as a synonym for 'player' here, not in the sense technical sense of the agent normal form as before.

subgame that follows that node).[159] While this is might not correspond intuitively to how we make 'choices' and maybe not how we would like to be as agents, this is how rational choice theory *defines* (preferential) choice.

Of course, we could just define some alternative decision rule that demands continuity between choices. But this would only be of interest if it could be linked to the criterion of rationality as maximisation in some plausible and consistent fashion. Any attempted re-definition has to specify what becomes of the *meaning* of a particular decision node if it does not represent a genuine opportunity to choose. The onus of addressing the following questions lies with the reformer of game theory:[160] When precisely are choices to be seen in an interdependent bundle of choices by a rational agent?[161] What is the exact relative normative authority of choices at a single decision point and choices over bundles of such decision points?

This brings us back to the concept of a disposition to choose. A disposition is a feature of a player which is externally recognisable and somewhat stable. It supposedly alters the maximising calculus of the agent, as described in Section 2. An agent with a disposition does not see each choice as particular. So what is the scope of a disposition; i.e. how many choices does it bind? Does a disposition determine choice at each node or is there reconsideration?

I believe that trying to find an answer to these questions poses an enormous challenge, because, within the framework of game theory, a choice node *signifies* an opportunity to take one action or another. The only restraint on this liberty lies in the (rational) requirement of expected utility maximisation. If we want to introduce a two-tier approach according to which rational agents first decide how to bundle choices and then proceed to maximise their preference-satisfaction, we have to introduce a sufficiently rich formal framework.[162] It has to accommodate two levels of choosing and regiment the relations between these levels, without collapsing into the orthodox solution of modelling the bundling of choices as an option of pre-commitment. In a sense, the structure of the game (deciding over which nodes are real opportunities to choose and which ones are bound in plans or bundles) would have to be endogenous to

---

[159] Compare McClennen's (1990) definition of the condition SEP, as in Chapter 2. McClennen criticises the tacit acceptance of the assumption by Hammond (1988). More will be said in Section 6 of this chapter.

[160] These are not rhetorical questions, as I am not implying that these issues are impossible to resolve.

[161] "Bundle" is not a technical term here. I introduce it as a loose concept to indicate a series of choices, which are linked in the indicated way.

[162] Compare also my digression on Bratman (1987) below.

the game itself. And this new formal complexity would have to be justified by showing that it delivers additional intuitive appeal and that it is compatible with plausible solution concepts.

The normative authority of CM hinges on it being a maximisation principle, as far as I understand Gauthier. But then, if agents were maximising over dispositions, this would effectively bind choices. But a bound choice is not a choice at all, or at least should not be represented as the same thing. Conversely, if agents maximise over singular choices, dispositions are normatively irrelevant. In that respect, I share Nida-Rümelin's (1993) critique that CM might not be a maximisation principle after all: "But if dispositions do not bind the rational actor [as Nida-Rümelin takes Gauthier to hold], then the rational actor who acts in accordance with constrained maximization is consciously *not* maximizing – since the primary object of choice is the singular action" (p.56, italics in original). Gauthier therefore has to specify and justify when maximisation is a criterion of rationality and when it is not.

Also it seems that the notion of a *conditional* disposition to choose is a hybrid between a concept of rational choice and the notion of a psychological (non-maximising) effect on decisions. Gauthier cannot afford such a hybrid concept, when it comes to arguing for the plausibility of the notion of translucency, as will be addressed soon.

Interpersonal separability and mutual disinterest

Somewhat analogous to the conceptual feature of intrapersonal separability game theory's formal structure, as it is, carries with it the feature of *inter*personal separability.[163] This is a *normative* concept in the sense that players' maximands are exclusively their own utilities (even though these might be defined as other-regarding etc.). It is also a *causal* notion in the sense that players can only influence their own choices. Possibly these assumptions are implicit to some amended version of game theory. Maybe one such amended theory could model the unit of agency as endogenous, for example. (I will come back to this point.) But Gauthier himself seems to endorse separability as both a normative and a causal principle:

[163] Generally this need not be an assumption about people, as players could also represent groups of people etc. But the reading of "interpersonal" as "between people" seems adequate here.

> "I suggest that the principle rests on the capacity each person has for normative motivation, determining herself to act on what she takes to be reasons. [...] This account affords no rationale for the *interpersonal integration* of persons into a single normative unit, [...]." (1993b, p. 181, italics in original)

Furthermore, Gauthier defines agents to be mutually unconcerned in the sense that they take "no interest in one another's interests" (MbA, p. 100). Having said that, Gauthier (1993) also indicates that separate normative agents can, and should, coordinate themselves in the relevant contexts. What is important about this coordination, however, is that it comes posterior to the individual perspective. That is to say, the individual treats his action as the sole object of his choice (while recognising the strategic interdependence of the optimisation problem); conversely, the individual sees his action as subject only to *his* choice.[164]

It is all the more surprising then that Gauthier introduces[165] a necessary condition of rational choice in games which requires Pareto optimality:

> "A person acting interdependently acts rationally only if the expected outcome of his action affords each person with whom his action is interdependent a utility such that there is no combination of possible actions, one for each person acting interdependently, with an expected outcome which affords each person other than himself at least as great a utility, and himself a greater utility." (1990, p. 227)

But the concept of such "agreed" or "interdependent" action is in conflict with the concept of mutual unconcern. In some games Pareto-optimality might be a plausible way to select a unique equilibrium among multiple Nash-equilibria.[166] But in these cases the feature of Pareto optimality only *informs* the agent that the opponent has no incentive to deviate from an agreement. It contradicts the orthodox conception of agency in rational choice theory, however, to claim that social welfare as such *motivates*

---

[164] Note that if this were not the case it would be hard, e.g., to make sense of the formal representation of a two-person game. Compare the 'Symmetry Fallacy' in Section 5.

[165] He does so in a less concise way in MbA.

[166] Consider the following two-player coordination game.

		2	
		l	r
1	t	5, 5	0, 0
	b	0, 0	1, 1

or *justifies* choice, especially if this excludes individual utility-maximisation, as in the prisoner's dilemma.

Translucency

The two mentioned kinds of separability undermine the plausibility of the translucency assumption. I can think of two interpretations of translucency. The first would be some direct connectedness between the minds of agents. If this were granted, an agent could either directly tell the other's mind state, or he could directly infer something about the other's mind state by introspection. This would amount to evidential reasoning, according to which one can take one's own decision as some form of evidence about what the other player will do. My intuition, however, simply cannot follow this line of reasoning. I will say more on this in Section 5.[167]

The second option seems to be some (external) recognition of the opponent's disposition, which displays somehow. While I will not address the epistemological problems in depth, the plausibility of someone's type being 'written on his face' seems to lie in the presumption of some connection between emotional states and bodily signals. But dispositions were not supposed to be such stable psychological conditions, but rather the subject of rational thought. The idea that outcomes of rational deliberation are translucent like entrenched character traits, however, seems to fly in the face of lay psychology.

Having said that, acting honestly according to one's disposition might involve a different mental process than, e.g., faking a certain disposition in the sense of rational deceit. It might well be that the first kind of process can be recognised by others through some sort of "mind-reading" as it is suggested by Hurley (2005a). Her argument is that so-called mirror-neurons, which 'fire' not only simultaneously to own actions, but also when observing the same behaviour in others, could enable such outguessing the decision procedures underlying the behaviour of others (instead of just outguessing the behaviour directly). I am not competent to comment on the scientific aspects of this claim. However, I see a fundamental difficulty in fitting psychological notions such as dispositions into a rational choice framework. I will come back to both Hurley's approach and my critique later.

[167] For a discussion of evidential reasoning in both decision theory and game theory, compare e.g. Hurley (1991).

Preferences and counter-preferential choice

In MbA, Gauthier endorses a form of causal expected utility theory. He recognises utility as a unique measure of preference and treats this measure as subjective and exogenous. It is not subject to any deliberation about ends and, *a fortiori*, it is not subject to an objective value judgement.[168] Rationality is defined as the maximisation of utility. However, Gauthier states that the *revealed* preference interpretation of utility "leaves no conceptual space between preference and choices." (1993b, 186) This subsection deals with the question of how much of such conceptual space is needed to defend his theory and how much space, if any, is methodologically permissible.

The standard behaviourist position on this matter is clear. As was discussed in Chapter 1, revealed preference only captures coherent, maximising choice behaviour exhibited by an agent.[169] Hence utility is not a standard that provides ground to test for (or indeed prescribe) maximising behaviour. Such a view renders equilibria tautologies. But that, as Binmore points out, only pleases game theorists, who as "closet mathematicians" (1993, p. 137) see no problem in promulgating such tautologies at all.

Gauthier is still at least three long steps away from this position. The first step consists in the fact that Gauthier conceives of preferences as reasons to, not merely measures of, choice.[170] The second step is that, in the case of the 'truly' rational agent, these reasons do not directly determine choice, but determine it via commitment or plans, i.e. through CM. The third step is that so-understood preferences are the basis for the legitimate definition of value. The discussion of the third step exceeds the focus of the discussion at hand.[171] The first and second steps are crucial for my analysis. While I follow Gauthier's first step, I disagree with the second. Let me consider them in turn.

Gauthier thinks that if he intends to drive a wedge between the concept of straightforward maximisation and rationality, he needs to posit a notion of maximisation with further content. "But," he adds, "revealed preference is too impoverished a conception to provide this further content." (MbA, p. 27) While I believe this to be true, it obviously only amounts to an argument against revealed preference if we state an additional premise to the effect that this further content is somehow indispensable. And indeed Gauthier holds that further content is needed "If reason and value are to play

[168] Gauthier adds the normative constraint that only those preferences count which are informed and considered. This has little or no implication for the point at hand.
[169] Compare Binmore (1993), p. 136; Luce and Raiffa (1957), p. 32.
[170] Compare Gauthier (1988), p. 192, and Gauthier (1993b), pp. 185/6.
[171] I am aware that if Gauthier's project is to get any mileage, some (weaker) form this claim has to hold.

normative roles in the framework for understanding human action afforded by the theory of rational choice." (MbA, p. 27) Kurt Baier points out that this defence is, strictly speaking, false because revealed preference still leaves room for the imperative "Choose in such a way that your choice is capable of maximising interpretation." (1988, p. 28)

But this minimal normative role of preference cannot be what Gauthier is after. What he needs to establish for his purposes is that preferences give coherent reason to choose one action over another. Gauthier needs to *assume* his agent's preferences to be transitive and complete. Under these assumptions it makes sense to normatively require that the agent choose consistently. This way, preferences could play a non-trivial normative role.[172] I accept the refutation of a radical revealed preference interpretation.[173] But Gauthier still needs more conceptual space between preference and choice to accommodate CM. This brings me to Gauthier's second step.

> "What underlies constrained maximization is the recognition that such a [fully integrated] person would be less successful if she were to take her reasons for acting exclusively and directly from her utilities, than if she were to include among her reasons other considerations only indirectly related to her utilities." (1993b, p. 185)

Gauthier needs this conceptual gap between choice and preference to distinguish CM from straightforward maximisation. But it poses two problems. The first is that without a clear formal definition of when and how preferences determine the action of a rational player in this indirect way, rationality would cease to be a useful concept.

The second problem is that Gauthier introduces a methodological asymmetry by supposing the adoption of a mode of maximisation (a plan for that matter) to be preferentially constrained, but the plan-execution to be potentially counter-preferential. Finkelstein (2001) observes that Gauthier, contrary to his declared aim, essentially presents a two-tier account of choosing. At the first stage, an agent forms some sort of intention (i.e. a disposition) on future choices, at the second stage, he moves on to particular choices and sticks to his prior intention, or not. The intentions are formed such as to maximise expected utility. The story goes that the rational agent will stick to

[172] Unfortunately, Gauthier's discussion of this point seems to miss a step. He focuses on requirements for suitable preferences (such as congruence of expressed/attitudinal and behavioural preferences; and that preferences be informed, experienced and considered) without establishing first in how far preferences can play a non-trivial normative role at all.
[173] Compare Chapter 1.

these intentions when it comes to particular choices and not fall for what presents itself as a rational temptation in that situation, i.e. an opportunity to deviate from the intention to increase the expected payoff. Finkelstein asks why in the particular choice a different merit of rationality (e.g. a non-preferential one) should hold: "[...] why should choice over a series of choices be structurally different from the choice on a single occasion?" (2001, p. 63). To my knowledge, Gauthier does not give a satisfactory answer.

Digression on Bratman

Having said that, the integration of dispositions into the rationalistic picture of agent deserves a fair go. Bratman (1987) points at a more complex picture of planning, intending and choosing, which might seem to offer a 'third way' between following an irrevocable disposition and reconsidering action completely from moment to moment. He draws a picture of agents forming longer term plans in order to (intentionally) influence events beyond the present and to coordinate with fellow agents. Bratman furthermore believes that the process of reconsidering plans itself incurs costs and therefore suggests pursuit of a plan as a default procedure.[174] Plans should only be re-considered very selectively, where the criteria for re-consideration might or might not be deliberative. Notably however, Bratman explicitly refutes CM or resolute choice (as will be discussed below) as rational procedures.

In a more recent paper, Bratman (1999) addresses the Toxin puzzle and the problem of reciprocity represented in the finitely iterated prisoner's dilemma.[175] He agrees with Gauthier and McClennen that both 'paradoxes' are based on the orthodox view that (1) one cannot intend something now which one knows one cannot rationally intend later and (2) that in choice situations past intentions do not matter. However, he does not agree with their solutions, according to which one should follow through with a plan in these cases. So in some of the 'paradoxical' cases, Bratman defends sophisticated choice, which demands an adoption of only those plans we will prefer to execute at any future point, given my current information. While Bratman generally refutes a two-tier approach (as an alternative to his planning/re-consideration picture of agency), and therefore both resolute and sophisticated choice as general representations

[174] These costs are neither to be confused with a notion of opportunity costs of choosing one strategy rather than another, nor with the notion of expected opportunity costs (i.e. risk) of precommitment.
[175] Bratman points out that these cases are not at the centre of his theory. He believes they should not dominate the debate about planning. (1998, p. 56)

of rational agency,[176] he argues that the motive not to regret a plan in the future is rationally decisive in the Toxin and the prisoner's dilemma cases. In fact, he advocates the classical 'solution' of external precommitment in cases of an anticipated prisoner's dilemma. (1998, p. 62)

Let me make two remarks about lessons learnt from Bratman vis-à-vis CM. Firstly, if Gauthier *were* to adopt Bratman's richer framework, he would have to expand the semi-formal structure of his argument significantly. The different ways of planning and reconsidering and their respective costs for instance would demand a considerably richer (formal) system of representation. Secondly, it is not clear whether Gauthier could defend his conclusions on the basis of this richer framework. So Bratman's theory, as it stands, does not offer a third way.

5. Which argument does the job?

I argued in the previous sections that Gauthier is not always perfectly clear about which methodological move exactly is supposed to do the work in showing a normatively sound way out of non-cooperative equilibrium. This limits the fruitfulness of an exegesis of Gauthier's argument. So I ask myself in this section which possible argumentative routes Gauthier *could* take and whether any of these is both new and suitable. To make this exercise less tiring I shall only sketch those routes which turn out to be 'old news.'

Changing the game

A game in extensive form is defined, among other things, by its payoff structure and the decision nodes. If these are altered, the nature of the game is altered unless the new representation is strategically equivalent. As Binmore (1993, 1994) tirelessly points out, once the game is defined, the background assumptions and the applied solution concept *determine* the equilibrium of the game. While a great challenge might lie in defining a game such as to represent the real-world situation of interest, once the game is defined and the search for a solution begins, appealing to some notion that has not been taken account of in the formal structure of the assumptions, would be to get the definition of a

[176] Bratman believes that such a two-tier account of rational agency is in tension with the fact that "as time goes by we are located differently to our plans" (1999, p. 72) and that our causal control changes with our temporal location.

game in game theory wrong.[177] There are signs that Gauthier commits exactly that fallacy.

If Gauthier's notion of a disposition, for instance, is supposed to be some credible way for an agent of committing himself to a strategy at zero cost, then this can be perfectly represented in both the strategic and the extensive form of game; and it should.[178] Binmore (1993, pp. 138 ff) discusses this in detail. As a principle of methodological parsimony, there seems to be no point in inventing new rules for game theory if the what is supposed to be captured by these new rules can also be represented by orthodox game theory with analytical rigour.

A similar critique holds in the case of accounting for psychological factors like regret, internal sanctioning or a preference for cooperation per se, as hinted at in the previous section. If these are to play a role in the analysis, they should be captured in the preferences.[179] More precisely, either they can be represented in the payoffs or we cannot capture them in the framework of game theory at all.

Gauthier (1993b), however, explicitly excludes the possibility of reformulating CM as straightforward maximisation of payoff combined with internal sanctions for non-cooperative behaviour. The constraint is supposed to lie in the process of maximisation not in the maximand. This is crucial to Gauthier's account because he wants to avoid assuming moral preferences.

But note that even if Gauthier wanted to endorse process-regarding preferences, this move would (unlike modelling commitment) obscure the intuition of a consequentialist representation of decisions. Interpretations of games might become obscure.[180] Furthermore, it is yet to be seen whether it is possible, even in principle, to incorporate an intrinsic wish to overcome myopic choice into the preferences.[181]

Finally note that if Gauthier retreated to the methodological strategy of changing the game, this would mean dissolving the prisoner's dilemma rather than solving it.

---

[177] Psychological games as discussed in Chapter 3 would be an exception. This is if they can be established as coherent.

[178] It is often assumed that commitment incurs some cost, e.g. that of the risk of inflexibility. Compare McClennen (1998), for example.

[179] This is so far as this is conceptually coherent. Compare Chapters 2 and 3.

[180] Compare my discussion in Chapter 3, Section 6.

[181] It appears as if Gauthier is considering this approach when he writes: "Maximization itself imposes conditions on preference" (MbA, p. 38). Towards the end of MbA, when Gauthier reflects on the conceptual limits of *homo oeconomicus*, Gauthier wants to fight this impression by insisting that the constrained maximiser "does not allow herself to be disadvantaged or worsened by tuistic concerns, but she does not seek to take advantage of or better herself in relation to others by seeking to circumvent constraints non-tuistically justified" (MbA, p. 329).

Relaxing the background assumptions

The assumptions needed to derive the non-cooperative equilibrium in the repeated prisoner's dilemma are quite strong. A number of authors have argued for a relaxation of one or another of these to explain or to justify cooperation as rational action.

If we relax for instance the assumption that the number of games played is a known finite number, the famous Folk Theorems apply.[182] They imply – among an enormous range of further equilibria – that if the players believe, in each round, with sufficiently high probability that there will be a next round to the game at any given point of time, there is a subgame-perfect Nash equilibrium of strategies such that players always cooperate.

Other authors suggest relaxing the assumption of common knowledge of rationality.[183] Strictly speaking this relaxation can be subdivided into relaxing either the assumption of rationality of players or of their common knowledge of it (without relaxing rationality itself). I discuss the necessary conditions for the applicability of backward induction in Chapter 4.

Still another way to support a cooperative equilibrium could be to relax the assumption of perfect information about payoffs. In the (one-shot) prisoner's dilemma this would have to be imperfect information about each player's own payoffs, because here non-cooperation is the dominant strategy, i.e. optimal independent of what the other player does.

The point here is, again, not to explain any of these relaxations further, nor to look for evidence that Gauthier pursues any of them. The point is rather to argue that if Gauthier did base his argument on any of these changes of assumptions he would merge with game theoretical orthodoxy.

Redefining rationality, preferences and agency

Quite a different thing from relaxing the rationality assumptions, as just discussed, is to *alter* the notion of rationality. In fact, if this were doing the work in Gauthier's argument his approach would not collapse into other replies to the puzzle posed by the prisoner's dilemma.

But the onus lies with Gauthier to show that his new notion of rationality has greater *normative* appeal and can at the same time be conceptually and formally

[182] For a general version see Fudenberg & Maskin (1986).
[183] Compare e.g. Bicchieri (1989) or Pettit & Sugden (1989).

embedded in a revised game theoretical framework. It is crucial that the formal framework is modified in such a way that we can still make sense of the prisoner's dilemma as the starting point of the discussion about morality in the formal terms. Also the new version of game theory has to allow for coherent solution concepts.

Gauthier is explicit about his intention to change the notion of maximisation. Let me suppose that this move is supposed to do the work in rationalising cooperation. So is this a modification of game theory that lives up to the standards as just outlined? This is an open question. So let us remind ourselves what conceptual modifications are implied by CM.

Firstly, the interpretation of preferences changes. Preferences are no longer the direct determinants of rational choice but they determine choice only via an additional instance, labelled dispositions. But as I have argued before, their mediating role is not specified. Again, the challenge is to present a clear and convincing demarcation of the role of dispositions against that of single decisions, and to formalise that role rigorously.

Secondly, the notion of rationality changes for the same reasons. Rational choice no longer extends to decisions at each decision node, to use the formal term, but it rather extends to dispositions. This implies that at some decision nodes it may no longer count as rational to maximise utility, in fact it may no longer be rational to consider preferences in the choice at all. Gauthier needs to be much more precise about when and why exactly such counter-preferential choice counts as rational.

This leads to the third change implied. The definition of agency has become more complex. Gauthier's agent no longer chooses actions or strategies for action but dispositions.[184] The dispositions however are complex themselves, because they are modes of maximisation, namely constrained or straightforward maximisation. So effectively, Gauthier's agent has two levels of choice. The higher-level concerns modes of the lower level choice, the lower level choice concerns actions.

Game theory (as it is) cannot accommodate this complex picture. It is a representation of decision making which is not rich enough to distinguish first and second order choice; i.e. points at which dispositions are picked and points at which dispositions are taken into action, while still remaining choices in some sense.[185]

---

[184] At least this is supposed to be the case in the prisoner's dilemma. It is not clear whether something like CM applies to games other than the prisoner's dilemma, and if so, which.

[185] This notion of partly determined choice is not necessarily incoherent as will be seen in the next subsection. There is, of course, a semantic issue about what "choice" means under the assumption that it is in some way dependent on dispositions to choose. This will not be discussed.

Neither can game theory represent anything like a distinct rationality of coordinated choosing, because by definition of individual rationality, a choice directly depends on the expected payoff. The opponent player's presence might affect which strategy is the rational one to choose, but that does not change the criterion of what *counts* as rational.

But none of this is written in stone. It might be possible, for instance, to extend game theory in a coherent and rigorous fashion so as to model the unit of agency itself as endogenous. Regan (1980) has taken first steps to explore this possibility for coordination games. Since then the idea has been picked up several times, e.g. in Sugden (1991, p. 777), Hurley (1989) and Bacharach (2005).

Hurley (2005) elaborates on the idea quite extensively. Drawing on her theory on the importance of mirror neurons in (human) interaction, as mentioned earlier, she argues that a number of 'mind-reading' individuals can form a single unit of agency in certain social contexts (like the prisoner's dilemma) should this be instrumentally expedient from the individual perspective. Hurley sees an analogy between such adaptive "social heuristics" (p. 25) and Gigerenzer's (1999) concept of "simple heuristics" which are sensitive to (not primarily social) decision environments. As for plausible decision procedures used by members of agency-groups, she refers to Howard (1988) and Danielson (1992). As I will discuss later, these authors show, among other things, that there are decision procedures which cooperate in the prisoner's dilemma on the condition of encountering an equally structured programme as an 'opponent.'

I sympathise with Hurley's view that human beings sometimes form groups that can be attributed some form of collective agency, which dominates or even replaces individual agency.[186] However, I am not sure whether this can be captured by modelling the size of the unit of agency in a game as endogenous. I see (at least) two problems. The first is that one needs to justify why the normative unit of agency should remain individualistic even though the causal unit of agency has moved to group level. (Remember that the motivation for forming a group was the individual payoff.) The greater problem, however, is that the concept of an agent adopting a decision procedure, which is programmed on principles other than utility maximisation (although such programmes might turn out to be successful in terms of payoff) [187], seems in conflict with rational choice theory. I will come back to this point soon.

[186] Having said that, I find the ontology of collective agency extremely problematic.
[187] Compare Danielson (1992).

Changing the notion of agency might be a possible path for supporting some of Gauthier's results. But it is yet to be spelled out how this can be done. Also, any such attempt to rigorously formulate additional rules allowing for endogenous units of agency must be shown not to render the original puzzle, e.g. the prisoner's dilemma, which first motivated the expansion of rules, unintelligible as a puzzle or dilemma at all.

Non-deliberative analysis: Translucency, dispositions and evolution

In two discussions of Hobbes' challenge of the Foole, Gauthier concludes that Hobbes' best reply to the Foole is that if he refuses to adhere to justice (defined as cooperation) he will be "unfit to be a partner in society" (Gauthier 1993a, p. 32). In his own theory (MbA, Chapter 6), Gauthier bases a part of his argument for the rationality of cooperation on a calculus that describes agents as choosing dispositions, which are translucent. Cooperation is to be depicted as a stable equilibrium of some evolutionary process. I will now argue that this kind of reasoning is fundamentally distinct from what I will refer to as the deliberative (or rationalistic) approach, which underlies orthodox game theory.

First let me introduce two fundamentally different ways to apply game theory in moral analysis, i.e. rationalistic (or eductive) and evolutionary accounts. According to rationalistic accounts, the only thing that binds the agent in his choice, given all information about a game and payoffs, is his rationality. This makes it possible to read optimal strategies as hypothetical imperatives for the deliberating agents. Evolutionary accounts, on the other hand, do not start from the assumption that agents are rational.

Evolutionary game theory defines agents as picking strategies according to *some* decision procedure. Some well-defined algorithm or selection process defines how successfully decision procedures spread by reproducing themselves. Reproductive success depends on payoffs. So here payoff maximisation of a decision mechanism is defined as a condition of reproductive success, but it is not a given assumption about agents in the analysis.

Crucially, the question that motivates such analyses is genealogical: How is it that certain decision mechanisms come to exist? It is not a meaningful question to ask what an agent *should* do in these models.

In large parts of MbA it seems as if CM is intended to be a rationalistic notion. And I think that it has to be, if Gauthier's theory is to deliver a normative theory and not merely a genealogy of morals. But in some passages of MbA, Gauthier seems to argue

that choosing to be a constrained maximiser is effectively to decide against maximising behaviour at some later (or at least logically posterior) decision point and instead to dispose oneself to cooperate *on the condition* that the opponent is equally disposed. Sufficient translucency of these dispositions supposedly yields the cooperative equilibrium. Can this be captured in a rationalistic account?

Firstly, it has to be shown that the notion of such equilibrium is coherent. Secondly, it has to be shown that it is compatible with the rationalistic approach to claim that an agent should choose a non-maximising decision procedure. An argument for the second point will also have to defeat the claim that it would be possible and rational to counterfeit cooperative decisions.[188]

As to the issue of coherency, the putative problem is of course the self-referential structure of the disposition suggested. After all, both agents mutually specify their disposition depending on the other agent's choice. Holly Smith (1991) argues that the outcome of an encounter of two such agents is indeterminate between joint defection and joint cooperation. But John Howard (1988) shows that this need not be the case. His proof by counterexample, as further developed by Danielson (1991), is the presentation of a pair of equal computer programmes, which refer to each other. The programmes initiate cooperation if and only if their opponent programme is identified as symmetric. This solves what Talbott (1998) labels the '*special* indeterminacy problem' by showing that there exists at least *one* set of mutually conditioned cooperative dispositions which reach a cooperative equilibrium. Talbott argues that this solution cannot be extended so as to show that there can be a programme that definitely cooperates with *any* decision procedure that conditionally cooperates, no matter how it is programmed exactly. While I believe that this *general* problem of indeterminacy in itself poses a real problem to Gauthier's intuitive account, I will not address this problem here.[189]

Instead I would now like to argue that Howard's equilibrium of programmes is in conflict with rational choice theory. If we interpret Howard's (or Danielson's) computer programmes as playing the prisoner's dilemma, it would be false to say that they are choosing their strategies rationally. The definition of the decision rule –

[188] Strictly speaking, it has to be shown, thirdly, that if mutual conditional dispositions to cooperate are an equilibrium, and if it is a rational equilibrium, that this is the *only* (or the only relevant) rational equilibrium. It is not necessary here to address this last criterion.
[189] Note also that Howard's programme only applies to the prisoner's dilemma.

interpreted as an emotional rigidity, for instance – *replaces* utility-maximisation.[190] In rationalistic game theory, however, the rationality assumption functions as the starting point of the analysis. In such a theory, there is no coherent way to argue that rational agents could or should choose to give up rational choice in favour of some decision procedure, e.g. Howard's programme in the prisoner's dilemma, unless this is modelled explicitly as a further choice in the game.

So we might define a *new* game in which rational agents can choose between entering the prisoner's dilemma or simply adopting a disposition and therefore giving up any further choice in the rational sense. Once again, this would be to change the game played.[191] But it shows that there is *some* rigorous way to represent choice over decision procedures within rationalistic game theory. Unfortunately, this representation is equivalent to one in which agents (each separately) have the choice to cease to have any further choice for other reasons, like binding explicit contracts. This might be seen as an impoverished representation of what it means for an agent to choose to act according to a disposition as opposed to how it would feel to fulfil a contract. But at least it is not in conflict with the rational choice framework.

Apart from the purely methodological concerns, I find it intuitively hard to imagine a person who can engage in a rational choice between two or more dispositions, or a disposition on the side and maximising behaviour on the other. It is hard to picture an agent as acting out of a disposition, when a moment before he had a choice not to have any disposition at all. He would be able to rationally choose to have a psychological feature constraining his rational choosing. It might well be that humans developed dispositions which are triggered in social environments (consisting of further humans who are moved by dispositions). But such decision procedures seem to be alternatives, and probably more successful ones, to rational deliberation, not the result of rational deliberation.

It is also hard to make intuitive sense of the additional assumption of translucency, if dispositions are conceptualised as rationally chosen. As laid out earlier, the assumption of translucency seems to be plausible if we presume an agent who chooses with some sort of psychological constraint and signals this constraint via body language, for instance. The rational agent, whose constitution allows considering all

[190] Danielson (1991) is aware of these discrepancies between his and Gauthier's stance. He remarks that "any appeal to emotional rigidities in humans sits uncomfortable in Gauthier's rational choice framework with its appeal to perfect agents" (p. 317).
[191] Compare Danielson (2001).

choices, might be able to signal and communicate as well. But for him these signals are subject to choice. This means that rationally chosen dispositions, if there were such things, would not directly produce signals. They would not inadvertently reveal themselves. Instead, as Sayre-McCord (1991) argues, rationally motivated agents would attempt to deceive their opponents by sending misleading signals about their 'disposition,' if doing so were expedient.

Now Hurley (2005) argues that there will be an evolutionary "arms race" (p. 20) between ways of wrongly signalling dispositions and ways of detecting such deceit. I find this a plausible hypothesis. But independently of that, I do not believe that real persons can first deliberate freely which disposition to have and then, once they have made up their mind, be disposed in one way or another as if they never had a choice over their dispositions. I do not believe that such a strategic and potentially repeated 'return to innocence' is an option for real persons in the short-run. So my viewpoint on this subject is a qualified version of Hollis' (1996) comment:

> "Alternatively [a player trying to escape the prisoner's dilemma] could place the key out of his own psychological reach by, for instance, changing his dispositions, so that he is no longer disposed to steal a march on anyone who will play fair with him. [...] As with self-deception he must consciously monkey with his own motivations and than pretend to forget doing so. It is not clear how this can be done. [...]" (p. 116)

The symmetry fallacy and counterfactual reasoning

There is a last candidate for an argument doing the work in supporting CM as rational. This is Gauthier's repeated reference to the "equal rationality" between agents. "Since Jane and John are similarly characterized in those respects relevant to rationality, what is rational for Jane must, other things equal, be rational for John. [...] What is rational for Jane and John alike is to accept a bargain that is no less advantageous than it is for the other." (1988, p. 187) In his article *Between Hobbes and Rawls*, Gauthier points out that his rational and cooperative agent "does not subordinate her reason, as to a Hobbesian sovereign, but, through the laws of nature, *coordinates her reason* with that of her fellows." (1993a, p. 38, my italics) What should we make of these lines of reasoning? The first, rather uncharitable, interpretation would be to say that Gauthier latently introduces some sort of Kantian rationality, so that agents are defined to act

only so that they could wish that their choice of strategy would actually be every opponents' choice as well. I discussed this in Section 3.

A second interpretation would be to accuse Gauthier of the "fallacy of the twins" or the symmetry fallacy, which was first committed by Rapoport (1966).[192] Consider the following version of the fallacy: [193]

(1) Equally rational agents necessarily choose symmetrically in all symmetric games.

(2) As a symmetric game, the prisoner's dilemma must have a symmetric outcome, either $\{C, C\}$ or $\{D, D\}$.

(3) Both agents prefer $\{C, C\}$ to $\{D, D\}$.

---

(4) In the prisoner's dilemma, equally rational agents play $\{C, C\}$.

There are several flaws in this argument. For one thing the generalisation "in all symmetric games" renders (1) false. Binmore gives the counterexample of the Battle of the Sexes game.[194]

But even a weaker version of this premise, such as

(1a) (Equally) rational agents necessarily choose symmetrically in some games, e.g. the prisoner's dilemma.

would be incompatible with the meaning of a game. If equal rationality of players necessarily implied symmetry between players' actions, the game to be analysed would degenerate into a one-player game, or at least a non-strategic game.

Let us consider this point in counterfactual terms. To defend (1a) would be to claim that under the assumption of mutual knowledge of rationality each player would know that no matter what choice he makes, the opponent would move equally, no

---

[192] Compare Binmore (1993), p. 137.

[193] This is my reformulation. Binmore offers a less explicit version. $\{C, C\}$ stands for mutual cooperation, $\{D, D\}$ for mutual defection.

[194]

		2	
		l	r
1	t	-1,-1	2, 0
	b	0, 2	-1,-1

Binmore (1994, p. 204) points out that in this game the only symmetric equilibrium, which is in mixed strategies, is Pareto-dominated by an equilibrium in pure strategies. While Pareto efficiency is not a relevant criterion for determining rational equilibria in classical game theory, it seems to be crucial in Gauthier's own argument.

matter whether this would be rational to do for either one. So each player's counterfactual reasoning (about deviating from whatever strategy is rational for them to chose) would treat the belief in symmetrical choice as more robust than the belief in rational choice of both players, or in fact any other belief. This would supposedly be correct reasoning. Consequently, there would only be two possible outcomes under the assumption of mutual knowledge of rationality, $\{C, C\}$ and $\{D, D\}$. It is hard to spell out the "rational" reasoning of players in these games, but it would be something like: 'No matter what I do, my opponent will do the same. So let me choose the symmetric choice set that I prefer.'

This is not the place to return to an elaborate discussion of counterfactuals and solution concepts in games.[195] But I insist that the above lines of reasoning are profoundly misguided. If the symmetry of choice were supposed to be *derived* from the assumption of mutual knowledge of rationality, it would be confused to hold that the latter would be less robust in imagining counterfactual deviation from $\{C,C\}$. But only such a strange presumption would render the subjunctive conditional 'If one player played a strictly dominated strategy, the other would as well' true. Whatever else might motivate the symmetry claim, it cannot be derived from the assumption of mutual (or even common) knowledge of rationality.

Furthermore, Gauthier points out that players are to be defined as separate normative units. He would probably also endorse the assumption that players' choices are causally separable.[196] All players might be governed by the same rationality, but this does not mean that they govern each other in some direct manner. Note that if they could, we might not even need strategic analysis. Game theory would be the wrong tool of analysis. Furthermore, Gauthier's entire project of deriving morals individualistically would be undermined.

All in all, the fallacy of the twins could only be justified if symmetric rationality were redefined as radically incompatible with the usual notion of rationality (which might or might not be symmetric between agents). In a way, this redefinition can be described as even more radical than switching to a Kantian notion of rationality, because the Kantian rational agent, while reasoning with an interpersonally constrained

[195] Compare Chapter 4.
[196] Compare Gauthier (1993b), p. 181, as the quoted in Section 4.

set of acceptable motives, at least does not choose under the presumption that his choice will determine other agents' actions.[197]

6. Resolute choice

McClennen once classified his take on the problem of cooperation as being in "spirit" (1988, p. 117) compatible with Gauthier's approach. In this section, I will consider his notion of resolution as a potential conceptual defence (or alternative) to CM, mostly referring to his more recent arguments (1998, 2001) and occasionally to his most systematic work on the issue, "Rationality and Dynamic Choice" (1990).

McClennen agrees with Gauthier that it is rational for agents to cooperate in one-shot and repeated prisoner's dilemmas. His argument draws on an alleged analogy between certain *inter*personal and *intra*personal decision problems. More specifically, he considers the following reading of the problem of Ulysses and the Sirens (Figure 5.1) and the prisoner's dilemma with prior outside option (Figure 5.2).[198]

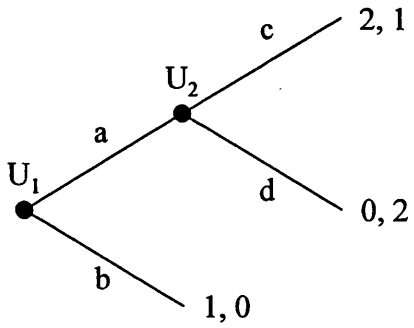


Figure 5.1

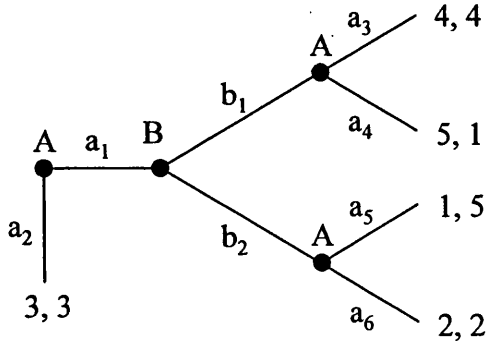


Figure 5.2

In the Ulysses problem, McClennen identifies three kinds of (more or less rational) choosers and respective outcomes. The 'myopic' chooser is the naïve version of Ulysses who sets sail for Ithaca without further precautions, ignoring that he will be tempted by Sirens later. U_1 (Ulysses before the temptation) chooses *a* with an intention to choose *c* next. But once the Sirens call, tempted Ulysses, U_2, chooses *d* and sails into doom. The 'sophisticated' chooser is a version of Ulysses endowed with the gift of

[197] The reader is asked to excuse this sloppy treatment of Kant.

[198] Both are borrowed from McClennen (1998). Notational amendments were made. Note that the payoffs vectors in Figure 5.1 represent the payoffs of U_1 and U_2 respectively. Analogously, the vectors in Figure 5.2 represent the payoffs for *A* and *B*.

anticipation. In his case, U_1 chooses b, binding himself to the mast (which incurs costs of precommitment), because he knows that otherwise the second decision would lead to tragedy. This is usually the version of Ulysses that game theorists label rational.

The 'resolute' chooser, finally, makes two *kinds* of decisions. The first is to adopt the plan to make the choice-sequence $\{a, c\}$. The second is to follow through with the plan. Resolute Ulysses can do so because he violates the (intrapersonal) separability condition of rationality.[199] The condition does not imply that players should ignore information learned at previous nodes; but it does imply that they should not see themselves as influenced by previous choices per se. McClennen (1990) argues that such violations of separability can be rational from a pragmatic viewpoint; he generally questions the status of separability as a condition of rationality. Resolute Ulysses, e.g., can be sure to reap the highest payoff. At least the situation presents itself this way in terms of pre-temptation preferences.[200] This is why McClennen calls this chooser the rational one. I come back to the issue of separability soon.

First, let me consider the alleged interpersonal analogue from Figure 5.2. Here a myopic pair of players, A and B, will behave as follows. A chooses a_1, entering a variant[201] of the prisoner's dilemma. Subsequently B will play his dominant strategy b_2. A will then choose a_6, leaving both players with a payoff of 2. If both players were sophisticated, A would choose the precommitment a_2. If both players were resolute on McClennen's account, however, A would choose a_1 knowing that both players chose the plan to mutually cooperate. This plan would be executed and both players would get higher payoffs than their myopic or sophisticated doppelganger. As in the Ulysses case the resolute choice implies a violation of the separability condition.

McClennen, is as ambiguous about what argument exactly is supposed to count as the defence of resolute choice as Gauthier is in his defence of CM.[202] And I will not

---

[199] Again, compare McClennen's (1990) definition of the condition SEP, as in Chapter 2.

[200] I will come back to the question of whether the criterion for rationality should be measured in terms of the pre-temptation preference order.

[201] McClennen calls it "a sequential form of the Prisoner's Dilemma." (1998, p. 27) To become a real prisoner's dilemma the two nodes for player A following player B's decision would have to merge into one information set. Note that since this is not the case, A could condition his second choice on the observable action of B. This might change the problem for Gauthier's analysis, since translucency on B's part might not be a crucial assumption any more. McClennen does not differentiate here.

[202] Consider, for instance, two lines of reasoning each of which supposedly provides crucial support for the justification and feasibility of resolute choice in different sources. (1) In early arguments it seems as if endogenous preferences for the resolute plan, triggered by the adequate decision context and the exogenous preferences, lead to the resolute outcome (e.g. McClennen (1988), p.111; (1990), p. 215). It is neither clearly defined how and when these endogenous preferences arise, nor in what sense deviating from the so-determined plan (in favour of exogenously preferred strategies) is still an option at all. Note

attempt to systematically evaluate the normative justifiability of resolute choice. Instead, I shall shortly comment on the status of the separability assumption and then focus on the analogy between intra- and interpersonal choice situations.

Separability as a rationality condition

Is separability a condition of rationality? I will not try to answer this question here. But a few words are due on McClennen's (1990) extensive treatment of that question. First of all, I appreciate McClennen's point that it would be hasty to dismiss the idea that rational agents could bind themselves to plans just because the choices necessary for executing that plan are not *feasible*. McClennen correctly argues that if feasibility is defined in terms of rationality, it is unfounded to dismiss separability as a rationality condition on the grounds that it violates feasibility.[203]

Secondly, McClennen argues that separability might be in conflict with another assumption, namely that all motives of action can be expressed in outcomes (and utilities), even modal or process-dependent ones.[204] The problem is that one cannot meaningfully separate a subtree from a greater game tree if the outcome-specifications on some of the leafs of the subtree contain modal features which refer to choices that precede the root of the subtree. I share this concern. Having said that, one could reply that this critical point does not threaten separability as a rationality assumption for meaningfully defined game trees, but rather raises concerns about the scope of problems that can be captured as games. Discussing this further would bring us back to the topic of Chapter 2.

The central point that McClennen brings forward against separability as a rationality assumption, however, is a pragmatic argument based on the success of the resolute chooser in terms of payoffs.[205] It is the weakest of McClennen's arguments, however. For one thing, it hinges on the plausibility of an admissibility criterion introduced by Levi (1980), which is not uncontroversial. But more importantly, it seems

---

that if in certain games reconsideration of a resolute plan is not an option at all, the representation of the game should effectively shrink, because certain strategies would be excluded by assumption and certain decision nodes would vanish. (2) In a recent, work McClennen (2001, p. 199) seems to claim that rational players (in the resolute sense) will choose the Pareto optimal outcome. The claim seems to be that this is the equilibrium *because* it has this social optimality characteristic. I have dealt with this suggestion in the subsection on interpersonal separability and agreed action in Section 4.

[203] Compare McClennen (1990), p. 14.

[204] Compare McClennen (1990), pp. 144 ff, 206 ff. The conflict is with the assumption labelled IRJ in Chapter 2.

[205] Compare McClennen (1990), pp. 183 ff.

to be building on an ambivalent interpretation about what a choice node represents. Sometimes it seems to represent the opportunity – and the necessity – to choose, sometimes it does not. In other words, one could argue that separability is a sensitive assumption simply in virtue of (1) the fact that game theory is defined as a consequentialist (and hence strictly forward-looking) theory[206] and (2) the convention that each node signifies a choice. But then again these definitions are not written in stone either, as discussed earlier.

This is not the place for developing this critique of McClennen's (1990) fully. So I settle for suggesting that some of my critical points against Gauthier could be brought forward against McClennen in a modified form.[207]

Analogies between intrapersonal and interpersonal conflicts

Let me focus on the McClennen's analogy between intrapersonal and interpersonal choice situations instead. It seems that the conceptual appeal of his redefinition of rationality in both kinds of situations hinges on whether the lesson learned from the virtues of resoluteness in the intrapersonal problem actually informs a solution to the interpersonal problem. So is there such an overarching intuition of resoluteness?

When addressing this point, it is instructive to analyse Gauthier's temporary attempt to integrate the notion of CM into the resolute choice framework. Consider the following quote by Gauthier in which he endorses McClennen's approach:

> "Resolute planning requires the rejection of separability as a condition on dynamic choice; in the context of a plan, it may be rational to choose what it would be irrational to choose otherwise [...] Thus even in a one-shot Prisoner's Dilemma, resolute choosers who can exchange credible assurances will plan to co-operate – and the credibility of their assurances will of course be in part dependent on their resoluteness." (1996, p. 237)

Some months after this statement Gauthier diverges from McClennen's approach again. He now believes that "the conditions under which resolute choice is rationally feasible, mistakenly assimilates the intrapersonal to the interpersonal problem, [...]" (1997, p. 2) In fact, Gauthier believes that the reasons to be resolute differ substantially in both cases. He can think of two reasons to be resolute in interpersonal choice – "one based

[206] That is, it is at least consequentialist in form. Compare Chapter 3, Section 2.
[207] Compare Orr (2004) for a detailed critique of McClennen's approach. It exhibits some structural parallels to my critique of Gauthier.

on the role of mutually beneficial interactions in realizing what ever concerns she has, the other based on the character of her relations with at least some other persons." (1997, p. 18) But, Gauthier argues, no such reasons matter in the intrapersonal case. The posterior self will not regard the prior self as an end in itself and it "will not regret having had concerns [it] no longer embraces."

Gauthier offers intuitive support for this claim in form of an odd version of the Ulysses example: As a boy, Mark finds girls disgusting. Mark also has the gift of resolute choice. Anticipating that he, like the older boys he observes, will nevertheless fall for girls as a teenager if no measures are taken, young Mark resolutely decides never to mingle with girls at all. Clearly, Gauthier concludes, Mark would make a huge mistake. Intuitions we have about deliberation tell us that Mark did badly in being resolute.

Gauthier further reasons that while in interpersonal contexts, such as prisoner's dilemmas, resolute choice is rational, the same does not hold in most cases of preference change of a single agent. He argues that there is reason to be resolute in intrapersonal contexts only if on some higher order account the prior agent's preference can be judged as "vanishing point preferences" which should take precedence over posterior, "proximate" preferences (1997, p. 20). [208] The cases thus described are known as cases of weakness of the will.

Gauthier's departure from McClennen's account brings before our eyes that McClennen's notion of resoluteness is ambiguous. On an intuitive level, it mixes the intuitions we have about intra- and interpersonal decision problems so that it becomes hard to properly understand either. In intrapersonal problems, there is the motive of trying to overcome (or at least limit) weaknesses of the will or erratic preference changes in order to be able to develop long term plans and realise them etc.[209] Resoluteness here means for one version of the self to overcome the other.[210] In

[208] Gauthier goes into more detail about a threshold which separates vanishing point and proximate preferences. The idea behind such a threshold is that even from a higher-order perspective giving into some temptation, i.e. proximate preferences might actually be the best way to go.

[209] This might lead to 'dictatorship of the present' in terms of preferences. One problem of such a dictatorship was exhibited by the example of Mark.

[210] This is to contradict McClennen's explicit appeal that resoluteness can only be a rational decision principle if it is not the case that one self tyrannises the other (1998, p. 18). What he suggests is that resoluteness is preferable as a coordination of the two selves, making both better of (Compare Figure 5.1). This is based on pragmatic arguments (compare also 1990, chapter 5), that both selves could be exploited if the agent does not act resolutely. But note, firstly, that the point of such pragmatic arguments is to expose the vice of intransitive preferences, not to show a way to live with them. Secondly, the paradigm

interpersonal problems, there is the motive of entering stable relations with other agents in order to reap the fruits of cooperation. If I can make sense of resoluteness in this case at all, it certainly has nothing to do with one agent "overcoming" the other's preferences, but rather with reaching a socially acceptable equilibrium. Subsuming the two intuitions under the term "resolute" is to conflate them.

This conflation is easily committed because there are some parallels between both decision problems on the formal level. In both cases the second best solution is commitment. In both cases, a 'Pareto suboptimal' (presuming that it makes sense to use that expression for two personas of the same person) outcome is achieved under orthodox assumptions. In both cases this problem could be overcome by altering the assumptions, e.g. to the effect that some decision nodes do not really represent preferential choice. This is how far the formal analogy stretches.

7. Summary and concluding remarks

I believe that Gauthier's project of supporting morality as rational, in the sense of cooperation in finitely iterated prisoner's dilemmas, is not successful. He fails to show how to expand game theory in intuitively plausible and formally rigorous ways to allow for his conclusion. Such a project might not be impossible in principle. But it is a great challenge to 'overcome' the non-cooperative equilibrium in the finitely iterated prisoner's dilemma without simply assuming the problem to be a different one or dispensing with a rationalistic approach altogether. To be sure, I share many intuitions with both Gauthier and McClennen. For instance, I believe that, in most cases, people should try to escape Pareto suboptima. I also share the idea that a person usually does and should form plans and exercise his agency with some temporal thickness and prudence.

However, I expressed doubts about whether these intuitions are best captured as a gradual modification of the rationality notion in the finitely iterated prisoner's dilemma. I have argued that this framework, as it stands, carries a certain conceptual baggage concerning intrapersonal and interpersonal separability of rational choices. Unless game theory undergoes some sort of coherent axiomatic reform, this baggage is not compatible with some of the concepts that seem to drive Gauthier's argument, such

---

example of Ulysses is not about tempted Ulysses not feeling so bad about being resolute when compared with being bound to the mast; this only is implied by the specific formalisation of Figure 5.1.

as dispositions, translucency of dispositions and counter-preferential choice at choice nodes.

But then again, it is hard to specify what exactly Gauthier's argument for constrained maximisation is supposed to consist in. I considered some options and argued that none of them are convincing in their current form. These options were (1) changing the prisoner's dilemma into some other game; (2) relaxing the background assumptions of game of the original dilemma; (3) redefining rationality or preferences or the concept of (fixed) individual agency; (4) exchanging the rational choice framework for some naturalised, non-rationalistic mode of explaining cooperation; and finally (5) altering the way we hypothesise about unilateral deviation from equilibrium.

I would now like to make a remark challenging the motivation behind 'overcoming' the mutual-defection outcome of the prisoner's dilemma. The single most important heuristic virtue of the prisoner's dilemma is that it captures a conflict of a (normatively plausible) notion of instrumental rationality and the (also normatively plausible) notion of Pareto optimality. Since we want the framework of our normative reasoning to be consistent, we do not like that fact. So what are we to do? Gauthier and McClennen think that it is best to unify our normative intuitions by extending the underlying notion of instrumental rationality while preserving the rest of the framework. But this manoeuvre might distract from real lessons to be learnt.

The puzzle of the prisoner's dilemma is hypothetical: *If* we accept the game theoretic background assumptions and the specific structure of the prisoner's dilemma as appropriate to capture the 'essence' of a situation calling for cooperation, then a conflict arises. In other words, *within* a certain representation of certain circumstances, one (prima facie plausible) intuition about what we ought to do, namely to maximise utility, is self-effacing.[211]

There seem to be three candidates for lessons to be learned from this puzzle, as read normatively. We could question whether we really captured a normative intuition in the representational form of game theory (or rational choice theory more generally); we could accept the general framework and deny that the prisoner's dilemma is the right game to capture our intuitions about cooperation; or we could appreciate that we have contradictory normative intuitions. Gauthier and McClennen, however want to go a fourth way. They want to redefine our normative intuitions in such a way that optimality

[211] This expression is borrowed from Parfit (1984).

and equilibrium in the finitely iterated prisoner's dilemma coincide. At second sight, it seems not so clear that this is a reasonable aim at all.

Finally I would like to comment on the relation between the normative puzzle caused by the finitely iterated prisoner's dilemma and developments in the theories of bounded rationality as represented, e.g., by Gigerenzer (1999) and Gigerenzer and Selten (2001). Among other things, these theories transfer insights from cognitive sciences into economic theory. They challenge the idea that human actions are ruled by some universal optimisation principle. Instead human actions are pictured as ruled by heuristic decision tools which adapt to choice environments. As such tools turn out to be successful, the theory refuses to treat heuristics as second-best decision principles.

I outlined some paragraphs earlier that describing agents as following heuristics might be something fundamentally different from describing them within the explanatory framework of rational choice theory. In rational choice theory, maximisation is a principle of understanding agents' actions. As I argued in Chapter 1, this maximisation principle has (a form of) normative appeal. Naturalising agency, in the way that Gigerenzer and others do, sacrifices this appeal for a descriptive account of which behaviour evolves under evolutionary pressure. Non-maximising heuristics may have their own normative appeal because they turn out to be successful under certain circumstances. But this success does not imply a recommendation *in terms of* rational choice because, for a given choice situation (which defines units of agency preferences etc.), the heuristics-driven choice can be in conflict with the maximising choice.

The normative puzzle of the finitely iterated prisoner's dilemma, and indeed many other 'paradoxes' of game theory, however, take their origin in the normative frame of reference of rational choice theory. Generalising a point made earlier, I doubt that these 'paradoxes' of game theory can be grasped as posing normative puzzles once the agents are no longer conceptualised as being (normatively) bound to some consistent principle of decision (but instead described as applying contingently successful tools). This concern is of a categorical nature and deserves separate treatment elsewhere.

Summary and outlook

This thesis was aimed at exploring some of the potentials and limitations of game theory as a method in social sciences, both normatively and descriptively. The persistent disillusionment of social scientists regarding the fruitfulness of game theory (not in the evolutionary sense) seems to spring from two sources. One is the theoretic struggle of finding solutions to games; the other is the practical struggle of applying games and game theoretic assumptions to real human interaction. I am convinced that crucial progress in both areas can be achieved by putting more thought into the conceptual foundations of game theory. Game theorists should ask themselves what game theoretic formalism intends to represent, how well it manages to do so, and what game theory can represent in principle. In that respect, the fate of game theory also depends on whether we can make more sense of games.

Let me first summarise each chapter's main results with special regard to the reoccurring themes. Next, I will point at overarching conclusions of the thesis. Finally, some suggestions for further research shall be made.

In Chapter 1, I argued that utility in game theory could not be interpreted as consistent choice because such an interpretation does not allow for a distinction between preference and beliefs. It fails to support meaningful hypothetical reasoning about situations that are not the actual equilibrium. This in turn undermines the justification of equilibrium play. A better notion of preference and utility in games also needs to make sense of the instrumental normativity that is inherent to the fundamental concept of a rational choice of strategy. I suggested interpreting utility as 'effective preference.' This was to draw a line between choice and preference, so as to allow for counterfactual as well as normative reasoning, while trying to preserve as the causal efficacy of the preference-concept. So Chapter 1 already touched on almost all crucial themes of this thesis such as counterfactuals in games, outcome individuation and the normative dimension of the rationality assumption.

Chapter 2 addressed an attack on the consistency of consequentialism as a foundational concept in decision and game theory. The attack aimed at exposing an incompatibility between the principle of rational justifiers in outcome individuation, normal-form/extensive-form coincidence and separability in decision trees. I argued that this trilemma dissolves if outcomes are fully specified, even regarding process-

dependent properties, before conclusions are being drawn from normal-form equivalence of extensive forms. This specific defence of consequentialism does not vindicate its legitimacy in general however. In fact, there are recalcitrant concerns regarding the specification of process-dependent outcomes as such. First, agents with process-dependent preferences might be made worse off simply by being given additional choices. This can only be avoided by introducing a substantial restriction to outcome-individuation. The second concern with process-dependent properties of outcomes is that they make it hard make sense of utilities in games in the sense of Savage's (1954) expected utility theory.

Chapter 3 raised further concerns regarding the process-dependence of outcomes and consequentialism in games. Theories about psychological games, which attempt to redefine games and solutions to games with belief-dependent payoffs, were questioned. The problems arise in attempting to justify solution concepts for such games, like psychological Nash equilibrium. The core problem here is a mutual endogeneity of payoffs and beliefs in such equilibrium obscuring its underlying intuition that there be no incentive for unilateral deviation, which is to be grasped in terms of players' hypothetical reasoning. Furthermore, I argued that psychological games could hardly capture intuitions about gratitude, disappointment and fairness. This, however, challenges the need for modelling psychological games. I raised three specific concerns concerning our intuitions: (1) Temporally extended learning plays no role in psychological game models. (2) Game theory as rational choice theory might be incompatible with modelling feelings that we have about each other's mode of choosing. (3) It is hard to grasp what it means to face process-dependence of utilities on top of an already complex utility definition that captures, e.g., other-regarding motives. So Chapter 3 shed some new light on the significance of the rationality assumption. Like the first two chapters, it also makes inferences from the significance of hypothetical reasoning in games.

Such reasoning stood in the centre of Chapter 4. It addressed the relations between epistemic assumptions, hypothetical reasoning and solution concepts in games. I argued with Stalnaker (1996) that the partitional representation of common knowledge of rationality fails to address the possibility that all players commonly and correctly believe that everyone will do the rational thing at all actual decision points but would cease to believe so should someone act otherwise. But Stalnaker's own approach requires very strong assumptions for supporting backward induction. I argued for

weaker assumptions supplemented with a restriction on what kinds of belief revision policies are permissible. In the context of the overall thesis, however, the central contribution of Chapter 4 was to explore the complexity of hypothetical reasoning in games, which all other chapters draw on.

In Chapter 5, I discussed Gauthier's modification of the orthodox notion of rationality as maximisation. This brought me back to the notion of separability of decisions in games, as first brought up in Chapter 2. Gauthier's normative concept of constrained maximisation is motivated by our puzzlement about the mutually defective equilibrium in finitely iterated prisoner's dilemmas. I argued, however, that the framework of game theory, as it stands, implies conceptual limitations concerning intrapersonal and interpersonal separability of rational choices, unless it undergoes rigorous reform. Such a reform might be possible. As it is, however, rationalistic game theory leaves no conceptual room for Gauthier's concepts of dispositions, translucency and counter-preferential choice at choice nodes. But then again, it is hard to specify what exactly Gauthier's argument for constrained maximisation is supposed to consist in. I have considered a number of exegetical options and expressed scepticism about all of them. One, for instance, builds on an implausible presumption about players' hypothetical reasoning in games. Another builds on an unclear interpretation of the meaning of decision points; it is an insufficiently specified relaxation of the separability assumption. A third possible approach seems to call for a hybrid between evolutionary and eductive game theory whose normative appeal remains obscure. Arguing that Gauthier suggests treating the unit of agency endogenously, finally, might be the most charitable interpretation. But it calls for further specification.

Turning to overarching conclusions now, the above chapters indicate that the limits of game theory as a tool of the social sciences, positive and normative, go deeper than the mere practical problems of application. Not only is it extremely difficult to specify what games are played in real human interaction and which epistemic assumptions are most applicable etc. It is not even clear whether game theory can in principle capture all human interaction of a strategic nature. This is for two interrelated reasons. Firstly, game theory and its solutions concepts seem to imply certain constraints on what can be captured in a game at all. Secondly, the methodological properties of game theory force the theorist to capture some aspects of choice differently than they 'naturally' present themselves. This limits the intuitive value added

of game theoretic explanation, especially when it comes to supporting the *understanding* of interaction.

As for the first kind of constraint, the dubious status of equilibria in psychological games might be the clearest example. Modelling a certain belief-dependence of utilities does not seem to be compatible with the notion of Nash equilibrium. It is intuitively not obvious whether and when exactly anyone faces such utilities. But if *some* players' payoffs were belief-dependent, there would be a limit to (game theoretic) equilibrium analysis. Also, even the definition of process-dependent outcomes might be limited if game theory were to make use of certain expected utility theories in assigning payoffs.

As for the second kind of limitation, I would like to mention some examples of problems. For one, our intuitions might fail us when process-dependent properties of an interactive choice situation are captured as part of the consequences of that situation. Accordingly, it might be hard to recognise process-relative motives of an agent, e.g. deontological principles or rule following, if they are captured as part of a consequence-regarding maximisation. Also, it is quite hard to appreciate and intuitively grasp an altruistic motive of an individual if this is represented as maximisation of other-regarding utility. Finally, a combination of these unnatural representations in a single game might render such a game completely cryptic when it comes to making sense of the players' reasoning. In Chapter 3, for example, I mentioned the case of someone maximising utilities which are not only other-regarding but also belief-dependent (relative to the game defined by the other-regarding payoffs).

But it is also a crucial insight of my discussion that methodologists should not 'throw out the baby with the bath water' when assessing the limits of game theory. Chapter 2, for example, explored a confused critique of some fundamental principles of consequentialism in decision theory and game theory. Formal consequentialism ought not to be dismissed for the wrong reasons. Furthermore, Chapter 5 offers two lessons. Firstly, one should not call for an amendment of the definition of rationality *because* of the puzzling outcome of the finitely iterated prisoner's dilemma, unless it is certain that the puzzle is actually *about* the finitely iterated prisoner's dilemma, not some other game. Secondly, if it can be argued that a game theory is genuinely in conflict with our normative intuitions, a reform of game theory should preserve its initial methodological advantage, namely rigour.

I would now like to point out two paths the game theorist can take if he is certain that his theory has inherent limits as a form of analysis. He can either explicitly limit the scope of game theory, or he can try to normatively 'intervene' by instructing the agents how to reason. Take the potential failure of formal consequentialism, for example. If it systematically fails to conform to a certain type of social phenomena, the theorist can either dismiss this type of situation as non-analysable or he can try to manipulate the situation by demanding, for instance, players to adhere to a substantial form of consequentialism. I do not offer a way to decide between these options here. But it certainly depends on whether, and in which sense, the game theorist takes game theory to be a normative theory.

Finally, let me point out a number of specific, potentially interesting areas of further research in the light of my thesis. One is to further investigate whether one could define games with belief-dependent payoffs supported by a dynamic framework, which is needed to model surprise, disappointment, gratefulness, etc. The follow-up question is whether a plausible solution concept could be applied to such games, and which.

Also, one could expand on the existing discussion about probabilities in games. One aim could be finding a plausible unifying interpretation of probability applicable in all contexts in which probability in games is referred to, such as the definition of expected utility, players' reasoning about mixed equilibria, and possibly belief-dependent payoffs.

Another challenge would be to attempt fully formalising an endogenous choice of the unit of agency within a given game. This formalisation would have to specify, among other things, under which well-defined circumstances players switch the unit of agency and what new solution principles this implies. The specific formal extensions of the existing game theoretic framework would have to be normatively justified. In particular, it would have to be argued how far a player can have an individual motivation to enhance the unit of agency but no individual motive to then deviate from this unit.

Last and most generally, one could further pursue the discussion on the concern raised at the end of the last chapter. Can frugal heuristics be understood as a principle of rational choice (in some integrated sense) or is it a categorical alternative?

References

Adams, E. (1970): 'Subjunctive and Indicative Conditionals' in *Foundations of Language* (6).

Aumann, R. J. (1974): 'Subjectivity and Correlation in Randomized Strategies' in *Journal of Mathematical Economics* (1), 67-96.

Aumann, R. J. (1976): 'Agreeing to Disagree' in *Annals of Statistics* (4), 1236-1239.

Aumann, R. J. (1987): 'Correlated Equilibria as an Expression of Bayesian Rationality', *Econometrica* (55), 1-18.

Aumann, R. J. (1995): 'Backward Induction and Common Knowledge of Rationality' in *Games and Economic Behaviour* (8), 6-19.

Aumann, R. J. (1998): 'On the Centipede Game' in *Games and Economic Behaviour* (23), 97-105.

Aumann, R. and L. J. Savage (1971): 'Letter from Robert Aumann to Leonard Savage and Letter from Leonard Savage to Robert Aumann' in *Essays on Economic Decisions under Uncertainty*, J. H. Dreze (ed.), Cambridge University Press, Cambridge, 1987.

Bacharach, M. O. L. (1987): 'A Theory of Rational Decision in Games' in *Erkenntnis* (27), 17-55.

Bacharach, M. O. L. (1994): 'The Epistemic Structure of a Theory of a Game' in *Theory and Decision* (37), 7-48.

Bacharach, M. O. L., L.-A. Gérard-Varet, Ph. Mongin and H. S. Shin (eds.) (1997): *Epistemic Logic and the Theory of Games and Decisions*. Kluwer, Amsterdam.

Bacharach, M. O. L. and S. Hurley (1991): *Foundations of Decision Theory: Issues and Advances*. Blackwell, Oxford.

Baier, K. (1988): 'Rationality, Value and Preference' in *Social Philosophy and Policy* (5), 17-45.

Battigalli, P. (1996): 'Strategic Rationality Orderings and the Best Rationalization Principle' in *Games and Economic Behaviour* (13), 178-200.

Ben-Porath, E. (1997): 'Rationality, Nash Equilibrium and Backward Induction in Perfect Information Games' in *Review of Economic Studies* (64), 23-46.

Bernheim, B. D. (1984): 'Rationalizable Strategic Behavior' in *Econometrica* (52), 1007-1028.

Bicchieri, C. (1988): 'Strategic Behaviour and Counterfactuals' in *Synthese* (76), 135-169.

Bicchieri, C. (1989): 'Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge' in *Erkenntnis* (30), 69-85.

Binmore, K. (1987/1988): 'Modeling Rational Players I and II' in *Economics and Philosophy* (3), 179-214 and (4), 9-55.

Binmore, K. (1993): 'Bargaining and Morality' in *Rationality, Justice and the Social Contract: Themes from Morals by Agreement*, D. Gauthier and R. Sugden (eds.), Hemel Hempstead, Harvester Wheatsheaf.

Binmore, K. (1994): *Game Theory and the Social Contract Volume I.* MIT Press, Cambridge (MA).

Binmore, K. and H. Shin (1992): 'Algorithmic Knowledge in Game Theory' in *Knowledge, Belief and Strategic Interaction*, C. Bicchieri and M. L. Dalla Chiara (eds.). Cambridge University Press, Cambridge (MA).

Bolker, E. D. (1966): 'Functions Resembling Quotients of Measures' in *Transactions of the American Mathematical Society* (124), 292-312.

Bonanno, G. (1991): 'The Logic of Rational Play in Games of Perfect Information' in *Economics and Philosophy* (7), 37-65.

Bradley, R. (2005): 'Axiomatic Utilitarianism and Probability Homogeneity', *Social Choice and Welfare*, forthcoming, Kluwer Academic Publishers.

Brandenburger, A. (1992): 'Knowledge and Equilibrium in Games' in *Journal of Economic Perspectives* (6), 83-101.

Brandenburger A. and E. Dekel (1985): *Hierarchies of Beliefs and Common Knowledge*. Mimeo, Harvard University.

Bratman, M. E. (1987): *Intention, Plans and Practical Reasoning*. Harvard University Press, Cambridge (MA).

Bratman, M. E. (1998): 'Following Through With One's Plans: A Reply to David Gauthier' in *Modeling Rationality, Morality, and Evolution*, P. A. Danielson (ed.), Oxford University Press, Oxford.

Bratman, M. E. (1999): 'Toxin, Temptation, and the Stability of Intention' in *Faces of Intention*, Cambridge University Press, Cambridge.

Broome, J. (1990): 'Bolker-Jeffrey Expected Utility Theory and Axiomatic Utilitarianism' in *Review of Economic Studies* (57), 477-502.

Broome, J. (1991a): *Weighing Goods*. Blackwell, Oxford.

Broome, J. (1991b): 'Can a Humean be moderate?' in *Value, Welfare and Morality*, R. G. Frey and C. Morris (eds.), Cambridge University Press, pp. 51-73.

Broome, J. (1991c): "Utility" in *Philosophy and Economics* (7), 1-12.

Broome, J. (1999): 'Normative Requirements' in *Ratio* (12), 398-419.

Broome, J. (2002): 'Practical Reasoning' in *Reason and Nature: Essays in the Theory of Rationality*, J. L. Bermúdez and A. Millar (eds.), Clarendon Press, Oxford.

Broome, J. (1993): 'Can a Humean Be Moderate' in *Value, Welfare, and Morality*, R.G. Frey and C. W. Morris (eds.), Cambridge University Press, Cambridge.

Clausing, T. (2004): 'Belief Revision in Games of Perfect Information' in *Economics and Philosophy* (20), 89-115.

Danielson, P. A. (1991): 'Closing the Compliance Dilemma: How It's Rational to Be Moral in a Lamarckian World' in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals By Agreement*, P. Vallentyne (ed.), Cambridge University Press, Cambridge.

Danielson, P. A. (1992): *Artificial morality: virtuous robots for virtual games.* Routledge, New York.

Danielson, P. A. (2001): 'Which Games Should Constrained Maximizers Play?' in *Practical Rationality and Preference: Essays for David Gauthier*, C. W. Morris and A. Ripstein (eds.), Cambridge University Press, Cambridge.

Debreu, G. (1959): 'Topological Methods in Cardinal Utility Theory' in *Mathematical Methods in the Social Sciences*, 1959, K. J. Arrow, S. Karlin and P. Suppes (eds.), Stanford University Press, pp. 16-26.

Diamond, P. A. (1967): 'Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment' in *The Journal of Political Philosophy* (75), 765-766.

Ellsberg, D. (1954): 'Classic and Current Notions of 'Measurable Utility'' in The Economic Journal (64), 528-556.

Elster, J. (1979): *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge University Press, Cambridge.

Finkelstein, C. (2001): 'Rational Temptation' in *Practical Rationality and Preference: Essays for David Gauthier*, C. W. Morris and A. Ripstein (eds.), Cambridge University Press, Cambridge.

Friedman, M. (1953): *Essays in Positive Economics.* Chicago University Press, Chicago.

Fudenberg, D. and E. Maskin (1986): 'The Folk Theorem in Repeated Games with Discounting and Incomplete Information', *Econometrica (54)*, 533-554.

Gärdenfors, P. (1988): *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge (MA).

Gauthier, D. (1969): *The Logic of Leviathan.* Clarendon Press, Oxford.

Gauthier, D. (1979): 'Thomas Hobbes: Moral Theorist' in *The Journal of Philosophy (76)*, 547-559.

Gauthier, D. (1986): *Morals by Agreement.* Oxford University Press, Oxford.

Gauthier, D. (1988): 'Morality, Rational Choice, and Semantic Representation: A Reply to My Critics' in *Social Philosophy and Policy* (5), 173-221.

Gauthier, D. (1990): 'The Incomplete Egoist' in *Moral Dealing*, Cornell University Press, Ithaca (NY).

Gauthier, D. (1993a): 'Between Hobbes and Rawls' in *Rationality, Justice and the Social Contract: Themes from Morals by Agreement*, D. Gauthier and R. Sugden (eds.). Hemel Hempstead, Harvester Wheatsheaf.

Gauthier, D. (1993b): 'Uniting Separate Persons' in *Rationality, Justice and the Social Contract: Themes from Morals by Agreement*, D. Gauthier and R. Sugden (eds.). Hemel Hempstead, Harvester Wheatsheaf.

Gauthier, D. (1996): 'Commitment and Choice: An Essay on the Rationality of Plans' in *Ethics, Rationality, and Economic Behaviour*, F. Farina, F. Hahn and S. Vanucci (eds.). Clarendon Press, Oxford.

Gauthier, D. (1997): 'Resolute Choice and Rational Deliberation: A Critique and a Defence' in *Noûs* (31), 1-25.

Geanakoplos, J. (1992): 'Common Knowledge' in *Journal of Economic Perspectives* (6), 53-82.

Geanakoplos, J., Pearce, D. and E. Stacchetti (1989): 'Psychological Games and Sequential Rationality' in *Games and Economic Behavior* (8), 56-90.

Gigerenzer, G. and R. Selten (eds.) (2001): *Bounded Rationality: The Adaptive Toolbox.* MIT Press, Cambridge (MA).

Gigerenzer, G., Todd, P., and the ABC Research Group (1999): *Simple Heuristics that Make us Smart.* Oxford University Press, New York.

Greenberg, M. S. and D. Frisch (1972): 'Effect of Intentionality on Willingness to Reciprocate a Favor' in Journal *of Experimental Social Psychology* (8), 99-111.

Güth, W., Schmittberger, R. and B. Schwarze (1982): 'An Experimental Analysis of Ultimatum Bargaining' in *Journal of Economic Behavior and Organization* III, 367-88.

Halpern, J. Y. (1999): 'Hypothetical Knowledge and Counterfactual Reasoning' in *International Journal of Game Theory* (28), 315-330.

Halpern, J. Y. (2001): 'Substantive Rationality and Backward Induction' in *Games and Economic Behavior* (37), 425-435.

Hammond, P. J. (1988): 'Consequentialist Foundations of Expected Utility', *Theory and Decision* (25), 25-78.

Hampton, J. (1998): *The Authority of Reason.* Cambridge University Press, Cambridge.

Hargreaves Heap, S. and Y. Varoufakis (2004): *Game Theory: A Critical Text.* Routledge, London.

Harsanyi, J. C. (1967-68): 'Games of Incomplete Information Played by Bayesian Players' in *Management Science* (14), 159-182, 320-334, 486-502.

Harsanyi, J. C. (1982): 'Subjective Probability and the Theory of Games: Comments on Kadane and Larkey's Paper' in *Management Science* (28), 120-124.

Hausman, D. M. (2000): 'Revealed Preference, Belief and Game Theory' in *Economics and Philosophy* (16), 99-115.

Hausman D. M. (2005a): 'Consequentialism and Preference Formation in Economics and Game Theory', forthcoming in a supplemental issue of *Philosophy.* (The page numbers in my references refer to a draft copy.)

Hausman D. M. (2005b): 'Sympathy, Commitment and Preference' in *Economics and Philosophy* (21), 33-50.

Hintikka, J. (1962): *Knowledge and Belief.* Cornell University Press, Ithaca.

Hobbes, T. (1651a): *Leviathan.* Reprinted by Penguin Books, London, 1968.

Hobbes, T. (1651b): *The English Works of Thomas Hobbes of Malmesbury.* Collected and edited by Sir W. Molesworth, Bohn, London, 1839.

Hollis, M. (1996): *Reasons in Action: Essays in the Philosophy of Social Science.* Cambridge University Press, Cambridge.

Houthakker, H. S. (1950): 'Revealed Preference and Utility Function' in *Economica* (17), 159-174.

Howard, J. (1988): 'Cooperation in the Prisoners' Dilemma' in *Theory and Decision* (24), 203-213.

Hubin, D. C. (2001): 'The Groundless Normativity of Instrumental Rationality' in *The Journal of Philosophy* (98), 445-468.

Hume, D. (1741): *A Treatise of Human Nature*. Reprinted by L.A. Selby-Bigge, Clarendon Press, Oxford, 1960.

Hurley, S.L. (1989): *Natural Reasons*. Oxford University Press, New York.

Hurley, S.L. (1991): 'Newcomb's Problem, Prisoners' Dilemma, and Collective Action' in *Synthese* (86), 173-196.

Hurley, S.L. (1994): 'A New Take from Nozick on Newcomb's Problem and Prisoners' Dilemma' in *Analysis* (54), 65-196.

Hurley, S.L. (2003): 'The Limits of Individualism Are Not the Limits of Rationality' in *Behavioral and Brain Science* (26), 164-165.

Hurley, S.L. (2005): 'Social Heuristics That Make Us Smarter: Instrumental Rationality, Collective Activity, and Mind-Reading' in *Philosophical Psychology*, forthcoming.

Jeffrey, R.C. (1983). *The Logic of Decision, 2nd edition*. University of Chicago Press, Chicago.

Kadane, J. B. and D. Larkey (1982): 'Subjective Probability and the Theory of Games' in *Management Science* (28), 113-120.

Kadane, J. B. and D. Larkey (1983): 'The Confusion of Is and Ought in Game Theory' in *Management Science* (29), 1365-1379.

Kahneman, D. and A. Tversky (1979): 'Prospect Theory: An Analysis of Decision Under Risk' in *Econometrica* (47), 263-291.

Kant, I. (1785): *Groundwork of the Metaphysics of Morals*. Translated by A. Zweig, Oxford University Press, Oxford, 2002.

Kohlberg, E. and J.-F. Mertens (1986): 'On the Strategic Stability of Equilibria' in *Econometrica* (54), 1003-1037.

Korsgaard, C. (1997): 'The Normativity of Instrumental Reason' in *Ethics and Practical Reason*, G. Cullity and B. Gaut (eds.), 215-254.

Kripke, S. (1963): 'Semantical Considerations on Modal Logic' in *Acta Philosophica Fennica* (16), 83-94.

Kreps, D. and R. Wilson (1982): 'Sequential Equilibria' in *Econometrica* (50), 863-894.

Levi, I (1980): *The Enterprise of Knowledge*. MIT Press, Cambridge (MA).

Levi, I. (1991): 'Consequentialism and Sequential Choice' in *Foundations of Decision Theory*, M. O. L. Bacharach and S. Hurley (eds.), Blackwell.

Lewis, D. (1973): *Counterfactuals*. Blackwell, Oxford.

Lismont, L. and Ph. Mongin (1994): 'On the Logic of Common Belief and Common Knowledge' in *Theory and Decision* (37), 75-106.

Little, I. M. D. (1949): 'A Reformulation of the Theory of Consumers' Behaviour' in *Oxford Economic Papers* (1), 90-99.

Luce, D. and H. Raiffa (1957): *Games and Decisions*. Wiley, New York.

Mandler, M. (2001): 'A Difficult Choice in Preference Theory: Rationality Requires Transitivity or Completeness, But Not Both' in *Practical Rationality and Preference: Essays for David Gauthier*, C. W. Morris and A. Ripstein (eds.). Cambridge University Press, Cambridge.

McClennen, E.F. (1988): 'Constrained Maximization and Resolute Choice' in *Social Philosophy and Policy* (5), 17-45.

McClennen, E.F. (1990): *Rationality and Dynamic Choice: Foundational Explorations.* Cambridge University Press, Cambridge.

McClennen, E.F. (1998): 'Rationality and Rules' in *Modeling Rationality, Morality, and Evolution*, P. A. Danielson (ed.). Oxford University Press, Oxford.

McClennen, E.F. (2001): 'The Strategy of Cooperation', in *Practical Rationality and Preference: Essays for David Gauthier*, C. W. Morris and A. Ripstein (eds.). Cambridge University Press, Cambridge.

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

Nash, J. F. (1951): 'Noncooperative Games' in *Annals of Mathematics* (54), 289-295.

Nida-Rümelin, J. (1993): 'Practical Reason, Collective Rationality and Contractarianism' in *Rationality, Justice and the Social Contract: Themes from Morals by Agreement*, D. Gauthier and R. Sugden (eds.). Hemel Hempstead, Harvester Wheatsheaf.

Orr, S. (2004): 'Resolute Choice and Instrumental Reason: A Critique of McClennen', mimeo (accessible on http://www.ucl.ac.uk/~ucesswo/ [status June 2005]).

Parfit, D. (1984): *Reasons and Persons*. Oxford University Press, Oxford.

Pearce, D. G. (1984): 'Rationalizable Strategic Behavior and the Problem of Perfection' in *Econometrica* (52), 1029-1050.

Pettit, Ph. and R. Sugden (1989): 'The Backward Induction Paradox' in *The Journal of Philosophy* (86), 169-182.

Rabin, M. (1993): 'Incorporating Fairness into Game Theory and Economics' in *American Economic Review* (83), 1281-1303.

Rabinowicz, W. (1998): 'Grappling with the Centipede: Defence of Backward Induction for BI-terminating Games' in *Games and Economic Behavior* (14), 95-126.

Rapoport, A. (1966): *Two-Person Game Theory*. University of Michigan Press, Ann Arbor.

Rawls, J. (1971): *A Theory of Justice*. Harvard University Press, Cambridge (MA).

Regan, D. (1980): *Utilitarianism and Co-operation*. Clarendon Press, Oxford.

Reny, P. (1992a): 'Rationality in Extensive Form Games' in *Journal of Economic Perspectives* (6), 103-118.

Reny, P. (1992b): 'Common Knowledge and Games with Perfect Information' in *Knowledge, Belief and Strategic Interaction*, C. Bicchieri and M. L. Dalla Chiara (eds.). Cambridge University Press, Cambridge (MA).

Reny, P. (1992c): 'Backward Induction, Normal Form Perfection and Explicable Equilibria' in *Econometrica* (60), 627-649.

Samet, D. (1996): ' Hypothetical Knowledge and Games with Perfect Information' in *Games and Economic Behavior* (17), 230-251.

Samuelson, P. A. (1938): 'A Note on the Pure Theory of Consumers' Behaviour' in *Economica* (5), 61-71.

Samuelson, P. A. (1952): 'Probability, Utility, and the Independence Axiom' in *Econometrica* (20), 670-678.

Savage, L. J. (1954): *The Foundations of Statistics*. Wiley, New York.

Sayre-McCord, G. (1991): 'Deception and Reasons to Be Moral' in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals By Agreement*, P. Vallentyne (ed.). Cambridge University Press, Cambridge.

Selten, R. (1975): 'Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games' in *International Journal of Game Theory* (4), 25-55.

Selten, R. and U. Leopold (1982): 'Subjunctive Conditionals in Decision and Game Theory' in W. Stegmüller, W. Balzer, W. Spohn (eds.), *Studies in Contemporary Economics (Vol. 2): Philosophy of Economics*. Springer, Berlin.

Sen, A. (1971): 'Choice Functions and Revealed Preference' in *Review of Economic Studies* (38), 307-317. Reprinted in *Choice Welfare and Measurement*, Blackwell, Oxford, 1982.

Sen, A. (1973): 'Behaviour and the Concept of Preference' in *Economica* (40), 241-259. Reprinted in *Choice Welfare and Measurement*, Blackwell, Oxford, 1982.

Sen, A. (1976): 'Rational Fools' in *Scientific Models of Man*. Oxford University Press, Oxford. Reprinted in *Choice Welfare and Measurement*, Blackwell, Oxford, 1982.

Sen, A. (1997): 'Maximization and the Act of Choice' in *Econometrica* (65), 745-779.

Shin, H.S. (1992): 'Counterfactuals and a Theory of Equilibrium in Games' in *Knowledge, Belief and Strategic Interaction*, C. Bicchieri and M. L. Dalla Chiara (eds.). Cambridge University Press, Cambridge (MA).

Smith, H. (1991): 'Deriving Morality from Rationality' in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals By Agreement*, P. Vallentyne (ed.). Cambridge University Press, Cambridge.

Stalnaker, R. (1968): 'A Theory of Conditionals' in *Studies in Logical Theory, American Philosophical Quarterly* (Monograph Series, 2), Blackwell, Oxford, 98-112.

Stalnaker, R. (1994): 'On The Evaluation of Solution Concepts' in *Theory and Decisions* (37), 49-73.

Stalnaker, R. (1996): 'Knowledge, Belief and Counterfactual Reasoning in Games' in *Economics and Philosophy* (12), 133-163.

Stalnaker, R. (1998): 'Belief Revision in Games: Forward and Backward Induction' in *Mathematical Social Sciences* (36), 31-56.

Stalnaker, R. (1999): 'Extensive and Strategic Forms: Games and Models for Games' in *Research in Economics* (53), 293-319.

Skyrms, B. (1994): 'Adams Conditionals' in *Probability and Conditionals: Belief Revision and Rational Decision*, E. Eells and B. Skyrms (eds.). Cambridge University Press, Cambridge (MA).

Skyrms, B. (1998): 'Subjunctive Conditionals and Revealed Preference' in *Philosophy of Science* (65), 545-574.

Sugden, R. (1991): 'Rational Choice: A Survey of Contributions from Economics and Philosophy' in *The Economic Journal* (101), 751-785.

Suppes P., D. Davidson and J. C. C. McKinsey (1955): 'Outlines of a Formal Theory of Value' in *Philosophy of Science* (22), 140-160.

Talbott, W. J. (1998): 'Why We Need a Moral Equilibrium Theory' in *Modeling Rationality, Morality, and Evolution*, P. A. Danielson (ed.). Oxford University Press, Oxford.

Valentyne, P. (1993): 'The Connection between Prudential Goodness and Moral Permissibility', *Journal of Social Philosophy* (24), 105-28.

Verbeek, B. (2001): Consequentialism, Rationality and the Relevant Description of Outcomes' in *Economics and Philosophy* (17), pp. 181-205.

Von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.