# Newcomb's Problem

by
Marion Ledwig

Philosophical Dissertation

Submitted to the
Department of Philosophy,
Faculty of Philosophy,
University of Constance,

in the January of 2000

# Table of Contents

For my Mother

In Memory of Peter Lanz (✝ 1997) and
my Father (✝ 1996)

I want to thank the Evangelische Studienwerk e. V. Villigst for funding my dissertation with a three year scholarship.
I would like to thank my philosophical teachers Wolfgang Spohn and Gottfried Seebaß.
Furthermore, I wish to thank Phil Dowe, Richard Jeffrey, James M. Joyce, Henry Kyburg, David Lewis, Alfred Schramm, Brian Skyrms, and especially Andreas Blank, Wlodek Rabinowicz, and J. Howard Sobel for very helpful comments.

# Introduction

## *Why Should We Look at Newcomb's Problem?*

Practical philosophy can be divided into moral theory, political philosophy, aesthetics, anthropology, and the theory of practical rationality. The latter has normative and descriptive aspects. While normative rationality theory tries to do justice to our normative intuitions, principles, and arguments by building theories which state what is rational in the concept-explicatory sense, descriptive rationality theory deals with what is actually rational in the sense that it describes how people actually act in various situations.

Nozick (1969) confronted normative rationality theory with a decision-theoretic problem, namely Newcomb's problem, which addresses two opposing intuitions of normative rationality. These intuitions lead to two normative rational decision theories which lead to two opposing solutions to Newcomb's problem. In such a case it is unclear which normative rational decision theory one should follow. One ought to make a decision with regard to the appropriate rational decision theory, but one doesn't know in advance which criterion is appropriate for such a decision. In this way Newcomb's problem is a foundational problem within practical philosophy. Newcomb's problem is a test for the adequacy of normative rational decision theories, and a solution to Newcomb's problem should yield criteria for the decision between opposing rational decision theories.

## *The Structure of Newcomblike Problems*

Newcomb's problem is one of a whole family of decision-theoretic problems which may be called Newcomblike problems due to their structural similarity. Before turning to Newcomb's problem itself, let us make clear the basic structure of Newcomblike problems. A simple Newcomblike problem is the patsy-gene-puzzle introduced by Joyce:

> "There is a 'patsy gene'! It causes its carriers both to have an interest in decision theory and to make lousy choices. If you have this gene, you can resign yourself to a life of misery and exploitation, since that is sure to be your lot: You will never get your money's worth. ...
>
> As it turns out you are doing the one thing that is known to provide a foolproof test. Research has shown that this very book provides a

> completely reliable way of determining whether or not a person has the sucker gene. Anyone who chooses to read as far as page 4 has the gene; anyone who stops on page 3 lacks it. Research also shows that people who turn to page 4 tend to enjoy reading the book ..., and reading it does not have any deleterious side effects. You are on page 3 now. Think carefully about what you should do next." (Joyce in press, p. 2)

Besides being one of the funniest and wittiest introductions to the problems of normative rational decision theory this purely hypothetical example is a Newcomblike problem. Newcomb's problem which will be introduced in chapter 1.1 is just one of them.

The patsy-gene-puzzle exhibits the common structure of all Newcomblike problems: On the one hand there is probabilistic dependence between the possible actions of the decision maker (e. g. not turning to page 4; turning to page 4) and the possible states of the world (e. g. lacking the patsy gene; having the patsy gene), so that the possible actions of the decision maker are reliable evidence that certain possible states of the world obtain. In the case of the patsy gene-puzzle not turning to page 4 is reliable evidence of lacking the patsy gene, and turning to page 4 is reliable evidence of having the patsy gene. On the other hand the possible states of the world are causally independent of the possible actions of the decision maker, so that under each possible state of the world one possible action (e. g. turning to page 4) strongly dominates, that is it is better than, the other possible action (e. g. not turning to page 4). Thus Newcomblike problems show that there is a conflict between probabilistic dependence and causal independence, and the question arises which decision is the rational one.

This conflict is at the same time a conflict between two schools of thought, namely evidential decision theories and causal decision theories. While evidential decision theorists take account of the probabilistic dependence, causal decision theorists take account of the causal independence. Yet there is also a third school of thought which stresses the decision maker's perspective in Newcomblike problems in contrast to evidential and causal decision theories which on first sight don't stress the decision maker's perspective. The decision maker's perspective could make a difference to what factors are viewed as causes in Newcomblike problems and therefore could lead to different solutions to Newcomblike problems. Furthermore, one could even claim that evidential decision theories and causal decision theories stress a certain perspective of the decision maker, so that causal decision theories and evidential decision theories are special cases of the third school of thought. This classification of the solutions to

Newcomb's problem in evidential decision theories, causal decision theories, and other proposals which stress the decision maker's perspective is my innovation.

## *An Overview of the Following Chapters*

The common structure of Newcomblike problems makes clear the importance of Newcomb's problem and gives the frame in which the following chapters, which yield a systematic account of the literature on Newcomb's problem, are placed. In chapter 1 the general history of Newcomb's problem, its history within philosophy, Newcomb's problem itself, variations of Newcomb's problem, and Newcomblike problems are presented, so that by means of the latter three a systematic account of Newcomb's problem is provided. Furthermore, decision-theoretic and game-theoretic foundations for analysing Newcomb's problem are given. The main part of the chapter is concerned with Nozick's earliest treatment of Newcomb's problem. Nozick (1969) analyses Newcomb's problem as a conflict between Jeffrey's (1965) evidential principle of maximising conditional utility and the principle of strong dominance. On this basis he treats Newcomb's problem as one of a whole family of Newcomblike problems. Nozick provides no unifying solution to Newcomb's problem or Newcomblike problems, but distinguishes for each problem two cases in which either the principle of strong dominance or the principle of maximising conditional utility holds.

The main aim of this chapter is to present and evaluate Nozick's approach to Newcomb's problem. In contrast to Nozick (1969) it will be shown that Newcomb's problem can be described more precisely as a conflict between the principle of maximising conditional utility and the principle of strong dominance with causal independence. Furthermore, the intuitions, which Nozick uses for backing up his two arguments in Newcomb's problem and which nobody has considered more closely, are investigated in more detail. As a result only the time intuition cannot be criticised, whereas the other intuitions can be questioned for their adequacy. With regard to Nozick's analysis of Newcomblike problems I will make clear that Sobel's (1990) analysis of Newcomblike problems is superior to Nozick's. Moreover, Nozick's (1969) reasons for treating Newcomb's problem as a Newcomblike problem will be attacked, especially Nozick's claim that the only difference between Newcomb's problem and the 2-possible-fathers-1-son-case and the prisoner's dilemma, which are two Newcomblike problems, consists in the illusion of causal influence in Newcomb's problem; the common knowledge assumption in Newcomb's problem can be made responsible for the illusion of causal influence, though. I will point out that Spohn (1978, p. 182) rightly

criticises Nozick (1969) for not providing a unifying solution to Newcomb's problem. Furthermore, in opposition to Nozick I will defend the view that Nozick's 2-possible-fathers-1-son-case and the prisoner's dilemma aren't simple, clear cut cases which have a clear and universally agreed decision-theoretic solution, so that it is problematical to use them, as Nozick does, for establishing a certain solution to Newcomb's problem as rational.

Yet some interesting results with regard to Newcomb's problem are also obtained as a by-product of the introductions into rational decision theory and into game theory: If one conceives a 1-shot Newcomb's problem as a game with two decision makers, there is a unique Nash equilibrium which is not Pareto-efficient. Therefore Newcomb's problem isn't only a paradox in rational decision theory, but also in game theory. A 1-shot Newcomb's problem seems to be a game against nature, whereas a finitely iterated Newcomb's problem with the same decision maker is a game with two decision makers. As a game against nature one partition of the possible states of the world suggests itself as the correct partition in Newcomb's problem. Thus the partitions of the possible states of the world in Newcomb's problem are considered more closely than in the existing literature on Newcomb's problem - Sobel (1986) is an exception in this respect. Furthermore, the predictor's high reliability in Newcomb's problem, which is essential for one of Nozick's two arguments, becomes irrelevant for providing a solution to the 1-shot Newcomb's problem, if one considers Newcomb's problem as a game against nature.

Chapter 1 is supplemented by an appendix which contains statements of some Newcomblike problems.

Chapter 2 deals with evidential decision theories and their solutions to Newcomb's problem. Differences among evidential decision theories obtain, because evidential decision theorists differ in what they count as evidence and differ in which kind of probability concept they favour. Furthermore, evidential decision theories propose different solutions to Newcomb's problem. This chapter is structured in the following way: The chapter begins with the first of the four evidential decision theories of Richard Jeffrey. Jeffrey's (1965) logic of decision, which provides the basis for all other evidential decision theories under consideration, advocates the principle of maximising conditional utility and is completely subjective. Furthermore, it takes the decision maker's possible actions as evidence of the possible states of the world. Becoming aware of Newcomb's problem as a counterexample to his logic of decision Jeffrey saw the need to revise his evidential decision theory which actually led to three

revisions, namely to Jeffrey's (1983) ratificationism, his probabilism (Jeffrey 1988), and his decision kinematics (Jeffrey 1996). Jeffrey's (1983) ratificationism, which recommends the principle of ratifiability and which is also completely subjective, takes the decision maker's final decisions as evidence of the possible states of the world, whereby final decision means the decision which is reached at the end of the deliberational process. Jeffrey's (1988) probabilism, which adheres to the principle of maximising final conditional utility with final credences as estimates of chances, takes the decision maker's possible actions as evidence of the possible states of the world. Jeffrey's distinction between final decisions and possible actions as evidence of the possible states of the world is relevant for a solution to Newcomb's problem. For if one takes final decisions as evidence of the possible states of the world, one conditions by final decisions in the formula for calculating the utility of a possible action, and if one takes possible actions as evidence of the possible states of the world, one conditions by the performance of the possible actions in the formula for calculating the utility of a possible action. Jeffrey's (1996) decision kinematics proposes the rigidity condition as a rationality constraint on the decision maker's credences, which is violated in Newcomb's problem, so that Newcomb's problem is no decision problem. Thus the important question arises and has to be answered whether Newcomb's problem is a well-defined decision problem at all.

Each theory is subsequently criticised and evaluated: Evidential decision theories are inadequate rational decision theories. For either they provide wrong solutions to Newcomb's problem, like Jeffrey (1965) and Jeffrey (1996), or they provide right solutions to Newcomb's problem, but their reasons for coming to this solution are inadequate, like Jeffrey (1983), Jeffrey (1988), and Eells (1981, 1982, 1985).

Jeffrey's logic of decision is regarded as inadequate even by Jeffrey (1983, 1988, 1996), because it gives the wrong recommendation in Newcomb's problem. And this wrong recommendation results mainly from two defects of Jeffrey's logic of decision: (1) Causation doesn't figure as a primitive term in his theory. (2) Every partition of the possible states of the world is permitted. Thus to get a right recommendation for Newcomb's problem at least these two defects have to be overcome.

Jeffrey's ratificationism proposes the right solution to Newcomb's problem. Yet Jeffrey's ratificationism demands too much self-knowledge from the decision maker in Newcomb's problem. For the decision maker is supposed to know what kind of person he will be when he has made his decision. Furthermore, Jeffrey's theory is limited in its

applicability. For example, there are decision problems in which possible actions are better than final decisions as evidence of the possible states of the world, so that Jeffrey's ratificationism cannot be applied to them. Moreover, Jeffrey's theory allows the possibility of two interpretations of ratificationism in cases in which no decision is ratifiable or all decisions are ratifiable. Finally, Jeffrey's ratificationism is as uneconomical as causal decision theory, for it introduces a new primitive term, namely final decisions.

Jeffrey's (1988) probabilism proposes the right solution to Newcomb's problem. Yet Jeffrey's formula for calculating the final conditional utility of a possible action cannot be right, because it uses final credences for possible actions. As we will see in chapter 3.6 Spohn's principle is valid, so that the decision maker shouldn't assign credences to his possible actions. Furthermore, Jeffrey's solution to Newcomb's problem depends on how large $n$ is in the formula for the calculation of the utility of a possible action, which opens up the possibility of arbitrariness on the side of the decision maker. Moreover, Jeffrey's probabilism demands too much self-knowledge from the decision maker in Newcomb's problem. For in Jeffrey's probabilism the decision maker has to anticipate that his initial preference differs from his final preference. Finally, Jeffrey's theory doesn't specify the decision maker's initial credence distribution, so that the following two interpretations of Jeffrey's probabilism can arise: First, there are no constraints on the decision maker's initial credence distribution, so that the problem of the priors applies to the decision maker's distribution. Second, there are constraints on the decision maker's initial credence distribution, so that the constraints have to be specified and problems with the constraints have to be dealt with.

Jeffrey's decision kinematics proposes the wrong solution to Newcomb's problem. For the predictor just has to be a little bit better than chance for Newcomb's problem to arise. Thus Newcomb's problem is a well-defined decision problem. Furthermore, if we don't allow credences for possible actions, then the rigidity condition cannot be violated in Newcomb's problem. Thus Jeffrey's conclusion that Newcomb's problem is no decision problem because of the violated rigidity condition doesn't hold.

Finally, towards the end of chapter 2 Eells' (1981, 1982, 1985) proposal of a common cause-solution to Newcomb's problem will be considered. Eells' (1981, 1982, 1985) evidential decision theory, which is closer to Jeffrey's (1965) logic of decision than to Jeffrey's later theories, advocates the principle of maximising conditional utility and analyses Newcomb's problem as a common cause-situation.

Again, I will try to make clear the advantages and disadvantages of Eells' solution to Newcomb's problem: Although Eells (1981, 1982, 1985) proposes the right solution to Newcomb's problem, his Eells' theory is limited in its applicability. For it only applies to ideal rational decision makers. Furthermore, Eells' proposal is dangerously self-referential. For the decision maker has to know his relevant wants and beliefs and has to know whether they will lead him to his possible actions. Moreover, Eells' approach can be compared with Jeffrey's (1983) ratificationism and with Price's (1986) proposal of decision maker probabilities. For both Eells' theory and Jeffrey's theory rely on the gap between making a decision and performing the corresponding possible action. Although Eells (1981, 1982, 1985) proposes a common cause for Newcomb's problem, he doesn't specify any common cause for Newcomb's problem. Finally, Eells assumes that the decision maker can assign credences to his possible actions, which the decision maker shouldn't do.

Chapter 3 takes a closer look at causal decision theories and their solutions to Newcomb's problem. These rational decision theories differ as to their specific theories of causality, their subjective or objective interpretation of probability, and as to their reasons for coming to the same solution in Newcomb's problem. Gibbard and Harper (1978) defend a subjective causal decision theory which is based on a counterfactual theory of causation using would-subjunctive conditionals. Skyrms' (1980, 1982, 1984) causal decision theory, which is subjective, is founded on a probabilistic theory of causation. Sobel (1986) advocates a causal decision theory which is an advancement of Jeffrey's logic of decision - in the sense that it takes Jeffrey's logic of decision as a starting-point and that it modifies Jeffrey's logic of decision - and which functions with a probabilistic counterfactual theory of causation using might-subjunctive conditionals. Furthermore, Sobel's theory is objective. Lewis' (1981a) causal decision theory, which is an advancement of Gibbard and Harper's causal decision theory, has as its basis a probabilistic counterfactual theory of causation which uses would-subjunctive conditionals. Moreover, Lewis' theory is completely subjective. Spohn (1978) defends in his causal decision theory a probabilistic theory of causation. Yet his solution to Newcomb's problem is mediated by the principle of dominance and Spohn's (1977, 1978) principle which is a rationality constraint on the decision maker's credences and which says: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts.*" (Spohn 1977, p. 114).

In this chapter, after each presentation there is criticism and evaluation of the respective theories: Causal decision theories do provide the right solution to Newcomb's problem. Yet their reasons for coming to this solution are not always adequate.

Whereas Gibbard and Harper (1978) don't provide a logic of would-subjunctive conditionals, but rely on the decision maker's intuitions instead, so that their theory is limited to decision makers who have clear intuitions with regard to would-subjunctive conditionals, Lewis (1973) provides a logic of would-subjunctive conditionals. Sobel (1986) doesn't provide a logic of might-subjunctive conditionals, so that his theory is limited to decision makers who have clear intuitions with regard to might-subjunctive conditionals. Therefore Lewis' causal decision theory is to be preferred to Gibbard and Harper's and Sobel's causal decision theories with regard to this point. While Gibbard and Harper's (1978) causal decision theory is formulated in terms of ultimate possible outcomes, Skyrms (1985) proposes a causal decision theory which works for proximate possible outcomes. Because proximate possible outcomes demand less knowledge from the decision maker than ultimate possible outcomes, Skyrms' causal decision theory is to be preferred to Gibbard and Harper's causal decision theory in this respect. Gibbard and Harper's (1978) causal decision theory presupposes the validity of conditional excluded middle which is open to two objections which can't be completely overcome. In opposition to this Sobel (1986) and Lewis (1981a) don't presuppose the validity of conditional excluded middle. Whereas Gibbard and Harper's (1978) causal decision theory is limited to decision makers who believe in deterministic worlds, Skyrms', Sobel's, Lewis', and Spohn's causal decision theories are even meant to work for decision makers who believe in indeterministic worlds. Gibbard and Harper's (1978) causal decision theory is limited to decision makers with non-backtracking intuitions, although Newcomb's problem arouses backtracking intuitions. In opposition to Gibbard and Harper (1978) Skyrms (1984) can account for backtracking intuitions of the decision maker.

Skyrms' (1980, 1982, 1984) causal decision theory is the rational decision theory which provides so far the most adequate solution to Newcomb's problem. Skyrms' (1984) causal decision theory is completely subjective, so that it isn't limited to decision makers who believe in chances. In opposition to this Sobel's (1986) causal decision theory relies on practical chance conditionals with an objectivist understanding of chance, so that it is limited to decision makers who believe in chances. Gibbard and Harper's (1978) causal decision theory is completely subjective, and Lewis' (1981a) causal decision theory is completely subjective, so that their causal decision theories

aren't limited to decision makers who believe in chances. In Skyrms' (1984) causal decision theory unconditional credences are used for calculating the utility of a possible action. Thus Spohn's (1977, 1978) principle is not violated. Because Gibbard and Harper (1978), Sobel (1986), and Lewis (1981a) use subjunctive conditionals in their causal decision theories, they don't violate Spohn's principle either. In Skyrms' (1984) causal decision theory the possible states of the world are partitioned rationally, so that Skyrms' causal decision theory works for in this respect fully rational decision makers. Skyrms' (1984) causal decision theory isn't formulated in terms of subjunctive conditionals - although his theory is compatible to the latter because of his (Skyrms 1984) Bayesian theory of conditionals -, so that his theory isn't limited in its applicability in this respect. In opposition to this Gibbard and Harper's (1978), Sobel's (1986), and Lewis' (1981a) causal decision theories are formulated in terms of subjunctive conditionals, so that decision makers who have weak intuitions with regard to subjunctive conditionals are not accounted for.

Sobel's (1986) causal decision theory is much too complicated from an economical point of view of theory building, so that it shouldn't be recommended as a rational decision theory. The following features make Sobel's theory so complicated: (i) Practical conditionals which provide the basis for Sobel's practical chance conditionals; (ii) conditional chances in the calculation of the utility of a possible action; (iii) the distinction between causally possible, open, and possibly open; (iv) the distinction between natural partitions, sufficiently fine partitions, and sufficiently exclusive partitions. Sobel's (1986) practical conditionals which provide the basis for his practical chance conditionals are vague. Sobel uses practical chance conditionals, but doesn't provide a logic for these conditionals; he just states that the vagueness of these conditionals should be resolved appropriately to contexts of decision which allows for a lot of arbitrariness. Sobel's (1986) distinction between possible actions which are causally possible, open, and possibly open can be applied to Newcomb's problem. Yet if the decision maker believes that his possible actions are not even possibly open for him, he cannot calculate the utility of a possible action in Sobel's causal decision theory.

Rabinowicz (1982) rightly claims that Lewis' (1981a) suggestion to apply rational decision theory not only to fully rational decision makers, but also to partly rational decision makers must have some limits. For we cannot allow the decision maker to be as irrational as he wants to be. Horgan (1981) correctly claims that the comparative overall similarity among possible worlds in Lewis' (1981a) causal decision theory is inherently vague. Yet Horgan's backtracking resolution of vagueness isn't

appropriate either. Furthermore, Lewis doesn't deny Horgan's claim, but provides his standard resolution of vagueness.

Spohn's (1978) solution to Newcomb's problem can be criticised, because Spohn applies the principle of dominance to Newcomb's problem, although the principle of dominance is restricted in its range, so that it cannot be applied to all decision problems. Yet Spohn's (1977, 1978) principle is valid for acts and for probabilistic acts.

A rational decision theory has to demand from the possible states of the world to be well-specified, which leads to an exclusion of the evidential partition of the possible states of the world in Newcomb's problem. For an evidential partition of the possible states of the world, like $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$., refers to the earlier possible state of the world, when the prediction was made, and it refers to the later possible state of the world, when the predicted possible action takes place. While the former possible state of the world already obtains, when the decision maker decides, it is logically possible that the latter possible state of the world doesn't turn out as predicted, so that the possible state of the world isn't well-specified. To be more precise the possible state of the world is over-specified. For it already determines a future possible state of the world as true.

Chapter 4 deals with proposals which stress the decision maker's perspective. These proposals differ in their solutions to Newcomb's problem. Meek and Glymour's (1994) proposal to distinguish between conditioning and intervening provides an explanation why evidential and causal decision theorists differ in their recommendations in Newcomblike problems; evidential decision theorists in contrast to causal decision theorists don't view the will as a causal factor. Nozick's (1993) proposal of the combination of various decision principles leaves it to the decision maker to determine what rationality is, that is which decision principles to use and how much to weigh them. Kyburg's (1980, 1988) proposal to maximise properly epistemic utility makes clear which role freedom of will plays in decision-making. Furthermore, Kyburg emphasises that first-person and third-person perspective deliberating about present decisions, and first-person perspective deliberating about past decisions, make a difference for recommendations in decision-making or for evaluations of decisions. My own proposal to view Newcomb's problem as a game against nature provides a foundation for Skyrms' (1980, 1982, 1984) causal decision theory.

Each proposal will be criticised and evaluated: Other proposals which stress the decision maker's perspective may provide right solutions to Newcomb's problem (Meek

and Glymour 1994; Nozick 1993; Kyburg 1980, 1988; my own proposal). Yet they may also provide wrong solutions to Newcomb's problem (Meek and Glymour 1994; Nozick 1993).

Although Meek and Glymour's (1994) proposal provides a good explanation why evidential and causal decision theorists differ in their recommendations in Newcomblike problems, their theory doesn't provide a criterion for deciding which position is the rational one, so that Meek and Glymour's theory is inadequate as a rational decision theory. Nozick's (1993) proposal is inadequate as a normative rational decision theory. It may be adequate as a descriptive rational decision theory. For in Nozick's theory it is only up to the decision maker to determine what rationality is. Kyburg's (1980, 1988) proposal is insofar inadequate as a rational decision theory as it holds only, if one assumes absolute freedom of will formation. And absolute freedom of will formation is given, if the decision maker can decide independently from all previous factors, what his will is going to be. This dependency is a central weakness of Kyburg's theory.

Meek and Glymour's (1994) and Nozick's (1993) proposals in opposition to Kyburg's (1980, 1988) proposal allow for a lot of arbitrariness on the decision maker's side. For within Meek and Glymour's proposal the decision maker can view his possible actions as interventions in one case and his possible actions as non-interventions in another case; within Nozick's proposal arbitrariness on the decision maker's side is possible, because the decision maker doesn't have to be fully rational; moreover, he can introduce new decision principles, can weigh them differently, and/or can assign different symbolic meanings to his possible actions. Furthermore, Meek and Glymour's (1994) theory in contrast to Nozick's (1993) and Kyburg's (1980, 1988) theories is limited to decision makers who have so much self-knowledge as to view their possible actions either as interventions or as non-interventions.

A further insight of this chapter is that in opposition to Nozick's (1993) approach in Meek and Glymour's (1994) and Kyburg's (1980, 1988) theories different perspectives, namely first-person perspective and third-person perspective both deliberating about present decisions, and first-person perspective deliberating about past decisions, make a difference for recommendations in decision-making or for evaluations of decisions. Moreover, in contrast to Kyburg's (1980, 1988) proposal Meek and Glymour's (1994) and Nozick's (1993) proposals allow for the violation of Spohn's (1977, 1978) principle which is a disadvantage of their theories. In opposition to Gibbard and Harper's (1978), Sobel's (1986), and Lewis' (1981a) causal decision theories Meek

and Glymour's (1994) and Kyburg's (1980, 1988) proposals work without subjunctive conditionals, whereas Nozick's (1993) proposal leaves it unspecified which causal decision theory the decision maker should favour - if he favours causal decision theories at all -, so that Meek and Glymour on the one side and Kyburg on the other side don't have to provide a logic of subjunctive conditionals, whereas in Nozick's case it is unclear whether he has to.

In contrast to Kyburg's (1980, 1988) proposal within Nozick's (1993) proposal the decision maker can be incoherent, if he uses decision principles which exclude each other, and inconsistent over time, if he uses different decision principles at different times and weighs them differently at different times. In Meek and Glymour's (1994) proposal the decision maker cannot be incoherent, but he can be inconsistent over time. For with regard to the former he doesn't use decision principles which exclude each other, and with regard to the latter the decision maker can view his possible actions as interventions at one time and as non-interventions at another time.

In comparison with all other rational decision theories Nozick's (1993) proposal is the most unprecise of all which is a central weakness of his theory. For while all other rational decision theories tell the decision maker how to calculate the utility of a possible action, Nozick's theory doesn't tell the decision maker how to calculate the symbolic utility of a possible action, if the decision maker wants to use the symbolic decision principle. Furthermore, Nozick doesn't specify which causal decision principle the decision maker should use, if the decision maker wants to use the causal decision principle.

Kyburg's (1980, 1988) theory is economical, because in opposition to all other rational decision theories it doesn't need possible states of the world or types of possible states of the world. As a result and in opposition to all causal decision theories Kyburg's theory doesn't have to determine the correct partition of the possible states of the world. Yet in opposition to all other rational decision theories Kyburg's theory has to answer what an appropriate reference class is. Kyburg's theory is economical, because in opposition to all causal decision theories his theory doesn't have causation as a primitive term. Kyburg's theory is even more economical than Meek and Glymour's (1994) and Nozick's (1993) proposals in this respect. For in Meek and Glymour's (1994) proposal causation figures as a primitive term, if the decision maker views his possible actions as interventions. In Nozick's (1993) theory causation figures as a primitive term, if the decision maker favours the causal decision principle.

Skyrms' (1980, 1982, 1984) causal decision theory, which is supported by my own proposal, is to be preferred to the other proposals of Meek and Glymour (1994), Nozick (1993), and Kyburg (1980, 1988). For Skyrms provides the right solution to Newcomb's problem, whereas the other proposals may provide the wrong solution to Newcomb's problem (Meek and Glymour 1994; Nozick 1993). Furthermore, it is a central weakness of Kyburg's (1980, 1988) proposal that it only holds, if one assumes absolute freedom of will formation, whereas Skyrms' causal decision theory isn't limited in this respect.

# Chapter 1

# The Starting-Point

## 1.1 Newcomb's Problem

### *General History*

Newcomb's problem was invented by the physicist Dr. William Newcomb from the Livermore Radiation Laboratories in California in 1960 (Nozick 1969, p. 143), when he was thinking about the prisoner's dilemma[1] (Gardner 1973, p. 104). Robert Nozick (1969, p. 143) heard of Newcomb's problem by way of Professor Martin D. Kruskal of the Princeton University Department of Astrophysical Sciences in 1963, and made it public in 1969 by writing his famous article "Newcomb's Problem and Two Principles of Choice". Martin Gardner popularised Newcomb's problem as a mathematical game in the *Scientific American* in 1973. Since its first publication Newcomb's problem led to a "Newcombmania", as Levi (1982) termed it, not only in philosophy, but also in economy (for example Frydman, O'Driscoll, and Schotter 1982; Broome 1989), politics (for example Brams 1975; Rapoport 1975), psychology (for example Rapoport 1975; Shafir and Tversky 1992), and theology (for example Factor 1978; Horne 1983; Craig 1987). The "Newcombmania" spread to Germany in 1978 by Wolfgang Spohn's dissertation, and after a slow start seems to flourish now (for example Pothast 1980; Lenzen 1997).

### *History within Philosophy*

Newcomb's problem led to the development of causal decision theories (for example Gibbard and Harper 1978; Spohn 1978; Lewis 1981a; Skyrms 1982; Sobel 1986) and other decision-theoretic proposals (for example Kyburg 1980; Nozick 1993), because evidential decision theories (for example Jeffrey 1965) didn't seem to provide

[1] In the literature one finds "prisoner's dilemma" (for example Nozick 1993; Pettit 1988; Shafir and Tversky 1992; Sobel 1985a) and "prisoners' dilemma" (for example Brams 1975; Hurley 1991; Lewis 1979a; Nozick 1969; Sobel 1991; Snow 1985). Unfortunately the authors don't explain their different usage. Their usage can be explained in the following way, though: With regard to "prisoner's dilemma" authors want to emphasise that one prisoner faces a dilemma; with regard to "prisoners' dilemma" authors want to emphasise that two prisoners each face a dilemma. Because "prisoner's dilemma" respectively "prisoners' dilemma" is just a name and nothing really depends on the different usage, I will for convenience sake and as long as I don't refer to a particular author use "prisoner's dilemma".

the correct solution to Newcomb's problem. This challenge was answered by more elaborate evidential decision theories (for example Eells 1981; Jeffrey 1983, 1988), so that after some time the situation was "deadlocked" (Lewis 1981a, p. 5) and a "stalemate" (Horgan 1985) was reached. Some philosophers tried to overcome this by mediating between evidential and causal decision theories (for example Meek and Glymour 1994). Another reaction to Newcomb's problem within the area of rational decision theory was to claim that Newcomb's problem is not a decision problem at all (for example Gardner 1973; Schlesinger 1974; Cargile 1975; Mackie 1977; Jeffrey 1996). By another line of thought Schlesinger (1974) tried to show that Newcomb's problem provides a proof for human freedom of will, which led to many objections (for example Cargile 1975; Benditt and Ross 1976; Locke 1978; Gallois 1979; Hudson 1979; Pothast 1980; Ben-Menahem 1986; Weintraub 1995; Ledwig 1997). To a large part independently from this Newcomb's problem raised anew the question of compatibility between determinism and freedom of will (for example Mackie 1977; Locke 1979; Gallois 1981; Leslie 1991; Fischer 1994). Last but not least the connections between Newcomb's problem and other problems of rationality were explored like Gideon's paradox (for example Bar-Hillel and Margalit 1985), Fisher's problem and Simpson's paradox (for example Wagner 1991), the toxin puzzle (for example Kavka 1983; Bratman 1987; Gauthier 1994), the prediction paradox and Moore's problem (for example Sorensen 1986; Olin 1988; Goldstein 1993), the time inconsistency problem (for example Frydman, O'Driscoll, and Schotter 1982; Broome 1989), the prisoner's dilemma (for example Nozick 1969, 1993; Brams 1975, 1983; Lewis 1979a; Snow 1985; Sorensen 1985; Sobel 1985a, 1991; Pettit 1988; Hurley 1991; Shafir and Tversky 1992), Parfit's (1984) problem of the hitch-hiker (for example Barnes 1997), etc.[2]

## *Newcomb's Problem*

Newcomb's problem is described by Robert Nozick (1969, pp. 114-115) in the following way:

> "Suppose a being in whose power to predict your choices you have enormous confidence. (One might tell a science-fiction story about a being from another planet, with an advanced technology and science, who you know to be friendly, etc.) You know that this being has often correctly predicted your choices in the past (and has never, so far as you know,

[2]For a statement of these problems, paradoxes, and puzzles see the appendix of chapter 1.

made an incorrect prediction about your choices), and furthermore you know that this being has often correctly predicted the choices of other people, many of whom are similar to you, in the particular situation to be described below. One might tell a longer story, but all this leads you to believe that almost certainly this being's prediction about your choice in the situation to be discussed will be correct. There are two boxes, (B1) and (B2). (B1) contains $ 1000. (B2) contains either $ 1000000 ($ M), or nothing. What the content of (B2) depends upon will be described in a moment. ... You have a choice between two actions: (1) taking what is in both boxes (2) taking only what is in the second box. Furthermore, and you know this, the being knows that you know this, and so on: (I) If the being predicts you will take what is in both boxes, he does not put the $ M in the second box. (II) If the being predicts you will take only what is in the second box, he does put the $ M in the second box. ... The situation is as follows. First the being makes its prediction. Then it puts the $ M in the second box, or does not, depending upon what it has predicted. Then you make your choice. What do you do?"[3]

## 1.2 Decision-Theoretic Foundations for Analysing Newcomb's Problem

### *The Primitive Terms*

In the decision situation of Newcomb's problem the decision maker[4] has to decide[5] between two possible actions[6]. In the decision situation of Newcomb's problem there are also two possible states of the world[7]. The possible actions and the possible states of the world get a time index *t* starting with number 1, so that lower numbers

[3]"If the being predicts that you will consciously randomise your choice, e. g., flip a coin, or decide to do one of the actions if the next object you happen to see is blue, and otherwise do the other action, then he does not put the $ M in the second box." (Nozick 1969, p. 143) According to Gardner (1973, p. 108) Newcomb and Kruskal recommend taking only the second box.

[4]In the literature of rational decision theory the terms "decision maker", "decider", and "agent" are used for the person who decides without making any distinction between these three terms.

[5]Decision theorists use the terms "decide" and "choose" interchangeably. According to Bittner (1992) one may distinguish between "decide to" and "decide that".

[6]In rational decision theory the terms "possible actions", "actions", "alternatives", "acts", and "options" are used for designating the possibilities the decision maker can decide for.

[7]Instead of using the term "possible states of the world" decision theorists also use the terms "states", "circumstances", and "conditions".

indicate earlier times and equal numbers indicate the same time. The possible actions and the possible states of the world in Newcomb's problem result in four possible outcomes[8]. Hence like every decision situation Newcomb's problem can be described by the set of possible actions, by the set of possible states of the world, and by the set of possible outcomes. What a possible action, a possible state of the world, and a possible outcome is, and how they are related to each other are very important questions, which decision theorists have answered differently.[9] For present purposes it is sufficient to appeal to the intuitive understanding of these terms and of their relationships.

Yet two points should be made clear in the beginning. First, in a decision problem of rational decision theory[10] exactly one possible action will be decided for and exactly one possible state of the world will obtain. The set of possible actions respectively the set of possible states of the world should thus be partitions which consist of incompatible possible actions respectively which consist of incompatible possible states of the world. Second, it should be clear which possible outcomes a given possible action/possible state of the world pair will produce.

In what follows the decision situation will be denoted by $D$, the decision maker by $X$, the set of possible actions by $A$, and the particular possible actions in this set by $a_1, a_2, ..., a_n$. The set of possible states of the world will be denoted by $S$, and the particular possible states of the world by $s_1, s_2, ..., s_m$. Finally, $O$ stands for the set of possible outcomes, and $o_{11}, o_{12}, ..., o_{nm}$ for the particular possible outcomes. In this way we obtain a matrix formulation of Newcomb's problem à la Savage (1954/1972).[11]

---

[8]In the literature of rational decision theory the terms "possible outcomes", "outcomes", "possible consequences", and "consequences" are used interchangeably.

[9]Joyce (in press, chapter 2) gives a very valuable discussion of Savage's (1954/1972) and Jeffrey's (1965) explications of the terms "possible action", "possible state of the world", and "possible outcome".

[10]Decision theorists also speak of "decision theory", "Bayesian decision theory", and "rational choice theory" instead of using the term "rational decision theory".

[11]The reason that I decided for a matrix formulation à la Savage (1954/1972) is that it provides a conceptual division of labour by distinguishing between possible actions, possible states of the world, and possible outcomes (cf. Joyce in press, p. 49).

## *The Matrix Formulation of Newcomb's Problem*

Taking Nozick's original formulation of the problem the decision situation of Newcomb's problem can be represented by the following decision matrix:

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | $o_{11}$: \$ 1,000 | $o_{12}$: \$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | $o_{21}$: \$ 0 | $o_{22}$: \$ 1,000,000 |

Figure 1. Decision matrix for Newcomb's problem of the possible outcomes $o_{11}, o_{12}, o_{21}, o_{22}$ which result from combining the possible actions $a_1, a_2$ with the possible states of the world $s_1, s_2$.

Nevertheless this doesn't seem to be the only possible matrix formulation of Newcomb's problem. For whereas the possible actions are usually given in decision situations - in Newcomb's problem, for example, the possible actions are already given - , the possible states of the world aren't equally clearly settled. This is also the case in Newcomb's problem. Some decision theorists (cf. Sobel 1988c) claim that the decision situation of Newcomb's problem can also be represented by the following decision matrix, where the possible states of the world are $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$. As a result of the different possible states of the world, the possible outcomes change, too.[12]

[12]Yet in chapter 1.3, in the section on the classification of Newcomb's problem in game theory, we will see that this partition isn't the correct partition for a 1-shot Newcomb's problem as a game against nature.

| | $s_1$: The predictor has predicted correctly at $t_1$. | $s_2$: The predictor hasn't predicted correctly at $t_1$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | $o_{11}$: \$ 1,000 | $o_{12}$: \$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | $o_{21}$: \$ 1,000,000 | $o_{22}$: \$ 0 |

Figure 2. Decision matrix for Newcomb's problem of the possible outcomes $o_{11}, o_{12}, o_{21}, o_{22}$ which result from combining the possible actions $a_1, a_2$ with the possible states of the world $s_1, s_2$.

Therefore it seems to be that for decision problems like Newcomb's problem rational decision theory has two problems to solve: First, which possible actions and which possible states of the world should figure in decision situations. Second, after having specified the possible states of the world which possible action should the decision maker decide for.[13]

## *(Subjective) Utilities/Objective Utilities and Credences/Chances*

From the viewpoint of rational decision theory two main factors determine our decisions[14]: (1) Our wants[15]; (2) our beliefs[16]. Whereas our wants are represented by

[13]As Joyce (in press, p. 72) has observed, and I agree with him, decision theorists usually just consider the second problem. With regard to the first problem Joyce (in press, p. 71) rightly claims that the deliberation process in decision-making is often a refinement process, so that the decision maker starts with a set of possible actions, a set of possible states of the world, and a set of possible outcomes, then goes on to consider other sets which provide a fuller and more realistic picture of the world, and finally, when decision-making is at hand, has to settle for one set of possible actions, one set of possible states of the world, and one set of possible outcomes. With regard to Newcomb's problem it seems to be that the deliberation process isn't a refinement process, but rather a conflict between two or even more on first sight equally plausible sets of possible states of the world (cf. MacCrimmon and Larsson 1979, pp. 391-392). This situation can also be viewed as a challenge to the question: Is there a decision principle which specifies to which set of possible states of the world it applies (cf. Sobel 1988c)? Or is there a decision principle which gives right answers for the decided for set of possible states of the world (cf. Joyce in press)?

[14]Kaufmann (1996) suggests that a third component is added to the belief-want model of rational decision theory, namely prospective intentions, because the standard model of rational decision-making may conflict in dynamic decision-making situations with the general rationality condition of dynamic consistency, whereas the extended model can resolve the conflict. Although I think that Kaufmann is right in his claim, I think that the belief-want model is sufficient to deal with a 1-shot Newcomb's problem. For a 1-shot Newcomb's problem in opposition to an iterated Newcomb's problem isn't a dynamical decision-making situation. Because I just want to find a solution for a 1-shot Newcomb's problem, and because Kaufmann (1996, p. 225) also admits that

the (subjective) utilities[17] or objective utilities[18] of the possible outcomes of our decisions, our beliefs are represented by the credences[19] or chances[20] of the possible outcomes of our decisions. The symbol for utility will be $u$, for objective utility will be $ou$, for credence will be $c$, and for chance will be $ch$. The utilities of the possible outcomes of our decision will be denoted by $u(o_{ij})$. The utilities of the possible outcomes of our decisions will be ranked in the following way: $u_1$ stands for the lowest utility, $u_2$ for the next higher utility, and $u_{n \cdot m}$ is the highest utility[21]. The same conventions hold for objective utilities. Credences of possible states of the world will be denoted by $c(s_j)$, chances of possible states of the world by $ch(s_j)$.

According to Jeffrey (1965, 1983) a rationality constraint for the utilities of the decision maker is the "desirability axiom", which says for any propositions $X$ and $Y$: If $c(X \cap Y) = 0$ and $c(X \cup Y) \neq 0$, then

$$u(X \cup Y) = \frac{c(X)u(X) + c(Y)u(Y)}{c(X) + c(Y)}.$$

---

some philosophers think that intentions can be reduced to a combination of wants and beliefs à la Audi (1986), I will not deal with the question whether an extended model of rational decision theory is needed.

[15]Instead of "wants" decision theorists also use the terms "wishes" and "desires". Yet Bittner (1992) claims that one has to distinguish between wishes and wants: Whereas wishes don't lead to actions, wants sometimes do lead to actions.

[16]Instead of "beliefs" one also finds the term "information".

[17]One also finds the terms "subjective value" and "desirability" instead of the term "utility" in the sense of "subjective utility". Spohn (1978, p. 36) uses the term "subjective value" rather than the term "utility", because "utility" implies egoism and materialism which aren't implied by the term "subjective value". According to Jeffrey (1983, p. 21) the term "utility" has misleading philosophical associations stemming from Bentham's hedonism. Thus he uses the term "desirability" which is not misleading insofar it refers to what the agent in fact desires. Kusser (1989, p. 14) criticises all three terms, because they are not directly connected to the wants of a person, and because they can easily lead to misinterpretations: "Utility" can refer to usefulness, "subjective value" can refer to what a subject should want, and "desirability" can refer to a mere possibility to want something. Since all the terms are problematical, I just picked one of them.

[18]Mellor (1995, pp. 82-83) is, as far as I know, the only one to defend objective utilities in rational decision theory. He uses the term "objective value" instead of the term "objective utility". How (subjective) utilities are related to objective utilities is a question which decision theorists haven't dealt with, as far as I know. For more details cf. footnote 36.

[19]Instead of the term "credence" decision theorists also use the terms "subjective probability", "personal probability", "judgemental probability", and "degree of belief".

[20]Decision theorists also use the terms "single case chances", "objective probabilities", and "propensities" instead of the term "chances". According to Lewis (1980) there are two different kinds of interpreting probability, namely credence or degree of belief, and chance or propensity, whereas according to Skyrms (1992) there are three different kinds of interpreting probability, namely degree of belief, relative frequency, and chance.

[21]The number $n \cdot m$ derives from the fact that there are $n$ different possible actions and $m$ different possible states of the world, so that $n \cdot m$ possible outcomes result.

In other words: The utility of a disjunction of incompatible propositions is a weighted average of the utilities of the incompatible ways in which it can come true, where the weights are the credences of those ways.

The credences of the decision maker should follow the axioms of the mathematical probability calculus, that is the axioms of Kolmogorov (1933)[22], for otherwise decision makers have incoherent credences[23] and therefore are in a position to face a betting situation which has become known as "dutch book"[24]. The equivalent should hold for chances.

The concepts of credence or chance led to a classification of decisions into three different categories (cf. Luce and Raiffa 1957, p. 13): (1) If the decision maker can ascribe a credence or a chance of 0 or 1 to one of the possible outcomes, then the decision maker faces a decision under certainty. (2) If the decision maker can ascribe a credence or a chance of $0<x<1$ to his possible outcomes, then the decision maker faces a decision under risk. (3) If the decision maker doesn't know which credences or chances to ascribe to his possible outcomes, so that he could equally ascribe any credence or chance to his possible outcomes, then the decision maker faces a decision under uncertainty.[25]

---

[22]The axioms of Kolmogorov (1933) are:
(1) The axiom of non-negativity: $P(A) \geq 0$, that is an event A has a probability of greater than or equal to 0.
(2) The axiom of normalisation: $P(A \cup \neg A) = 1$, that is the conjunction of an event A with an event $\neg A$ has a probability of 1.
(3) The axiom of finite additivity: If two events A and B are disjoint, that is $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.
The axioms of Kolmogorov (1933) can also be found in, for example, Joyce (in press, p. 11) and Skyrms (1980, p. 155, 1984, p. 22).

[23]It is a minimal requirement of rationality to be coherent.

[24]According to Kennedy (1995, p. 213) a dutch book is a bet or a combination of bets whereby the bettor will suffer a net loss regardless of the outcome of the bet. Moreover, Ramsey (1931) and de Finetti (1937) argued that consistent or coherent credences should satisfy the mathematical probability calculus, so Skyrms (1992, pp. 374-375); and a bettor is coherent, if no dutch book can be made against him.

[25]According to Hargreaves Heap (1992) the distinction between decision under risk and decision under uncertainty can be attributed to Keynes (1921, 1936) and Knight (1921). Hargreaves Heap points out that the probabilities which the decision maker ascribes to his possible outcomes can be subjective or objective, whereas on the one hand Joyce (in press, p. 16) claims that the probabilities are objective and on the other hand Meggle (1985, p. 417) claims that they are subjective. In my opinion Hargreaves Heap is right, for I don't see any reason why one should restrict the decision maker to one sort of probability. Neither Hargreaves Heap nor Joyce and Meggle provide reasons for their point of view, though.

## *Maximising Principles*[26]

According to Savage (1954/1972) the utility[27] of a possible action of the decision maker can be calculated as follows (the utility of a possible action $a_i$ will be denoted by $U(a_i)$):

$$U(a_i) = \sum_{j=1}^{m} c(s_j)u(o_{ij}).$$

That is the utility of a possible action $a_i$ is the sum of the weighted utilities of the possible outcomes $u(o_{ij})$ where the weights are the credences of the possible states of the world $c(s_j)$.

> Savage (1954/1972) claims that the decision maker should use the **principle of maximising utility**: In a given decision situation $D$ the decision maker $X$ should decide for a possible action $a_i$ with maximal utility.[28]

Savage's (1954/1972) principle of maximising utility was modified by Jeffrey (1965), because he thinks that the possible actions of the decision maker can influence his credences of the possible states of the world (cf. Gärdenfors and Sahlin 1988, p. 9). To use Jeffrey's example (1983, p. 4) a dinner guest believes that his host starts cooking after he arrives, that his host can cook chicken or beef, that he has to provide the wine, that he has a red wine and a white wine, but is just able to bring one of them, and that his host will cook whatever suits the wine. Therefore it seems to be that the dinner guest can influence his credences of the possible states of the world by what possible action he decides for. These considerations led Jeffrey to introduce the notion of the conditional utility of a possible action which is calculated in the following way ("|" means "conditional on" or "given".): If $c(a_i) > 0$, then

---

[26]I start with Savage's (1954/1972) principle of maximising utility, because his rational decision theory besides Ramsey's (1931) provides the basis for Jeffrey's (1965) rational decision theory which decision theorists took as a starting-point to discuss Newcomb's problem. I will not present Ramsey's (1931) rational decision theory. For its influence with regard to solutions to Newcomb's problem is only marginal.

[27]In the literature of rational decision theory one can also find the terms "subjective value", "subjective expected utility", "expected utility", "estimated desirability", and "desirability" instead of the term "utility".

[28]According to Slote (1995) to maximise utility can be criticised. For to satisfice utility may be acceptable, too, where to "satisfice" means "to choose or do the good enough rather than the most or the best" (Slote 1995, p. 712). Although the dispute between maximisers and satisficers is very interesting, I will not enter it, because the solutions to Newcomb's problem all rely on maximising principles, and because a solution to Newcomb's problem by means of satisficing is obvious, namely taking B2. Furthermore, the decision maker in Newcomb's problem wants to get as much money as he can, which speaks against satisficing as a utility principle in Newcomb's problem.

$$CU(a_i) = \sum_{j=1}^{m} c(s_j \mid a_i) u(o_{ij}).$$

Jeffrey (1965) demands that the decision maker should adopt the **principle of maximising conditional utility**: In a given decision situation *D* the decision maker *X* should decide for a possible action $a_i$ with maximal conditional utility.

## *The Principles of Dominance*

Another relevant decision principle[29] in rational decision theory is the principle of dominance. To be correct there are actually two principles: There is the principle of strong dominance and the principle of weak dominance. Being aware of this difference decision theorists usually don't distinguish between them. First, I will present the definition of strong dominance and the principle of strong dominance.

**Definition of strong dominance**: The possible action $a_i$ dominates all of the other possible actions in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ strongly if and only if the possible action $a_i$ leads to higher utilities of the possible outcomes $o_{11}$, $o_{12}$, ..., $o_{nm}$ in comparison with all of the other possible actions in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ for every possible state of the world $s_1, s_2, ..., s_m$. That is $u(o_{ij}) > u(o_{kj})$ for all $k = 1, ..., i\text{-}1, i\text{+}1, ..., n$ and all $j = 1, ..., m$. ("\" means "without".)

**Principle of strong dominance**: In a given decision situation *D* the decision maker *X* should decide for the strongly dominant possible action $a_i$, if there is one.

The definition of weak dominance and the principle of weak dominance are the following:

**Definition of weak dominance**: The possible action $a_i$ dominates all of the other possible actions in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ weakly if and only if the possible action $a_i$ leads at least in one case to a higher utility of the possible outcomes $o_{11}, o_{12}, ..., o_{nm}$ in comparison with all of the other possible actions in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ for the considered possible states of the world $s_1, s_2, ..., s_m$, and leads in the rest of the cases to equal utilities of the possible outcomes $o_{11}, o_{12}, ..., o_{nm}$ in comparison with all

[29]There are many more relevant decision principles in rational decision theory, for example, the maximin rule, the minimax regret rule, the optimism-pessimism rule, the principle of insufficient reason (cf. Resnik 1987). As far as I know, nobody has used these rules for solving Newcomb's problem, though. Therefore I will not deal with them in my dissertation.

of the other possible actions in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ for the considered possible states of the world $s_1, s_2, ..., s_m$. That is $u(o_{ij}) > u(o_{kj})$ for at least one $k = 1, ..., i\text{-}1, i\text{+}1, ..., n$ and all $j = 1, ..., m$, and $u(o_{ij}) = u(o_{kj})$ for the rest of the $k$'s $= 1, ..., i\text{-}1, i\text{+}1, ..., n$ and all $j = 1, ..., m$.

**Principle of weak dominance**[30]: In a given decision situation $D$ the decision maker $X$ should decide for the weakly dominant possible action $a_i$, if there is one.

The principles of dominance are restricted in their range, for they can only be applied, when the decision maker believes that the possible actions of the decision maker don't causally influence the possible states of the world, or the possible actions of any other decision maker. This can be illustrated for the principle of strong dominance[31] by the Israel-and-Egypt-game of Bar-Hillel and Margalit (1972): Israel can decide for one of two possible actions: $a_1$: Israel remains in the occupied territories at $t_1$. $a_2$: Israel withdraws from the occupied territories at $t_1$. Egypt can decide for one of two possible actions: $a_1$: Egypt makes war at $t_2$. $a_2$: Egypt keeps peace at $t_2$. Israel's task is to make a rational decision.

| | $a_1$: Egypt makes war at $t_2$. | $a_2$: Egypt keeps peace at $t_2$. |
|---|---|---|
| $a_1$: Israel remains in the occupied territories at $t_1$. | $u_2$ | $u_4$ |
| $a_2$: Israel withdraws from the occupied territories at $t_1$. | $u_1$ | $u_3$ |

Figure 3. Matrix of the utility indices for Israel.

Figure 3 shows that Israel's possible action to remain in the occupied territories strongly dominates Israel's other possible action to withdraw from the occupied territories, for $u_2 > u_1$ and $u_4 > u_3$. However, if Israel believes that withdrawal leads to peace and remaining leads to war, then Israel will prefer withdrawal to remaining and will reject the strongly dominant possible action. This example makes clear that the principle of

[30]Weak dominance is sometimes also called "semi-dominance".

[31]Of course it can also be illustrated for the principle of weak dominance with a slightly altered example.

strong dominance cannot be applied, when the decision maker believes that the possible actions of the decision maker causally influence the possible states of the world, or the possible actions of any other decision maker. The principle of strong dominance can be applied, when the decision maker believes that the possible states of the world, or the possible actions of any other decision maker are causally independent of the possible actions of the decision maker.

## *Maximising Principles and the Principles of Dominance with Probabilistic and Causal Independence*

Gibbard and Harper (1978) and Sobel (1988c) claim that the principle of dominance with probabilistic independence is a corollary of an evidential maximising principle and that the principle of dominance with causal independence is a corollary of a causal maximising principle.[32] This can be seen for the principle of strong dominance by the following:

Suppose one starts with Savage's (1954/1972) principle of maximising utility. In case of the decision maker's belief in probabilistic or causal independence of the possible states of the world from the possible actions of the decision maker Savage's (1954/1972) formula for the utility of a possible action of the decision maker, $U(a_i) = \sum_{j=1}^{m} c(s_j)u(o_{ij})$, can take the following two forms (cf. Gibbard and Harper 1978):

(1) $U(a_i) = \sum_{j=1}^{m} c(s_j \mid a_i)u(o_{ij})$, which is the definition of Jeffrey's (1965) (evidential) conditional utility,

(2) $U(a_i) = \sum_{j=1}^{m} c(a_i \mathbin{\square\!\!\rightarrow} s_j)u(o_{ij})$, which is the definition of Gibbard and Harper's (1978) (causal) *U*-utility in Jeffrey's (1965) notation. ("$a_i \mathbin{\square\!\!\rightarrow} s_j$" means "if I were to do $a_i$, then $s_j$ would obtain".)

Savage's (1954/1972) formula can take form (1), because if the decision maker believes that the possible states of the world are probabilistically independent of the possible actions of the decision maker, then $c(s_j) = c(s_j|a_i)$ and one obtains formula (1) by substituting $c(s_j|a_i)$ for $c(s_j)$ in Savage's (1954/1972) formula. Savage's (1954/1972) formula can take form (2), because if the decision maker believes that the possible states of the world are causally independent of the possible actions of the decision

[32]The distinction between evidential and causal decision theories will be explained more thourougly in the next section.

maker, then $c(s_j) = c(a_i \boxdot\!\!\rightarrow s_j)$ and one obtains formula (2) by substituting $c(a_i \boxdot\!\!\rightarrow s_j)$ for $c(s_j)$ in Savage's (1954/1972) formula.

Obviously the principle of strong dominance with probabilistic independence gives the same recommendation as the principle of maximising the utility of form (1), if $c(s_j) = c(s_j|a_i)$. For in this case formula (1) reduces to Savage's formula; and if one, for example, compares the utilities of two possible actions with each other, $U(a_1) = \sum_{j=1}^{m} c(s_j)u(o_{1j})$ and $U(a_2) = \sum_{j=1}^{m} c(s_j)u(o_{2j})$, and neglects the weights, $c(s_j)$, because they equal each other in $U(a_1)$ and $U(a_2)$, then the sum of the utilities of the possible outcomes with regard to $a_1$ can be compared with the sum of the utilities of the possible outcomes with regard to $a_2$, which is a very similar procedure as in dominance reasoning.[33] And if the sum of the utilities of the possible outcomes with regard to $a_1$ is greater than the sum of the utilities of the possible outcomes with regard to $a_2$,[34] then the principle of maximising utility recommends $a_1$, which is the same result as in dominance reasoning. The equivalent holds for the principle of strong dominance with causal independence and the principle of maximising the utility of form (2).[35]

## *Evidential and Causal Decision Theories*

In the last section I showed that Savage's (1954/1972) formula for the utility of a possible action of the decision maker, $U(a_i) = \sum_{j=1}^{m} c(s_j)u(o_{ij})$, can take two forms. By means of this example I want to illustrate the fact that there are basically two kinds of rational decision theories, namely evidential decision theories and causal decision theories. For even if the utilities of the possible outcomes in Savage's formula cannot get an evidential or a causal interpretation, the probabilities in Savage's formula can get an evidential or a causal interpretation. This can be seen by the following: If one substitutes $c(s_j)$ by $c(s_j|a_i)$, then this reflects the fact that the decision maker believes that there is either probabilistic independence or probabilistic dependence between the

---

[33]The difference to dominance reasoning consists in the fact that in dominance reasoning one compares the utilities of the possible outcomes with each other for each possible state of the world.

[34]The difference to dominance reasoning consists in the fact that the utilities of the possible outcomes with regard to $a_1$ have to be greater than the utilities of the possible outcomes with regard to $a_2$ for each possible state of the world in the case of strong dominance.

[35]If one additionally assumes that $c(s_j) > 0$ for all $j = 1, ..., m$, then the principle of weak dominance with probabilistic independence respectively the principle of weak dominance with causal independence is a corollary of the principle of maximising the utility of form (1) respectively a corollary of the principle of maximising the utility of form (2).

possible states of the world and the decision maker's possible actions, so that the decision maker believes that his possible actions are either no evidence/signs or evidence/signs for possible states of the world. And if one substitutes $c(s_j)$ by $c(a_i \mathbin{\square\!\!\rightarrow} s_j)$, then this reflects the fact that the decision maker believes that there is either causal independence or causal dependence of the possible states of the world from the decision maker's possible actions, so that the decision maker believes that his possible actions are either no causes or causes for possible states of the world. Rational decision theories which take account of the evidential relations a decision maker believes to hold between possible states of the world and possible actions are called evidential decision theories, whereas decision theories which take account of the causal relations a decision maker believes to hold between possible states of the world and possible actions are called causal decision theories.

An example for an evidential decision theory is Jeffrey's (1965) rational decision theory. For he uses $c(s_j|a_i)$ in his formula for the utility of a possible action of the decision maker. An example for a causal decision theory is more difficult to find before Newcomb's problem was invented. For although Savage's (1954/1972) formula for the utility of a possible action of the decision maker can get an evidential reading by substituting $c(s_j)$ by $c(s_j|a_i)$ or a causal reading by substituting $c(s_j)$ by $c(a_i \mathbin{\square\!\!\rightarrow} s_j)$, his own formula is neither on the evidential nor on the causal side. Yet if one looks at his terminology, Savage uses the term "consequence" instead of the term "possible outcome", which is significant. For "consequence" often just means "effect", so that a causal reading suggests itself (cf. Joyce in press, p. 48). Thus the classification of rational decision theories into evidential decision theories and causal decision theories relies on two criteria, namely which relations it takes account of and which terminology it uses.

Given the distinction between causal and evidential decision theories we can say that the principle of dominance with probabilistic independence belongs to evidential decision theory, whereas the principle of dominance with causal independence belongs to causal decision theory. For the principle of dominance with probabilistic independence takes into account whether or not there are evidential relations between the possible states of the world and the possible actions of the decision maker, and the principle of dominance with causal independence takes into account whether or not there are causal relations between the possible states of the world and the possible actions of the decision maker. Furthermore, because the principle of dominance with probabilistic independence can also be viewed as a corollary of Jeffrey's (1965)

principle of maximising conditional utility, and because Jeffrey's (1965) rational decision theory is an evidential decision theory, the principle of dominance with probabilistic independence is part of evidential decision theory, too. The equivalent holds for the principle of dominance with causal independence and causal decision theory.

## *Descriptive and Normative Rational Decision Theory*

Rational decision theory can be descriptive or normative (Eells 1982, p. 24, Spohn 1993). Therefore the principles of maximising utility, of maximising conditional utility, and of dominance can be also read descriptively or normatively.[36] Gärdenfors and Sahlin (1988, p. 5) explain that a rational decision theory is descriptive, when it describes how decision makers actually decide in various decision situations, and that a rational decision theory is normative, when it gives rules what a decision maker should decide for in various decision situations. While a descriptive interpretation of rational decision theory is prevalent among psychologists, a normative interpretation of rational decision theory is prevalent among economists and among philosophers who are interested in the foundations of rational decision theory (Gärdenfors and Sahlin 1988, p. 6).

With regard to the descriptive interpretation of rational decision theory Eells (1982, p. 33) points out that the question is not whether decision makers consciously use a particular decision-making principle while making a decision. According to Eells (1982, pp. 33-34) the question is rather whether decision makers generally decide in such a way as if they consciously use the particular decision-making principle. Considering the normative interpretation of rational decision theory Eells (1982, pp. 32-

[36]Mellor (1995, p. 81) uses the term prescriptive instead of the term normative. Mellor (1995, p. 81) believes that the principle of maximising conditional utility is subjective, descriptive, and deterministic. It is descriptive, for he believes Ramsey (1990, p. 69) is right in claiming that "we act in the way we think most likely to realise the objects of our desires, so that a person's actions are completely determined by his desires and opinions ... seems ... a useful approximation to the truth". Mellor (1991a, chapter 16) himself proposes an objective rational decision theory, and he (Mellor 1991a, pp. 276-277) claims that one needs objective utilities and objective probabilities to make prescriptive sense of the principle of maximising conditional utility. In my opinion, however, to make prescriptive sense of the principle of maximising conditional utility one doesn't need objective utilities and objective probabilities, if utilities and credences can be conceived as approximations of objective utilities and objective probabilities. Furthermore, it is unclear to me why one can't have subjective wants and beliefs and still make a rational decision given these subjective wants and beliefs, so that the respective decision principle prescribes a certain possible action as rational given these subjective wants and beliefs. Or as Eells (1982, p. 30) puts it the principle of maximising conditional utility is "as applicable to the deliberation of a monster as it is to that of a saint".

33) distinguishes between a weak and a strong normative interpretation. While the weak normative interpretation only tells the decision maker which decision principle to use, the strong normative interpretation also tells him how to get optimal probability and utility assignments. While Eells (1982, pp. 31-32) rejects objective probabilities and objective utilities as optimal probability and utility assignments, for the decision maker generally doesn't know them, and for it is questionable that they exist at all, Mellor (1991a, chapter 16) defends both objective probabilities and objective utilities. Furthermore, Lewis (1981a), Skyrms (1984), and Sobel (1986) defend chances in their causal decision theories. Therefore we shouldn't reject from the very beginning that optimal probability and utility assignments could be identified with objective probability and utility assignments.

Other elements of rational decision theory can also be treated descriptively or normatively. For which possible actions and which possible states of the world figure in decision situations can be given a descriptive or a normative interpretation, too. Possible actions and possible states of the world are used descriptively in a rational decision theory, when it describes which possible actions and possible states of the world the decision maker actually uses. Possible actions and possible states of the world are used normatively in a rational decision theory, when it gives rules which possible actions and possible states of the world the decision maker should use. All in all the following points can be treated descriptively or normatively: The decision situation, the probability and utility assignments, the decision principle, and the assignment of the possible actions and the possible states of the world. This seems to suggest that rational decision theories can be ordered along a continuum ranging from purely descriptive rational decision theories to purely normative rational decision theories.

Gärdenfors and Sahlin (1988, p. 6) point out, though, that the distinction between descriptive and normative rational decision theories is slippery, because normative rational decision theories are tested against our intuitive understanding of rational decisions, which is not questioned anymore, and our behaviour may be influenced by existing normative rational decision theories. This situation can also be observed in the case of Newcomb's problem. Newcomb's problem led to the development of rational decision theories which were on the normative side of the spectrum and which were tested against unquestioned intuitions.[37] For in Newcomb's problem two opposing arguments are on first sight equally plausible, which rest on two

[37]Swain (1988) is the only one I know who investigates and questions intuitions in Newcomb's problem.

normatively interpreted decision-theoretic principles, viz. the principle of maximising conditional utility and the principle of dominance, and which are backed up by several unquestioned intuitions. We will come back on this point.

## *The Ontology of Rational Decision Theory*

The concepts of rational decision theory have received a wide variety of ontological interpretations. Savage regards a possible state of the world as "a description of the world leaving no relevant aspect undescribed" (Savage 1954/1972, pp. 9 and 13), a possible outcome as "anything that might happen to a person" (Savage 1954/1972, p. 13), and a possible action as an abstract function from possible states of the world to possible outcomes.[38] Jeffrey (1965) modified Savage's rational decision theory by construing possible actions, possible states of the world, and possible outcomes as propositions. This is an improvement, because it yields a unified ontology for rational decision theory, and because in opposition to the former treatment one can apply standard logical operations to propositions. Because of these advantages from Jeffrey (1965) onwards most decision theorists adhered to the view that the objects of beliefs and wants of the decision maker are propositions. Nevertheless this view has to struggle with problems originating from the ontology of propositions.

Lewis (1979b) and Spohn (1997) questioned this view. For taken as sets of possible worlds propositions are not always appropriate objects of attitudes, whereas properties seem to be always appropriate (Lewis 1979b). Spohn (1997, p. 305) illustrates Lewis' (1979b) point of view in the following more conventional way[39]: A content of belief is a set of triples $\langle w, s, t \rangle$, where $w$ is a possible world, $s$ is an object, and $t$ is any time at which the object $s$ exists in $w$. From this Spohn (1997) derives his account of content of beliefs: A content of belief is a set of quadruples $\langle w, s, t, d \rangle$, where $w$ is a possible world, $s$ is a subject, $t$ is a time, and $d = \langle d_1, d_2, ... \rangle$ is a finite or infinite sequence of objects, which exist in $w$. Lewis (1979b) uses de se attitudes, Spohn (1997) de se and de nunc attitudes to argue for their respective accounts.

---

[38]On first sight I found it odd to identify possible actions with abstract functions from possible states of the world to possible outcomes. Yet Joyce (in press, p. 50) supports Savage (1954/1972) by claiming that acting often sets up a functional relationship between possible states of the world and possible outcomes. For example, a woman who decides to walk to work on a cloudy day makes her happiness a function of the weather.

[39]This account is more conventional than Lewis' (1979b) own account, because in opposition to Lewis it doesn't presuppose that each possible being exists only in one possible world, so Spohn (1997, p. 305, footnote 22).

Perry (1979, p. 3) gives an example to illustrate a de se attitude: Suppose you are shopping in a supermarket. After a while you notice that there is a sugar trail on the ground. You follow the trail to let the customer causing the trail know what is happening. You don't catch anybody, though, and then you recognise that you yourself leave the sugar trail behind. After that you rearrange the torn sack in your cart. The point of the story is: As long as you think that someone else causes the sugar trail, you don't change your behaviour. When you recognise that you are responsible, you change your behaviour. The following de se attitude corresponds to the story: "I am responsible for the sugar trail!". Perry (1979, p. 4) also gives an example to illustrate a de nunc attitude: A professor who wants to go to a department meeting at noon sits motionless in his office at that time. Suddenly he starts to move, so the question is: What explains his action? Perry's answer is: A change in belief. He believed all the time that the department meeting starts at noon; then he came to believe that it starts now. The following de nunc attitude corresponds to the story: "The meeting I want to go to starts now!"

The connection between de se attitudes and properties as appropriate objects of attitudes is: For Lewis (1979b) the word "proposition" means a set of possible worlds, that is a region of logical space; and the word "property" means the set of exactly those possible beings that have the property in question. Lewis uses two arguments. The first of them shows that propositions are not always appropriate objects of attitudes. If the decision maker lacks a de se attitude like in the example above, he can locate himself in logical space, in the sense that he is in a position to self-ascribe properties of inhabiting such-and-such a kind of world, but cannot locate himself in ordinary space, in the sense that he isn't in a position to self-ascribe properties where in the world he is, while if the decision maker has a de se attitude in the case above, he self-ascribes the property of being the cause of the sugar trail; this property doesn't correspond to any proposition, so Lewis (1979b), for there are worlds where some decision makers have the property in question and others don't have it. Lewis' second argument shows that when propositions are appropriate objects of attitudes, properties are appropriate, too. Lewis (1979b) claims that to any set of worlds there corresponds the property of inhabiting some world in that set; furthermore, he claims that if a property corresponds to any proposition, it corresponds to exactly one. Therefore when propositions are appropriate objects of attitudes, properties are appropriate, too. The equivalent holds for de nunc attitudes. Moreover, one can easily derive the connection between de se and de nunc attitudes and Spohn's account of content of beliefs from the foregoing.

In his 1981a, however, Lewis uses propositions instead of properties to formulate his causal decision theory. Lewis justifies his position in the following way (1981a, p. 7):

> "... a credence function over possible worlds allows for partial belief about the way the world is, but not for partial beliefs about who and where and when in the world one is. Beliefs of the second sort are distinct from those of the first sort; it is important that we have them; however they are seldom very partial. To make them partial we need either an agent strangely lacking in self-knowledge, or else one who gives credence to strange worlds in which he has close duplicates. I here ignore the decision problems of such strange agents."

Thus Lewis doesn't use properties, but propositions in his causal decision theory, because he doesn't want to consider decision makers who are strangely lacking in self-knowledge or who give credences to strange worlds in which they have close duplicates.

I have to acknowledge that the content of beliefs and wants in rational decision theory is an important topic. Yet solving this problem is another task and leaving it unsolved will not cause too much harm, for Newcomb's problem doesn't refer to agents in particular who lack de se-knowledge and de nunc-knowledge. I will use propositions as the content of beliefs and wants. By doing this I will conform to the mainstream in rational decision theory and will avoid considering decision makers who lack de se- and de nunc-knowledge.

## *Propositions in Rational Decision Theory*

The most widely held view is that a proposition is an abstract object to which a person is related by a belief, want, or other psychological attitude; furthermore, a proposition is typically expressed by a that-clause preceded by a psychological verb like "believe", "want", etc. (Wagner 1995). Wagner (1995, p. 658) explains that the psychological states in question are called propositional attitudes and illustrates the relation between propositional attitude and proposition by the following example: "When I believe that snow is white I stand in the relation of believing to the proposition that snow is white." Wagner (1995) continues that a proposition can be a common object for various attitudes of various persons. Moreover, Wagner (1995) makes the relation between sentences and propositions clearer: Because a sentence expressing an attitude is also taken to express the associated proposition, the proposition is the shared meaning

of this sentence and all its synonyms in every language. This summarises the traditional doctrine of propositions (Wagner 1995).

In rational decision theory and with regard to Newcomb's problem propositions are conceived in the following way: Jeffrey's (1983, pp. 64-65) position is similar to the traditional doctrine of propositions. He maintains that propositions either are linguistic entities or have affinities to certain linguistic entities. We name propositions by putting the word "that" or the phrase "the proposition that" before the corresponding declarative sentences (Jeffrey 1983). Thus the relation between a sentence and the proposition it expresses is similar to the relation between direct and indirect quotation. Sentences that have the same meaning express the same proposition (Jeffrey 1983). Later in the text Jeffrey (1983, pp. 83-85) defends the view that possible actions can be conceived as propositions. For possible actions can often be described by declarative sentences. And if possible actions are characterised with sufficient accuracy[40] by declarative sentences, we can identify the possible actions with the propositions that the sentences express, Jeffrey (1983) claims, so that a possible action is a proposition which is within the decision maker's power to make true. Jeffrey (1983, pp. 65-66) uses italic capitals, like "*A*", "*B*", "*C*", ..., to stand for his propositions. I will take over this usage.

Gibbard and Harper (1978, p. 125) state that propositions can take the form: "If I were to do $a_i$, then $o_j$ would happen.". Or to say it in a matrix formulation: "If I were to do $a_i$, then $o_{ij}$ would happen.". Gibbard and Harper (1978) call such a proposition a counterfactual. In the literature counterfactuals are also called subjunctive conditionals. Gibbard and Harper (1978) form subjunctive conditionals with the connective $\Box\!\!\rightarrow$; and they denote "if I were to do $a_i$, then $o_j$ would happen" by "I do $a_i \Box\!\!\rightarrow o_j$ happens". Or to say it in a matrix formulation: "If I were to do $a_i$, then $o_{ij}$ would happen." is denoted by "I do $a_i \Box\!\!\rightarrow o_{ij}$ happens.". In this dissertation I will use a matrix formulation and denote "if I were to do $a_i$, then $o_{ij}$ would happen" by "$a_i \Box\!\!\rightarrow o_{ij}$".

Yet Gibbard and Harper (1978) also use propositions à la Jeffrey (1983). How can we reconcile this? As far as I know there is no a priori reason why that-clauses can only contain declarative sentences like in Jeffrey's (1983) case. That-clauses can also contain subjunctive conditionals like in Gibbard and Harper's (1978) case, or they can contain indicative conditionals.[41] Indicative conditionals take the following form in a

---

[40]Unfortunately Jeffrey (1983) doesn't give a criterion, when possible actions are characterised with sufficient accuracy by declarative sentences.

[41]I would like to thank Howard Sobel for drawing my attention to this fact. In personal communication Howard Sobel wrote to me (29th of August, 1998): "Subjunctive conditionals are

matrix formulation: If I do $a_i$, then $o_{ij}$ will happen. This will be denoted by "$a_i \triangle\!\!\rightarrow o_{ij}$".[42]

A completely different point of view is taken by Lewis (1979b; 1981a). For Lewis (1979b, pp. 514-515; 1981a, p. 6) means a set of possible worlds by the word "proposition". He thinks that propositions in the traditional sense might rather be regarded as sentential meanings (cf. Lewis 1970). Because the nature of possible worlds is highly controversial (cf. Adams 1995), and because Lewis' (1981a) causal decision theory is an advancement of Gibbard and Harper's (1978) causal decision theory, I will follow the traditional doctrine of propositions amended by the fact that that-clauses can contain declarative sentences, indicative conditionals, and subjunctive conditionals.

## 1.3 Game-Theoretic Foundations for Analysing Newcomb's Problem[43]

Because Newcomb's problem hasn't been studied much in game theory, and because Newcomb's problem is difficult to classify in game theory, I will introduce the reader to game theory by a more typical game, namely the prisoner's dilemma[44], which is easy to classify in game theory. My reasons for looking at Newcomb's problem from

---

a kind of proposition. There is nothing special about their probabilities. For example, for incompatible propositions, p and q, P(p v q) = P(p) + P(q). p and q are any propositions. They can (one or both) be subject-predicate propositions, generalisations, material conditionals, subjunctive conditionals. It doesn't matter."

[42]As far as I know there is no standard for the symbolism of indicative conditionals. The use of the symbol $\triangle\!\!\rightarrow$ is my suggestion.

[43]According to Spohn (1994, p. 198) game theory is a specialisation of rational decision theory, so that a game situation can be described by the set of possible actions of a first decision maker, the set of possible actions of another decision maker, or of the other decision makers, the set of possible outcomes of the decision maker, and the set of possible outcomes of another decision maker, or of the other decision makers. Skyrms (1990a, p. 145) also defends the view that game theory should be embedded in rational decision theory, but stresses the process of rational deliberation.

[44]According to Campbell (1985, p. 3) the prisoner's dilemma was formulated for the first time around 1950 by the social psychologist Merill M. Flood and the economist Melvin Dresher and was formalised for the first time by the game theorist Albert W. Tucker. Campbell (1985, p. 3) states that Flood and Dresher used the prisoner's dilemma for testing a theorem in game theory. Because of being a game the prisoner's dilemma has been extensively studied in game theory (for example Luce and Raiffa 1967). Campbell (1985, p. 3-4) points out that the prisoner's dilemma has been studied outside of philosophy, for example, by psychologists (for example Rapoport and Chammah 1965), political scientists (for example Taylor 1976), and evolutionary theorists (for example Axelrod and Hamilton 1981).

the perspective of game theory are that it could provide a unique solution to Newcomb's problem and other Newcomblike problems.

## *The Prisoner's Dilemma*

Jeffrey (1983, p. 15) describes one version of the prisoner's dilemma in the following way:

> "Two men are arrested for bank robbery. Convinced that both are guilty, but lacking enough evidence to convict either, the police put the following proposition to the men and then separate them. If one confesses but the other does not, the first will go free (amply protected from reprisal) while the other receives the maximum sentence of ten years; if both confess, both will receive lighter sentences of four years; and if neither confesses, both will be imprisoned for one year on trumped-up charges of jaywalking, vagrancy, and resisting arrest."

## *The Primitive Terms*

The decision situation of the prisoner's dilemma consists of two decision makers[45]. Each decision maker has to decide between two possible actions, which are the same for both decision makers, so that the set of possible actions is the same for both decision makers. The set of decision makers will be denoted by $X$, and the particular decision makers will be denoted by $x_1, x_2, ..., x_o$. The possible actions of decision maker $x_1$ and the possible actions of decision maker $x_2$ get one time index $t$. The possible actions of decision maker $x_1$ and the possible actions of decision maker $x_2$ result in possible outcomes for decision maker $x_1$ and in possible outcomes for decision maker $x_2$. The wants of decision maker $x_1$ respectively $x_2$ are represented by the utilities or objective utilities of $x_1$'s possible outcomes respectively $x_2$'s possible outcomes. The utilities of the possible outcomes will be denoted by $u(o_{ij})$. The equivalent holds for objective utilities. The utilities of the possible outcomes will be ranked in the following way: $u_1$ is the lowest utility, $u_2$ is the second lowest utility, ..., and $u_{n \cdot m}$ is the highest utility.[46] The equivalent holds for objective utilities. The beliefs of decision maker $x_1$ respectively $x_2$ are represented by the credences or chances of

---

[45]In game theory the term "decision maker" is not used, but the terms "person" and "player".

[46]The number $n \cdot m$ derives from the fact that there are $n$ different possible actions of the decision maker and $m$ different possible actions of the other decision maker, so that $n \cdot m$ possible outcomes for each decision maker result.

$x_1$'s respectively $x_2$'s possible outcomes. In comparison to Newcomb's problem in the decision situation of the prisoner's dilemma the possible states of the world are replaced by the possible actions of decision maker $x_2$, and the possible outcomes, utilities or objective utilities, utility indices, and credences or chances for decision maker $x_2$ are added. Otherwise everything remains the same. The decision situation of Jeffrey's instantiation of the prisoner's dilemma can be represented by the following decision matrix:

| | | $x_2$ | |
|---|---|---|---|
| | | $a_1$: I confess at $t_1$. | $a_2$: I don't confess at $t_1$. |
| $x_1$ | $a_1$: I confess at $t_1$. | 4 year sentence, 4 year sentence | 0 year sentence, 10 year sentence |
| | $a_2$: I don't confess at $t_1$. | 10 year sentence, 0 year sentence | 1 year sentence, 1 year sentence |

Figure 4. Decision matrix for Jeffrey's instantiation of the prisoner's dilemma of the possible outcomes for decision maker $x_1$ (first item of the possible outcomes), which result from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$, and of the possible outcomes for decision maker $x_2$ (second item of the possible outcomes), which result from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

## *Classification of the Prisoner's Dilemma in Game Theory*

In game theory the prisoner's dilemma is classified as a non-strictly competitive, non-co-operative, 2-person, non-zero sum game. Games are non-strictly competitive, if the decision makers' interests aren't entirely opposed (Resnik 1987, p. 126). Games are non-co-operative, if the decision makers can't communicate with each other to co-ordinate their possible actions (Resnik 1987, p. 126); furthermore, games are classified as 2-person games, if each of the two decision makers has to decide between his possible actions. Games are classified as non-zero sum games, if the utilities of the possible outcomes of the decision makers at the corresponding possible outcomes don't sum to 0 (Brams 1983, p. 180).[47]

[47] According to Resnik (1987, p. 128) 2-person non-strictly competitive games and 2-person non-zero sum games are the same. For in any 2-person non-strictly competitive game decision maker $x_1$ and decision maker $x_2$ don't have strictly opposing preferences, so that the utility function of

## *Solutions of the Prisoner's Dilemma in Game Theory*

According to Hardin (1995, p. 292) most solutions in game theory are about possible outcomes, that is they postulate which possible outcomes or range of possible outcomes are game-theoretically rational; yet some solutions are also about strategies, that is they postulate which plans are game-theoretically rational.

With regard to the prisoner's dilemma the solutions are about strategies and about possible outcomes. For Nozick (1969) applies the principle of strong dominance and the principle of maximising conditional utility to the prisoner's dilemma, which leads to the recommendation of certain possible actions, that is strategies, in the prisoner's dilemma. With regard to solutions which postulate certain possible outcomes as game-theoretically rational Brams (1983) uses the concept of a Nash equilibrium to defend a certain solution to the prisoner's dilemma, and Resnik (1987, p. 151) applies the concept of Pareto-optimality to the prisoner's dilemma. In the following I will show how the different principles and concepts apply to the prisoner's dilemma. Let's start with the principle of strong dominance.

According to Nozick (1969) each decision maker can apply the principle of strong dominance to the prisoner's dilemma. For $a_1$ strongly dominates $a_2$ for decision maker $x_1$, and $a_1$ strongly dominates $a_2$ for decision maker $x_2$. If both decision makers apply the principle of strong dominance, they will both confess at $t_1$ and will both end up with a 4 year sentence.

According to Brams (1983, p. 54) the decision makers' shared possible outcome of a 4 year sentence represents a Nash equilibrium which he explains in the following way (Brams 1983, p. 182): "In a normal-form game, a Nash equilibrium is an outcome from which no player would have an incentive to depart unilaterally because he would do (immediately) worse, or at least not better, if he moved."[48] Applying this to the prisoner's dilemma we get the following result: If decision maker $x_1$ changes from $a_1$ to $a_2$, and decision maker $x_2$ stays with $a_1$, decision maker $x_1$ gets a worse possible outcome. If decision maker $x_2$ changes from $a_1$ to $a_2$, and decision maker $x_1$ stays with $a_1$, decision maker $x_2$ gets a worse possible outcome. This seems to suggest

---

decision maker $x_1$ isn't the negative utility function of decision maker $x_2$. Thus decision maker $x_1$'s utility for a particular outcome $o_{ij}$ isn't the negative of decision maker $x_2$'s utility for this particular outcome $o_{ij}$, and their utilities don't sum to 0.

[48] According to Brams (1983, p. 182) a normal-form game is a game, in which it is supposed that the decision makers decide independently between their possible actions.

that if all decision makers follow the principle of dominance, they will reach Nash equilibria.[49]

Can the concept of a Nash equilibrium be applied to Newcomb's problem?[50] I think yes. Let's see why. If we suppose that the possible states of the world in Newcomb's problem can also be conceived as possible actions of the predictor, then Newcomb's problem becomes a game with two decision makers. According to Skyrms (1990b, p. 44) this is a justifiable move for the following reason:

> "The inclusion of acts in the Boolean algebra over which probabilities are defined is an innovation which may provoke varying reactions. ... However, one may argue that this feature makes the system attractive for dealing with sequential decision problems. In such problems, the choice of an option may change its status over time from consequence to act to part of the state of the world, and each change goes with an appropriate updating of subjective probability."

Furthermore, we have to suppose that the predictor gets some payoff out of the game. Although in the original version of Newcomb's problem no such payoff is indicated, we can simply suppose that the decision maker's gain is the predictor's loss, so that the utilities of the possible outcomes of the predictor are as in the following decision matrix:

| | $a_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $a_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | \$ 1,000, -\$ 1,000 | \$ 1,001,000, -\$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | \$ 0, \$ 0 | \$ 1,000,000, -\$ 1,000,000 |

Figure 5. Decision matrix for Newcomb's problem of the possible outcomes for the decision maker (first number of the possible outcomes), which result from combining

[49]How game-theoretic terms are related to decision-theoretic terms is a very interesting question which cannot be answered here, though. Skyrms (1990a) tries to transfer the equilibrium concept from game theory to rational decision theory and uses it for modelling the dynamics of rational deliberation.

[50]I wish to thank Luc Bovens for discussion of this point.

the possible actions of the decision maker with the possible actions of the predictor, and of the possible outcomes for the predictor (second number of the possible outcomes), which result from combining the possible actions of the decision maker with the possible actions of the predictor.

In this way Newcomb's problem becomes a strictly competitive, non-co-operative, 2-person, zero sum game. The Nash equilibrium for this game results in the combination of the decision maker's possible action $a_1$ with the predictor's possible action $a_1$. For if the decision maker changes from $a_1$ to $a_2$ and the predictor stays with $a_1$, the decision maker gets a worse possible outcome; and if the predictor changes from $a_1$ to $a_2$ and the decision maker stays with $a_1$, the predictor gets a worse possible outcome. Therefore the decision maker should take both boxes.

The Nash equilibrium solution of the prisoner's dilemma is not Pareto-efficient.[51] Binmore (1992, p. 177) defines Pareto-efficiency in the following way: "Something is feasible if it is possible to select it. An outcome $x$ in a feasible set $X$ is *Pareto-efficient* ... if there is no other outcome $y$ in $X$ that all the players like at least as much as $x$ and some players like more than $x$." Applying this definition to the prisoner's dilemma we get the following result: The shared possible outcome of 4 is not Pareto-efficient, because the shared possible outcome of 1 is better for decision maker $x_1$ and for decision maker $x_2$, and not worse for decision maker $x_1$ and for decision maker $x_2$. The shared possible outcome of 1 is Pareto-efficient, though.

By contrast the Nash equilibrium solution of Newcomb's problem is not Pareto-efficient, because the shared possible outcome of 0 is better for decision maker $x_1$ and for decision maker $x_2$, and not worse for decision maker $x_1$ and for decision maker $x_2$.

---

[51]The term "Pareto" comes from the Italian sociologist Vilfredo Pareto (Binmore 1992, p. 177). In the literature the terms "Pareto-efficient", "Pareto-optimal", which are properties, and "Pareto-superior", which is a relation, are used almost interchangeably as can be seen by the following definitions: While Resnik (1987, p. 151) defines Pareto-optimality in the following way: "An outcome associated with $(u_1, u_2)$ is *Pareto optimal* just in case for no pair of values $(v_1, v_2)$ associated with outcomes of the game, do we have either $v_1 > u_1$ and $v_2 \geq u_2$ or $v_2 > u_2$ and $v_1 \geq u_1$.", Davis (1970, p. 118) defines it like that: "An outcome is Pareto optimal if there is no other possible agreement that enables both players to do better *simultaneously*.". Brams (1983, p. 183) gives the definition: "An outcome is Pareto-inferior if there exists another outcome that is better for some player(s) and not worse for all the other player(s). If there is no such other outcome, the outcome in question is Pareto-superior." According to Binmore (1992, p. 177) the term "Pareto-efficient" is to be preferred to the term "Pareto-optimal", because "it suggests that a Pareto-efficient point cannot be improved on. But it is Pareto-efficient for a mother to give all the cookies in her cooky jar to one of her children, leaving the others with nothing. No child can then have its situation improved without making another worse off. However, nobody would wish to claim that the mother's decision is necessarily socially optimal."

The shared possible outcome of 0 is Pareto-efficient, though. Therefore with regard to Pareto-efficiency the decision maker should take B2.

Nozick (1969) claims that each decision maker can apply the principle of maximising conditional utility to the prisoner's dilemma. Nozick (1969, p. 131) illustrates this by modifying the prisoner's dilemma in the following way:

> "... suppose ... there are two inherited tendencies, one to perform the dominant action ..., and one to perform the other action. Either tendency is genetically dominant over a possible third inherited trait. Persons (I) and (II) are identical twins, who care only about their own payoffs ..., and know that their mother had the neutral gene, one of their two possible fathers had only the gene to perform the dominant action, and the other had only the gene not to perform the dominant action. Neither knows which man was their father, nor which of the genes they have. Each knows, given the genetic theory, that it is almost certain that if he performs the dominant (dominated) action his brother will also. ... suppose ... the theory tells us and them that given all this information ..., the correlation between their actions holds as almost certain, and also given *this* additional information, it holds as almost certain, etc. ... . Suppose brother I argues: 'If I perform the dominant action then it is almost certain$_1$ that I have that gene, and therefore that my brother does also, and so it is almost certain$_2$[16] that he will also perform the dominant action and so it is almost certain$_2$ that I will get 4. Whereas if I perform the dominated action, for similar reasons, it is almost certain that my brother will also, and hence it is almost certain that I will get 8. So I should perform the dominated action'."

According to Nozick (1969) the argument for not-confessing for decision maker $x_1$ and for decision maker $x_2$ is:

**Premise 1**: If I confess at $t_1$, I and my brother have the confessing gene with almost certainty$_1$, so that my brother also confesses at $t_1$ with near certainty$_2$ and I get 4 with near certainty$_2$.

**Premise 2**: If I don't confess at $t_1$, I and my brother have the non-confessing gene with almost certainty$_1$, so that my brother also doesn't confess at $t_1$ with near certainty$_2$ and I get 8 with near certainty$_2$.

**Conclusion**: Therefore I shouldn't confess at $t_1$.

Yet Nozick (1969) doubts that this argument works; for what decision maker $x_1$ does, doesn't causally affect what decision maker $x_2$ does.

## *Classification of Newcomb's Problem in Game Theory*

According to Resnik (1987, p. 121) rational decision theory can be distinguished from game theory in the following way: Whereas in rational decision theory the decision of just one decision maker plays an active role in determining his possible outcomes, in game theory the decisions of at least two decision makers play an active role in determining his possible outcomes. This distinction, however, isn't quite correct, for there are also games against nature.[52] Brams (1983, p. 181) defines games against nature in the following way: "A game against nature is a game in which one player is assumed to be 'nature', whose choices are neither conscious nor based on rational calculation but on chance instead." For Brams (1983, pp. 41-65) games against nature are classified as decision-theoretic problems and not as game-theoretic problems, though, so that Resnik's distinction seems to be correct after all. By classifying Newcomb's problem as a 1-person game against nature Brams (1983, pp. 41-65) evaluates Newcomb's problem as a decision-theoretic problem. Unfortunately Brams doesn't give a justification for classifying Newcomb's problem as a 1-person game against nature. And it is problematic to view the predictor as nature, whose decisions are based on chance. For the predictor's predictions are determined in some other way as we will see in chapter 1.4.

Newcomb's problem can be conceived as a game against nature, if one conceives nature as something fixed and determinate, that is if one stresses that the predictor had already made his prediction, and that he has put $ 1,000,000 in B2 or not, when the decision maker is faced with Newcomb's problem. In this perspective the decision maker is confronted with two possible states of the world and not with two possible actions of the predictor. Yet Brams (1983, p. 50) partitions the possible states of the world differently in Newcomb's problem, viz. in terms of the predictor's correctness and the predictor's incorrectness. Thus it remains difficult to make sense of Brams' claim that Newcomb's problem is a game against nature, if games against nature are explicated in the way he does it.

Newcomb's problem can also be conceived as a game with two decision makers. Brams' (1983, p. 51) objection that - whatever the decision maker does - the two possible states of the world, namely that the predictor is correct or incorrect, occur with the same relative frequency, doesn't force us to conclude that the predictor cannot decide freely which prediction to make. It all depends on what a free decision is, and if

[52] Resnik (1987, p. 122) regards games against nature as games of chance.

you explicate freedom of decision in such a way that the decision-finding process is unhindered, you can conceive Newcomb's problem as a game with two decision makers. The different times of decision-making, namely that the predictor decides at an earlier time than the decision maker, shouldn't be a reason for conceiving Newcomb's problem not as a game with two decision makers; there are a lot of games, for example, chess, where the decision makers have to decide at different times. The different objects of decision-making, namely that the predictor has to decide which prediction to make, and the decision maker has to decide which possible action to perform, shouldn't be a reason for conceiving Newcomb's problem as not a game with two decision makers; there are a lot of games, for example, the Israel-and-Egypt-game of Bar-Hillel and Margalit (1972), where the decision makers have to decide about different objects.

Yet one objection remains to view Newcomb's problem as a game with two decision makers, namely in a game each decision maker usually receives a payoff. But in Newcomb's problem it is unclear what the predictor's utilities of his possible outcomes are. We simply don't know whether the predictor wants the decision maker to win as much as possible or not, whether the predictor wants to be correct or not, or whether the predictor wants to punish greedy decision makers, who decide for taking both boxes, or not, or whether a combination of any of these three factors holds. Yet even if we don't know what the predictor's utilities of his possible outcomes are, this doesn't mean that he doesn't have any payoffs. Surely he has some payoffs, and for simplicity's sake we can just suppose as in the case of the Nash equilibrium solution what the predictor's payoffs are. The conclusion is that no sharp distinction between games against nature and games with two decision makers can be drawn for Newcomb's problem, that is Newcomb's problem can either be conceived as a 1-person game against nature or as a game with two decision makers.

But perhaps this situation changes, if we compare a 1-shot Newcomb's problem with a finitely iterated Newcomb's problem with the same decision maker. Sorensen (1985) claims that in a finitely iterated Newcomb's problem with the same decision maker the decision maker tries to build up a reputation as a B2 taker as we will see in chapter 1.6. But why should the decision maker try to build up a reputation as a B2 taker, if he views the predictor as something belonging to the possible states of the world? The only way the decision maker can make sense of his reputation building behaviour is to view the predictor as playing an active role in determining the decision maker's possible outcomes. Moreover, the active role of the predictor is stressed by the following: Sorensen (1985) postulates that in a finitely iterated Newcomb's problem with

the same decision maker the predictor's last prediction is affected by his opinion of the decision maker's decision-making tendencies. That is the predictor will put $ 1,000,000 in B2 in the last play, if he believes that the decision maker has the tendency to take B2; and the predictor will put $ 0 in B2 in the last play, if he believes that the decision maker has the tendency to take both boxes. Thus the predictor's future action is affected by the predictor's opinion of the decision maker's decision-making tendencies. Therefore a finitely iterated Newcomb's problem with the same decision maker seems to be a game with two decision makers, namely the decision maker and the predictor.

In opposition to that in a 1-shot Newcomb's problem the decision maker doesn't try to build up a reputation as a B2 taker, for one cannot build up a reputation in one single play. Furthermore, the predictor had already made his prediction in a 1-shot Newcomb's problem, when the decision maker decides, that is the predictor's prediction had already become a part of the possible states of the world, when the decision maker decides (except in case backwards causation takes place in Newcomb's problem). Therefore a 1-shot Newcomb's problem seems to be a game against nature.

If a 1-shot Newcomb's problem is a game against nature, the following partition of the possible states of the world suggests itself as the correct partition: $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. Thus Newcomb's problem should be represented by the following decision matrix:

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | $o_{11}$: $ 1,000 | $o_{12}$: $ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | $o_{21}$: $ 0 | $o_{22}$: $ 1,000,000 |

Figure 6. Decision matrix for Newcomb's problem of the possible outcomes $o_{11}, o_{12}, o_{21}, o_{22}$ which result from combining the possible actions $a_1, a_2$ with the possible states of the world $s_1, s_2$.

The reason why this is the correct partition is: If a 1-shot Newcomb's problem is a game against nature and nature is conceived as something fixed and determinate, then the predictor's prediction is fixed and determinate, when the decision maker decides. Thus the predictor's prediction cannot be changed anymore to correspond to the predictor's high reliability and to the probabilistic dependence between the predictor's prediction and the decision maker's decision. In this way the predictor's high reliability and the probabilistic dependence become irrelevant to the solution of Newcomb's problem. And partitions of the possible states of the world which stress the predictor's correctness, like $s_1$: The predictor has predicted correctly at $t_1$ vs. $s_2$: The predictor hasn't predicted correctly at $t_1$, turn out to be wrong.[53]

One problem remains, namely is Newcomb's problem in Nozick's original formulation a finitely iterated Newcomb's problem or a 1-shot Newcomb's problem? Horne (1983, p. 219) has rightly observed that Nozick's (1969) story of Newcomb's problem stresses the following: The decision maker knows that the predictor has often correctly predicted the decisions of other decision makers in Newcomb's problem, so that Newcomb's problem could be viewed as a finitely iterated Newcomb's problem with different decision makers. Furthermore, the decision maker knows that the predictor has often correctly predicted the decision maker's decisions in other decision situations in the past. But even if Newcomb's problem can be considered as a finitely iterated Newcomb's problem with different decision makers, the decision maker himself has no reason to view his decision situation as a finitely iterated Newcomb's problem. His decision situation just consists of a 1-shot Newcomb's problem. Furthermore, the decision maker who is only once confronted with Newcomb's problem has no reason to build up a reputation as a B2 taker. And even if the predictor has been reliable with regard to the decision maker in other decision situations, this doesn't include that the decision maker has had an incentive to build up a reputation during these other decision situations. For Nozick's (1969) story about Newcomb's problem doesn't contain any information about whether correct predictions were connected to payoffs for the decision maker in these other decision situations. Thus it seems to be that the decision maker views Newcomb's problem as a 1-shot Newcomb's problem and not as a finitely iterated Newcomb's problem.

---

[53]I wish to thank Wlodek Rabinowicz and Howard Sobel for discussion about the different partitions of the possible states of the world. Furthermore, I would like to thank Wlodek Rabinowicz for pointing out to me the irrelevance of the predictor's high reliability.

## 1.4 Nozick's Two Opposing, but Equally Plausible Arguments for Newcomb's Problem

According to Nozick (1969) in the decision situation of Newcomb's problem two opposing, but equally plausible arguments arise, which lead to different decisions.[54] While the argument for the decision to take B2 (= 1-box-argument) is based on the principle of maximising conditional utility, the argument for the decision to take both boxes (= 2-boxes-argument) is based on the principle of strong dominance.

### *Nozick's (1969) 1-Box-Argument*

**Premise 1**: If I take the content of both boxes at $t_3$, the predictor had predicted this with high reliability at $t_1$ and has put \$ 0 in B2 at $t_2$, so that I get \$ 1,000 with near certainty.

**Premise 2**: If I take the content of B2 at $t_3$, the predictor had predicted this with high reliability at $t_1$ and has put \$ 1,000,000 in B2 at $t_2$, so that I get \$ 1,000,000 with near certainty.

**Conclusion**: Therefore I should take the content of B2.

Assuming that the credences of the decision maker are:

---

[54] As Nozick (1969, pp. 134-135) points out other arguments come up, (a) if the decision maker believes that there is backwards causation operating, or (b) if the decision maker believes that the predictor's high predictability is due to his looking into the future. Nozick (1969), however, doesn't want the decision maker to look at Newcomb's problem in this way. The decision maker should assume that the predictor observes the decision maker, before he is confronted with Newcomb's problem, examines him with complicated techniques, etc., and uses his theoretical knowledge to predict - on the basis of the decision maker's diagnosed state - the decision maker's decision. The question of backwards causation operating in Newcomb's problem and the question of how the predictor operates will not be treated at length in this dissertation. While the latter topic will be shortly presented later in chapter 1.6, the former question has been treated in the literature in the following way: Mackie (1977) precludes backwards causation in Newcomb's problem, but doesn't give a reason why. Swain (1988) and Horgan (1981) don't want to presuppose backwards causation either. Dowe, Oakley, and Rosier (forthcoming), however, claim that if we suppose backwards causation, then the decision to take B2 in Newcomb's problem is rational. Cargile (1975) goes even further in regarding probabilistic backwards causation as coherent. Cargile (1975) regards Newcomb's problem as underdetermined, though. Anglin (1981) defends the logical possibility of backwards causation and applies backwards causation to Newcomb's problem. And Schmidt (1998) takes backwards causation to establish that a variation of Newcomb's problem and therefore also Newcomb's problem can be realised, in the sense that it is a well-defined decision problem. Locke (1978, 1979) takes a completely different position, which is objected by Gallois (1979, 1981), because he regards backwards causation as irrelevant to Newcomb's problem. (A critique of Schmidt (1998) and a review of the literature of backwards causation and Newcomb's problem can be found in Ledwig (forthcoming).)

$c(s_1|a_1) = 0.9$,
$c(s_1|a_2) = 0.1$,
$c(s_2|a_2) = 0.9$,
$c(s_2|a_1) = 0.1$,

and assuming that the decision maker's utilities of his possible outcomes grow linearly with his possible outcomes, the conditional utilities for his possible actions $a_1$ and $a_2$ are calculated as follows:

$CU(a_1)$ = (0.9)(1,000) + (0.1)(1,001,000),
= 101,000,
$CU(a_2)$ = (0.1)(0) + (0.9)(1,000,000),
= 900,000.

Because $CU(a_2)$ is greater than $CU(a_1)$, the principle of maximising conditional utility recommends to take B2.

## *Nozick's (1969) 2-Boxes-Argument*

**Premise 1**: The predictor had already made his prediction at $t_1$ and has put \$ 1,000,000 in B2 at $t_2$ or not.

**Premise 2**: Suppose the predictor has put \$ 1,000,000 in B2 at $t_2$:
If I take the content of both boxes at $t_3$, I get \$ 1,001,000.
If I take the content of B2 at $t_3$ , I get \$ 1,000,000.

**Premise 3**: Suppose the predictor has put \$ 0 in B2 at $t_2$:
If I take the content of both boxes at $t_3$, I get \$ 1,000.
If I take the content of B2 at $t_3$, I get \$ 0.

**Premise 4**: Under each supposition I get \$ 1,000 more, if I take the content of both boxes instead of taking the content of B2.

**Conclusion**: Therefore I should take the content of both boxes.

Assuming that the utility of \$ 1,000 is greater than the utility of \$ 0 for the decision maker and that the utility of \$ 1,001,000 is greater than the utility of \$ 1,000,000 for the decision maker, the possible action of taking both boxes strongly dominates the possible action of taking B2, so that the principle of strong dominance recommends to take both boxes.

Whereas the 1-box-argument relies on the fact that the decision maker believes that the predictor's prediction is probabilistically dependent on the decision maker's decision, the 2-boxes-argument relies on the fact that the decision maker believes that the predictor's prediction is causally independent of the decision maker's decision. The

latter fact just means that the decision maker's beliefs fulfil the principles of dominance's condition of application. And the principles of weak or strong dominance with probabilistic independence cannot be applied to Newcomb's problem, because the decision maker believes that the possible states of the world are probabilistically dependent on the possible actions of the decision maker, while the principle of strong dominance with causal independence can be applied, because the decision maker believes that the possible states of the world are causally independent of the possible actions of the decision maker. Thus to be more accurate than Nozick (1969) Newcomb's problem can be described as a conflict between the principle of maximising conditional utility and the principle of strong dominance with causal independence.

Nozick's analysis of Newcomb's problem made the conflict between evidential decision theory (for example Jeffrey 1965) and causal decision theory (for example Savage 1954/1972) explicit, and enhanced a development of evidential decision theories (for example Jeffrey 1983; Eells 1981), causal decision theories (for example Gibbard and Harper 1978; Lewis 1981a; Skyrms 1982; Spohn 1978; Sobel 1986), and other proposals (for example Nozick 1993; Kyburg 1980).

## 1.5 Nozick's Intuitions Backing Up his Two Arguments

Nozick (1969) backs up his two arguments by a number of intuitions[55]. First, I will present two intuitions behind the 1-box-argument which I shall call the intuition of indifference and the betting intuition, and then I will present two intuitions behind the 2-boxes-argument which I shall call the time intuition and the intuition of the well-wishing friend[56]:

### *The Intuition of Indifference for the 1-Box-Argument*

**Premise 1**: Many persons were confronted with Newcomb’s problem.

---

[55]I wish to thank Michael Wolff for drawing my attention to the intuitions behind Newcomb's problem.

[56]Nozick (1969) supports his 1-box-argument and his 2-boxes-argument by overall considerations. For he doesn't distinguish between the intuition of indifference and the betting intuition and between the time intuition and the intuition of the well-wishing friend. Yet I think it is better to distinguish between these intuitions, so that each intuition can be checked for its adequacy, and so that relevant factors for a solution to Newcomb's problem can be distinguished from irrelevant factors.

**Premise 2**: While the persons who had taken B2 won $ 1,000,000, the persons who had taken both boxes won $ 1,000.

**Premise 3**: There is no reason that I differ from these persons with regard to predictability.

**Conclusion**: Therefore I should take B2.

This intuition can be criticised. For premise 2 is false with regard to Newcomb's problem. To be true, premise 2 should say the following: While most of the persons who had taken B2 won $ 1,000,000, most of the persons who had taken both boxes won $ 1,000. Premise 3 can be criticised, too. For the decision maker could have good reasons to believe that he belongs to the minority of persons where the predictor fails with regard to his predictions. The decision maker, for example, has observed that he and the minority of persons where the predictor fails have a special character trait which the majority of persons where the predictor is successful lacks. Furthermore, there are no other significant differences between him and the minority of persons where the predictor fails on the one side and the majority of persons where the predictor is successful on the other side. Thus it would be best for him to take both boxes.

Eells (1985) makes a similar point. For he distinguishes between type-*A* beliefs and type-*B* beliefs. While type-*A* beliefs are beliefs of the decision maker about a randomly chosen rational decision maker, type-*B* beliefs are beliefs of the decision maker about himself. Eells (1985) claims that the decision maker should use type-*B* beliefs and not type-*A* beliefs for calculating his conditional utility à la Jeffrey (1965). Eells' (1985) theory will be discussed in chapter 2.6. Kyburg (1980) distinguishes in a similar way between stochastic probability and properly epistemic probability. While stochastic probability refers to the probability of a possible action to take B2 respectively to take both boxes in Newcomb's problem, properly epistemic probability refers to the probability of his possible action to take B2 or to take both boxes. Kyburg (1980) claims that the decision maker should use properly epistemic probabilities for calculating his utility. I will deal with Kyburg's proposal in chapter 4.4.

## *The Betting Intuition for the 1-Box-Argument*

This intuition comes in two versions:

**Version 1:**

**Premise**: I know that a third person shares the indifference intuition with me.

**Conclusion**: Therefore I know that he is rationally justified to bet with me, giving high odds, that I get $ 1,000, if I take the content of both boxes.

**Version 2:**

**Premise 1**: I have the indifference intuition.

**Premise 2**: I take both boxes, and the announcement of my possible outcome is delayed.

**Conclusion**: Therefore it is rational for me to bet with a third person, giving high odds, that I get $ 1,000.

In both versions it is of no relevance whether there is a betting partner, for the decision maker just has to ask himself whether he wants to take both boxes and therefore wants to decide against a rationally justified bet.

The betting intuition can be criticised in both versions. For premise 1 in both versions relies on the truth of the indifference intuition which is problematical as we have already seen.

## *The Time Intuition for the 2-Boxes-Argument*

**Premise 1**: The predictor had already made his prediction at $t_1$, has put $ 1,000,000 in B2 at $t_2$ or not and has left the scene.

**Premise 2**: This has happened a week ago, a year ago, etc.

**Premise 3**: B1 is transparent, so that I see $ 1,000 in B1.

**Premise 4**: $ 1,000,000 have already been in B2 for a week, for a year, etc. or not, although I can't see what is the case.

**Premise 5**: It doesn't matter whether I see the content of B2 or not, for the content of B2 has been fixed a week ago, a year ago, etc. and will remain the same.

**Conclusion**: Therefore I should take the content of both boxes.

This intuition opens up a very interesting area of investigation. For the question arises what role does the time factor play in Newcomb's problem. Furthermore, it can be asked whether the time factor is necessary and sufficient for establishing the causal independence of the predictor's prediction from the decision maker's decision in Newcomb's problem, or whether the causal independence can be established by other means.[57] I think that the causal independence of the predictor's prediction from the decision maker's decision can be established by other means than the time factor. For it is possible that the predictor makes his prediction at the same time (or even at a later time) as the decision maker makes his decision with a delayed opening of the respective boxes, and that the predictor's prediction isn't causally influenced by the decision

[57] I wish to thank Alfred Schramm for drawing my attention to this question.

maker's decision. Yet the best way to strengthen the decision maker's belief in the causal independence of the predictor's prediction from the decision maker's decision is by letting the predictor's prediction happen before the decision maker's decision in time.[58]

## *The Intuition of the Well-Wishing Friend for the 2-Boxes-Argument*

**Premise 1**: The predictor has put $ 1,000,000 in B2 a week ago, a year ago, etc. or not.

**Premise 2**: While I can't see the content of B2, for the box is non-transparent on my side, a well-wishing friend of mine has been sitting on the other transparent side of the box, since the predictor has put $ 1,000,000 in B2 or not.

**Conclusion 1**: Therefore my well-wishing friend has been looking at $ 1,000,000 in B2 for a week, for a year, etc. or not.

**Premise 3**: If the money is there, it will not suddenly disappear, and if the money isn't there, it will not suddenly appear; both cases being independent of what my decision looks like.

**Premise 4**: My well-wishing friend hopes, if $ 1,000,000 are in B2, and if $ 1,000,000 aren't in B2, that I take both boxes.

**Premise 5**: I know the hopes of my well-wishing friend, for I can take his perspective.

**Conclusion 2**: Therefore I should take both boxes.

**Premise 6**: In principle the presence of my well-wishing friend is of no importance, for I know standing alone in front of the boxes that he would hope that I would take both boxes.

**Conclusion 3**: Therefore I should take both boxes.

Also this intuition can be criticised.[59] Segerberg (1993, p. 276) points out that the decision maker cannot take the third-person perspective, which is the normal perspective of any observer, while making a decision. With regard to the intuition of the well-wishing friend this means that premise 5 and premise 6 have to be rejected, so that conclusion 2 and conclusion 3 have to be rejected, too.[60] Segerberg (1993, p. 276) justifies his view: If the decision maker tries to decide whether to do $a_1$, he could reason in either of the following two ways, which Segerberg considers as strange: (1) The probability of my doing $a_1$ is 0.95, therefore I will do $a_1$. (2) The probability of my

[58] An exception to this rule are decision makers who believe in backwards causation.

[59] I would like to thank Rüdiger Bittner and Andreas Blank for discussion of this point.

[60] Segerberg (1993) doesn't apply his claim to the intuition of the well-wishing friend, though.

doing $a_1$ is 0.95, therefore I will not do $a_1$. The only explanation that Segerberg (1993) gives with regard to this point is that the oddity stems from the decision maker taking the third-person perspective. The difference between the two perspectives is that from the third-person perspective the decision maker is seen from the outside, that is as part of the possible states of the world, so Segerberg (1993, p. 264), whereas from the first-person perspective the decision maker is seen from the inside. Unfortunately Segerberg doesn't give any more arguments for his position; yet he seems to imply that the decision maker cannot deliberate, if he views himself as part of the possible states of the world which I regard as reasonable.[61]

I am not sure whether the decision maker always deliberates from a first-person perspective. For why shouldn't the decision maker reason as follows: I have a friend who is very rational. If he were in my position, he would do $a_i$. Because I also want to act rationally, I take his perspective while deliberating and do $a_i$. One could object to that in the following way: What the decision maker does, is to take over the perspective of his friend and not to take the perspective of his friend. For the decision maker to take over a perspective means to make it a part of his own perspective, whereas for the decision maker to take the perspective of his friend means to change from his own perspective to the perspective of his friend. Thus Segerberg seems to be right after all.

## 1.6 Variations of Newcomb's Problem

Nozick (1969, 1993) considers variations of Newcomb's problem for the following reasons: (1) He wants to find out which factors besides the conditional utility argument and the dominance argument are relevant for creating Newcomb's problem and whether these factors have influence on the force of these arguments (Nozick 1969, pp. 140-141). (2) He wants to show that decision makers don't have complete confidence in the decision principle which they favour in Newcomb's problem (Nozick 1993, pp. 44-45). Spohn (1978, p. 179) puts it in slightly different words: That both decisions are prima facie equally plausible and that decision makers can be put in a predicament with regard to their decisions depends on two factors, namely that the predictor is very reliable and that the amount of money in B1 is small, but not negligible

[61] In the end Segerberg (1993, p. 276) adds that the distinction between the first- and the third-person perspective can explain why the high reliability of the predictor is irrelevant for providing a solution to Newcomb's problem without giving any further argument. Perhaps Segerberg thinks that it is useless for the decision maker to take the perspective of the predictor, because by doing so the decision maker takes a third-person perspective.

in comparison to the possible high amount of money in B2. Yet in the end Spohn (1978, p. 182) criticises Nozick (1969) for not providing a unifying solution to Newcomb's problem and its variations.[62]

On first sight the following factors, which can also be combined, are of relevance for creating Newcomb's problem: (1) Different grades of predictability, (2) different amounts of money in B1 and B2, (3) different numbers of iterations with the same decision maker, and (4) further changes of conditions in Newcomb's problem. I will deal with each of them in turn.

## *Different Grades of Predictability*

According to Nozick (1969) one obtains variations of Newcomb's problem, if the decision maker assigns different grades of predictability to the predictor. Nozick (1969) considers the following two cases:

(1) The decision maker assigns a predictability of $c = 0.6$ to the predictor;

(2) the decision maker assigns a predictability of $c = 1$ to the predictor.

In the first case Nozick (1969) supposes that every decision maker will take both boxes, although the conditional utility for taking B2 is greater than the conditional utility for taking both boxes. Nozick (1969) doesn't give a justification for his supposition, but it can be easily provided. For if the decision maker assumes that the predictability of the predictor is slightly better than chance, the decision maker also assumes that the predictor is very likely to make a wrong prediction, so that it is irrational to rely on his predictability for decision-making. This case shows that the principle of maximising conditional utility doesn't hold in general, for it gives irrational recommendations, if the decision maker assigns a predictability of $c = 0.6$ to the predictor. Furthermore, it shows that the principle of maximising conditional utility isn't sufficient for generating a forceful argument for the 1-box-solution. Crucial for yielding such an argument is that the decision maker assigns a high or even a 100% predictability to the predictor and that he applies the principle of maximising conditional utility. For Newcomb's problem to arise as a conflict between the principle of

[62]Spohn can be criticised for not providing a unifying solution to Newcomb's problem and its variations, too. For Spohn (in press) demands that in the finitely iterated Newcomb's problem the decision maker should take B2, whereas in the 1-shot Newcomb's problem the decision maker should take both boxes. If one, however, conceives the finitely iterated Newcomb's problem as a game with two decision makers and the 1-shot Newcomb's problem as a game against nature, so that one actually has two different games at hand, the criticism against Spohn loses its substance.

maximising conditional utility and the principle of strong dominance the decision maker has to assign a high, but not a 100% predictability to the predictor as can be seen by the second case, so Nozick (1969).

In the second case the decision maker argues as follows, so Nozick (1969):

**Premise 1**: I know that, if I take both boxes, I will get $ 1,000.

**Premise 2**: I know that, if I take B2, I will get $ 1,000,000.

**Conclusion**: Therefore I should take B2.

By arguing in this way Nozick (1969) claims that in the extreme case in which the decision maker assigns a predictability of $c = 1$ to the predictor the principle of strong dominance doesn't work anymore, so that for the principle of dominance to function, one needs that the predictability assignment of the decision maker is very high, but not 100%. Causal decision theorists (for example Gibbard and Harper 1978; Sobel 1988a), however, claim that the 2-boxes-argument is sound in the case of a 100% predictability assignment of the decision maker, which I will deal with in chapter 3.2.

The 100% predictability assignment of the decision maker led to another interesting area of investigation. Schlesinger (1974, 1976, 1977a, 1977b) interpreted the 100% predictability assignment of the decision maker as infallibility of the predictor and tried to derive a proof of human freedom of will out of Newcomb's problem, which led to many objections (for example Ben-Menahem 1986; Benditt and Ross 1976; Cargile 1975; Gallois 1979; Hudson 1979; Ledwig 1997; Locke 1978; Pothast 1980; Weintraub 1995). To a large part independently from Schlesinger's proof and its criticism Newcomb's problem raised anew the question of compatibility between determinism and human freedom of will (for example Fischer 1994; Gallois 1981; Leslie 1991; Locke 1979; Mackie 1977) and the question of compatibility between omniscience and human freedom of will (for example Craig 1987; Factor 1978; Horne 1983). Although the proof of human freedom of will and the compatibility questions are of utmost importance, their connections to Newcomb's problem are not so important. For it is very improbable that Newcomb's problem accomplishes which better candidates have not accomplished, namely to prove human freedom of will and to solve the compatibility questions.[63] Furthermore, Schlesinger's proof, its criticism, and the compatibility questions are only marginally relevant for solving Newcomb's problem. I have treated Schlesinger's proof and its criticism elsewhere (Ledwig 1997) and will not come back to these issues and the compatibility questions here.

---

[63] I wish to thank Eike von Savigny for an illuminating remark in this direction.

## *Different Amounts of Money in B1 and B2*

Not only by changing the predictability assignment of the decision maker one obtains variations of Newcomb's problem, but also by changing the amount of money in both boxes. Nozick (1993) considers the following two cases:

(1) B1 contains $ 1;

(2) B1 contains $ 900,000.

In the first case Nozick (1993) doubts whether the decision makers who follow the principle of strong dominance still want to adhere to it; but he doesn't justify his doubt. A justification, however, can be easily provided, for the amount of money the decision maker could gain by following the principle of strong dominance would be just $ 1 in the case there is $ 1,000,000 in B2 and in the case there is $ 0 in B2. In my opinion the decision maker still could adhere to the principle of strong dominance, for he/she could argue in the following way: My utility of $ 1,000,000 equals my utility of $ 1,000,001, and my utility of $ 0 equals my utility of $ 1, so that I cannot apply the principle of strong dominance, and therefore don't have to reject the principle of strong dominance.

In the second case Nozick (1993) questions whether the decision maker who follows the principle of maximising conditional utility still wants to adhere to it. He doesn't give an argument for this claim, though. Nozick (1993) seems to suggest that the decision makers are unwilling to change their decision from taking B2 in the original version of Newcomb's problem to taking both boxes in this variation of the problem. Actually I don't see why the decision makers should be unwilling to change their decision from taking B2 to taking both boxes, if the decision situation changes so drastically. Furthermore, the principle of maximising conditional utility recommends this change. Therefore I doubt Nozick's claim that the decision makers who follow the principle of maximising conditional utility give it up, if $ 1,000 are substituted by $ 900,000 in B1. Moreover, I question Spohn's (1978) position that decision makers can be put in a predicament with regard to their decisions in this case. For the decision makers could argue that they didn't make any commitments with regard to their decisions, if the decision situation changes. As a consequence the conflict between the principle of strong dominance and the principle of maximising conditional utility would disappear, for both principles would recommend to take both boxes in this case.

From both cases we can draw the following conclusions: The amount of money in B1 besides the conditional utility argument and the dominance argument is relevant

for creating Newcomb's problem as a conflict between these arguments; in particular Spohn (1978) is right in claiming that the amount of money in B1 has to be small, but not negligible in comparison to the possible high amount of money in B2 to make both decisions prima facie equally plausible. Yet on second sight causal decision theorists (for example Spohn 1978; Sobel 1988a) propose that the amount of money in B1 doesn't matter in making the decision to take both boxes rational.

## *Different Numbers of Iterations with the Same Decision Maker*

Other variations of Newcomb's problem are obtained, if the situation is repeated *n* times with the same decision maker where *n* can range from 1 to infinity. I will not consider infinitely repeated Newcomb's problems with the same decision maker, for decision makers, that is human beings, can only make a finite number of decisions. Sorensen (1985), for example, considers finitely iterated Newcomb's problems with the same decision maker and finitely iterated prisoner's dilemmas with the same two prisoners.[64] He claims that these finitely iterated Newcomb's problems and prisoner's dilemmas are paradoxical, because the principle of strong dominance opposes other principles, and because the decision makers are considered as epistemically static and not as epistemically dynamic.

Sorensen's (1985) solution to the finitely iterated Newcomb's problems with the same decision maker and to the finitely iterated prisoner's dilemmas with the same two prisoners is obtained by focusing on the role of reputation building, that is by viewing the decision makers as epistemically dynamic. In a finitely iterated Newcomb's problem with the same decision maker Sorensen (1985) claims that the predictor's last prediction is affected by his opinion of the decision maker's decision-making tendencies. Thus in Newcomb's problem played two times with the same decision maker, the decision maker may decide to take B2 in the first play to influence the predictor's last prediction. And the more iterations Newcomb's problem with the same decision maker consists of, the more the decision maker can build up a reputation. According to Sorensen (1985) the decision maker should not build up a reputation as a both boxes taker, but as a B2

[64]Sorensen (1985) just says that he considers finitely iterated Newcomb's problems and prisoner's dilemmas; he doesn't distinguish between finitely iterated Newcomb's problems with the same decision maker and finitely iterated Newcomb's problems with different decision makers. The same holds for finitely iterated prisoner's dilemmas. Yet the only way to make sense of the reputation building behaviour is by supposing that Sorensen considers finitely iterated Newcomb's problems with the same decision maker and finitely iterated prisoner's dilemmas with the same two prisoners.

taker. For in the former case the predictor will be influenced to put $ 0 in B2, whereas in the latter case the predictor will be influenced to put $ 1,000,000 in B2.

This variation of Newcomb's problem shows that besides the conditional utility argument and the dominance argument the number of iterations of Newcomb's problem with the same decision maker is relevant for creating Newcomb's problem. The more Newcomb's problem is iterated with the same decision maker, the more the decision maker can try to build up a reputation and thereby can try to influence the predictor. Yet in the 1-shot Newcomb's problem the decision maker isn't able to build up a reputation and therefore cannot influence the predictor. To factor out the decision maker's reputation building tendencies in providing a solution to Newcomb's problem, I will just consider a 1-shot Newcomb's problem in this dissertation.

## *Further Changes of Conditions in Newcomb's Problem*

Further variations of Newcomb's problem are obtained, if new conditions are added, for example, a third possible action is added (Skyrms 1984, p. 67), or if old conditions are changed, for example, the predictor works by seeing into the future (Nozick 1969, pp. 134-135), or, for example, the predictor is substituted by a computer (Horne 1983, p. 220).

Skyrms (1984) varies Newcomb's problem: Suppose B1 and B2 are made of glass, but B2 is covered with a black velvet cloth. Because the decision maker is undecided, the predictor offers him a third possible action, namely that he can look under the cloth before he decides. Nevertheless the predictor remains very reliable. According to Skyrms the decision maker should argue as follows in this variation of Newcomb's problem: If I look under the cloth, then I will take both boxes. For I prefer $ 1,001,000 to $ 1,000,000 and $ 1,000 to $ 0.

Nozick (1969) shortly considers a variation of Newcomb's problem: The predictor works by seeing into the future, so that he actually sees what the decision maker will do. If the predictor sees that the decision maker will take B2, the predictor puts $ 1,000,000 in B2. And if the predictor sees that the decision maker will take both boxes, the predictor puts $ 0 in B2. Therefore the decision maker should take B2. This variation of Newcomb's problem shows that besides the conditional utility argument and the dominance argument the way the predictor works is relevant for creating Newcomb's problem. Yet Nozick (1969) doesn't want the predictor to work like this.

Horne (1983) poses the following variation of Newcomb's problem: The predictor is substituted by a diagnostic machine. Thus if the machine had diagnosed that

the decision maker will take B2, the machine has put $ 1,000,000 in B2; and if the machine had diagnosed that the decision maker will take both boxes, the machine has put $ 0 in B2. Suppose the machine had already made his prediction and has put $ 1,000,000 in B2, then the decision maker argues as follows: If I take the content of both boxes, I get $ 1,001,000. If I take the content of B2, I get $ 1,000,000. And suppose the machine had already made his prediction and has put $ 0 in B2, then the decision maker argues as follows: If I take the content of both boxes, I get $ 1,000. If I take the content of B2, I get $ 0. Under each supposition the decision maker gets $ 1,000 more, if he takes the content of both boxes instead of taking the content of B2. Therefore he should take the content of both boxes.

Yet I will not consider these changes of conditions in Newcomb's problem with regard to their influence on the conditional utility argument and the dominance argument, for this leads us too far away from the original version of Newcomb's problem.

## *Different Grades of Predictability and Different Amounts of Money in B1 and B2*

One gets variations of Newcomb's problem by combining different values of the following three factors: The predictability assignment of the decision maker, the amount of money in B1, and the amount of money in B2. Sobel (1988a), for instance, considers a case, in which the decision maker assigns a predictability of $c = 1$ to the predictor[65], in which $ 1,000,000 are in B1, and in which the amount of money in B2 and its conditions are unchanged. According to Sobel (1988a) the decision maker has to argue in the following way:

**Premise 1**: I have a chance to get $ 2,000,000, if I decide to take both boxes.

**Premise 2**: I have nothing to lose, if I decide to take both boxes.

**Conclusion**: Therefore I should take both boxes.

Sobel (1988a) continues his argumentation by changing the amount of money in B1 from $ 1,000,000 to $ 999,999 and 99 cents. Sobel (1988a) doubts that the decision maker continues arguing in the following way:

**Premise**: In this case I lose a cent.

---

[65]Sobel (1988a) claims that the decision maker's predictability assignment of $c = 1$ to the predictor can be interpreted in two ways: (1) The predictor is unerring, so that he factually makes no wrong predictions. (2) The predictor isn't able to make wrong predictions.

**Conclusion**: Therefore I should take B2 in all cases in which I lose something and should take both boxes in all cases in which I don't lose anything.

For according to Sobel (1988a) no decision maker wants to argue that a monetary difference makes a difference for a rational decision. Therefore a difference of $ 999,000, so that $ 1,000 are in B1, shouldn't make a difference either. If, however, you want that a monetary difference makes a difference for a rational decision, you would have to justify it adequately, although Sobel doesn't see how this could be done.

In my opinion, however, a monetary difference can make a difference for a rational decision. For one can imagine situations in which one cannot afford to lose a cent. Suppose you bought a house in the most expensive area of Constance. It cost you $ 2,000,000. Luckily you saved up $ 1,000,000, so that you can already pay half of your debt; but unfortunately the person who has sold you the house has a mean character. He wants you to pay the rest of the debt at the end of the month; if you cannot fulfil his demands till then, he will kill you. But fortunately a generous being offers you Sobel's variation of Newcomb's problem except that $ 999,999 and 99 cents are in B1 instead of $ 1,000,000, and you have no other money left; even your friends cannot give you anything, for they are broke themselves. In this case I have to ask you: Do you really want to risk your life for the minute possibility[66] to get almost $ 2,000,000 by taking both boxes?

This seems to suggest that the utility of a possible action is determined by other factors than the utilities of the possible outcomes as well. Gärdenfors and Sahlin (1988) argue in this direction. They claim that the utility of a possible action doesn't only depend on the utilities of the possible outcomes, but also on other factors like the decision maker's attitude towards risk (Arrow 1971; Pratt 1964) and the amount of possible change from the decision maker's reference point (Kahneman and Tversky 1988). Gärdenfors and Sahlin (1988) report that in the economic literature a decision maker is risk averse, if he prefers the monetary utility of a gamble to the gamble itself, and a decision maker is risk prone, if he has the opposite preferences. If you are highly risk averse in this variation of Newcomb's problem, if the utilities of the possible actions are determined by the utilities of the possible outcomes and by the degree of risk

[66]The decision maker's predictability assignment of $c = 1$ can be interpreted in different ways (cf. Hubin and Ross 1985; Ledwig 1997; Locke 1978; Sobel 1988a; and chapter 3.2), so that the decision maker, for example, can attribute the predictability of $c = 1$ to chance which opens up the possibility to win almost $ 2,000,000 by taking both boxes.

aversion, and if you follow the principle of maximising conditional utility, then you should take B2. Yet the question remains whether it is rational to be risk averse or risk prone.

Kahneman and Tversky (1988, p. 199) explain the concept of a reference point:

> "When we respond to attributes such as brightness, loudness, or temperature, the past and present context of experience defines an adaptation level, or reference point, and stimuli are perceived in relation to this reference point ... . The same principle applies to non-sensory attributes such as health, prestige, and wealth. The same level of wealth ... may imply abject poverty for one person and great riches for another - depending on their current assets."

Furthermore, they claim that the relationship between the utility of a possible action and the reference point is as follows (Kahneman and Tversky 1988, p. 200):

> "Strictly speaking, value should be treated as a function in two arguments: the asset position that serves as a reference point, and the magnitude of the change (positive or negative) from that reference point."

To make it more precise the utility of a possible action is determined by the possible changes from the reference point rather than by the utility of the possible outcomes. By means of the above given variation of Newcomb's problem I will shortly illustrate the effect of different reference points: If you are in debt by $ 1,000,000, your reference point is - $ 1,000,000, and if you take part in the above mentioned variation of Newcomb's problem, your possible positive changes from the reference point are + $ 999,999 and 99 cents (in case you take both boxes, and the predictor has predicted this, and he has put $ 0 in B2), + $ 1,999,999 and 99 cents, (in case you take both boxes, and the predictor has predicted that you take B2 only, and he has put $ 1,000,000 in B2) and + $ 1,000,000 (in case you take B2, and the predictor has predicted this, and he has put $ 1,000,000 in B2). If you already own $ 2,000,000, your reference point is + $ 2,000,000, and if you take part in this variation of Newcomb's problem, your possible changes from the reference point are as in the previous case. Whereas in the first case all three possible positive changes are very valuable for you, because a possible positive change of + $ 999,999 and 99 cents almost evens your debt, a possible positive change of + $ 1,999,999 and 99 cents makes you almost a millionaire, and a possible positive change of + $ 1,000,000 evens your debt, in the second case, however, all three possible positive changes are not so valuable for you, because a possible positive change of + $ 999,999 and 99 cents is just almost a half of what you already own, a possible positive

change of + $ 1,999,999 is almost all of what you already own, and a possible positive change of + $ 1,000,000 is just a half of what you already own.

This variation of Newcomb's problem shows that besides the conditional utility argument and the dominance argument the decision maker's attitude towards risk (Arrow 1971; Pratt 1964) and the amount of possible change from the decision maker's reference point (Kahneman and Tversky 1988) are relevant for creating Newcomb's problem. With regard to the decision maker's attitude towards risk I claim the following: If the decision maker's deliberation is dominated by his risk averseness in Newcomb's problem, he will take B2. For the predictor's high reliability almost guarantees that there is $ 1,000,000 in B2, if he takes B2, and that there is $ 0 in B2, if he takes both boxes. Furthermore, if the decision maker is very risk averse, he doesn't want to lose the guarantee to win $ 1,000,000 by taking both boxes. Therefore he takes B2. If the decision maker's deliberation is dominated by his risk proneness in Newcomb's problem, he will take both boxes. For if the decision maker is very risk prone, he doesn't care whether he loses the guarantee to win $ 1,000,000 by taking both boxes. He prefers the chance to win more, namely $ 1,001,000, by taking both boxes. Therefore he takes both boxes. To factor out the decision maker's attitude towards risk in dealing with Newcomb's problem, we have to suppose that the decision maker's attitude towards risk is located on the middle of the scale going from risk averseness to risk proneness.

With regard to the amount of possible change from the decision maker's reference point I maintain the following position: If in Newcomb's problem the decision maker's deliberation is dominated by the amount of possible change from his reference point, which is, for example, - $ 1,000,000, then he takes B2. For if the decision maker is very poor, he cares for thousands, but even more so he cares for millions. Thus it matters more to him to gain millions than to gain thousands. Because the predictor's high reliability almost guarantees that there is $ 1,000,000 in B2, if he takes B2, and that there is $ 0 in B2, if he takes both boxes, and because it matters more to the decision maker to gain a million than to gain nothing, he wants to make sure that he gains a million. Therefore he takes B2. One can object to that, though. For the decision maker's main aim could also be to prevent the possibility of getting only $ 0, which leads to a 2-boxes-solution. If in Newcomb's problem the decision maker's deliberation is dominated by the amount of possible change from his reference point, which is, for example, + $ 1,000,000, then he takes both boxes. For if the decision maker is very rich, he doesn't care for thousands, but even more so he doesn't care for millions. Thus it doesn't matter more to him to gain millions than to gain thousands. Because the predictor's high

reliability almost guarantees that there is $ 1,000,000 in B2, if he takes B2, and that there is $ 0 in B2, if he takes both boxes, and because it doesn't matter more to the decision maker to gain a million than to gain nothing, he doesn't want to make sure that he gains a million. Therefore for him a 1-box-solution has no advantage over a 2-box-solution. To factor out the amount of possible change from the decision maker's reference point, we have to suppose that the decision maker is neither very poor nor very rich, but is located in the middle of the scale going from being poor to being rich.

The decision maker's attitude towards risk and the amount of possible change from the decision maker's reference point touch upon the tasks of rational decision theory (cf. chapter 1.2, in the section on the matrix formulation of Newcomb's problem). For the second task, namely which possible action the decision maker should decide for, includes the task to find out which probabilities and utilities should be used in calculating the utility of a possible action. This latter task, so Jeffrey (1996), belongs to the area of substantive rationality in opposition to the area of structural rationality.[67] Jeffrey (1996, p. 4) explains the difference between substantive and structural rationality:

> "Bayesian decision theory is said to represent a certain structural concept of rationality. This is contrasted with substantive criteria of rationality (Kahneman 1996) having to do with the aptness of particular probability and utility functions to particular predicaments. ... What remains when all substantive questions of rationality are set aside is bare logic, ... . A complete set of substantive judgements would be represented by a Bayesian *frame*, i. e., a probability measure *pr* defined on a Boolean algebra of subsets of a space Ω (the 'propositions'), a utility function *u* defined on Ω, and an assignment of elements of the Boolean algebra as values to '*A*', '*B*' etc. ... . In a Bayesian logic of decision, Bayesian frames represent all possible answers to substantive questions of rationality; ... . On this view, consistency - bare structural rationality - is simply representability in the Bayesian framework."

Unfortunately neither Jeffrey (1996), nor Albert (1998) and Kahneman (1996) give a more detailed account of what substantive rationality and structural rationality is. My impression is that the difference between substantive rationality and structural rationality consists in the following: While substantive rationality deals with the question

[67] Having the same difference in mind Albert (1998, pp. 29-34) distinguishes between substantive and formal rationality and Kahneman (1996, p. 203) between substantive and logical rationality.

whether the contents of the decision maker's wants, beliefs, and decisions are rational, structural rationality deals with the question whether the decision maker's wants, beliefs, and decisions are consistent with each other. With regard to Newcomb's problem both Nozick's (1969) 1-box-argument and his 2-boxes-argument seem to be consistent, yet the question remains whether each argument is substantively rational.

## 1.7 Newcomblike Problems[68]

### *Criteria for Newcomblike Problems*

Newcomb's problem belongs to the class of Newcomblike[69] problems which according to Nozick (1969, p. 132) can be identified as such, if they fulfil the following two criteria: (1) A dominant possible action exists; (2) the possible actions have no causal influence on the possible states of the world, and the probabilities of the possible states of the world given the possible actions differ per action from each other (cf. also Spohn 1978, p. 181). In my opinion these two criteria just reflect the principle of dominance and its condition of application, namely that the decision maker believes that the possible states of the world are causally independent of the possible actions of the decision maker and that the decision maker believes that the possible states of the world are probabilistically dependent on the possible actions of the decision maker. To judge whether these two criteria are sufficient let's have a look at another classification.

According to Sobel (1990, p. 225) Newcomblike problems contain the following four elements:

> "*Element A* In a Newcomblike problem the agent would be sure that certain features relevant to the values of outcomes of his actions, for example, whether or not there is *$M* in the second box (Newcomb's Problem), or whether or not the other prisoner is going to confess (Prisoner's Dilemma), are *causally independent* of his possible actions and are things he can in no way influence.

---

[68] For statements of some Newcomblike problems see the appendix of chapter 1.

[69] Nozick (1969) doesn't name them "Newcomblike problems". As far as I know the term stems from Sobel (1990). This name is justified, because in the literature on Newcomb's problem its solution is almost always motivated by trying to show that Newcomb's problem belongs to a class of other problems which can be solved easily in decision-theoretic terms. In the literature this class of other problems is termed "Newcomblike problems" (Sobel 1990), "Newcomb situations" (Eells 1985, p. 189), and "Newcomb problems" (Eells 1985, p. 188; Lewis 1979a).

> *Element B.* These features would be for him *epistemically* or *evidentially dependent* on his actions, so that news of these actions would provide him with signs, with evidence for and against these features.
>
> *Element C.* It is maintained, largely on the basis of Element A, that a certain action would be uniquely choiceworthy and rational.
>
> *Element D.* It is maintained, largely on the basis of Element B, that the evidential desirability of this action is exceeded by that of some other action that is open to the agent. Given Element C, this entails that the choiceworthy action in the case is *not* the action news of which would be most welcome."[70]

Sobel (1990) claims that a lot of different problems belong to the class of Newcomblike problems (cf. the appendix of chapter 1) and that each of them is a coherent challenge of evidential decision theories.[71] According to Sobel (1990, p. 225) the variety of Newcomblike problems can be explained in the following way:

> "Problems vary in the grounds provided in them for these elements. They vary in how they purport to secure the dependencies and independencies in Elements A and B, and in the character of explicit or implicit arguments for Elements C and D. Important differences among Newcomblike problems relate mainly to Elements B and C, but there is variety in the other elements as well."

Sobel (1990) makes these short remarks more precise:

> Grounds for the causal independence may consist in the following:

The decision maker believes that the feature which is causally independent of his possible actions is already fixed at decision time, because this feature relates to the past (for example, to a past prediction) or it relates to an already determined future (for example, to the development of a disease whose causes are already in place).[72]

---

[70]Unfortunately Sobel (1990) doesn't explain what "largely" in Element C and Element D means. He doesn't say which other factors and to what extent other factors are relevant for recommending a particular possible action as rational. Nozick's (1969) classification is much clearer in this respect.

[71]In opposition to Nozick (1969) Sobel's (1990) classification entails that every variation of Newcomb's problem must be considered as a Newcomblike problem, too. This is an advantage of Sobel's classification, for it is a simplification of cases.

[72]Unfortunately Sobel (1990) doesn't explain what "the feature" is. Furthermore, Sobel (1990) doesn't explain how the feature relates to the past, or how it relates to an already determined future. On p. 226 Sobel (1990) gives an example for the feature, namely that there is $ 1,000,000 in B2 in Newcomb's problem.

Grounds for the epistemic dependence may be:

(1) Prediction cases: The decision maker believes that the feature (for example, the feature that there is $ 1,000,000 in B2 in Newcomb's problem) depends on the predictor's prediction, so that news of a possible action would be a sign that it had been predicted and that the feature was present or not.

(2) Causal cases:

(a) The decision maker believes that there is one cause for the feature and one cause for his possible actions and that both causes are either present or absent. For example, in Nozick's (1969, p. 125) 2-possible-fathers-1-son-case the son believes it is possible that he has inherited a gene for a disease from his father who also carries a gene for intellectualism.

(b) The decision maker believes that there is a common cause for the feature and his possible actions. For example, in Fisher's problem (for example Jeffrey 1981, p. 476) a gene independently causes smoking and cancer.

(c) The decision maker believes that the feature causes his possible actions. For example, in the popcorn problem (Sobel 1986, pp. 411-413) a signal flashing on the screen when and only when there is popcorn registers subliminally and causes the decision maker to go for popcorn.

(3) Similar decision processes: If the feature is a possible action, the decision maker could believe that similar processes lead to the other possible action. For example, in the prisoner's dilemma (Jeffrey 1983, p. 15) the prisoner could believe that the other prisoner will reason as he does, even if the prisoner is in doubt about his own reasoning and decision.

(4) Signs of character: If the feature is a possible aspect of the decision maker's character (for example, charisma in Solomon's problem, Gibbard and Harper 1978, pp. 135-136), his possible action could be a sign and not a cause for possible states of the world or possible outcomes.

Yet cases 1, 3, and 4 can be reduced to the causal cases by means of Reichenbach's (1956, p. 157) common cause principle which can take the following simplified form[73]: "*If an improbable coincidence has occurred, there must exist a common cause*." For in case 1 the improbable coincidence that the decision makers

[73]I wish to thank Wolfgang Spohn for drawing my attention to the fact that cases 1, 3, and 4 can be reduced to the causal cases by means of Reichenbach's (1956) common cause principle. Reichenbach (1956, p. 163) also states a more complicated version of the common cause principle which I don't need for present purposes.

take B2 respectively take both boxes and that $ 1,000,000 are in B2 respectively $ 0 are in B2 in Newcomb's problem can be explained by a common cause, namely, for example, by the decision makers' non-greediness or greediness. This non-greediness or greediness causes on the one hand the predictions to take B2 respectively the predictions to take both boxes which cause the predictors to put $ 1,000,000 in B2 respectively to put $ 0 in B2; on the other hand this non-greediness or greediness causes the decision makers to take B2 respectively to take both boxes. In case 3 the improbable coincidence that both prisoner confess respectively don't confess in the prisoner's dilemma can be explained by a common cause, namely by a similar decision process. And in case 4 the improbable coincidence that just behaviour and unsuccessful revolts occur repeatedly together in Solomon's problem can be explained by a common cause, namely by the kings' charisma. Therefore the epistemic dependence can be explained in causal terms and can be reduced to causal cases.

With regard to the arguments for elements C and D Sobel (1990, p. 227) states:

> "*Arguments for the Choiceworthinesses of Bad News Actions*. Choiceworthiness of bad news actions are made plausible in the best-known Newcomblike problems by dominance arguments, but this device is not a feature of every Newcomblike problem. Some use the more widely applicable common sense that an action is choiceworthy if it would probably have the best consequences."

Unfortunately Sobel doesn't go into much detail here and continues in this cryptic style (Sobel 1990, p. 227):

> "*Arguments for the Desirability of Another Action.* There is little variation in how relative desirabilities of actions are established. Most problems use assumptions concerning cardinalities of desirabilities of possible outcomes. One class of exceptions are probability-of-one problems. In these, ordinal relations of desirabilities for outcomes suffice."

I think that Sobel's (1990) four elements together with his further specifications for identifying Newcomblike problems should for the following reasons be preferred to Nozick's (1969) two criteria: (1) At the bottom of Newcomb's problem lies the conflict between the probabilistic dependence and the causal independence. And in opposition to Nozick (1969) Sobel (1990) has rightly observed that the probabilistic dependence and the causal independence can be established by different means. (2) Newcomb's problem would remain a problem, if dominance reasoning were not applicable. For as we have seen in chapter 1.2, in the section on the maximising principles and the

principles of dominance with probabilistic and causal independence, the principle of dominance with probabilistic independence and the principle of dominance with causal independence are corollaries of respective maximising principles (cf. Gibbard and Harper 1978; Sobel 1988c). Thus even if dominance reasoning with causal independence were not applicable in Newcomb's problem, the respective maximising principle would still be applicable and would recommend a certain possible action as rational. Thus Nozick's (1969) first criterion is too narrow for Newcomblike problems, and Sobel (1990) is right when he claims that dominance reasoning is not a feature of every Newcomblike problem. Furthermore, Sobel (1990, p. 227) provides the popcorn problem as a Newcomblike problem in which dominance reasoning isn't applicable.

In the subsequent pages and chapters we will consider in more detail the elements of Newcomblike problems and the grounds for the causal independence and the epistemic dependence. Furthermore, the necessary and sufficient conditions for establishing the causal independence and the epistemic dependence for Newcomblike problems and in particular for Newcomb's problem will be investigated more thoroughly there.

## *Nozick's Reasons for Treating Newcomb's Problem as a Newcomblike Problem*

Nozick (1969) presents two Newcomblike problems besides Newcomb's problem, namely the 2-possible-fathers-1-son-case and the prisoner's dilemma. While the prisoner's dilemma is introduced in chapter 1.3, in the section on the prisoner's dilemma, the 2-possible-fathers-1-son-case will be introduced here (Nozick 1969, p. 125):

> "*P* knows that *S* or *T* is his father, but he does not know which one is. *S* died of some terrible inherited disease, and *T* did not. It is known that this disease is genetically dominant, and that *P*'s mother did not have it and that *S* did not have the recessive gene. If *S* is his father, *P* will die of this disease; if *T* is his father, *P* will not die of this disease. Furthermore there is a well-confirmed theory available, let us imagine, about the genetic transmission of the tendency to decide to do acts which form part of an intellectual life. This tendency is genetically dominant. *S* had this tendency (and did not have the recessive gene), *T* did not, and *P*'s mother did not. *P* is now deciding whether (a) to go to graduate school and then teach, or (b)

to become a professional baseball player. He prefers (though not enormously) the life of an academic to that of a professional athlete."

Nozick (1969) states that the utility matrix for the 2-possible-fathers-1-son-case of the possible outcomes is:

| | $s_1$: S is P's father. | $s_2$: T is P's father. |
|---|---|---|
| $a_1$: P goes to graduate school and teaches then. | $u(o_{11})$: -20 | $u(o_{12})$: 100 |
| $a_2$: P becomes a professional baseball player. | $u(o_{21})$: -25 | $u(o_{22})$: 95 |

Figure 7. Utility matrix for the 2-possible-fathers-1-son-case of the possible outcomes.

Nozick (1969) assumes that the well-confirmed theory contains the following: P probably has the tendency to decide for an academic life, if he decides to lead an academic life; if P doesn't decide to lead an academic life, however, he probably hasn't the tendency to decide for an academic life. Because P has this tendency, if and only if S is his father, we get the following probabilities:

$c(s_1|a_1) = 0.9$,
$c(s_2|a_1) = 0.1$,
$c(s_1|a_2) = 0.1$,
$c(s_2|a_2) = 0.9$.

On the basis of the principle of strong dominance P should decide for an academic life, so Nozick (1969). For leading an academic life strongly dominates making a career as a professional baseball player. On the basis of the principle of maximising conditional utility P should decide for the career of a professional baseball player. For $CU(a_1) > CU(a_2)$:

$CU(a_1) = 0.9(-20)+0.1(100)$,
$= -8$,
$CU(a_2) = 0.1(-25)+0.9(95)$,
$= 83$.

According to Nozick (1969) the last recommendation cannot be right. For who is P's father is already determined, and P's decision for the career of a professional baseball player doesn't make it more probable that S is P's father and that P will die of this

disease. Therefore the principle of strong dominance gives the right recommendation in the 2-possible-fathers-1-son-case. Nozick (1969) generalises this result and states that the principle of dominance should be applied to all Newcomblike problems and therefore also to Newcomb's problem.

Nozick (1969) claims that the decision-theoretically relevant similarities and differences between Newcomb's problem on the one hand and the 2-possible-fathers-1-son-case and the prisoner's dilemma on the other hand are: While in Newcomb's problem and in the other two Newcomblike problems the decision maker cannot causally influence the possible states of the world or the possible actions of the other decision maker, in Newcomb's problem the decision maker has the illusion to causally influence the possible states of the world, and in the other two Newcomblike problems the decision maker hasn't the illusion to causally influence the possible states of the world or the possible actions of the other decision maker. Nozick (1969) maintains that the difference between Newcomb's problem and the other two Newcomblike problems is of no importance, so that no other decision principle has to be applied in Newcomb's problem.

In the following Nozick (1969, p. 139) tries to explain what is responsible for the illusion of causal influence in Newcomb's problem and why this illusion is of no importance:

> "Thus I wish to claim that Newcomb's example is less clear than the others because
> (a) in it the explanation of the state's obtaining refers to the action (though this reference occurs in a nonextensional belief-context)
> and that
> (b) the conditions of the problem prevent one obvious way of refuting the teleologist's view, in this case, which view depends upon the truth that generally if *y* is part of the explanation of *x*, then *y* influences *x*."[74]

In the following I hope to illuminate these rather short remarks. With regard to (a) two points have to be made:

---

[74]Nozick (1969) doesn't name any authors who write about teleology. Furthermore, Nozick (1969, pp. 137-138) only states the following about the teleologist's point of view: "... in the simple case where someone goes to the refrigerator to get an apple, it is not the apple's being there when he gets there which caused him to go, or which (partly) explains his actions, but rather his beliefs about an apple's being there. ... To show that the apple's being there does not influence the person's actions, but rather it is his beliefs about the apple's being there that do, they usually argue that even if the apple were not there, so long as the person had the beliefs, he would act in the same way."

First, Nozick's (1969) claim that in Newcomb's problem the explanation of the possible state of the world refers to the possible action of the decision maker is questionable. For this depends on how one conceives the possible states of the world. To see this let's consider two possible partitions of the possible states of the world in Newcomb's problem:

(1) $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. With regard to this partition of the possible states of the world Nozick's claim is right. For if you ask the predictor why he put $ 1,000,000 in B2 respectively $ 0 in B2, he answers because I believe that the decision maker will take B2 respectively both boxes.

(2) $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$. With regard to this partition of the possible states of the world Nozick's claim is wrong. For if you ask the predictor why he predicted correctly respectively didn't predict correctly, it wouldn't make sense to answer because I believe that the decision maker will take B2 respectively both boxes.

Furthermore, the question arises whether the 2-possible-fathers-1-son-case and the prisoner's dilemma differ from Newcomb's problem in this respect, so that in the 2-possible-fathers-1-son-case and in the prisoner's dilemma the explanation of the possible state of the world respectively of the possible action of the other prisoner doesn't refer to the possible action of the decision maker. While the 2-possible-fathers-1-son-case differs from Newcomb's problem with regard to this point, the prisoner's dilemma doesn't differ from Newcomb's problem in this respect in my opinion. For there is no predictor in the 2-possible-fathers-1-son-case, and if one asks why S respectively T is the decision maker's father, the answer doesn't refer to the possible actions of the decision maker. But in the prisoner's dilemma the other prisoner can be conceived as a predictor of the prisoner's possible action, and if one asks the other prisoner why he confesses respectively doesn't confess, it is possible that he answers because I believe that the prisoner confesses respectively doesn't confess. His reason for doing so might consist in the fact that he believes in similar decision processes in the case of him and the prisoner. Thus the explanation of the possible action of the other prisoner in the prisoner's dilemma can refer to the possible action of the prisoner.

Moreover, one can argue that in the prisoner's dilemma the illusion of causal influence is greater than in Newcomb's problem. For the best way to strengthen the prisoner's belief in the causal independence of the other prisoner's possible action from

his possible action is by letting the other prisoner's possible action happen before the prisoner's possible action in time. Yet this is not given in the prisoner's dilemma; both prisoners decide at the same time. And the best way to strengthen the decision maker's belief in the causal independence of the predictor's prediction from the decision maker's decision in Newcomb's problem is by letting the predictor's prediction happen before the decision maker's decision in time, which is the case in Newcomb's problem. Therefore the illusion of causal influence can be greater in the prisoner's dilemma than in Newcomb's problem.

Second with regard to (a), Nozick (1969, p. 138) claims that one feature of non-extensional belief-contexts is: The existence of x doesn't follow from "P believes that ... x ...", and the truth of p doesn't follow from "P believes that p". Nozick (1969) believes that this feature is relevant for Newcomb's problem and for the apple-in-the-refrigerator-example which is the following: An apple being in the refrigerator doesn't causally influence the decision maker to go to the refrigerator, but the decision maker's belief of an apple being in the refrigerator does. Nozick (1969) presents an argument to show that the above mentioned feature of non-extensional belief-contexts is relevant for the apple-in-the-refrigerator-example:

**Premise 1**: The apple isn't in the refrigerator.

**Premise 2**: If the decision maker had the belief of the apple being in the refrigerator, the decision maker would act as if the apple were in the refrigerator.

**Premise 3**: The existence of x doesn't follow from "P believes that ... x ...", that is the existence of the apple in the refrigerator doesn't follow from "The decision maker believes that there is an apple in the refrigerator.".

**Conclusion**: Therefore the decision maker could have the belief of an apple being in the refrigerator without there being an apple in the refrigerator.

According to Nozick (1969) this shows that the apple doesn't causally influence the decision maker's possible action. Therefore the explanation for the decision maker's possible action should be the same in the case of an apple being in the refrigerator and the decision maker's belief of an apple being in the refrigerator, and in the case of an apple not being in the refrigerator and the decision maker's belief of an apple being in the refrigerator.

According to Nozick (1969) the parallel argument for Newcomb's problem is the following: It is possible that the predictor believes that the decision maker takes B2, although he actually takes both boxes. This shows that the decision maker's possible action doesn't causally influence whether $ 1,000,000 are in B2 or not, but that the

predictor's belief of the decision maker's possible action causally influences whether $ 1,000,000 are in B2 or not.

With regard to (b), the thesis that the conditions of Newcomb's problem prevent the decision maker from refuting the teleological point of view, Nozick's (1969) claim is: It is questionable whether the predictor could believe that the decision maker takes B2 even if the decision maker takes both boxes. For one of the conditions of Newcomb's problem consists in the predictor's high predictability, so that if the predictor believes that the decision maker takes both boxes, then with high probability he takes both boxes, and if the predictor believes that the decision maker takes B2, then with high probability he takes B2. According to Nozick (1969) this condition of Newcomb's problem implies that the predictor's beliefs don't have the feature of non-extensional belief-contexts which is: The decision maker takes B2 respectively both boxes doesn't follow from "The predictor believes that the decision maker will take B2 respectively both boxes.". For if the predictor's predictions are very reliable, then the following is almost certainly the case: If the predictor believes that the decision maker will take B2 respectively both boxes, then the decision maker will take B2 respectively both boxes. Yet Nozick (1969) maintains that the predictor's beliefs have the feature of non-extensional belief-contexts. But why should the conditions of Newcomb's problem, that is the predictor's high reliability, prevent the decision maker to refute the teleological point of view which is: The decision maker's possible action doesn't causally influence whether there is $ 1,000,000 in B2 or not, but the predictor's beliefs about the decision maker's possible action causally influence whether there is $ 1,000,000 in B2 or not? Unfortunately Nozick (1969) doesn't answer this question.

In my opinion there is another factor which could be made responsible for the illusion of causal influence in Newcomb's problem and which on second sight is of no relevance for a solution to Newcomb's problem. Nozick (1969, p. 115) explicitly states that there is common knowledge[75] among the decision maker and the predictor:

> "Furthermore, and you know this, the being knows that you know this, and so on: (I) If the being predicts you will take what is in both boxes, he does

[75]According to Matsuhisa and Kamiyama (1997, pp. 389-390) two persons (1 and 2) have common knowledge of an event, if both know it, 1 knows that 2 knows it, 2 knows that 1 knows it, 1 knows that 2 knows that 1 knows it, etc. Kamiyama (forthcoming) reports that Lewis (1969) introduced the intuitive notion of common knowledge to the philosophical community and that Aumann (1976) formalised it in a set-theoretical setting. Thus the common knowledge assumption that the structure of games (the rules of the game, the rationality of the players, their payoff functions, etc.) is common knowledge among the players and which was implicit in game theory was made explicit (Kamiyama forthcoming).

not put the $ M in the second box. (II) If the being predicts you will take only what is in the second box, he does put the $ M in the second box."

The common knowledge assumption could be made responsible for the illusion of causal influence in Newcomb's problem. For the common knowledge assumption in Newcomb's problem suggests that Newcomb's problem is a game with two decision makers and not a game against nature. And in a game with two decision makers it could be possible that the decision maker could causally influence the predictor, whereas in a game against nature it is impossible for the decision maker to causally influence the predictor, if the predictor is conceived as nature and nature is fixed and determinate. Yet in chapter 1.3, in the section on the classification of Newcomb's problem in game theory, I argued that the 1-shot Newcomb's problem could best be viewed as a 1-person game against nature. Thus there is no way for the decision maker to causally influence the predictor in the 1-shot Newcomb's problem.

## 1.8 A Critique of Nozick's Position in 1969

Nozick (1969) distinguishes between two times three cases in rational decision theory: (a) A dominant possible action exists; (b) no dominant possible action exists. (1) The possible actions have a causal influence on the possible states of the world, and the probabilities of the possible states of the world given the possible actions differ per action from each other; (2) the possible actions have no causal influence on the possible states of the world, and the probabilities of the possible states of the world given the possible actions differ per action from each other; (3) the possible actions have no causal influence on the possible states of the world, and the probabilities of the possible states of the world given the possible actions don't differ per action from each other. Nozick (1969) claims that in (2a), which is the case in Newcomb's problem, the principle of dominance is adequate, that in (1a), (1b) and (3b) the principle of maximising conditional utility is adequate, that in (3a) both principles are adequate, and that in (2b) another principle is adequate, which Nozick (1969, p. 133) doesn't specify exactly. Spohn (1978) criticises that Nozick (1969) doesn't provide a justification for recommending these different decision principles for these different cases. Furthermore, Spohn (1978, p. 181) rightly observes that Nozick (1969) omits the following case without giving a reason for doing so: The possible actions have a causal influence on the possible states of the world, and the probabilities of the possible states of the world given the possible actions don't differ per action from each other. Thus for

reasons of completeness Nozick should actually distinguish between two times four cases instead of two times three cases. Moreover, Spohn (1978, p. 182) reproaches Nozick (1969) for not providing a unifying solution to Newcomb's problem. For Nozick (1969) maintains that in Newcomb's problem two cases have to be distinguished: (1) If the decision maker assigns a predictability of $c = 1$ to the predictor, the principle of dominance doesn't work anymore, so that Nozick (1969) recommends the decision maker to take B2. (2) If the decision maker assigns a predictability of $c < 1$ to the predictor, the principle of dominance works, so that Nozick (1969) recommends the decision maker to take both boxes.

Nozick (1969) takes the 2-possible-fathers-1-son-case and the prisoner's dilemma as decision-theoretically clear cut cases, but as Price (1986, p. 195) has rightly observed the prisoner's dilemma has not a clear and universally agreed correct answer. Moreover, in the prisoner's dilemma there are two decision makers, so that one also has to deal with collective rationality in opposition to individual rationality, which is sufficient for Newcomb's problem, if one conceives Newcomb's problem as a 1-shot game against nature. Collective rationality in the prisoner's dilemma leads to not-confessing for both prisoners, for that results in the best possible outcome for both of them together. Yet an intersubjective utility isn't necessarily a subjective utility. And what is collectively rational isn't necessarily individually rational.

Furthermore, the causal structure in the 2-possible-fathers-1-son-case is on first sight ambiguous in my opinion. Sobel (1990) believes that the causal structure in the 2-possible-fathers-1-son-case is the following: The decision maker believes that there is one cause for the disease and one cause for intellectualism and that both causes are either present or absent. Yet I think that the causal structure is the following: If S is P's father, then S is the common cause for P's disease and P's intellectualism; if T is P's father, then T is the common cause for P's non-disease and P's non-intellectualism. But probably the following compromise is correct: If S is P's father, then S is the common cause for transmitting the disease gene to his son, which causes the son's disease, and for transmitting the intellectualism gene to his son, which causes the son's intellectualism; if T is P's father, then T is the common cause for transmitting the non-disease gene to his son, which causes the son's non-disease, and for transmitting the non-intellectualism gene to his son, which causes the son's non-intellectualism.

Thus Nozick can be criticised in the following way: If he wants to establish that Newcomb's problem should be treated like other Newcomblike problems decision-theoretically, these other Newcomblike problems should be simple, clear cut cases

which have a clear and universally agreed correct answer. But Nozick's 2-possible-fathers-1-son-case and the prisoner's dilemma don't satisfy that demand. As we will see in the subsequent chapters there are better cases (for example, Fisher's problem, Skyrms 1984, p. 65).

Despite all this criticism Nozick (1969) has to be praised. For he introduced a lot of relevant distinctions with regard to Newcomb's problem, e. g. the distinction between probabilistic dependence and causal independence, the distinction between Jeffrey's (1965) principle of maximising conditional utility and the principle of dominance. Furthermore, he made clear that other factors (for example, different grades of predictability) could be of relevance for a solution to Newcomb's problem. Moreover, he provided a solution to Newcomb's problem by drawing analogies to other Newcomblike problems. Thus he opened up a whole area of investigation which all the subsequent decision theorists had to fill out or had to discuss.

## 1.9 Summary

After having given an introduction into rational decision theory by means of Newcomb's problem and into game theory by means of the prisoner's dilemma, the following can be said about Newcomb's problem:

(1) If one conceives a 1-shot Newcomb's problem as a game with two decision makers, there is a unique Nash equilibrium which recommends the decision maker to take both boxes and which is not Pareto-efficient. Therefore Newcomb's problem isn't only a paradox in rational decision theory, but also in game theory.

(2) A 1-shot Newcomb's problem seems to be a game against nature, whereas a finitely iterated Newcomb's problem with the same decision maker is a game with two decision makers. As a game against nature the following partition of the possible states of the world is the correct partition in Newcomb's problem: $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. Furthermore, the predictor's high reliability is irrelevant for providing a solution to the 1-shot Newcomb's problem.

Nozick's (1969) analysis of Newcomb's problem and its evaluation yield first results:

(1) Newcomb's problem can be described as a conflict between the principle of maximising conditional utility and the principle of strong dominance with causal independence.

(2) The time intuition for the 2-boxes-argument cannot be criticised, whereas the other intuitions can be questioned for their adequacy.

(3) Sobel's (1990) analysis of Newcomblike problems is to be preferred to Nozick's (1969) analysis.

(4) Nozick's (1969) reasons for treating Newcomb's problem as a Newcomblike problem can be attacked, especially Nozick's claim that the only difference between Newcomb's problem and the 2-possible-fathers-1-son-case and the prisoner's dilemma consists in the illusion of causal influence in Newcomb's problem. Furthermore, the common knowledge assumption can be made responsible for the illusion of causal influence in Newcomb's problem.

(5) Spohn (1978, p. 182) rightly criticises Nozick (1969) for not providing a unifying solution to Newcomb's problem. Furthermore, Nozick's 2-possible-fathers-1-son-case and the prisoner's dilemma aren't simple, clear cut cases which have a clear and universally agreed decision-theoretic solution, so that it is problematical to use them for establishing a certain solution to Newcomb's problem as rational.

## *Outlook*

In the following three chapters I will present and evaluate the decision-theoretic solutions to Newcomb's problem. While chapter 2 deals with evidential decision theories starting with Jeffrey's (1965) logic of decision, chapter 3 deals with causal decision theories beginning with Gibbard and Harper's (1978) *U*-utility. In chapter 4 other proposals will be looked at, which either can be conceived as mediations between evidential and causal decision theories (Meek and Glymour 1994) or as compromises between evidential and causal decision theories (Nozick 1993), or which cannot be classified in any of these terms (Kyburg 1980; my own proposal). Yet the main core of these other proposals is that they stress the decision maker's perspective.

I begin with evidential decision theories, because Nozick (1969) proposed Newcomb's problem as a conflict between the principle of maximising conditional utility and the principle of strong dominance. Thus Nozick set the stage in such a way that Jeffrey's (1965) evidential logic of decision stands against the principle of strong dominance which, as far as I know, nobody associated with causal decision theories at that time. Therefore evidential decision theories should get the first position in the

arrangement of chapters. Then causal decision theorists dealt with Newcomb's problem, the first ones were Spohn (1977, 1978) and Gibbard and Harper (1978). Therefore causal decision theories should get the second position in the arrangement of chapters.

Furthermore, by following this temporal order the simpler theories are set on first place.[76] While evidential decision theories lack causation as a primitive term in their theories, causal decision theories contain causation as a primitive term.

Moreover, by following this temporal order two crucial points with regard to Newcomb's problem, namely the probabilistic dependence between the predictor's prediction and the decision maker's decision and the causal independence of the predictor's prediction from the decision maker's decision, correspond to the main focus of the respective chapters. For evidential decision theories give the probabilistic dependence more weight than the causal independence, whereas causal decision theories give the causal independence more weight than the probabilistic dependence.

I end with the other proposals, because in the temporal order they are the last ones to deal with Newcomb's problem. Especially Meek and Glymour's (1994) and Nozick's (1993) approaches wouldn't be possible without both evidential and causal decision theories. For Meek and Glymour (1994) try to mediate between evidential and causal decision theories, and Nozick (1993) tries to build a compromise between evidential and causal decision theories. Furthermore, for my own proposal I needed to evaluate all other solutions to Newcomb's problem first. Therefore I found it best to place the other proposals after the evidential and the causal decision theories.

## Appendix: Some Newcomblike Problems

According to Sobel (1990) the following problems are Newcomblike problems: Newcomb's problem (Nozick 1969, pp. 114-115; cf. chapter 1.1, in the section on Newcomb's problem), a variation of Newcomb's problem (Jeffrey 1983, p. 15), the prisoner's dilemma (for example Jeffrey 1983, p. 15; cf. chapter 1.3, in the section on the prisoner's dilemma), Fisher's problem (for example Skyrms 1984, p. 65; cf. chapter 3.3), the popcorn problem (Sobel 1986, pp. 411-413), uncle Albert's problem (Skyrms

[76]To put the simpler theories first is desirable, because it is in agreement with preference practices in the sciences: Sklar (1995, p. 612) explains that one reason for choosing the most plausible theory is the simplicity of the theory besides conformity of the theory with the observational data.

1982, p. 700; cf. chapter 3.3), Jones' problem (Gibbard and Harper 1978, p. 137), Solomon's problem (Gibbard and Harper 1978, pp. 135-136; cf. chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 4), the pregnant women's problem (Price 1986, p. 196), and Nozick's (1969, p. 125) 2-possible-fathers-1-son-case (cf. chapter 1.7, in the section on Nozick's reasons for treating Newcomb's problem as a Newcomblike problem). Skyrms (1984) mentions Fisher's problem, Newcomb's problem, and the prisoner's dilemma as Newcomblike problems; Skyrms (1980) claims that the eggs-benedict-for-breakfast-problem (cf. chapter 2.6), Fisher's problem, the Calvinists' problem (cf. chapter 3.3), and Newcomb's problem are Newcomblike problems. In 1982 Skyrms adds uncle Albert's problem to the Newcomblike problems. Lewis (1981a) counts Newcomb's problem, the eggs-benedict-for-breakfast-problem (cf. Skyrms 1980, p. 129), Fisher's problem (cf. Skyrms 1984, p. 65), the loafing-weak-heart-problem (cf. Lewis 1981a, pp. 9-10), and the prisoner's dilemma (cf. Lewis 1979a) as Newcomblike problems.[77]

Jeffrey's (1983, p. 15) variation of Newcomb's problem is the following:

> "A preternaturally good predictor of human choices does or does not deposit a million dollars in your bank account, depending on whether he predicts you will reject or accept the extra thousand he will offer you just before the bank reopens after the weekend. Would it be wise on Monday morning to decline the bonus?"

This is a description of the popcorn problem (Sobel 1986, pp. 411-413):

> "I am in a cinema watching a movie. I want very much to have some popcorn, though I would not want to miss part of the movie for nothing, as I would do if I went to the lobby and found that there was no popcorn there to be bought. However, I would *really* hate to learn that I had passed up an opportunity to get some popcorn, as I will have done if though there is popcorn to be had I do not go for it. ... I am nearly sure that the popcorn vendor has sold out and closed down, and that there is no popcorn in the lobby to be bought. And so, ..., I have decided not to go for popcorn, and I am nearly sure that I *will* not go for popcorn. Still, I am also nearly sure that in *this* theatre, when and only when there *is* popcorn in the lobby to be bought, the signal
>
> POPCORN!!!

[77] Sobel (1990), Skyrms (1980, 1982, 1984), and Lewis (1981a) don't claim that their lists are complete, though.

> is flashed on the screen repeatedly, though at a speed that permits only subliminal, unconscious awareness. Furthermore, ..., I consider myself to be a highly *suggestible* person. Indeed, ..., I am nearly sure that I will change my mind and *go* for popcorn if and only if I am influenced by this subliminal signal to do so, and so if and only if there is popcorn in the lobby to be *had*! ... As has been said, I am nearly certain that there is no popcorn in the lobby and that I am not going to the lobby ... to buy some, and I am nearly certain of the conjunction of these things. Furthermore, ..., while I think it is very unlikely that I will go for popcorn, and *thus* very unlikely that I will go for popcorn in the circumstance in which there is popcorn there, ..., I think it is *much more* unlikely that I will go for popcorn though there is *none* there to be bought, ... . I can hardly *imagine* how *that* conjunction could come true, for I am nearly sure that its truth would involve my changing my mind and going for popcorn unprodded by that sometimes flashing light, and despite the absence of those sometimes present subliminal suggestions. And *that*, given my conviction that there is no popcorn there to be bought, is an eventuality which I consider 'as near to impossible as makes no difference'."

Gibbard and Harper (1978, p. 137) view Jones' problem as a modern alternative to Solomon's problem; furthermore, Jones' problem is apt for all those who mistrust charisma as a character trait:

> "Robert Jones, rising young executive of International Energy Conglomerate Incorporated ... and several other young executives have been competing for a very lucrative promotion. ... Jones learns ... that all the candidates have scored equally well on all factors except ruthlessness. ... Jones [before the promotion decision is announced] must decide whether or not to fire poor old John Smith. ... Jones knows that [his] ruthlessness factor ... accurately predicts his behaviour in just the sort of decision he now faces."

Price (1986, p. 196) describes the pregnant women's problem:

> "Many physiological states produce quite specific symptomatic effects on choice behaviour. Pregnancy, for example, often affects what a woman chooses to eat. Any such physiological effect is a potential source of a medical Newcomb problem. All else that is needed is that the attitudes of the person concerned about the possession of the underlying physiological

> state and the performance of the symptomatic act should conflict. ... If the agent is someone who knows that pregnancy tends to make her decline garlic, then the problem will arise if either she likes garlic (ceteris paribus) but wants to be (now) pregnant; or doesn't like garlic (ceteris paribus) and doesn't want to be pregnant. Clearly she shouldn't refuse the delectable garlic in order to be already pregnant, in the former case; or eat the distasteful garlic in order not to be, in the latter."

The loafing-weak-heart-problem (Lewis 1981a, pp. 9-10) is:

> "Suppose you like ... loafing when you might go out and run. You are convinced, contrary to popular belief, that these pleasures will do you no harm at all. (Whether you are right about this is irrelevant.) But also you think you might have some dread medical condition: ... a weak heart. If you have it, there's nothing you can do about it now and it will probably do you a lot of harm eventually. In its earlier stages, this condition is hard to detect. But you are convinced that it has some tendency, perhaps slight, to cause you to ... loaf. So if you find yourself indulging, that is at least some evidence that you have the condition and are in for big trouble. But is that any reason not to indulge in harmless pleasures?"

Other candidates for Newcomblike problems are the following problems of rationality, for they were compared with Newcomb's problem: Gideon's paradox (for example Bar-Hillel and Margalit 1985), Simpson's paradox (for example Wagner 1991), the toxin puzzle (for example Bratman 1987; Gauthier 1994; Kavka 1983), the prediction paradox and Moore's problem (for example Goldstein 1993; Olin 1988; Sorensen 1986), the time inconsistency problem (for example Broome 1989; Frydman, O'Driscoll, and Schotter 1982), and Parfit's (1984) problem of the hitch-hiker (for example Barnes 1997).

Here is a description of Gideon's paradox (Bar-Hillel and Margalit 1985, p. 139):

> "A decision maker, DM, is given a choice between two gifts, $G_1$ and $G_2$. DM prefers $G_1$ to $G_2$ (for example, let $G_1$ = \$1000 and $G_2$ = \$0). Just as he is about to choose, a perverse bystander, PB, promises DM that if DM will choose irrationally, then PB will reward him by giving him R. DM prefers R to both $G_1$ and $G_2$ (for example, let R = \$1,000,000). The dilemma into which PB's announcements plunges DM can be spelled out informally as follows: I prefer $G_1$ to $G_2$. So it would be irrational of me to settle for $G_2$ when I could have $G_1$. Hence, I should choose $G_1$. On the

other hand, an irrational choice will entitle me to receive R, which I prefer to $G_1$. So I will ultimately be better off by choosing $G_2$ than by choosing $G_1$. In other words, it might be rational for me to choose $G_2$ after all, even though I prefer $G_1$. But if so, the choice of $G_2$ *wouldn't* entitle me to R, now would it? So what would be the point of choosing it? In short, it seems that any choice I make is rational if and only if it is irrational."

This is Simpson's paradox (Otte 1985, pp. 110-111):

"Simpson's paradox is that any correlation in a population can be reversed in the subpopulations by finding a third variable that is properly correlated with the other variables. Thus positive relevance can become negative relevance or independence, negative relevance can become positive relevance or independence, and independence can become positive or negative relevance."

Kavka's (1983, pp. 33-34) toxin puzzle looks like the following:

"You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. ... The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you *intend* to drink the toxin tomorrow afternoon. He emphasises that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. ... All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin."

The prediction paradox can be described along the following lines (Olin 1988, p. 111):

"The prediction paradox is typically stated in terms of a surprise test. A teacher announces to his student that there will be one test next week, and that the exact day of the test will be a surprise. The student objects that this is impossible, reasoning as follows: 'The surprise test cannot be given on Friday. For, if it were, then on Thursday evening, I would expect it the next day; so it would not be a surprise. Nor can it be held on Thursday. For, given that Friday is ruled out, on Wednesday evening I would be able to predict that the test would be held the next day; hence it would not be a

surprise. Similarly for each of the remaining days.' But, of course, the surprise test can be given - on Tuesday, for example."

According to Sorensen (1986, p. 506) this is Moore's problem:

"Moore's problem is the problem of explaining the oddity of sentences like

(4) It is raining but I do not believe it.

Despite the consistency of (4) our immediate reaction is that it cannot be consistently believed."

A description of the time inconsistency problem is the following (Broome 1989, p. 221):

"What happens in an economy depends on people's expectations of what is going to happen, because their expectations influence their actions. A theory of the behaviour of the economy therefore requires a theory of people's expectations. In recent years a theory known as 'rational expectations' has become popular. The rational expectations theory says that people's expectations are pretty much correct. What people expect to happen is what will happen, apart from some purely random errors. Suppose the government would like to increase the level of employment in the economy. One way of doing so is to expand the money supply. However, according to a common argument ..., this method will only work if the expansion is unexpected. If people expect it, they will themselves act in a way that cancels out the beneficial effects of the expansion. The only result will be inflation. On the other hand, if people expect the money supply to expand and it does not, the result will be a recession. The government's predicament is summarised in Table 2.

| | Expand predicted | Keep constant predicted |
|---|---|---|
| Expand | Inflation (third best) | Increased employment (best) |
| Keep constant | Recession (worst) | No change (second best) |

*Table 2*"

Barnes (1997, pp. 2-3) quotes Parfit's (1984, p. 7) problem of the hitch-hiker:

"Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger and the only other driver near. I manage

> to stop you, and I offer you a great reward if you rescue me. I cannot [pay] you now, but I promise to do so when we reach my home. Suppose next that I am *transparent*, unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away. Suppose, finally, that I know myself to be never self-denying. If you drive me to my home, it would be worse for me if I gave you the promised reward. Since I know that I never do what will be worse for me, I know that I shall break my promise. Given my inability to lie convincingly, you know this too. You do not believe my promise, and therefore leave me stranded in the desert. This happens to me because I am never self-denying. It would have been better for me if I had been *trustworthy*, disposed to keep my promises even when doing so will be worse for me. You would then have rescued me."

In my opinion Dummett's (1964, pp. 348-349) example of the dancing chief is a Newcomblike problem, for there is probabilistic dependence between the chief's dancing and the young men's prior bravery, and there is causal independence of the young men's prior bravery from the chief's dancing. Dummett's example of the dancing chief is:

> "Suppose a tribe has the following custom which is practised every two years: The young men of the tribe are sent on a lion hunt to prove their manhood. They travel for two days, hunt lions for two days, and spent two days travelling back. Observers are sent with them to report to the chief of the tribe whether the young men have been brave. While the young men are away, the chief performs dances which are intended to cause the young men to be brave. To people outside the tribe it seems odd that the chief continues dancing, while the young men are on their return journey, for either the young men have already been brave or not. The chief, however, argues that he doesn't know whether they have been brave or not. Furthermore he has evidence from his past experience that his young men will have been brave, if he dances, and that his young men will not have been brave, if he doesn't dance. Because the chief can decide whether he dances, he believes that it is also up to him whether his young men will have been brave."

Schramm (unpublished) claims that the following forecast lottery is a Newcomblike problem:

"Agent *N* doesn't know what the weather forecast for a particular (past or future) day *d* said or will say, nor has she any other evidence for what the weather actually was or will be on day *d*. But *N* knows the forecast's average success rate which is, say, 75 times right out of 100 predictions. In short, lacking any further evidence, *N*'s trust in the weather forecast being right (*R*) for any day *d* can be expressed by her personal probability $p(R) = .75$ (and, therefore, $p(\neg R) = .25$).

*N* is offered to choose between two lotteries *A* or $\neg A$, where declining acceptance of *A* means acceptance of lottery $\neg A$ and vice versa, i. e., *N*'s options to choose are exhaustive between these two lotteries:

*A*: You receive an advance payment of one Thousand and, in addition, if the forecast for day *d is wrong*, you get one Million, otherwise nothing.

$\neg A$: You receive no advance payment and, if the forecast for day *d is right*, you get one Million, otherwise nothing."

# Chapter 2

# Evidential Decision Theories

## 2.1 Introduction

### *Why Should We Look at Evidential Decision Theories?*

After Nozick's article on Newcomb's problem appeared in 1969, evidential decision theories were opposed by causal decision theories (for example Gibbard and Harper 1978; Lewis 1981a; Skyrms 1980; Spohn 1978). Yet in the mean time even causal decision theorists (for example Sobel 1986; Joyce in press) have come to see the virtues of Jeffrey's (1965) logic of decision. Sobel (1986, p. 429) says:

> "I propose, in fact, to retain Jeffrey's theory, not of course as a theory of rational choice or of preference for propositions as facts, but as a theory of rational preferences for propositions as items of news."

Likewise Joyce (in press, pp. 5-6) writes:

> "I will show how to express Jeffrey's Equation and Stalnaker's Equation as instances of a general *conditional expected utility theory* ... . Evidential decision theorists gain something too, though, for one of the main morals I wish to draw is that there is a deep sense in which Jeffrey's theory is exactly right: *all value is a kind of news value* even if not all kinds of news value are relevant to the choice of actions."

Moreover, current approaches of rational decision theory try to mediate between evidential and causal decision theories (Meek and Glymour 1994) or try to make a compromise between evidential and causal decision theories (Nozick 1993). Therefore it still seems a worthwhile enterprise to look at evidential decision theories for a solution to Newcomb's problem.

### *Order of Presentation*

The most prominent evidential decision theories are[78]:

---

[78]Joyce (in press, p. 48) points out that Savage's (1954/1972) rational decision theory is consistent with either a causal or a non-causal interpretation (cf. chapter 1.2, in the section on evidential and causal decision theories). Thus Savage's rational decision theory could be the first evidential decision theory. Yet as far as I know no one else claims this; moreover, it is

(1) Jeffrey's (1965) logic of decision,

(2) Jeffrey's (1983) ratificationism,

(3) Jeffrey's (1988) probabilism,

(4) Jeffrey's (1996) decision kinematics, and

(5) Eells' (1981, 1982, 1985) common cause-theory.

I will present and criticise these theories in this order for the following reasons: I have set Jeffrey's (1965) logic of decision on first position, because all of the other four proposals are more or less based on it, and because Jeffrey's logic of decision is the earliest proposal of the five proposals. I'm aware that the chronological order demands that Eells' (1981, 1982, 1985) theory should be set on second place before Jeffrey's ratificationism of 1983, especially because Jeffrey's (1983) ratificationism was inspired by Eells' (1981, 1982) theory. Yet I think it is unwise to split up Jeffrey's proposals. For by doing so the development of Jeffrey's thought over the years would be interrupted in some way; furthermore, Jeffrey's aim to base all his different proposals on his logic of decision and thus to save his logic of decision more or less would be concealed to a certain degree. Therefore I found it best to place Eells' (1981, 1982, 1985) approach of the common cause after Jeffrey's proposals at the last position.

One could argue that Jeffrey's (1983) ratificationism, Jeffrey's decision kinematics (1996), and Eells' (1981, 1982, 1985) proposal of the common cause should be set forth in the chapter on causal decision theories. For Jeffrey (1983), Jeffrey (1996), and Eells (1981, 1982, 1985) make use of a general feature of causal chains for providing a solution to Newcomblike problems.[79] Yet to put these proposals in the chapter on causal decision theories would go against the intentions of the authors and against the main contents of their proposals. For Eells (1981, 1982, 1985) bases his solution to Newcomb's problem to such a great extent on Jeffrey's logic of decision that he even retains Jeffrey's (1965) principle of maximising conditional utility. The only

---

philosophical mainstream to consider Jeffrey's (1965) logic of decision as the first evidential decision theory.

[79]Price (1986, p. 203) claims that causal chains have the following general feature: Suppose $C$ is a common cause of $A$ and $B$, or suppose $B$ is the indirect cause of $A$ by means of causing $C$ which directly causes $A$, then knowledge that $C$ has occurred in both cases makes $B$ probabilistically irrelevant to $A$. This amounts to the screening assumption (cf. Rabinowicz 1988, pp. 411-412): $C$ screens off $B$ from $A$ if and only if $c(A|B\cap C) = c(A|C)$. Another way to formulate the screening assumption is found in Price (1986, p. 203): $C$ screens off $B/\neg B$ from $A$ if and only if $c(A|B\cap C) = c(A|\neg B\cap C)$; and $\neg C$ screens off $B/\neg B$ from $A$ if and only if $c(A|B\cap\neg C) = c(A|\neg B\cap \neg C)$.

modification Eells presents which is not in harmony with Jeffrey's logic of decision is the causal structure he proposes for Newcomblike problems. Jeffrey's (1983) only modifications to his logic of decision are his formula for calculating the utility of a possible action, his principle of ratifiability, and the screening assumption; otherwise everything remains the same. Furthermore, in Jeffrey's 1983-proposal the decision maker's final decision becomes evidence for possible states of the world, which is another argument for placing Jeffrey's ratificationism in the chapter on evidential decision theories. Jeffrey (1996) just adds the rigidity condition to his logic of decision; furthermore, he proposes a causal structure for Newcomblike problems and uses the screening assumption to analyse Newcomblike problems; otherwise everything else remains the same. Therefore I found it best to place Jeffrey's (1983) ratificationism, Jeffrey's (1996) decision kinematics, and Eells' (1981, 1982, 1985) proposal of the common cause in the chapter on evidential decision theories.

## *Similarities and Differences between Evidential Decision Theories*

The common characteristic of these evidential decision theories is that they are based on Jeffrey's (1965) logic of decision, which is a preference theory between propositions. Jeffrey's (1983) ratificationism, Jeffrey's (1988) probabilism, Jeffrey's (1996) decision kinematics, and Eells' (1981, 1982, 1985) common cause-solution are all revisions or improvements of Jeffrey's (1965) logic of decision. Therefore there is similarity in content between these evidential decision theories.

What distinguishes all five proposals from each other is:

(1) They propose different definitions of the utility of a possible action and therefore also propose different maximising principles.

(2) They add structure to Newcomb's problem.

(3) They add certain rationality constraints on the decision maker's credences or on the decision maker.

(4) They propose different solutions to Newcomb's problem.

**Ad 1**: Jeffrey (1965) and Eells (1981, 1982, 1985) propose the following definition of the utility of a possible action: If $c(a_i) > 0$, then

$$CU(a_i) = \sum_{j=1}^{m} c(s_j \mid a_i) u(o_{ij}).$$

The critical fact is that the decision maker should condition by the performance of the possible actions. In 1983 Jeffrey departs from his logic of decision and claims that the

decision maker should calculate the utility of a possible action as follows: If $c(da_i) > 0$, then

$$U(a_i|da_i) = \sum_{j=1}^{m} c(s_j | da_i) u(o_{ij}).$$ ("$da_i$" denotes the final decision to perform $a_i$.)

According to this formula the decision maker should condition by the final decision to perform the possible actions. In 1988 Jeffrey reverts to his logic of decision by postulating that the utility of a possible action should be calculated in the following way: If $c_n(a_i) > 0$, then

$$U_n(a_i) = \sum_{j=1}^{m} c_n(s_j | a_i) u(o_{ij}),$$

so that the decision maker should use his final credences and should condition by the performance of the possible actions. In 1996 Jeffrey doesn't introduce any definition of the utility of a possible action.

While Jeffrey (1965) and Eells (1981, 1982, 1985) claim that the decision maker should follow the principle of maximising conditional utility, Jeffrey (1983) departs from his logic of decision and defends the view that the decision maker should adopt the principle of ratifiability, that is the decision maker should make ratifiable decisions. And according to Jeffrey (1983, p. 16) "A ratifiable decision is a decision to perform an act of maximum estimated desirability relative to the probability matrix the agent thinks he would have if he finally decided to perform that act.". In 1988 Jeffrey reverts to his logic of decision by proposing that the decision maker should follow the principle of maximising final conditional utility. In 1996 Jeffrey doesn't deal with decision-making principles.

**Ad 2**: Eells (1982) claims that Newcomb's problem has the following causal structure: A common cause (cf. Reichenbach 1956) causes on the one hand by means of causing a prediction a possible outcome and causes on the other hand by means of causing some element of *R* a possible action. According to Eells (1985) *R* is the set of propositions describing the possible sets of beliefs and wants the decision maker might have in the respective decision situation. Eells (1981, 1982, 1985) claims that the common cause, which is responsible for the predictor's prediction and for the decision maker's decision, explains the high correlation between prediction and decision.

**Ad 3**: Jeffrey (1996) adds the rigidity condition as a constraint on the decision maker's credences.[80] The rigidity condition demands that conditional credences are

[80]In personal communication Jeffrey (1997) wrote to me: "The 'rigidity' condition makes the system as independent of the probabilities of acts as it needs to be. Rigidity is the condition of constancy of conditional probabilities of states given acts as unconditional probabilities of acts

constant. If the rigidity condition is satisfied, then generalised conditioning is possible. Jeffrey (1996) claims that Newcomblike problems are no decision problems, because too much information is given about the decision maker's conditional credences, so that the unconditional credences of the possible actions are already fixed, which is a problem. For in any decision problem where the possible outcomes are not known from the very beginning the decision maker's credences of the possible actions vary during deliberation until one possible action has reached a credence of 1 and is decided for.

Eells (1985) proposes that the decision maker conceives himself as an ideal rational decision maker who holds the following three beliefs: (1) I am certain what my credences and utilities are at decision time. (2) I am certain that I determine $a_i$ as the rational possible action if and only if I perform it. (3) I attribute the same credence to the following two propositions (= screening assumption): (i) I determine $a_i$ as the rational possible action given my credences and utilities at decision time and the presence of the common cause. (ii) I determine $a_i$ as the rational possible action given my credences and utilities at decision time and the absence of the common cause.

**Ad 4**: While Jeffrey (1965) proposes a 1-box-solution, Jeffrey (1983), Jeffrey (1988), and Eells (1981, 1982, 1985) propose a 2-boxes-solution. In 1996 Jeffrey believes that Newcomb's problem is no decision problem at all.[81]

Jeffrey's 1965-proposal cannot be called a solution to Newcomb's problem, for Newcomb's problem became known to the philosophical community by Nozick's article in 1969. Jeffrey's (1965) logic of decision has to be treated as a solution to Newcomb's problem, though, for it is the only decision-theoretic solution to Newcomb's problem which proposes the only 1-box-solution for Nozick's original version of Newcomb's problem.[82]

---

vary. This is necessary and sufficient for conditioning or generalised conditioning ('probability kinematics') to be the correct mode of updating, and is thus necessary and sufficient for 'desirability' (conditional expectation of utility given the act) to be the correct figure of merit for acts. It is precisely this condition which is violated in Newcomb Problems. ... I have come to think that Newcomb Problems are not really decision problems after all."

[81] Jeffrey (1996) isn't the first one to doubt that Newcomb's problem is no decision problem at all. For Mackie (1977, p. 223) concludes that Newcomb's problem "simply cannot occur", Cargile (1975, p. 238) claims that Newcomb's problem is "incredible", and Schlesinger (1974, p. 211) states that it is "impossible". While Cargile and Mackie criticise that Newcomb's problem isn't precise enough, Schlesinger tries to show that it is contradictory. In opposition to Cargile and Mackie Jeffrey (1996) claims that too much information is given in Newcomb's problem.

[82] Argumentative only 1-box-solutions (without providing a rational decision theory) for Nozick's original version of Newcomb's problem are presented over and over again, though (for example Bach 1987; Voizard 1996). For a discussion of Voizard's arguments in favour of taking B2 see Ledwig (1998, 1999a).

Jeffrey's (1983, 1988, 1996) later proposals explicitly deal with Newcomb's problem. In 1983, 1988, and 1996 Jeffrey considers Newcomb's problem as a counterexample to his logic of decision (Jeffrey 1965) and reacts in the following way to it: In 1983 he explicitly departs from his logic of decision by proposing a different definition of the utility of a possible action and by proposing a different maximising principle. Yet Jeffrey (1983, p. 20) already acknowledges that his ratificationism is faced with a counterexample, which differs from Newcomb's problem. In 1988 Jeffrey reverts to his logic of decision and claims that it is sufficient for giving a solution to Newcomblike problems, if Newcomblike problems are probabilised on two levels. On the first level is the unknown chance that the decision maker will take B2 or both boxes, whereas on the second level are the decision maker's credences. The decision maker needs his credences to find out what his chances are. In 1996 Jeffrey improves his logic of decision by adding the rigidity condition as a constraint on the decision maker's credences.

In opposition to Jeffrey Eells (1981, 1982, 1985) sticks to his 2-boxes-solution over the years. Yet Eells refines his proposal and claims that rational deliberation is a dynamic process (Eells 1984)[83], that common causes can lead to different probability structures (Eells and Sober 1986), namely to conjunctive forks (cf. Reichenbach 1956) or interactive forks (cf. Salmon 1978)[84], and that the Markov condition doesn't oppose the transitivity of causal chains (Eells and Sober 1983)[85]. Because these refinements don't change Eells' major points in his common cause-theory and don't change his solution to Newcomb's problem, I will not deal with them. Yet Eells' theory is one of the

---

[83]Eells (1984, p. 92) claims that the decision maker should continually calculate his utility of a possible action with appropriate alterations in his credences in the light of the results of previous calculations. Furthermore, Eells (1984, p. 92) proposes a theory of continual utility maximisation and adds certain rationality constraints on the decision maker's process of deliberation.

[84]Eells and Sober (1986, p. 228) report that Reichenbach's (1956) conjunctive fork is defined by the following four conditions:

(1) $P(A\cap B|C) = P(A|C)P(B|C)$,
(2) $P(A\cap B|\neg C) = P(A|\neg C)P(B|\neg C)$,
(3) $P(A|C) > P(A|\neg C)$,
(4) $P(B|C) > P(B|\neg C)$,

where $C$ is the common cause of $A$ and $B$, and where $C$ is prior in time to $A$ and $B$. To obtain an interactive fork, so Eells and Sober (1986, p. 229), one has to replace the first condition by the following condition:

(1') $P(A\cap B|C) > P(A|C)P(B|C)$.

[85]According to Eells and Sober (1983, p. 46) the Markov condition obtains, in case the state of the system at $t_3$ depends only on what happens at $t_2$, but not on what happens before.

most elaborated rational decision theories. Eells (1981) motivates his proposal for Newcomb's problem by drawing an analogy to a medicinal Newcomblike problem (the eggs-benedict-for-breakfast-problem by Skyrms 1980, p. 129) and two moral Newcomblike problems (Solomon's problem by Gibbard and Harper 1978, pp. 135-136, and Jones' problem by Gibbard and Harper, p. 137) which have a unique solution and therefore can be applied to Newcomb's problem. Normally causal decision theorists (for example Gibbard and Harper 1978) use Newcomblike problems as counterexamples of evidential decision theory; yet Eells (1981) tries to show that Newcomblike problems are on second sight no counterexamples of evidential decision theory.

After this short overview let's have a more closer and critical look at evidential decision theories and their solutions to Newcomb's problem.

## 2.2 Jeffrey's Logic of Decision[86]

### *Central Features of Jeffrey's Logic of Decision*

Being more opposed to Ramsey's (1931) and Savage's (1954/1972) classical rational decision theories than being an advancement of them Jeffrey's (1965) logic of decision has the following central features:

(1) Jeffrey's theory is non-causal.

(2) The decision maker can partition the possible states of the world in any way.

(3) The decision maker can assign credences to his possible actions.

(4) Jeffrey's theory establishes preferences between news items.

(5) Jeffrey's theory has a representation theorem.

(6) The decision maker should follow the principle of maximising conditional utility.

**Ad 1**: Jeffrey (1965) proposes a non-causal preference theory between propositions and takes "it to be the principal virtue of the present theory, that it makes no use of the notion of a gamble or of any other causal notion" (Jeffrey 1983, p. 157). So let's see what Jeffrey means, when he takes gambles to be a causal notion.

---

[86]The term logic derives from the fact that Jeffrey (1965) uses logical operations on propositions, namely denial, conjunction, and disjunction, instead of forming gambles, which operate in the theories of Ramsey (1931) and Savage (1954/1972).

In the following passage Jeffrey (1983, p. 156) makes his position more explicit by conceiving gambles as causal relationships:

"In general, a gamble of form

*A* if *C*, *B* if not

exists if there is a causal relationship between *C*, *A*, and *B*, in virtue of which *A* will happen if *C* does, and *B* will happen if *C* does not."

Jeffrey (1983, pp. 156-157) gives the following example to clarify his point of view: Suppose I offer you a bet to give you $ 1, if *C* will happen, and to get $ 1 from you, if *C* will not happen. You accept the bet. *A* is the proposition that you give me $ 1 at the time, when we learn whether *C* is true or false, and *B* is the proposition that I pay you $ 1 at the time, when we learn whether *C* is true or false. Then we have established a causal relationship between *C*, *A*, and *B*.

According to Jeffrey (1983) this relationship is as causal as the relationship between the proposition that the gas tank is empty and the proposition that the car doesn't go. Furthermore, Jeffrey (1983) claims that the causal relationship in the betting case depends on the mutual trust the betting partners have in each other and on the betting partners' ability to pay the other at the time in question.

Jeffrey's (1983) reason for proposing a non-causal preference theory is that if one conceives gambles as causal relationships, one can formulate bets which are difficult to rank in the preference ordering of propositions. Jeffrey (1983, p. 157) illustrates this by the following example: Suppose your preference ordering contains the proposition *A* that there will be a nuclear war next week, the proposition *B* that there will be fine weather next week, and the proposition *C* that this coin lands head up. Then it must also contain the gamble *A*, if *C* is the case, and *B*, if *C* is not the case, that is there will be a nuclear war next week, if this coin lands head up, and there will be fine weather next week, if this coin lands tail up. According to Jeffrey (1983) the decision maker isn't able to rank this gamble in his preference ordering of propositions. For suppose this gamble is in effect, then the decision maker has to revise his beliefs of the causes of war and weather in such a way that it is unclear which of his other beliefs will be affected by this revision.

Yet Jeffrey's (1965, 1983) theory isn't completely non-causal. Jeffrey (1983, p. 157) writes:

"The theory ... deals only with such causal relationships as the agent believes (rightly or wrongly) actually obtain among the propositions in his preference ranking. We do not need to know how he would revise his

> preference ranking if he believed in the existence of further causal relations which may be in serious conflict with the relationships he does believe in."

Thus Jeffrey excludes such gambles as the nuclear-war-fine-weather-gamble, but not every gamble. Jeffrey (1983, p. 156) allows gambles in which the possible outcomes are the truth and the falsity of the proposition on which the decision maker gambles.

**Ad 2**: In Jeffrey's (1965, 1983) preference theory the decision maker can partition the possible states of the world in any way. For no matter how the decision maker partitions the possible states of the world, the decision maker's calculation of the utility of a possible action yields the same result.[87] Thus Joyce (in press, p. 121) calls Jeffrey's preference theory "partition invariant".

**Ad 3**: Because Jeffrey (1983, p. 83) conceives possible actions as propositions, the decision maker can also assign credences to his possible actions. If the decision maker believes that a possible action is totally within his power to make true, the decision maker should assign $c = 1$ to this possible action (Jeffrey 1983, p. 85). Because Jeffrey's formula for the utility of a possible action applies only when $c(a_i) > 0$, the decision maker shouldn't assign $c = 0$ to one of his possible actions. If the decision maker believes that a possible action is totally not within his power to make true, the decision maker should assign a very low credence to this possible action, but not a credence of 0. In cases like this one Jeffrey (1983, p. 85) adopts the point of view that the decision maker can only try to make this possible action true.

Jeffrey distinguishes between "acts" (Jeffrey 1983, p. 83) and "probabilistic acts" (Jeffrey 1983, p. 177). With regard to both of them Jeffrey (1983, pp. 177-178) writes:

> "It may be that the agent decides to perform an act which is not simply describable as *making the proposition B true*, but must be described as changing the probabilities of two or more propositions ... from
>
> $prob\ B_1, prob\ B_2, ..., prob\ B_n,$
>
> to a new set of values,
>
> $PROB\ B_1, PROB\ B_2, ..., PROB\ B_n.$
>
> In the simplest cases, where $n = 2$, where $B_1$ is some good proposition $B$, and where $B_2$ is the bad proposition $\overline{B}$, we speak of the agent as *trying to make B true*; and where *PROB B* ... is very close to 1, we may speak of

[87] Joyce (in press, p. 121) provides the proof for this result.

> the agent as believing it to be in his power to make *B* happen if he chooses. ... Rather, to speak of the agent's trying to make *B* true is to speak of his performing an act of which he takes the net effect to be an increase in the probability of *B*. ... the agent may speak in this way without thereby assuming the existence of a proposition *E* in his preference ranking for which we have
>
> (11-9) $PROB\ E = 1$ $PROB\ B = prob(B/E)$
>
> where *PROB* is the agent's belief function after he decides to perform the act. Thus, trying to hit the bullseye may be an act without there being any proposition that plays *E* to hitting the bullseye's *B*, above."[88]

Thus in the case of acts the decision maker believes that a possible action is in his power to make true, if he wants to; whereas in the case of probabilistic acts the decision maker believes that a possible action isn't in his power to make true; he can only try to make it true, if he wants to.

**Ad 4**: That the utility of a possible action $a_1$ is greater than the utility of a possible action $a_2$ is to say that a possible action $a_1$ is higher in the decision maker's preference ordering than a possible action $a_2$ (Jeffrey 1983, p. 82). According to Jeffrey this means that the decision maker welcomes the news of $a_1$'s truth more than the news of $a_2$'s truth. By doing so Jeffrey (1983, p. 82) establishes preferences between news items.

Furthermore, this interpretation explains why the impossible proposition doesn't occur in any preference ordering of propositions, so Jeffrey (1983, pp. 82-83). For there could be no such news as

> "It will rain all day tomorrow in San Francisco and it will not rain all day tomorrow in San Francisco." (Jeffrey 1983, p. 83).

Moreover, this interpretation explains why the necessary proposition is in another sense no news, Jeffrey (1983, p. 83) claims. For it must be true that

> "It will rain all day tomorrow in San Francisco or it will not rain all day tomorrow in San Francisco." (Jeffrey 1983, p. 83).

That a proposition *A* is lower than the necessary proposition in the decision maker's preference ordering of propositions, means that *A* is bad news in the sense that no news is good news for the decision maker compared with the news that *A* is true (Jeffrey 1983, p. 83). And that a proposition *A* is higher than the necessary proposition in the

---

[88] *E* is the proposition: The agent tries to make *B* true. *prob B* is the probability of *B* before the agent has decided to perform *B*.

decision maker's preference ordering of propositions, Jeffrey (1983, p. 83) interprets as follows: Compared with the good news that $A$ is true no news is bad news for the decision maker. And that a proposition $A$ is ranked with the necessary proposition in the decision maker's preference ordering of propositions, means that the decision maker is indifferent to the news that $A$ is true (Jeffrey 1983, p. 83).

According to Jeffrey (1983, p. 84) the notion of preference relates to possible actions and to news items, so that this notion is active and passive. But if the decision maker deliberates whether to do $a_1$ or $a_2$ and if $a_1 \cap a_2$ is impossible, then there is no effective difference between asking whether he prefers $a_1$ to $a_2$ as a news item or as a possible action, for the decision maker makes the news (Jeffrey 1983, p. 84).

**Ad 5**: In 1978 Jeffrey (pp. 227-228) clarifies the framework of his logic of decision:

> "The logic of decision can be viewed as a theory with a binary relation term ('nonpreference') as its only primitive ... . A *model* of the theory is a pair $(u, P)$ where $P$ is a probability measure on a $\sigma$-field $M$ of 'events' ... and $u$ is an integrable function on the union, $W$, of $M$. A sentence of the theory is valid in a model iff true for all assignments to its free variables of sets in $M$, when for $E \in M$ we define $U(E)$ as the conditional expected utility.
>
> (1) $U(E) = \frac{1}{P(E)} \int_E u(w) \mathrm{d}P(w)$
>
> if $P(E) \neq 0$,
>
> $U(E) = U(W)$ if $P(E) = 0$,
>
> and we define nonpreference in terms of $U$:
>
> (2) $E \precsim F$ iff $U(E) \precsim U(F)$ for all $E, F \in M$. ...
>
> Call a relation $\precsim$ of nonpreference *representable* when there exist $u$ and $P$ for which (2) holds, with $U$ defined as in (1). Such a pair is then to *represent* $\precsim$."

Jeffrey (1978) lays down ten axioms, (A1) - (A10), for his logic of decision, which contain properties of the relation of non-preference and which are used in deriving Jeffrey's representation theorem from Bolker's (1965, 1966, 1967) representation theorem. (For the proofs see Jeffrey 1978.) Jeffrey's (1978, p. 230) representation theorem, which consists of an existence theorem and a uniqueness theorem, is:

> "(8) **Existence.** $\precsim$ *is representable if (A1)-(A10) hold*.

(9) **Uniqueness.** *If (U, P) represents the* $\preceq$ *of (8) then(V, P) does iff there exist real a, b, c, d where ad > bc and -d/c is outside the range of U on M, and for all E ∈ M we have*

$$V(E) = \frac{aU(E)+b}{cU(E)+d},\ Q(E) = P(E)(cU(E)+d).$$"

**Ad 6**: Jeffrey (1983, pp. 80-81) states that the decision maker's credences[89] and utilities should follow the three axioms of Kolmogorov (1933) and Jeffrey's (1965, 1983) utility axiom (cf. chapter 1.2, in the section on (subjective) utilities/objective utilities and credences/chances). These four axioms together with their consequences yield the same result as Jeffrey's formula for calculating the conditional utility of a possible action (Jeffrey 1983, p. 81); and the decision maker should calculate the conditional utility of a possible action in the following way: If $c(a_i) > 0$, then

$CU(a_i) = \sum_{j=1}^{m} c(s_j \mid a_i)u(o_{ij})$ (cf. chapter 1.2, in the section on the maximising principles).

Moreover, the decision maker should use the principle of maximising conditional utility (cf. chapter 1.2, in the section on the maximising principles). In Newcomb's problem calculating the respective conditional utilities for taking both boxes and for taking B2 result in a higher conditional utility for taking B2 than for taking both boxes. Therefore the decision maker should take B2 in Newcomb's problem.

## *Advantages and Disadvantages of the Central Features of Jeffrey's Logic of Decision*

**Ad 1**: In December 1981 Jeffrey added the following passage to his "The Logic of Decision Defended", *Synthese*, September 1981, which explains why Jeffrey prefers a non-causal decision theory[90]:

> "But of course, ratificationism is a revamping of the 1965 system - one in which causal imputations play no explicit role. Partisans of causal decision theory see that as a defect: causality is smuggled in the back door, in disguise. On the other hand, with causality unrepresented in the primitive basis, one can hope to devise a preference-theoretic account (perhaps a

[89]The decision maker's credence is descriptive and not normative, so Jeffrey (1983, p. 21), that is the decision maker's credence is "as it is, not as it should be". (For a discussion of descriptive vs. normative rational decision theories see chapter 1.2, in the section on descriptive and normative rational decision theory.)

[90]I would like to thank Howard Sobel for giving me the text, which Jeffrey added to copies of his 1981-article that he distributed.

definition) of what it is to impute causal dependencies in definite directions."

One further reason to prefer a non-causal decision theory to a causal decision theory is to prefer a simpler theory to a more complicated theory, if the simpler theory is as adequate as the more complicated theory for solving the problems in its domain. Jeffrey's logic of decision is simpler than causal decision theories, because it has fewer primitive terms than causal decision theories. Causation doesn't belong to the primitive terms in Jeffrey's logic of decision[91], while it belongs to the primitive terms in causal decision theories.

Yet causal decision theorists claim that their theories are more adequate than Jeffrey's logic of decision for solving the problems in their domain. Especially, causal decision theorists claim that only their theories give the right recommendations in Newcomblike problems. In my opinion causal decision theorists are right in claiming that their theories are more adequate than Jeffrey's logic of decision for solving decision problems. For if one looks at Newcomb's problem the intuitions which backed up the 1-box-argument, namely the indifference intuition and the betting intuitions, can both be criticised, while with regard to the intuitions which backed up the 2-boxes-argument, namely the time intuition and the intuition of the well-wishing friend, only the intuition of the well-wishing friend is easily criticised (cf. chapter 1.5).

The time intuition for the 2-boxes-argument stands up to now unrefuted, which is of some significance. For the time intuition exemplifies the decision maker's belief in the causal independence of the predictor's prediction from the decision maker's decision. In opposition to that the intuition of the well-wishing friend for the 2-boxes-argument exemplifies the decision maker's belief in the causal independence only insofar as the intuition of the well-wishing friend partly relies on the time intuition. The intuition of the well-wishing friend partly relies on another factor, namely the well-wishing friend, which is absent in the time intuition, so that the well-wishing friend factor can be made responsible and actually is made responsible (cf. chapter 1.5) for the refutation of the intuition of the well-wishing friend. Thus causation should figure as a primitive term in rational decision theory.

With regard to the intuitions backing up the 1-box-argument the following can be said: The indifference intuition exemplifies the decision maker's belief in the probabilistic dependence between the predictor's prediction and the decision maker's

[91]While in 1978 (p. 227) Jeffrey takes non-preference as the only primitive term in his logic of decision, in 1983 (p. 144) he takes preference-or-indifference as the only primitive term.

decision. In opposition to that the betting intuitions exemplify the decision maker's belief in the probabilistic dependence only insofar as the betting intuitions partly rely on the indifference intuition. The betting intuitions partly rely on another factor, namely the betting, which is absent in the indifference intuition, so that the betting factor could be made responsible for the refutation of the betting intuitions. Yet as we have already seen (cf. chapter 1.5) the indifference intuition is made responsible for the refutation of the betting intuitions. Thus the decision maker's belief in the probabilistic dependence between the predictor's prediction and the decision maker's decision is of no relevance for rational decision-making.

To support this conclusion let's have one more look at Nozick's (1969) 1-box-argument, and let's try to refute it:

**Premise 1**: If I take the content of both boxes at $t_3$, the predictor had predicted this with high reliability at $t_1$ and has put \$ 0 in B2 at $t_2$, so that I get \$ 1,000 with near certainty.

**Premise 2**: If I take the content of B2 at $t_3$, the predictor had predicted this with high reliability at $t_1$ and has put \$ 1,000,000 in B2 at $t_2$, so that I get \$ 1,000,000 with near certainty.

**Conclusion**: Therefore I should take the content of B2.

One way of refuting Nozick's 1-box-argument is by making the high reliability of the predictor irrelevant for providing a solution to Newcomb's problem.[92] And Segerberg (1993), although he doesn't apply his thoughts directly to Nozick's 1-box-argument, takes exactly this stance. Furthermore, he suggests that the distinction between the first- and the third-person perspective can explain the irrelevance. For the decision maker deliberates from a first-person perspective and not from a third-person perspective, which is the normal perspective of any observer. And if the decision maker deliberates from a first-person perspective, he cannot take the perspective of the predictor, a third person, during deliberation, and therefore cannot take the predictor's predictions into account (cf. chapter 1.5).

Thus the following conclusion suggests itself: Because the decision maker cannot take the perspective of the predictor during deliberation and therefore cannot take into account the predictor's predictions into his deliberation, Nozick's 1-box-argument, which relies on the decision maker taking into account the predictor's predictions, isn't sound. Because the situation is similar with regard to the intuitions

[92] I would like to thank Wlodek Rabinowicz for the thesis that the predictor's high reliability is irrelevant for decision-making in Newcomb's problem.

backing up the respective arguments, that is the intuitions for the 1-box-argument stand refuted, while the time intuition for the 2-boxes-argument stands unrefuted, causation should figure as a primitive term in rational decision theory.

With regard to Jeffrey's argument that gambles are causal relationships one can question whether Jeffrey's (1965) analogy between the empty gas tank, which caused the car not to go, and the giving of a dollar in a bet on *C*, which is being caused by the occurrence of *C*, holds true. For one could argue that the former case is governed by laws of nature, while the latter case isn't. Three replies are possible in this case[93]: First, causes shouldn't be analysed in terms of laws of nature (cf. Armstrong 1997, chapter 14). Second, human behaviour is governed by laws of nature, namely the laws of psychology (cf. Hume 1993, chapter 8). Third, the fact, for example, that the horse Morning Star on which I betted wins the race is the cause or at least a crucial causal factor of my receiving a big payout. Against the second reply psychologists would probably object that the science of psychology is still far away from providing laws of human behaviour, so that we can at best speak of normative regulations in the case of the bet. Yet the fact that science still isn't able to provide laws of psychology doesn't say that science isn't in principle able to provide laws of psychology. Thus this objection against the second reply isn't cogent. Therefore on second sight Jeffrey's analogy between the empty gas tank and the giving of a dollar in a bet seems to be sound.

**Ad 2**: Joyce (in press, p. 7) claims that causal decision theory suffers from the problem of partition dependence; thus causal decision theory can only be applied, when the decision situation is described in a very specific and detailed way. Yet Jeffrey's logic of decision should be commended for its partition invariance (Joyce in press, p. 122). For partition invariance enables the decision maker to apply Jeffrey's theory to small-world decision problems and to grand-world decision problems (Joyce in press, p. 121).

Following Savage (1954/1972) Joyce (in press, p. 73) distinguishes between grand-world decision problems and small-world decision problems. While a grand-world decision is marked by the fact that the decision maker takes all his possible actions into account and resolves the partition of the possible states of the world to the highest level of pertinent detail, a small-world decision is a coarsening of the grand-world decision. By means of refinement a small-world decision can become a grand-world decision, so Joyce (in press, p. 73); yet no human being is able to deliberate about a grand-world

[93]I wish to thank Phil Dowe for drawing my attention to these replies.

decision. For this amounts to deliberating about all the decisions the decision maker might make over his whole life. Thus the decision maker always faces more or less refined small-world decision problems, and Joyce (in press, p. 7) wants rational decision theory to apply to these.

In opposition to Joyce (in press) I like to claim that causal decision theory should be commended for its partition dependence and that on the contrary Jeffrey's (1965) logic of decision suffers from the problem of partition invariance. With regard to the former the question remains whether causal decision theories mark the right partitions. For causal decision theorists don't explicitly deal with the partition dependence of their theories and their theories only implicitly contain criteria for right and wrong partitions - an exception is Sobel (1986) with regard to both points -, so that it could easily be the case that causal decision theorists mark the wrong partitions. I will deal with this problem in chapter 3. With regard to the latter I provide an example[94] to show that partition invariance may be a problem. For if the decision maker can partition the possible states of the world in any way, he can partition the possible states of the world in Newcomb's problem in the following way, too:

| | $s_1$: I take the content of both boxes at $t_3$. | $s_2$: I take the content of B2 at $t_3$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | $o_{11}$: / | $o_{12}$: / |
| $a_2$: I take the content of B2 at $t_3$. | $o_{21}$: / | $o_{22}$: / |

Figure 8. Decision matrix for Newcomb's problem in which no possible outcomes $o_{11}$, $o_{12}$, $o_{21}$, $o_{22}$ result from combining the possible actions $a_1$, $a_2$ with the possible states of the world $s_1$, $s_2$.

Yet this partition of the possible states of the world makes no sense. For no possible outcomes result from the combination of the possible actions with the possible states of the world. Another example for a senseless partition of the possible states of the world in Newcomb's problem is the following[95]: $s_1$: I take the boxes with my left hand. vs. $s_2$: I take the boxes not with my left hand. For it should be of no relevance with which

[94]The example stems from Sobel (1986, pp. 427-428); Sobel uses it for a different purpose, though.

[95]The example comes from Sobel (1989, p. 81); again Sobel uses it for another purpose.

hand the decision maker takes the boxes. Moreover, in chapter 1.3, in the section on the classification of Newcomb's problem in game theory, we have already argued that there is only one correct partition of the possible states of the world in Newcomb's problem: $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. Therefore partition invariance is really a defect of Jeffrey's logic of decision.

**Ad 3**: I will evaluate this central feature that the decision maker can assign credences to his possible actions when dealing with Spohn's principle in chapter 3.6. Yet one thing seems obvious. In Newcomb's problem the decision maker decides for acts and not for probabilistic acts. For it would be an unnecessary complication for rational decision theory and the decision maker would be considered as being strange, if he believed that both possible actions were not totally within his power to make true. Furthermore, in chapter 3.6 I try to show that the decision maker shouldn't assign credences to his possible actions. Thus this central feature of Jeffrey's logic of decision can be criticised.

**Ad 4**: That Jeffrey's theory establishes preferences between news items can be criticised: Joyce (in press, p. 121) rightly claims that Jeffrey proposes "an odd way" to think about possible actions. For if the decision maker makes the news when he acts, Jeffrey's logic of decision requires the decision maker to seek the good by seeking goods news (Joyce in press). Furthermore, Gibbard and Harper (1978), Joyce (in press), and Lewis (1981a) stress the fact that seeking good news by doing a possible action is not the same as causally bringing about good news by doing a possible action. Or to say it in Lewis' (1981a, p. 8) words: "This seeking of good news may not seem so sensible, however, if it turns out to get in the way of seeking good results. And it does." Gibbard and Harper (1978) try to show by means of Solomon's problem how the decision maker's seeking of good news may get in the way of his seeking good results.

This is a description of Solomon's problem, in which Solomon has to decide whether it is rational to send for another man's wife (Gibbard and Harper 1978, pp. 135-136):

> "Solomon faces a situation like David's, but he, unlike David, has studied works on psychology and political science which teach him ... : Kings have two ... personality types, charismatic and uncharismatic. A king's degree of charisma depends on his genetic make-up and early childhood experiences and cannot be changed in adulthood. ... charismatic kings tend

> to be justly and uncharismatic kings unjustly. Successful revolts against charismatic kings are rare, whereas successful revolts against uncharismatic kings are frequent. Unjust acts ... do not cause successful revolts; the reason that uncharismatic kings are prone to successful revolts is that they have a sneaky, ignoble bearing. Solomon does not know whether or not he is charismatic; he does know that it is unjust to send for another man's wife."

With regard to Solomon's problem Gibbard and Harper (1978) claim the following: Solomon would welcome the news that he was about to abstain from his neighbour's wife; his reason for welcoming this news is not that he takes just possible actions any more likely to have desirable possible outcomes than unjust possible actions; his reason for welcoming this news is rather that he takes just possible actions to be a sign of charisma, which is a possible state of the world, and that charisma may causally bring about a wanted possible outcome. Yet Solomon should send for his neighbour's wife; for he knows that sending will causally bring about a wanted possible outcome, namely having the woman, and he knows that sending will not causally bring about a possible outcome he doesn't want to have, namely having a revolt. Thus Gibbard and Harper (1978) rightly conclude that the utility of a possible action shouldn't measure its welcomeness as news for a wanted possible state of the world, but its efficacy of causally bringing about a wanted possible outcome. And although the decision maker makes the news, he does not make all the news his possible action seems to grant (Gibbard and Harper 1978).

With regard to Newcomb's problem this means that although the decision maker would welcome the news that he was about to take B2, which is a sign that there is $ 1,000,000 in B2, taking B2 doesn't causally bring about that there is $ 1,000,000 in B2. Furthermore, although the decision maker wouldn't welcome the news that he was about to take both boxes, which is a sign that there is $ 0, taking both boxes doesn't causally bring about that there is $ 0 in B2. Yet the decision maker should take both boxes, for he knows that taking both boxes will causally bring about a wanted possible outcome, namely having the extra $ 1,000, and he knows that taking both boxes will not causally bring about a possible outcome he doesn't want to have, namely $ 0 in B2.

**Ad 5**: That Jeffrey's logic of decision has a set of axioms on rational preference and a representation theorem is an advantage of Jeffrey's theory. For it provides a logical foundation for his rational decision theory. Furthermore, all the other

rational decision theories, which propose a solution to Newcomb's problem, don't have a set of axioms on rational preference and a representation theorem. An exception is Skyrms' (1980) causal decision theory. Armendt (1986) provides a set of axioms on rational preference and a representation theorem for Skyrms' theory; Armendt's representation theorem is built on Fishburn's (1973) rational decision theory. But whereas decision principles are necessary for supplying solutions to Newcomb's problem, axioms on rational preference and representation theorems are not necessary for this purpose. Therefore I will not dwell on the topic of which axioms on rational preference and representation theorems are adequate for a rational decision theory, and whether Jeffrey's axioms and his representation theorem are adequate for a rational decision theory.

Joyce would probably object to the claim that axioms on rational preference and a representation theorem aren't necessary for providing a solution to Newcomb's problem. For in Joyce's (in press, p. 82) opinion a perfect representation theorem would lay down a set of axioms on rational preference which make certain local preferences rational and others irrational, and it would show that these axioms are individually necessary and jointly sufficient for the existence of the utility representation. Joyce postulates that this would make us understand what the global requirement to maximise utility demands at the level of individual preferences. Furthermore, it would make it possible for utility maximisers to justify their decisions by relying on the plausibility of the local axioms rather than by relying on the global utility principle itself or to offer a symbiotic justification of the local and global requirements by appealing to the independent plausibility of each. Thus axioms on rational preference and perfect representation theorems could provide a more thorough justification for a solution to Newcomb's problem than maximising principles.

Yet Joyce (in press, p. 82) concedes that no currently available representation theorem is perfect. For all currently available representation theorems guarantee representability by including axioms on rational preference that are not strictly required for the existence of the utility representation, Joyce claims. These non-necessary axioms are existential statements that concern the size and complexity of preference orderings; Suppes (1974, pp. 136-144) calls them structure axioms (cf. Joyce in press, p. 82). But there are also necessary axioms which Suppes (1974, pp. 136-144) calls axioms of pure rationality. According to Joyce these axioms are universally quantified statements which don't force preference orderings to be defined over large or complex sets of propositions and which cannot be violated if the wanted representation is to

exist. And although Joyce (in press, p. 229) provides a representation theorem for both evidential decision theory and causal decision theory, his representation theorem for evidential decision theory relies on two non-necessary structure axioms, and his representation theorem for causal decision theory is built on his representation theorem for evidential decision theory. Thus Joyce's representation theorems aren't perfect either. Therefore all the currently available representation theorems are not apt for providing a more thorough justification for a solution to Newcomb's problem than the maximising principles.

Besides not being perfect the following can be said about Jeffrey's (1965) and Armendt's (1986) representation theorems: In Jeffrey's theory (1983, chapter 6) the decision maker's credence function and utility function are determined by his preference ordering of propositions; yet the decision maker's preference ordering of propositions determines his utility function only up to a fractional linear transformation with a positive determinant and determines his credence function only within a certain quantisation. Thus Jeffrey's representation theorem is marked by non-uniqueness. In opposition to that Armendt's (1986) representation theorem is marked by uniqueness. For in Armendt's representation theorem the decision maker's utility function is unique up to a positive linear transformation, and the decision maker's credence function is uniquely determined. Thus with regard to uniqueness Armendt's representation theorem is to be preferred to Jeffrey's. Yet with regard to Newcomb's problem non-uniqueness respectively uniqueness doesn't make a difference for a solution to Newcomb's problem.

Furthermore, Armendt's (1986) representation theorem uses mitigators which according to Joyce (in press, p. 228) is a disadvantage. For in some cases there are no mitigators; furthermore, mitigators don't belong to the axioms of pure rationality, so Joyce. A mitigator is a possible action which is able to offset things which might occur, for example, a possible action whose performance can make a decision maker indifferent between the prospect of an asteroid destroying all life on earth in the next five minutes and the prospect of peace and prosperity for a millennium. Armendt's (1986) concept of a mitigator stems from an axiom of Fishburn's (1973) rational decision theory which says that for any two events $E$ and $F$ there is a possible action $A$, so that the decision maker prefers $A$ given $E$ to $A$ given $F$. But if you suppose that $E$ = "An asteroid hits the earth in the next five minutes and destroys all life on earth." and that $F$ = "There is peace and prosperity on earth for a millennium.", there is no $A$ (Joyce in press, p. 228). Therefore in some cases there are no mitigators, and Fishburn's axiom

is false. Moreover, Fishburn's axiom is a structure axiom which only deals with the size and complexity of preference orderings. Therefore Armendt's representation theorem can be criticised for using mitigators. In opposition to Armendt Jeffrey doesn't use mitigators in his representation theorem. Thus in this respect Jeffrey's representation theorem is to be preferred to Armendt's representation theorem. Whether a representation theorem uses mitigators is of no relevance for a solution to Newcomb's problem, though. For the decision maker just has to decide between two possible actions.

Whether non-uniqueness or mitigators is the greater disadvantage of a representation theorem is difficult to say. Yet both representation theorems, Jeffrey's and Armendt's, have their problems and are not perfect.

**Ad 6**: Spohn (1977) points out that Jeffrey's logic of decision leads to inconsistencies, when the decision maker is certain which possible action to decide for. For if there are two possible actions $A$ and $\neg A$ and the decision maker ascribes a credence of 1 to $A$, then according to Jeffrey's logic of decision $U(A) = U(A \cup \neg A)$ and $U(\neg A)$ is undefined. The problem consists in the fact that if the decision maker is aware of his preference for $A$ over $A \cup \neg A$, this preference reduces to indifference. Jeffrey describes this problem as a paradox (Jeffrey 1977, p. 136):

> "Suppose you think it within your power to make A true if you wish, that you prefer A to W, and that you are convinced that A is preferable to every other one of your options. Then P(A) = 1, for you know you will make A true. But then, setting B = -A in (3), we have U(W) = U(A), so that by (1) you are indifferent between A and W after all, in contradiction to the assumption that you prefer A to W."[96]

Jeffrey comments this passage: If the decision maker knows that he prefers $A$ to $W$, it doesn't have to be the case that $P(A) = 1$. For on the supposition that the decision maker decides to do $A$, the following can be the case, Jeffrey (1977) claims: The decision maker doesn't view making $A$ true to be under his control, and he doesn't view $A$ as the highest proposition in his preference ordering. Even if the decision maker can make $A$ true, it is possible that $P(A) \neq 1$ (Jeffrey 1977); a reason for this is that the decision maker can deem it possible that his preferences change before he performs $A$, so that $A$ isn't the highest proposition in his preference ordering at action time. Given

---

[96] $A$ is a subset of $W$ with regard to all possible worlds; (1) says that $U(A) > U(B)$, if $A$ is strictly preferred to $B$, and that $U(A) = U(B)$, if $A$ and $B$ are indifferent to each other; (3) says that $U(A \cup B)P(A \cup B) = U(A)P(A)+U(B)P(B)$, if $A \cap B = \emptyset$.

that everything is like in the cited passage above and that $P$ and $U$ mirror the preferences of the decision maker at action time, then $P(A) = 1$ and $U(A) = U(W)$, which is unproblematical (Jeffrey 1977). For the decision maker prefers $A$ to every other proposition until action time, and he knows that he will perform $A$. The indifference is a result of the fact that the decision maker is performing $A$ (Jeffrey 1977).

According to Jeffrey (1977) the paradox stems from an incompatibility of the following two conditions of Jeffrey (1965) with the preference relation: (i) The impossible proposition $\varnothing$ is supposed to share its rank with $W$ in every preference ordering, and (ii) $U_A(B)$[97] is supposed to equal $U(A \cap B)$ for every $B$ in the preference ordering, if $P(A) \neq 0$. If both conditions hold, the decision maker is indifferent between all propositions with positive credences, so that the paradox arises. Jeffrey solves the problem by having $\varnothing$ in every preference ordering and by separating $\varnothing$ from $W$, i. e. he drops condition (i) and keeps condition (ii). Jeffrey claims that the paradox arises only diachronically. For if the decision maker persists in being indifferent between $\varnothing$ and $W$ at a particular time, no problem arises. But if the decision maker insists that he is indifferent between $\varnothing$ and $W$ at all times, every conditioning is ruled out. An exception is conditioning by $A$, if $A$ and $W$ share the same rank before conditioning. Therefore Jeffrey drops (i). In my opinion this seems to be okay.

Joyce (in press, p. 231) points out another problem with Jeffrey's logic of decision: If the decision maker is certain that he will perform $A$, he must assign the same utility to the other possible actions $B$, $C$, $D$, etc., which are incompatible with $A$, so that the decision maker cannot compare the utilities of possible actions he is certain he will not perform. Thus the most the decision maker can say is that he prefers $A$ to the other possible actions. According to Joyce this can be criticised. For the decision maker's justification of which possible action to decide for often relies on the comparisons between the utilities of our possible actions. The decision maker, for example, can say the following: I am certain that I will do $A$, because I am aware that $A$'s utility is higher than $B$'s utility, that $B$'s utility is higher than $C$'s utility, that $C$'s utility is higher than $D$'s utility, etc., where $B$, $C$, and $D$ are possible actions the decision maker is certain he will not perform. In Jeffrey's logic of decision the decision maker can only say that $A$'s utility is higher than $B$'s utility, but he cannot make the other two statements, so Joyce (in press). Thus Jeffrey's formula for calculating the conditional

---

[97] "$U_A(B)$" means "$U(B)$ given $A$".

utility of a possible action must be rewritten, so that $U(B)$, $U(C)$, $U(D)$, etc. are well defined, even if $P(B) = 0$, $P(C) = 0$, $P(D) = 0$, etc.

Joyce solves this problem by using Renyi-Popper measures (cf. Spohn 1986) instead of Jeffrey's conditional probabilities in Jeffrey's formula for calculating the conditional utility of a possible action. If $P(A|B)$ is a Renyi-Popper measure, $P(A|B)$ can be well defined and positive, even if $P(B) = 0$. According to Spohn (1986, p. 69) a Renyi-Popper measure is a conditional probability, which is construed as a fundamental, and not as a derived concept. Because I am not sure how learning from experience, which I deem to be a part of our concept of rationality, is possible with regard to Renyi-Popper measures, I don't know whether Renyi-Popper measures are a solution to Jeffrey's problem. For one can condition Renyi-Popper measures, but one doesn't get a new Renyi-Popper measure after conditioning.[98] With regard to Jeffrey's conditional probabilities learning from experience is expressed by means of conditioning; furthermore, one gets well-defined probabilities after conditioning in Jeffrey's theory. Thus no problem arises for Jeffrey's logic of decision. Therefore even if Renyi-Popper measures solve the problem of probabilities given propositions having the probability of 0 in Jeffrey's logic of decision, some other problems like how to learn from experience may arise.[99]

But there is another way out of the dilemma: One could also argue that the decision maker should never assign a credence of 1 or 0 to a proposition. Or to say it in Lewis' (1981a, p. 14) words: "Absolute certainty is tantamount to a firm resolve never to change your mind no matter what, and that is objectionable." And this even holds for partly rational decision makers, so Lewis (1981a). For a decision maker, whether partly rational or fully rational, ascribes a particular credence to a proposition following a systematic pattern of ascriptions, which contains ascriptions to himself, ascriptions to others like him, actual ascriptions, and ascriptions which would have been, if events had gone differently. Lewis doubts that absolute certainties could be part of such a best pattern. Thus Jeffrey's formula for calculating the conditional utility of a possible action must not be rewritten.

---

[98]I would like to thank Wolfgang Spohn for drawing my attention to this fact. Spohn's solution to the problem of the Renyi-Popper measures is his theory of ordinal conditional functions (Spohn 1988).

[99]Another question with regard to Renyi-Popper measures is whether they fulfil the axioms of Kolmogorov (1933). According to Spohn (1986) this is the case.

A last problem with Jeffrey's (1965) formula for calculating the utility of a possible action is[100]: The more probable a possible action becomes, that is the higher the decision maker's credence for a possible action is, the lower the utility of that particular possible action becomes. This seems to run against our intuitions. For we actually think that the utility of a possible action should remain the same no matter what credence the decision maker ascribes to that particular possible action. But there is an explanation for this phenomenon. For Jeffrey's utility scale is an interval scale, so that the 0-point of the utility scale can change during deliberation. And as the decision maker's credence of a particular possible action reaches 1, some of the other possible actions under consideration reach a credence of 0, so that they drop out of the preference ordering and the 0-point of the utility scale changes. Thus the phenomenon that the more probable a possible action becomes, the lower the utility of that particular possible action becomes, is due to a change of the 0-point of the utility scale during deliberation.

The only relevance of this phenomenon for Newcomb's problem is that the more probable a possible action becomes like the more probable it becomes to take both boxes respectively to take B2, the lower the utility of the possible action to take both boxes respectively to take B2 becomes. And as the decision maker's credence of the possible action to take both boxes respectively to take B2 reaches 1, the possible action to take B2 respectively to take both boxes reaches a credence of 0, so that they drop out of the preference ordering and the 0-point of the utility scale changes.

## *A Critique of Jeffrey's Logic of Decision*

If the question is how to rationally decide between possible actions, it seems more plausible to rely on logic, which is used in Jeffrey's (1965) logic of decision, than on gambles, which are used in Ramsey's (1931) and Savage's (1954/1972) rational decision theories. For gambling in contrast to doing logic isn't generally associated with being rational. Thus intuitively Jeffrey's idea is a very good one.

Yet Jeffrey's logic of decision has some serious defects: (1) Causation doesn't figure in the theory as a primitive term. (2) Jeffrey's logic of decision is partition invariant. (3) The decision maker can assign credences to his possible actions. (4) Jeffrey's logic of decision establishes preferences between news items not

[100]I would like to thank Hans Rott for drawing my attention to this fact.

distinguishing between making good news and making good results. (5) Jeffrey's representation theorem isn't perfect; furthermore, it is marked by non-uniqueness.

But Jeffrey's logic of decision has also some advantages: (1) Jeffrey's logic of decision has a unified ontology. For it uses propositions as its primitive terms instead of using and distinguishing between possible actions, possible states of the world, and possible outcomes. (2) Jeffrey's representation theorem isn't marked by mitigators like Armendt's (1986) representation theorem.

Because Jeffrey's later proposals (Jeffrey 1983, 1988, 1996) and Eells' (1981, 1982, 1985) theory of the common cause are based on Jeffrey's (1965) logic of decision, the central features of Jeffrey's logic of decision (with the exception of feature 1, namely that Jeffrey's theory is non-causal, and the exception of feature 6, namely that the decision maker should follow the principle of maximising conditional utility) can also be found in his later proposals (Jeffrey 1983, 1988, 1996) and in Eells' (1981, 1982, 1985) theory. Thus the positive or negative criticism which I mentioned with regard to these features also holds with regard to Jeffrey's later proposals and Eells' theory.

With regard to Newcomb's problem Jeffrey's logic of decision provides a solution which is regarded as inadequate even by Jeffrey (1983, 1988, 1996). This solution is inadequate, because it gives the wrong recommendation in Newcomb's problem, namely to take B2. And this wrong recommendation results mainly from two defects of Jeffrey's logic of decision: (1) Causation doesn't figure as a primitive term in his theory. (2) Every partition of the possible states of the world is permitted. Thus to get a right recommendation for Newcomb's problem at least these two defects have to be overcome.

## 2.3 Jeffrey's Ratificationism

Newcomb's problem was considered by Jeffrey as a counterexample to his logic of decision. Jeffrey (1983) gives two reasons for this judgement: (1) He claims (1983, p. 18) that his principle of maximising conditional utility produces bad advice in Newcomb's problem, namely not to take the extra $ 1,000. (2) Jeffrey (1983, p. 20) points out that his logic of decision is wrong in cases in which decisions are seen as evidence of possible states of the world which the decided for possible actions cannot promote or prevent.

In Jeffrey (1983) he modified his rational decision theory by adopting ratificationism which in some aspects was inspired by Eells (1981, 1982). With regard to this new theory Jeffrey (1983, p. 23) writes:

> "My two sorts of probabilities - hypothetical on final choices, and otherwise - were inspired by Eells's type *B* and type *A* beliefs, but whereas Eells's contrast is between the agent's beliefs about his own predicament and about the predicaments of all people in his sort of trouble, mine is between the agent's conditional and unconditional beliefs about his own predicament. Although inspired by it, ratificationism differs from Eells's approach, and must be judged independently."

What is the central idea of ratificationism? Jeffrey (1983, p. 16) gives the following answer:

> "A ratifiable decision is a decision to perform an act of maximum estimated desirability relative to the probability matrix the agent thinks he would have if he finally decided to perform that act."

With regard to this definition, which has final decisions as a new primitive[101], Skyrms (1990b, p. 45) complains that it is imprecise within Jeffrey's system. Rabinowicz (1988, p. 411) suggests the following more precise definition:

> "... a decision to perform *A* is *ratifiable* iff the expected desirability of *A* given dA is at least as high as the expected desirability of any alternative action *B* given dA."[102]

The corresponding decision principle is that the decision maker should make a ratifiable decision (= principle of ratifiability) (Jeffrey 1983, p. 16).

## *The Concept of Ratifiability*

In the following I will try to find out whether Rabinowicz' (1988) interpretation of Jeffrey's definition of ratifiability is correct. Jeffrey (1983) motivates his ratificationism by claiming that the decision maker's decision changes his own credences, so that, for example, in one variation of the prisoner's dilemma the decision maker's credence for confessing ($a_1$) given that he decides to confess ($da_1$) gets close to 1, that is $c(a_1|da_1)$ = [1]. It shouldn't reach 1, for the decision maker should know

---

[101] Sobel (1988b, p. 116) and Rabinowicz (1988) point out that ratificationism has decisions as a new primitive.

[102] "dA" denotes the proposition that *A* is the possible action that the decision maker will finally decide to perform (Rabinowicz 1988, p. 410).

that it is always possible that he cannot carry out his decision (Jeffrey 1983, pp. 15-16).[103]

Jeffrey (1983) maintains that for Newcomblike problems the decision maker's credences for the possible states of the world given the possible actions, that is $c(s_j|a_i)$, are changed by the decision maker's decision. Jeffrey (1983) provides a variation of the prisoner's dilemma to exemplify how the credences of the decision maker change: The following figure presents the decision matrix of the prisoner's dilemma of the possible outcomes for decision maker $x_1$. According to the principle of strong dominance decision maker $x_1$ should confess, for the utility of a 4 year sentence is better than the utility of a 10 year sentence, and for the utility of no sentence is better than the utility of a 1 year sentence.

| | | $x_2$ | |
|---|---|---|---|
| | | $a_1$: I confess at $t_1$. | $a_2$: I don't confess at $t_1$. |
| $x_1$ | $a_1$: I confess at $t_1$. | 4 year sentence | 0 year sentence |
| | $a_2$: I don't confess at $t_1$. | 10 year sentence | 1 year sentence |

Figure 9. Decision matrix of the prisoner's dilemma of the possible outcomes for decision maker $x_1$ which result from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

According to the principle of maximising conditional utility decision maker $x_1$ shouldn't confess, if he sees his decisions as evidence of the other decision maker's ($x_2$) corresponding possible actions and therefore assigns high credences, that is p and q, to the corresponding decisions and possible actions (cf. figure 10).

[103]Rabinowicz (1988, p. 410) has rightly observed that one has to distinguish between not carrying out one's decision and carrying out another decision instead. For one can question that the decision maker can fail to carry out his decision by carrying out another decision instead.

| | | $x_2$ | |
|---|---|---|---|
| | | $a_1$: I confess. | $a_2$: I don't confess. |
| $x_1$ | $a_1$: I **decide** to confess. | **p** | 1-p |
| | $a_2$: I **decide** not to confess. | 1-q | **q** |

Figure 10. Credence matrix of the prisoner's dilemma for decision maker $x_1$ which results from combining the decisions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

For decision maker $x_1$ the utilities of his decisions are:

$U(da_1) = p(-4)+(1-p)0,$

$= -4p,$

$U(da_2) = (1-q)(-10)+q(-1),$

$= 9q-10.$

And $U(da_1) < U(da_2)$, if

$-4p < 9q-10,$

$4p > 10-9q,$

$p > 1/4(10-9q).$

Thus $U(da_1) < U(da_2)$, if p and q are both greater than 10/13 (for example).

If decision maker $x_1$ is nearly certain that his future possible actions are evidence of his corresponding past final decisions, that is if his credences look like the credences of figure 11, then p and q of figure 10 will approximate the credences, which decision maker $x_1$ assigns to the possible actions of decision maker $x_2$, whatever it is that decision maker $x_1$ does, so Jeffrey (1983, pp. 16-17). Thus for decision maker $x_1$ the utilities of his decisions approximate the utilities of his possible actions. Therefore decision maker $x_1$ shouldn't confess.

| | | $x_1$ | |
|---|---|---|---|
| | | $a_1$: I **decided** to confess. | $a_2$: I **decided** not to confess. |
| $x_1$ | $a_1$: I **will** confess. | **0.95** | 0.05 |
| | $a_2$: I **will** not confess. | 0.03 | **0.97** |

Figure 11. Credence matrix of the prisoner's dilemma for decision maker $x_1$ which results from combining the future possible actions of decision maker $x_1$ with his past final decisions.

But according to the principle of ratifiability the decision maker $x_1$ should confess, for on each hypothesis of his final decision the decision maker $x_1$ regards his possible actions as predictively irrelevant to the possible actions of decision maker $x_2$ (= screening assumption) and sees that during his deliberation, so that the credences are the same in both rows in figure 12 and figure 13.

| | | $x_2$ | |
|---|---|---|---|
| | | $a_1$: I confess. | $a_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | **p** | 1-p |
| | $a_2$: I don't confess. | **p** | 1-p |

Figure 12. Credence matrix of the prisoner's dilemma for decision maker $x_1$ on the hypothesis that his final decision is to confess which results from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

| | | $x_2$ | |
|---|---|---|---|
| | | $a_1$: I confess. | $a_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | 1-q | **q** |
| | $a_2$: I don't confess. | 1-q | **q** |

Figure 13. Credence matrix of the prisoner's dilemma for decision maker $x_1$ on the hypothesis that his final decision is not to confess which results from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

Whereas in the logic of decision the decision maker's possible actions are evidence of the possible states of the world, or of the possible actions of any other decision maker, ratificationism suggests that the decision maker's final decisions are evidence of the possible states of the world, or of the possible actions of any other decision maker, which the decision maker's possible actions cannot causally influence, so the decision maker believes. Therefore the utility of a possible action on the hypothesis of a final decision to perform a possible action is calculated in the following way ("$s_j$" denotes the possible states of the world, or the possible actions of any other decision maker depending on the respective decision problem): If $c(da_i) > 0$, then

$$U(a_i|da_i) = \sum_{j=1}^{m} c(s_j|da_i)u(o_{ij}).$$

For decision maker $x_1$ the utilities of his possible actions on the hypothesis of his final decision to confess are:

$U(a_1|da_1)$ = p(-4)+(1-p)0,
= -4p,
$U(a_2|da_1)$ = p(-10)+(1-p)(-1),
= -9p-1.

For decision maker $x_1$ the utilities of his possible actions on the hypothesis of his final decision not to confess are:

$U(a_1|da_2)$ = (1-q)(-4)+q(0),
= 4q-4,
$U(a_2|da_2)$ = (1-q)(-10)+q(-1),
= 9q-10.

These calculations yield the following recommendation: According to Jeffrey (1983) ratifiability is a classificatory notion, not a comparative one, so that the utilities of the decision maker's possible actions on the hypothesis of a final decision to perform a certain possible action should be compared with each other. Therefore $U(a_1|da_1)$ should be compared with $U(a_2|da_1)$, and $U(a_1|da_2)$ should be compared with $U(a_2|da_2)$. This results in: $U(a_1|da_1) > U(a_2|da_1)$, if p varies from 0 to 1. And $U(a_1|da_2) > U(a_2|da_2)$, if q varies from 0 to 1.

For finding out which definition of a ratifiable decision is in Jeffrey's (1983) sense adequate, consider Jeffrey (1983, pp. 19-20):

> "On the hypothesis that *that* option will finally be chosen, estimate the desirabilities of actually carrying it out, and of actually carrying out each of the alternatives. The option in question is ratifiable or not depending on whether or not the expected desirability of actually carrying it out (having chosen it) is at least as great as the expected desirability of actually carrying out each of the alternatives. ... suppose that among the performances you might choose none has higher estimated desirability on the hypothesis that *A* is chosen than *A* itself has. Then, and only then, is *A* ratifiable."

From this passage we can extract a definition of a ratifiable decision:

> A decision to perform the possible action $a_i$ is ratifiable, if and only if the utility of $a_i$ on the hypothesis that the decision maker's final decision is to perform the possible action $a_i$ is at least as high as the utility of any other possible action in $\{a_1, a_2, ..., a_n\}\backslash\{a_i\}$ on the hypothesis that the decision maker's final decision is to perform the possible action $a_i$.

This definition of a ratifiable decision agrees with Rabinowicz' (1988) interpretation of Jeffrey's definition of a ratifiable decision. Applying this definition to our example we get the following result: The decision maker's decision to confess is ratifiable, for the utility of confessing on the hypothesis that the decision maker's final decision is to confess is higher than the utility of not confessing on the hypothesis that the decision maker's final decision is to confess.

With regard to Newcomb's problem we just have to exchange $x_1$'s possible actions in Jeffrey's (1983) variation of the prisoner's dilemma $a_1$: I confess. vs. $a_2$: I don't confess. with the decision maker's possible actions in Newcomb's problem $a_1$: I take both boxes at $t_3$. vs. $a_2$: I take B2 at $t_3$. and $x_2$'s possible actions $a_1$: I confess. vs. $a_2$: I don't confess. with the possible states of the world in Newcomb's problem $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$, so that the decision maker's possible outcomes result as a combination of the possible actions and the possible states of the world in Newcomb's problem. Then we calculate the decision maker's utilities of his possible actions on the hypothesis of his final decision to take both boxes and apply the principle of ratifiability, so that the decision maker's decision to take both boxes results

as rational. For the utility of taking both boxes on the hypothesis that the decision maker's final decision is to take both boxes is higher than the utility of taking B2 on the hypothesis that the decision maker's final decision is take both boxes.

## *Problems of Jeffrey's Ratificationism*

Jeffrey (1983) acknowledges that his ratifiability approach has several weaknesses: First, there are decision problems where no decision is ratifiable; Jeffrey's green-eyed monster is such a case (Jeffrey 1983, p. 18): "Where the agent must choose one of two goods and see the other go to someone else, greed and envy my conspire to make him rue either choice."

Second, there are decision problems where all decisions are ratifiable; Jeffrey's triumph of the will-problem is such a case (Jeffrey 1983, p. 19): "A madly complacent agent could find all acts ratifiable because with him, choice of an act always greatly magnifies his estimate of its desirability - not by changing probabilities of conditions, but by adding a large increment to each entry in the chosen act's row of the desirability matrix." In these cases, in which the number of ratifiable decisions isn't exactly 1, the decision maker should re-evaluate his beliefs and wants before making a decision, so Jeffrey (1983).

Third, according to van Fraassen (cf. Jeffrey 1983, p. 20)[104] there are also decision problems where choiceworthy possible actions aren't ratifiable. For there are decision problems where possible actions are better than final decisions as evidence of possible states of the world, of possible actions of the other decision maker, or of possible actions of the other decision makers.

Van Fraassen's example (cf. Jeffrey 1983, p. 20)[105] is the situation in figure 12 and 13, where the credences of decision maker $x_1$ are the same in both rows. On each hypothesis of his final decision the decision maker $x_1$ regards his possible actions as predictively irrelevant to the possible actions of decision maker $x_2$ (= screening assumption) and sees that during his deliberation. But if other persons influence decision maker $x_1$ in such a way that he cannot carry out his decision, this influence might also work on decision maker $x_2$. Thus the credences of decision maker $x_1$ might differ in

[104] Jeffrey (1983) doesn't give a reference for van Fraassen's objection.

[105] A better example is in my opinion the following (Gärdenfors and Sahlin 1988, p. 338): A smoker decides on New Year's Eve to quit smoking, but on New Year's Day he finds himself lighting a cigarette. Thus in some cases the decision maker learns by his possible actions and not by his decisions what kind of person he is.

both rows in figure 12 and 13. This reflects the fact that the possible action of decision maker $x_1$ is better evidence for the possible action of decision maker $x_2$ than the final decision of decision maker $x_1$ to perform the possible action. Thus the decision maker's decision to confess is unratifiable, although it is the rational decision.

This last weakness of Jeffrey's ratificationism is the worst of all, for it goes against Jeffrey's (1983) doctrine of ratificationism, namely that the choiceworthy possible actions are the ratifiable decisions to perform. This problem has led Jeffrey to revise his preference theory a second time (Jeffrey 1988).

## *A Critique of Jeffrey's Ratificationism*

Jeffrey's ratificationism, which makes use of the gap between decision-making and performing one's possible action, can be compared with Eells' (1981, 1982, 1985) proposal of the common cause. For in Jeffrey's ratificationism the decision maker's decision and not his possible action already tells the decision maker which possible state of the world obtains, whereas in Eells' (1981, 1982, 1985) proposal of the common cause the decision maker's wants and beliefs and not his possible actions already tell the decision maker which possible states of the world obtains.[106] Furthermore, at decision time the decision maker's credence function has been modified by the decision maker's decision (= the tickle) in Jeffrey's ratificationism and by the decision maker's wants and beliefs (= the tickle) in Eells' (1981, 1982, 1985) proposal, so that the decision maker's calculation of the utility of a possible action agrees with the recommendations of causal decision theory.

According to Skyrms (1990b, p. 53) Jeffrey's ratificationism is valuable for opening up the following topics for investigation: (1) The status of equilibrium as a rationality concept, (2) the nature of the deliberational process, (3) the connections between rational decision theory and game theory, and (4) the consequences of self-knowledge for decision makers. With regard to (1) and (3) Skyrms states: Ratifiability is an equilibrium concept in rational decision theory.[107] Together with the common

---

[106]Thus Jeffrey's ratificationism is another sort of Tickle Defence (cf. chapter 2.6, in the section on a critique of Eells' position). Lewis' (1981a) term "Tickle Defence", which he uses for Eells' 1981) position, derives from the fact that in Eells' (1981) theory the common cause causes beliefs and wants which manifest themselves, for example, in tickling taste buds in Skyrms' (1980, pp. 128-129) eggs-benedict-for-breakfast-problem. This tickle already tells the decision maker which possible state of the world obtains, so that he doesn't have to wait for his decision to find out which possible state of the world obtains, Lewis (1981a) claims.

[107]Unfortunately Skyrms (1990b) doesn't explain what an equilibrium concept is. Wolfgang Spohn pointed out to me (personal communication from the 1st of June, 1999) that something is

knowledge assumption (cf. Aumann 1987) of game theory it generates the main equilibrium concepts of game theory (Skyrms 1990b, pp. 49-51). Ad (2) and (4) Skyrms (1990b) maintains: If the decision maker has so much self-knowledge that he foresees that no decision is ratifiable, that is if a decision maker anticipates his deliberational process which results in no ratifiable decision, he may proceed by applying Jeffrey's (1965) principle of maximising conditional utility. In my opinion this latter reasoning can be transferred to cases in which all decisions are ratifiable. Yet actually Jeffrey (1983) gives instructions what to do in cases in which no decision or all decisions are ratifiable. He says that the decision maker should reassess his wants and beliefs in these cases. Unfortunately Jeffrey doesn't tell us what to do in cases in which no decision or all decisions remain ratifiable after the decision maker's reassessment of his wants and beliefs.

With regard to (2) and (4) the following can be said independently from Skyrms: The ratifiability maxim demands from the decision maker that he should decide for the person he expects to be when he has made a decision. Thus in Jeffrey's ratificationism the decision maker is supposed to make his decision on the basis of his "would-be *after-choice* probabilities", so Rabinowicz (1985, p. 195). In opposition to Jeffrey's ratificationism the decision maker should decide for the person he is before he has made a decision in Savage's (1954/1972) rational decision theory. Thus the decision maker is supposed to make his decision on the basis of his "*pre-choice* probability assignments" in Savage's (1954/1972) rational decision theory (Rabinowicz 1985, p. 195).

Rabinowicz' distinction is very important. For it points out that the deliberational process could take several forms. Should the decision maker take his would-be after-choice probabilities as a basis for his decision, or should he take his pre-choice probabilities, or should he take both probabilities as a basis for his decision which seems the most reasonable account because of its completeness (cf. Rabinowicz 1985, p. 195)? Yet in my opinion the latter proposal doesn't seem to recommend itself, if one takes into account that an enormous amount of self-knowledge is demanded from the decision maker who is supposed to make his decision on the basis of his would-be after-choice probabilities and his pre-choice probabilities. But perhaps there is a way how

---

in an equilibrium state with itself if and only if it has reached a state from which it cannot be driven out by internal forces. Applying this to Jeffrey's (1983) ratificationism a ratifiable decision is in an equilibrium state with itself, given that this particular decision will be performed, if and only if there are no reasons for evaluating any other decision as the better one.

pre-choice probabilities becoming would-be after-choice probabilities could be modelled decision-theoretically.

Rabinowicz' distinction also shows that Jeffrey's ratificationism demands much more self-knowledge from the decision maker than Savage's rational decision theory. Furthermore, one could claim that there are also decision makers who don't have as much self-knowledge as is demanded from them in Jeffrey's ratificationism. For there are also dithery and self-deceptive decision makers.[108] Or to say it in other words: How should I know what kind of person I will be when I have made a decision? Especially, how should I know what kind of person I will be when I have made a decision in Newcomb's problem? Because the decision maker hasn't been exposed to Newcomb's problem before, because particular expectations need certain numbers of trials to build up (cf. Ledwig 1994), and because expectations can be wrong, it seems to me very difficult for the decision maker to know what kind of person he will be when he has made a decision in Newcomb's problem. Thus Jeffrey's ratificationism demands from the decision maker that he has a certain ability to anticipate his own decisions; as it seems to me this ability to anticipate his own decisions could in certain cases like, for example, in Newcomb's problem be too much to demand from the decision maker.

Another reason why Rabinowicz' distinction is valuable is that his distinction can be applied to other rational decision theories. And as a matter of fact all other rational decision theories which provide a solution to Newcomb's problem rely on pre-choice probability assignments of the decision maker.

Jeffrey's (1983) weaknesses of his ratificationism deserve some comments: With regard to the first weakness, namely that there are decision problems in which no decision is ratifiable, Rabinowicz (1988, p. 424) claims that completeness for a rational decision theory isn't required, although it would be better, if it could be had. Yet a real problem for Jeffrey's ratificationism would arise, if there were cases in which no decision were ratifiable, but a rational possible action existed.

With regard to the second weakness, namely that there are decision problems in which all decisions are ratifiable, one could maintain that this would be only insofar a weakness as there are cases in which one decision is rational, although all decisions are ratifiable. For there is nothing unusual with several decisions being ratifiable. In such a case the decision maker just has to decide for one of the ratifiable decisions. Rabinowicz (1985) would probably object to that. For this depends on how we interpret

---

[108] Cf. chapter 2.6, in the section on a critique of Eells' position.

Jeffrey's ratificationism. According to Rabinowicz (1985) there are two interpretations of ratificationism possible: (1) "An action available to the agent is choiceworthy iff (the decision to perform) it is ratifiable." (Rabinowicz 1985, p. 180) (2) "If, in a given case, the set of ratifiable actions available to the agent is neither empty nor involves a ratifiability conflict, then an action available to the agent is choiceworthy iff (the decision to perform) it is ratifiable." (Rabinowicz 1985, p. 181) The second interpretation arises, because one could claim that Jeffrey intended his ratificationism not to apply to cases in which no decision is ratifiable or all decisions are ratifiable. For in these cases there are either pathological wants and beliefs of the decision maker present or the possible states of the world are cruel (cf. Jeffrey 1983, p. 19). Thus in cases in which all decisions are ratifiable and one opts for the first interpretation of ratificationism, the decision maker just has to decide for one of the ratifiable decisions; while in cases in which all decisions are ratifiable and one opts for the second interpretation of ratificationism, the decision maker cannot decide for one of the ratifiable decisions. And in cases in which no decision is ratifiable and one opts for the first interpretation of ratificationism, the decision maker cannot decide for any possible action, because there is no ratifiable decision; while in cases in which no decision is ratifiable and one opts for the second interpretation of ratificationism, the decision maker cannot decide for any possible action, because the second interpretation of ratificationism doesn't apply to cases in which no decision is ratifiable.

With regard to the third weakness Sobel (1994, p. 71) has rightly observed that van Fraassen's prisoner's dilemma isn't very plausible. For Sobel rightly questions what kind of extraneous factors could prevent the decision maker from confessing when the decision maker has made a final decision to confess. In another connection Jeffrey (1983, p. 18) mentions death, or a non-fatal cerebral haemorrhage as examples for these extraneous factors; yet Sobel (1994) rightly deems it unplausible that the decision maker should think that these things would happen to him if and only if they happened to the other decision maker in the prisoner's dilemma. Nevertheless Sobel shows that there are more plausible examples for illustrating the third weakness of Jeffrey's ratificationism.

Rabinowicz (1985, p. 172) claims that Jeffrey's ratificationism has a fourth weakness. For its applicability has to be restricted in the following way: We have to exclude cases in which the hypothesis screens off not only the evidential utility of a given possible action, but also some of its causal utility. According to Rabinowicz (1985) this will happen whenever the possible action in question is expected to have certain

causal effects that would be brought about by the decision to perform it. Rabinowicz (1985) provides the following example to illustrate this point: Deciding to help my neighbour will make me feel good quite independently of whether I will manage to carry out this decision. In this case it is unreasonable to screen off this benefit. Yet the ratifiability maxim demands exactly this. Therefore Rabinowicz is right in claiming that Jeffrey's ratificationism has to exclude cases like this one from its range of applicability.

Another critical point with regard to Jeffrey's ratificationism is that Jeffrey introduces another primitive term, namely final decisions, which makes Jeffrey's ratificationism as complicated as causal decision theory with regard to the number of primitive terms. Thus for economical reasons Jeffrey's ratificationism isn't to be preferred to causal decision theory.

With regard to Newcomb's problem Jeffrey's ratificationism proposes the right solution (cf. chapter 1.3 and 2.2). Yet Jeffrey's ratificationism demands too much self-knowledge from the decision maker in Newcomb's problem. For the decision maker is supposed to know what kind of person he will be when he has made his decision, which is an unreasonable demand for some decision makers and/or for decision makers faced with Newcomb's problem. Furthermore, Jeffrey's theory is limited in its applicability because of the third and the fourth weakness. And actually we want to have a rational decision theory which is applicable in all cases. Moreover, Jeffrey's theory allows the possibility of two interpretations of ratificationism in cases in which no decision is ratifiable or all decisions are ratifiable. And in fact we demand from a rational decision theory that it is unambiguous. Finally, we demand from a rational decision theory that it is economical. Yet Jeffrey's ratificationism introduces a new primitive term, namely final decisions, so that it is as uneconomical as causal decision theory.

## 2.4 Jeffrey's Probabilism

Being aware of the shortcomings of his 1983-theory in 1988 Jeffrey reverted to his preference theory of 1965, but improved it in several ways. Jeffrey (1988, p. 249) writes:

> "In either case, the evidentiary decision theory of the first (1965) edition of *The Logic of Decision* seems to be satisfactory when the Newcomb problem is probabilised on two levels as here - once we see that it's final desirabilities (final preferences) that must rule."

Jeffrey (1988, p. 251) maintains that his probabilism shares the advantages of ratificationism, but avoids the disadvantages of van Fraassen's counterexample to ratificationism.

## *Motivations for Probabilism*

Jeffrey motivates his probabilism by claiming that there are two ways of considering judgements, namely dogmatism and probabilism. Whereas dogmatism states that judgements are and ought to be a matter of assertion and denial, probabilism states that judgements are and ought to be a matter of probabilising. Jeffrey's (1988) opinion is that probabilism is the correct way of considering judgements and illustrates this by the following example: When asked whether it will rain today, I cannot assert it or deny it, but can give a probabilistic judgement only, so that my credence for rain today is, for example, $c = 0.7$. Furthermore, this credence is action-guiding, so that probabilism is relevant for rational decision theory. Jeffrey continues by considering credences of $c = 1$ which correspond to infinite odds. If my credence for rain today is $c = 1$, this means that I would stake everything on its truth for the possibility of getting any, even the smallest gain. Jeffrey claims that we don't assign a credence of $c = 1$ to whatever we state as true, for when we state something as true we don't probabilise, but dogmatise. He illustrates this by the following example: If you ask me "What time is it?" and I answer "It is ten past two." after having looked at my watch and even though I acknowledge the possibility of having misread the time, of having a defect watch, or of other kinds of failure, then I'm dogmatising and not probabilising.[109]

Jeffrey gives the following characterisation of the concept of probabilising (Jeffrey 1988, p. 242):

> "... probabilising is here seen as an art that calls on subject matter dependent skills that need learning and polishing. The artefact is a probability distribution or a partial characterisation of one as meeting certain constraints."

[109]Unfortunately Jeffrey (1988) doesn't give any more justifications for his probabilism. One could stipulate that Jeffrey deems dogmatic beliefs inappropriate as a basis for a rational decision theory and/or that dogmatic beliefs are not action-guiding. But he never explicitly states that.

## *Probabilism and Newcomblike Problems*

This rather general statement will hopefully get a more specific meaning in an application of Jeffrey's theoretical ideas to a variation of the prisoner's dilemma which is construed à la Lewis (1979) as a combination of two Newcomb's problems: Decision maker $x_1$ is called Alma, and decision maker $x_2$ is called the Count. For Alma "$a_1$" respectively "$a_2$" denote "I confess." respectively "I don't confess.", and for the Count "$s_1$" respectively "$s_2$" denote "I confess." respectively "I don't confess.". Alma believes that she and the Count are so similar to each other that she and the Count decide to confess or decide not to confess by drawing with replacement from the same urn which contains an unknown portion of confessing and not-confessing tickets. Furthermore, she believes that the Count believes the same. According to Jeffrey (1988) the composition of the urn represents the decision makers' common final chance of confessing which is determined by the decision makers' common attitudes and similar ways of thinking about their common predicament.

Jeffrey (1988) supposes that Alma's initial credence distribution for the unknown portion $p$ of confessing tickets among all the tickets in the urn is, for example, the uniform distribution over the unit interval. On this supposition Alma's credences for the hypotheses $a_1$ and $a_1 \cap s_1$ are the following ("$c_0$" denotes Carnap's (1950/1962, appendix) function $m^*$ for a language with two individuals (Alma, the Count) and one primitive property (confessing)[110])[111]:

$$\begin{aligned} c_0(a_1) &= \int_0^1 p \, dp, \\ &= 1/2, \\ c_0(a_1 \cap s_1) &= \int_0^1 p^2 \, dp, \\ &= 1/3. \end{aligned}$$

---

[110]Carnap (1950/1962) gives in the appendix an outline of a quantitative system of inductive logic and states that $m^*$ is a $m$-function which serves as the basis of his system of inductive logic and which fulfils the following two conditions: $m^*$ is a symmetrical $m$-function, and $m^*$ has the same value for all structure-descriptions in $L_N$, which are the respective object languages. Carnap (1950/1962, p. 580) explains a $m$-function in the following way: A $m$-function is a "measure function assigning numerical values first to the ... state-descriptions, then to all sentences, representing an ... explicatum for ... probability$_1$ a priori". For more detail cf. Carnap (1950/1962).

[111]With regard to my questions why Jeffrey assumes that Alma's credence distribution for the unknown portion $p$ of confessing tickets among all the tickets in the urn is the uniform distribution over the unit interval and why Jeffrey uses Carnap's function $m^*$ Jeffrey responded (personal communication from the 22nd of September, 1998): "The assumption of a uniform distribution over the unit interval is equivalent to choice of Carnap's $m^*$. I made that assumption for illustrative purposes only: it is a simple assumption that has any plausibility."

From this Jeffrey (1988) calculates Alma's conditional credences and her unconditional credences for the four combinations of her possible actions with the Count's possible actions:

$c_0(s_1|a_1) = c_0(s_2|a_2) = 2/3$,
$c_0(a_1 \cap s_1) = c_0(a_2 \cap s_2) = 1/3$,
$c_0(a_1 \cap s_2) = c_0(a_2 \cap s_1) = 1/6$.

The latter four are represented in figure 14.

| | | $x_2$ | |
|---|---|---|---|
| | | $s_1$: I confess. | $s_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | 1/3 | 1/6 |
| | $a_2$: I don't confess. | 1/6 | 1/3 |

Figure 14. Credence matrix ($c_0$) of the prisoner's dilemma for decision maker $x_1$ which results from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

From this and from Alma's utilities of the possible outcomes (cf. figure 15) Jeffrey (1988) calculates Alma's initial conditional utilities of $a_1$ and $a_2$ on the hypotheses that she does and doesn't confess:

$U_0(a_1) = [1c_0(a_1 \cap s_1)+10c_0(a_1 \cap s_2)]/c_0(a_1) = 4$,
$U_0(a_2) = [0c_0(a_2 \cap s_1)+9c_0(a_2 \cap s_2)]/c_0(a_2) = 6$.

Because $U_0(a_2) > U_0(a_1)$, Alma should decide to not-confess, given that she takes her initial conditional utilities as a basis for her decision.

| | | $x_2$ | |
|---|---|---|---|
| | | $s_1$: I confess. | $s_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | 1 | 10 |
| | $a_2$: I don't confess. | 0 | 9 |

Figure 15. Decision matrix of the prisoner's dilemma of the possible outcomes for decision maker $x_1$ which result from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

Jeffrey (1988), however, believes that Alma's initial conditional utilities are irrelevant for decision-making, for in the credence matrix of figure 14 Alma is far from

decision; her credences for confessing and for not-confessing are equal to each other, they each amount to 0.5. Relevant for decision-making are Alma's final conditional utilities, so Jeffrey (1988).

In the following Jeffrey (1988) examines the case that Alma moves from her initial credence distribution $c_0$ in figure 14, which is derived from a uniform credence distribution for the chance $p$ that she will confess, to credence distributions $c_n$ that push her credence more towards $a_1$ as her credence distribution for $p$ pushes in the unit interval up to 1.

In particular Jeffrey (1988) supposes that at the end of Alma's deliberation her judgement about $p$ will be characterised by a normalised density function $f_n$ which is proportional to $p^n$[112]:

$$f_n(p) = (n+1)p^n.$$

In general her final density $f_n$ for the random variable $p$ determines Alma's final credences for her confessing, for the Count confessing, and for their both confessing in the following way:

$$c_n(a_1) = c_n(s_1) = \int_0^1 p f_n(p)\, dp = (n+1)/(n+2),$$

$$c_n(a_1 \cap s_1) = \int_0^1 p^2 f_n(p)\, dp = (n+1)/(n+3).$$

From this Jeffrey (1988) calculates her conditional final credences for the Count's possible actions given her possible actions, namely

$$c_n(s_1|a_1) = (n+2)/(n+3),$$

$$c_n(s_1|a_2) = (n+1)/(n+3),$$

$$c_n(s_2|a_1) = 1/(n+3),$$

$$c_n(s_2|a_2) = 2/(n+3),$$

and her other unconditional final credences which are presented in figure 17.

---

[112] Jeffrey's reason for this supposition consists in the fact that Alma's initial state has that form with $n = 0$, so that

$$f_0(p) = (0+1)p^0, \quad = 1.$$

| | | $x_2$ | |
|---|---|---|---|
| | | $s_1$: I confess. | $s_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | $p^2$ | $p(1-p)$ |
| | $a_2$: I don't confess. | $(1-p)p$ | $(1-p)^2$ |

Figure 16. Chances of the prisoner's dilemma for decision maker $x_1$ which result from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

| | | $x_2$ | |
|---|---|---|---|
| | | $s_1$: I confess. | $s_2$: I don't confess. |
| $x_1$ | $a_1$: I confess. | $(n+1)/(n+3)$ | $(n+1)/(n+2)(n+3)$ |
| | $a_2$: I don't confess. | $(n+1)/(n+2)(n+3)$ | $2/(n+2)(n+3)$ |

Figure 17. Unconditional credence matrix ($c_n$) of the prisoner's dilemma for decision maker $x_1$ which results from combining the possible actions of decision maker $x_1$ with the possible actions of decision maker $x_2$.

From this and from Alma's utilities of the possible outcomes (cf. figure 15) Jeffrey (1988) calculates Alma's final conditional utilities of $a_1$ and $a_2$:

$U_n(a_1) = 1c_n(s_1|a_1)+10c_n(s_2|a_1) = (n+12)/(n+3)$,

$U_n(a_2) = 0c_n(s_1|a_2)+9c_n(s_2|a_2) = 18/(n+3)$.

And $U_n(a_1)-U_n(a_2) = (n-6)/(n+3)$, which is positive for $n > 6$ and which approaches 1 as a limit as $n$ increases without bound, so that Alma should confess.

Jeffrey's (1988) formula for the calculation of the final conditional utility of a possible action is the following: If $c_n(a_i) > 0$, then

$$U_n(a_i) = \sum_{j=1}^{m} c_n(s_j|a_i)u(o_{ij}).$$

And decision maker $x_1$ should adopt the

> **principle of maximising final conditional utility**: In a given decision situation $D$ the decision maker $x_1$ should decide for a possible action $a_i$ with maximal final conditional utility.

Then Jeffrey (1988) considers the case that Alma decides to not-confess, so that at the end of Alma's deliberation her judgement about $p$ will be characterised by a density function $g_n$ which is proportional to $(1-p)^n$:

$g_n(p) = (n+1)(1-p)^n$.

In general her final density $g_n$ for the random variable $p$ determines Alma's final credences for her not-confessing, for the Count not-confessing, and for their both not-confessing in the following way:

$d_n(a_2) = d_n(s_2) = \int_0^1 p g_n(p)\, dp = (n+1)/(n+2)$,

$d_n(a_2 \cap s_2) = \int_0^1 p^2 g_n(p)\, dp = (n+1)/(n+3)$.

From this Jeffrey (1988) calculates her conditional final credences for the Count's possible actions given her possible actions, namely

$d_n(s_1|a_1) = 2/(n+3)$,

$d_n(s_1|a_2) = 1/(n+3)$,

$d_n(s_2|a_1) = (n+1)/(n+3)$,

$d_n(s_2|a_2) = (n+2)/(n+3)$,

and her other unconditional final credences which are identical with her unconditional final credences in figure 17 except that the entries for $a_1 \cap s_1$ and $a_2 \cap s_2$ are exchanged.

From this and from Alma's utilities of the possible outcomes (cf. figure 15) Jeffrey (1988) calculates Alma's final conditional utilities of $a_1$ and $a_2$:

$V_n(a_1) = 1[2/(n+3)]+10(n+1)/(n+3) = (10n+12)/(n+3)$,

$V_n(a_2) = 0[1/(n+3)]+9(n+2)/(n+3) = (9n+18)/(n+3)$.

And $U_n(a_1)$-$U_n(a_2)$ = $V_n(a_1)$-$V_n(a_2)$ = $(n-6)/(n+3)$, which is positive for $n > 6$ and which approaches 1 as a limit as $n$ increases without bound, so that the principle of maximising final conditional utility would recommend confessing.

Jeffrey (1988) draws the following conclusions: Alma's decision problem is a Newcomblike problem, because she considers her possible actions as evidence of the Count's possible actions. Alma believes that if she knew the chances, her possible actions and the Count's possible actions would be probabilistically independent, that is the chance that they would both confess would be the product of their separate chances for confessing. According to Jeffrey (1988) these separate chances are both $p$.[113]

---

[113] $p$ is a random variable for which Jeffrey (1988) discussed various densities of the beta form, that is densities which are proportional to $p^a(1-p)^b$ with $a = 0$, or $b = 0$, and $a+b = n$. When $b = 0$, the density is $f_n(p) = (n+1)p^n$, and when $a = 0$, the density is $g_n(p) = (n+1)(1-p)^n$. Jeffrey (1988) didn't provide reasons for selecting various densities of the beta form and for not selecting any other densities. Therefore it looks like as if other densities would have served Jeffrey's purposes as well.

Jeffrey (1988, p. 248) advances a 2-level-model for Newcomblike problems in which credences serve as estimates of chances:

> "Newcomb problems require two-level probability models. At the lower level of Alma's problem is the unknown chance *p* that she will confess. At the upper level are her possible judgemental probabilities ($P_n$, $Q_n$) for *A*, *AC*, etc. which are determined by her possible judgemental densities ($f_n$, $g_n$) for $p$. The two layers are needed because it's basically $p$ that she needs to make up her mind about. An acceptable 'final' density for *p* will determine her preferential choice, whether to confess or remain silent, in such a way that her probability for *A* is near 0 (choice of silence) or 1 (choice of confession) depending on whether her final desirability for *A* is less or greater than that for -*A*."[114]

Thus Jeffrey models Alma's prisoner's dilemma in the following way:

[114] "$P_n$" respectively "$Q_n$" are denoted by "$c_0$" respectively "$d_0$", and "*A*" respectively "-*A*" respectively "*AC*" are denoted by "$a_1$" respectively "$a_2$" respectively "$a_1 \cap s_1$".

| | | |
|---|---|---|
| **Start**: Alma's initial judgemental credence distribution for the unknown chance $p$ that she will confess:<br>for example, the uniform distribution over the unit interval, which is the following:<br>$f_0(p)=(0+1)p^0=1$. | → one derives | Alma's final judgemental credence distribution for the unknown chance $p$ that she will confess:<br><br>a normalised density function $f_n$ which is proportional to $p^n$:<br><br>$f_n(p)=(n+1)p^n$. |
| ↓ one derives | | ↓ one derives |
| Alma's initial credences:<br>$c_0(a_1)=\int_0^1 p\,dp=1/2$,<br><br>$c_0(a_1 \cap s_1)=\int_0^1 p^2\,dp=1/3$. | | Alma's final credences:<br>$c_n(a_1)=c_n(s_1)=\int_0^1 p f_n(p)\,dp=(n+1)/(n+2)$,<br>$c_n(a_1 \cap s_1)=\int_0^1 p^2 f_n(p)\,dp=(n+1)/(n+3)$. |
| ↓ one derives | | ↓ one derives |
| Alma's initial conditional credences:<br>$c_0(s_1\|a_1)=c_0(s_2\|a_2)=2/3$.<br><br>Alma's initial unconditional credences:<br>$c_0(a_1 \cap s_1)=c_0(a_2 \cap s_2)=1/3$,<br>$c_0(a_1 \cap s_2)=c_0(a_2 \cap s_1)=1/6$. | | Alma's final conditional credences:<br>$c_n(s_1\|a_1)=(n+2)/(n+3)$,<br>$c_n(s_1\|a_2)=(n+1)/(n+3)$,<br>$c_n(s_2\|a_1)=1/(n+3)$,<br>$c_n(s_2\|a_2)=2/(n+3)$.<br>Alma's final unconditional credences:<br>$c_n(a_1 \cap s_1)=(n+1)/(n+3)$,<br>$c_n(a_2 \cap s_2)=2/(n+2)(n+3)$,<br>$c_n(a_1 \cap s_2)=c_n(a_2 \cap s_1)=(n+1)/(n+2)(n+3)$. |
| ↓ one derives | | ↓ one derives |
| From this and from Alma's utilities of the possible outcomes (cf. figure 15) Jeffrey (1988) calculates Alma's initial conditional utilities of $a_1$ and $a_2$. As a result Alma shouldn't confess. | | **End**: From this and from Alma's utilities of the possible outcomes (cf. figure 15) Jeffrey (1988) calculates Alma's final conditional utilities of $a_1$ and $a_2$. As a result Alma should confess. |

Figure 18. Jeffrey's way of modelling Alma's prisoner's dilemma decision-theoretically.

If $n < 6$, both densities ($f_n$, $g_n$) are so flat that Alma sees a strong positive correlation between $a_1$ and $s_1$, so that confessing looks like the less desirable possible action. If $n < 6$, $c_n(a_1)$ and $d_n(a_1)$ are far from the extremes, so that it is implausible to believe that Alma has made a decision. If $n > 6$, both densities are so extreme that

Alma sees $a_1$ and $s_1$ as probabilistically independent, so that confessing looks like the more desirable possible action. And if $n > 6$, $c_n(a_1)$ and $d_n(a_1)$ can be extreme enough, so that it is plausible to believe that Alma has made a decision.

Because Jeffrey (1988) construes this variation of the prisoner's dilemma à la Lewis (1979) as a combination of two Newcomb's problems, so that Alma faces a Newcomb's problem and the Count faces a Newcomb's problem, it is easy to find out which decision is rational in Newcomb's problem according to Jeffrey's probabilism. One just has to exchange Alma's possible actions $a_1$: I confess. vs. $a_2$: I don't confess. with the decision maker's possible actions in Newcomb's problem $a_1$: I take both boxes at $t_3$. vs. $a_2$: I take B2 at $t_3$. and the Count's possible states of the world $s_1$: I confess. vs. $s_2$: I don't confess. with the possible states of the world in Newcomb's problem $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$, so that the decision maker's possible outcomes result as a combination of the possible actions and the possible states of the world in Newcomb's problem. Furthermore, one just has to proceed as in Alma's case, so that the decision to take both boxes ensues as the rational decision.

## *A Critique of Jeffrey's Probabilism*

Jeffrey's (1988) probabilism can be compared with Skyrms' (1984) causal decision theory (cf. chapter 3.3). For both Jeffrey (1988) and Skyrms (1984) probabilise Newcomblike problems on two levels. While in Jeffrey's (1988) theory credences for possible actions serve as estimates of chances for these possible actions, Skyrms (1984) calculates the utility of a possible action in terms of credence and chance. Yet in Skyrms' theory credences refer to possible states of the world and are used, when the decision maker doesn't know which possible state of the world obtains, and chances refer to possible outcomes; these chances can also be rewritten as credences of these possible outcomes. Furthermore, in Skyrms' (1984) causal decision theory chances for possible outcomes are used in the formula for calculating the objective utility of a possible action and in the formula for calculating the subjective utility of a possible action. In this respect the question arises whether the decision maker ever has epistemic access to chances? I doubt that, so that Skyrms' theory is for an idealised decision maker who has epistemic access to chances. In opposition to that in Jeffrey's (1988) probabilism the decision maker just tries to approximate the chances of his possible actions by his final judgemental credence distribution, so that Jeffrey's decision

maker isn't idealised in this respect. Thus Jeffrey's probabilism is to be preferred to Skyrms' causal decision theory with regard to what the theory epistemically demands from the decision maker.[115]

With regard to Jeffrey's (1988) probabilism I want to point out that Jeffrey uses final credences for possible actions in his formula for calculating the final conditional utility of a possible action. Furthermore, Jeffrey also uses initial credences for possible actions in his theory. Yet in chapter 3.6 I try to show that the decision maker shouldn't assign credences to his possible actions. Because initial and final credences are a special case of credences, my criticism against credences which I set forth in chapter 3.6 also applies to initial and final credences. Thus Jeffrey's formula for calculating the final conditional utility of a possible action cannot be right.

It is of some significance that Jeffrey's solution to the prisoner's dilemma and therefore also Jeffrey's solution to Newcomb's problem depends on how large $n$ = ... is. On the one hand this is a desirable result, because it could explain why some prisoners don't confess respectively some decision makers take B2, while some other prisoner's confess respectively some other decision makers take both boxes; yet on the other hand this is an undesirable result. For we actually want to have solutions which don't depend on the length of the deliberational process. Furthermore, Alma in the prisoner's dilemma and therefore also the decision maker in Newcomb's problem have to anticipate that her or his initial preference differs from her or his final preference. For otherwise why should they wait for their final preferences to appear as a result of their deliberational processes? Moreover, how should Alma in the prisoner's dilemma and therefore also the decision maker in Newcomb's problem anticipate her or his own final preference, given that she or he has never played the prisoner's dilemma respectively Newcomb's problem before? I don't see how they can do that. Thus Jeffrey's probabilism demands a lot of self-knowledge from Alma in the prisoner's dilemma and therefore also from the decision maker in Newcomb's problem. In my opinion the demanded amount of self-knowledge is unreasonable.

Jeffrey's probabilism doesn't specify any constraints which the decision maker's initial credence distribution has to fulfil[116], although Jeffrey (1988, p. 242) states that

---

[115]Yet as we will see in chapter 3.3, in the section on a critique of Skyrms' position, Skyrms' (1984) causal decision theory is in the ultimate analysis completely subjective, so that Jeffrey's probabilism and Skyrms' theory are on a par with regard to what the respective theories epistemically demand from the decision maker.

[116]To illustrate further problems with Jeffrey's probabilism I assume for the moment and for the sake of the argument that there is no problem with assigning credences to possible actions. Yet

"The artefact is a probability distribution or a partial characterisation of one as meeting certain constraints". Thus there are two alternatives how Jeffrey's probabilism can be interpreted: First, the decision maker's initial credence distribution doesn't have to fulfil any constraints. Second, the decision maker's initial credence distribution has to fulfil certain constraints. Let's consider the alternatives in turn.

If Jeffrey's probabilism is to be interpreted in such a way that there aren't any constraints on the decision maker's initial credence distribution, in the sense that it doesn't have to fulfil any constraints, then the problem of the priors[117] (cf. Earman 1992; Lambert and Brittan 1991) also applies to the decision maker's initial credence distribution. For the decision maker's initial credence distribution has to be determined in some way. Otherwise arbitrariness figures in rational decision theory which is undesirable.

According to Lambert and Brittan (1991, pp. 124-126) the problem of the priors is: When using Bayes theorem[118] the calculation of the posterior probability of the hypothesis H given evidence E depends on the following prior probabilities: P(E|H), P(H), and P(E). Yet Bayes theorem doesn't provide a method for determining these prior probabilities, so that we have to determine them by other means. Thus Bayesianism has to be supplemented. Lambert and Brittan (1991) report that there are two proposals what prior probabilities represent: (1) Prior probabilities represent our initial credence in the hypothesis under consideration. Thus subjective elements enter confirmation theory which is problematical. (2) Prior probabilities represent simple generalisations with regard to our previous experiences. This proposal leads to an introduction of confirmation by enumeration which has its problems.

---

as I already pointed out the decision maker isn't permitted to assign credences to his possible actions whether they be initial or final.

[117]The problem of the priors is prominent in and especially important for confirmation theory in the philosophy of science. Yet Earman (1992, pp. 90-92) points out that Jeffrey (1983, p. 194) explicitly considers the problem of 0-priors in his second edition of his logic of decision, so that the problem of the priors can be extended to rational decision theory. Jeffrey (1983, p. 194) claims that the decision maker's credence in a universal generalisation is supposed to be 0. For if the decision maker is willing to assign a positive credence to a universal generalisation, he is willing to learn from experience at such a great rate that he is jumping to conclusions which is irrational. With regard to Newcomb's problem this particular problem of 0-priors doesn't apply. For there are no universal generalisations in Newcomb's problem with which the decision maker has to deal.

[118]This is a statement of Bayes theorem (cf. Lambert and Brittan 1991, pp. 117-118): P(H|E) = P(E|H)P(H)/P(E), where H is a hypothesis, E is a piece of evidence, P(H|E) is the posterior probability of H given E, P(H) is the prior probability of H, P(E) is the probability of E, and P(E|H) is the likelihood of E given H.

Earman (1992, pp. 57-58) points out three defences why the assignment of prior probabilities is no problem for Bayesianism:

(1) As the evidence accumulates the differences in the prior probabilities "wash out" (cf. Earman 1992, chapter 6). Yet it remains a problem that the formal results apply only to the long run and not to the short and medium runs, so Earman (1992). And exactly this problem also applies to Jeffrey's probabilism in my opinion. For the recommendation to confess in the prisoner's dilemma respectively to take both boxes in Newcomb's problem only results, if $n$ is large enough. Furthermore, prior probabilities only converge to a particular posterior probability, if prior probabilities don't have the values 1 and 0.[119] With regard to this latter point Jeffrey's probabilism doesn't have any problems. For Jeffrey could claim that the decision maker always allows for the possibility that something may prevent the decision maker from carrying out his decision, so that he shouldn't assign a prior probability of 1 to a particular possible action; and Jeffrey could claim that the decision maker can always try to make a certain possible action true, so that he shouldn't assign a prior probability of 0 to a particular possible action.

(2) One can provide rules to fix reasonable initial credences. Yet Earman (1992) claims that up to now there is no rule which is able to cope with the wealth of information which the assignment of prior probabilities demands. Jeffrey (1988) doesn't provide any adequate rules either.

(3) There are plausibility considerations that can be used as guidelines for assigning prior probabilities. Yet Jeffrey (1988) doesn't provide any guidelines. Furthermore, in case one has such plausibility considerations, the following dilemma opens up (Suppe 1989, p. 399):

> "If standard inductive logic is intended to provide an analysis of that plausibility reasoning, then we have a vicious regress where each iteration of the Bayesian method requires a logically prior application; hence it is impossible to ever get the Bayesian method going. Hence standard inductive logic is an inadequate model of scientific reasoning about evidence and the evaluation of hypotheses. If, on the other hand, standard inductive logic does not provide an analysis of that plausibility reasoning, standard inductive logic is a critically incomplete, hence an inadequate

[119] For a thorough treatment of 0-prior probabilities see Earman (1992, chapter 4).

> model of scientific reasoning about evidence and the evaluation of hypotheses."

Therefore because Earman's (1992) three defences can be criticised, the assignment of prior probabilities remains a problem for both Bayesianism and Jeffrey's probabilism.

If Jeffrey's probabilism is to be interpreted in such a way that there are certain constraints on the decision maker's initial credence distribution, in the sense that it does have to fulfil certain constraints, the question arises whether there are any principles for constraining prior probabilities and whether they work. According to Earman (1992, p. 139) there are two reasons why such principles (for example, the principle of insufficient reason) don't work: (1) Different applications of these principles are possible, and the different applications can yield different results. Yet in my opinion there could be arguments that certain applications of these principles are the correct ones, so that certain results can be marked as the right ones. (2) Even if there was no ambiguity in the conditions of application of these principles, there would remain the problem that the conditions (for example, the condition of complete ignorance with regard to the unknown event) are rarely satisfied in real life. Earman (1992, p. 140) claims that a more realistic Bayesian would recognise the local and episodic character of problem-solving, so that the Bayesian should use different probability functions for different problem-solving contexts. Yet Earman hasn't given an a priori argument that with regard to every principle the conditions of application are never satisfied in real life. Thus it is still possible that there are principles whose conditions are satisfied in real life. Therefore Earman's arguments, why principles for constraining prior probabilities don't work, leave me undecided.

With regard to Newcomb's problem Jeffrey's (1988) probabilism proposes the right solution (cf. chapter 1.3 and 2.2). Yet Jeffrey's formula for calculating the final conditional utility of a possible action cannot be right, because it uses final credences for possible actions. Furthermore, Jeffrey's solution to Newcomb's problem depends on how large $n$ is which opens up the possibility of arbitrariness on the side of the decision maker. Moreover, Jeffrey's probabilism demands too much self-knowledge from the decision maker in Newcomb's problem. For in Jeffrey's probabilism the decision maker has to anticipate that his initial preference differs from his final preference. Furthermore, Jeffrey demands from the decision maker that he has to anticipate his final preference in a decision problem he never has encountered before, which is very much to demand. Finally, Jeffrey's theory doesn't specify the decision maker's initial

credence distribution, so that the following two interpretations of Jeffrey's probabilism can arise: First, there are no constraints on the decision maker's initial credence distribution, so that the problem of the priors applies to the decision maker's distribution. Second, there are constraints on the decision maker's initial credence distribution, so that the constraints have to be specified and problems with the constraints have to be dealt with, which hasn't been successfully done yet.

## 2.5 Jeffrey's Decision Kinematics

Jeffrey (1996) advances a fourth solution to Newcomb's problem by proposing his theory of decision kinematics. Jeffrey (1996, p. 5) explains the term "decision kinematics" in the following way:

> "In the design of mechanisms, kinematics is the discipline in which rigid rods and distortionless wheels, gears, etc. are thought of as faithful, prompt communicators of motion. The contrasting dynamic analysis takes forces into account, so that, e. g., elasticity may introduce distortion, delay, and vibration; but kinematical analyses often suffice, or, anyway, suggest relevant dynamical questions. That is the metaphor behind the title of this section and behind use of the term 'rigidity' below for constancy of conditional probabilities."

With Schmidt (1996) I believe that Jeffrey's (1996) rather general remarks can be interpreted in such a way that decision kinematics is a metaphor which stands for modelling the mental dynamics of the decision maker's conditional credences during deliberation. This will be made clearer in the remainder of this chapter.

### *Decision Kinematics and Rigidity*

Jeffrey (1996) goes on by explaining (1) what rigidity is and (2) why it is necessary for conditioning. With regard to the first point Jeffrey (1996) claims that according to the elementary probability calculus the following equation for conditioning (capital letters like *A*, *B*, *C*, etc. denote propositions)

$$c_{new}(A) = c_{old}(A|\text{data}),$$

is equivalent to the following two equations together

$$c_{new}(A|\text{data}) = c_{old}(A|\text{data}),$$

which expresses rigidity, that is constancy of conditional credences, and

$$c_{new}(\text{data}) = 1,$$

which expresses certainty of new credences of given data.

With regard to the second point Jeffrey (1996) proves that rigidity is necessary for conditioning by showing that certainty alone isn't sufficient for granting the equivalence with the equation for conditioning: If certainty of new credences of given data were sufficient, then for all $A$ the following equation would hold: $c_{new}(A) = c_{old}(A|A\cup\neg A) = c_{old}(A)$, which amounts to a never changing credence function. (For a second proof cf. Jeffrey 1996, p. 6.) And this consequence is unacceptable for Jeffrey, for he believes that the decision maker's credences of his possible actions vary during deliberation until the decision maker has decided for one particular possible action which then assumes a credence near 1.

Jeffrey, however, claims that conditioning isn't sufficient in all cases, for the rational effect of observations mustn't always be certainty of truth of data propositions. He therefore proposes two alternatives to conditioning, namely generalised conditioning of credences and generalised conditioning of factors. Jeffrey claims that if rigidity conditions were satisfied, that is if $c_{new}(A|C_i) = c_{old}(A|C_i)$ for all $i$, then the following two equations would hold:

$$c_{new}(A) = \sum_i c_{new}(C_i)c_{old}(A|C_i),$$

which Jeffrey denotes as generalised conditioning of credences, and

$$fac(A) = \sum_i fac(C_i)c_{old}(C_i|A),$$

which Jeffrey denotes as generalised conditioning of factors.[120] The significance of generalised conditioning of either form consists in the fact, so Jeffrey, that it allows for probabilistic responses to observations which don't change the decision maker's conditional credences, but which lead to new credences or factors for $C_i$.

Jeffrey (1996) doesn't want to consider the rational effect of observations in detail - except for showing that conditioning is in general insufficient -, he rather wants to consider the rational effect of deliberations which change the decision maker's credences in decision-making. He therefore shifts his attention to decision problems.

Jeffrey wants to show that Newcomblike problems are no decision problems. Jeffrey starts his proof by claiming that suppositions of causal influence show themselves not only in momentary features of credential states of mind, but also in evolving features of credential states of mind. He goes on by giving these rather general remarks the following more concrete shape: A consequence of the judgement that truth of one proposition stating that a cause is present promotes truth of another

[120] $fac(A)$ is the factor $c_{new}(A)/c_{old}(A)$, and $fac(C_i)$ is the factor $c_{new}(C_i)/c_{old}(C_i)$.

proposition stating that the cause's effect is present is that the regression coefficient β [121] is greater than 0, that is $c(\text{effect}|\text{cause})-c(\text{effect}|\neg\text{cause}) > 0$. With Arntzenius (1990) Jeffrey (1996) maintains that the decision maker can distinguish cause and effect from each other by rigidity relative to the partition {cause, ¬cause}, that is constancy of $c(\text{effect}|\text{cause})$ and of $c(\text{effect}|\neg\text{cause})$ while $c(\text{cause})$ varies. According to Jeffrey (1996) the variable "*c*" which ranges over a set of credence functions has to fulfil two conditions: (1) The regression coefficient is greater than 0, and (2) rigidity obtains.

With regard to Newcomblike problems and therefore also with regard to Newcomb's problem the rigidity condition is violated, Jeffrey (1996) claims, for too much information is given about the conditional credences, so that the unconditional credences for the two possible actions are fixed and cannot vary during deliberation anymore. Thus Newcomb's problem is overdetermined as a decision problem, and therefore is no decision problem. Jeffrey (1996, pp. 8-9) writes:

> "Newcomb problems ... seem ill-posed as decision problems because too much information is given about conditional probabilities, i. e., enough to fix the unconditional probabilities of the acts. We are told that there is an association between acts (making *A* true or false) and states of nature (truth or falsity of *B*) which makes acts strong predictors of states, and states of acts, in the sense that *p* and *q* are large relative to *p'* and *q'* - the four terms being the agent's conditional probabilities:
>
> $$p = P(B|A),\ p' = P(B|\neg A),$$
> $$q = P(A|B),\ q' = P(A|\neg B).$$
>
> But the values of these terms themselves fix the agent's probability for *A*[14]
>
> $$pr(A) = \frac{q \cdot p'}{q \cdot p' + (1-q) \cdot p}.$$
>
> Of course this formula does not fix *pr*(*A*) if the values on the right are not all fixed, but as decision problems are normally understood, values are fixed, once given."

Jeffrey (1996) makes his general claim that the rigidity condition is violated in Newcomblike problems and therefore in Newcomb's problem more precise by (1)

---

[121] The regression coefficient β of a random variable *Y* on another random variable *X* is obtained by the following equation: $\beta = \text{cov}(X, Y)/\text{var}(X)$, where $\text{cov}(X, Y) = E[(X-EX)(Y-EY)]$ and $\text{var}(X) = E(X-EX)^2$.

distinguishing between ordinary decision problems and Newcomblike problems and by (2) pointing out that while in ordinary decision problems the rigidity condition is applied to the possible actions of the decision maker, in Newcomblike problems the rigidity condition is applied to the deep possible states of the world.[122] (3) Jeffrey specifies Newcomb's problem even more by maintaining that there is a continuum of deep possible states of the world.[123]

With regard to the first point Jeffrey (1996) explains: While in ordinary decision problems possible actions $\pm A$ screen off deep possible states of the world $\pm C$ from plain possible states of the world $\pm B$ (cf. figure 19), in Newcomblike problems deep possible states of the world $\pm C$ screen off possible actions $\pm A$ from plain possible states of the world $\pm B$ (cf. figure 20).[124]

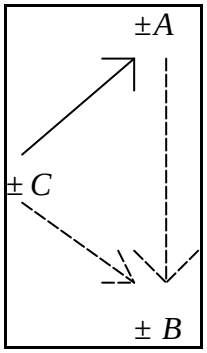


Figure 19. In ordinary decision problems possible actions $\pm A$ screen off deep possible states of the world $\pm C$ from plain possible states of the world $\pm B$. Solid/dashed arrows indicate stable/labile conditional probabilities.

[122]Unfortunately Jeffrey (1996) doesn't define "deep" possible states of the world. Jeffrey (1996, p. 11) just gives an example for a deep possible state of the world, namely "the sort of person I am", and that they work "from behind the scenes".

[123]Because Jeffrey (1996) doesn't specify what a continuum of deep possible states of the world in Newcomb's problem is, I asked him for a comment. Jeffrey gave me the following answer (personal communication from the 22nd of September, 1998): "... I don't have the article with me, but I guess I mean the states of the world to identify the 'objective chance'. If so, the continuum assumption is that this chance can assume any value in the unit interval. In the discrete (non-continuous) case it might only be able to assume (say) the 11 values 0, .1, .2, ..., .9, 1 -- perhaps because the objective chance is the chance of drawing a black ball at random from an urn in which 1/10th of the balls are black."

[124]Jeffrey (1996) also doesn't define "plain" possible states of the world.

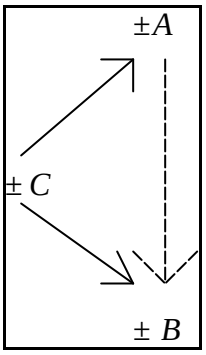


Figure 20. In Newcomblike problems deep possible states of the world $\pm C$ screen off possible actions $\pm A$ from plain possible states of the world $\pm B$. Solid/dashed arrows indicate stable/labile conditional probabilities.

Jeffrey's (1996) explication of the notion of "screening off" is: In general $\pm H$ screens off $F$ and $G$ from each other if and only if $c(F|G\cap H) = c(F|H)$, $c(G|F\cap H) = c(G|H)$, $c(F|G\cap\neg H) = c(F|\neg H)$, and $c(G|F\cap\neg H) = c(G|\neg H)$. That $\pm H$ screens off $F$ and $G$ from each other is implied by the following independence condition

$c(F\cap G|H) = c(F|H)c(G|H)$,

by the validity of the following equation according to the elementary probability calculus

$c(F|G\cap H) = c(F\cap G|H)/c(G|H)$,

and by the following rigidity condition

$c(F|H)$, $c(G|H)$, $c(F|\neg H)$, and $c(G|\neg H)$ are constant, while $c(H)$ varies.

## *A Critique of Jeffrey's Decision Kinematics*

Although Jeffrey's pertinacity to try to rescue his logic of decision has to be praised, Jeffrey's third modification of his logic of decision by proposing his decision kinematics can be criticised on the following grounds: Jeffrey's assumption that the decision maker's credences of his possible actions vary during deliberation until the decision maker has decided for one particular possible action which then assumes a credence near 1 is open to criticism. For as I try to show in chapter 3.6 the decision maker shouldn't assign credences to his possible actions. Furthermore, if we don't allow credences for possible actions, then the rigidity condition cannot be violated in Newcomb's problem. For if we don't allow credences for possible actions, we cannot claim anymore that there is too much information about the conditional credences in Newcomb's problem, so that the unconditional credences for the two possible actions are fixed and cannot vary during deliberation. Thus Jeffrey's conclusion that

Newcomb's problem is no decision problem because of the violated rigidity condition doesn't hold.

Yet Jeffrey (1996) is not the only one to claim that Newcomb's problem is no decision problem at all.[125] For Cargile (1975), Gardner (1973), Mackie (1977), and Schlesinger (1974) come to the same conclusion. Schlesinger's (1974) reason why Newcomb's problem is no decision problem is the following: Schlesinger doubts that one can predict free decisions.[126] For even if the decision maker was a person who had a strong tendency to take B2 and even if the predictor knew that, the decision maker could use his will power to resist this tendency and take both boxes instead. One could object to Schlesinger's position by claiming that a good predictor should be able to take into account the decision maker's usage of will power. Thus Schlesinger's conclusion that one cannot predict free decisions doesn't follow.[127]

Mackie's (1977, p. 223) reason why Newcomb's problem "in its intended interpretation ... simply cannot occur" is:

> "We simply cannot reconcile the requirements that the player should have, in a single game, a genuinely open choice, that there should be no trickery or backward causation, and that the seer's complete predictive success should be inductively extrapolable."

Mackie (1977) comes to this conclusion by considering several cases, for example, the predictor could be a very good psychologist, or a hypnotist, there could be sleight of hand operating on the decision maker's side, or backwards causation operating on the decision maker's side, etc., whereby the conditions of Newcomb's problem are made more precise.[128] Furthermore, Mackie (1977) classifies these cases in the following way: (1) Cases in which it is rational to take B2. (2) Cases in which it is rational to develop the character of a B2 taker. (3) Cases in which it is rational to develop the character of a both boxes taker. (4) Cases in which if the decision maker were free, it

---

[125]Peter Lanz pointed out to me the importance of the question whether Newcomb's problem is a decision problem at all.

[126]Gardner (1973) takes a similar stance. For he rhetorically asks (Gardner 1973, pp. 107-108): "Can it be that Newcomb's paradox validates free will by invalidating the possibility, in principle, of a predictor capable of guessing a person's choice between two equally rational actions with better than 50 percent accuracy?" I will not discuss Gardner's point of view here, because he doesn't present any arguments for his position.

[127]For a more detailed discussion of Schlesinger's (1974) position see Ledwig (1997).

[128]Cargile (1975) also considers several cases how the conditions of Newcomb's problem can be explained and comes to the conclusion that Newcomb's problem is underdetermined and an imaginary problem, which doesn't have to be scientifically possible. Because Mackie's (1977) arguments are much clearer than Cargile's, I will only deal with Mackie's arguments and cases.

would be rational to take both boxes. And (5) cases in which it is rational and the decision maker is free to take both boxes.

In order to understand how Mackie (1977, p. 223) comes to the conclusion that all these cases are "off-colour", let's consider Mackie's most reasonable case, which belongs to the fourth class, namely that the predictor is a very good psychologist. According to Mackie (1977, p. 219) such a predictor works in the following way: The predictor is able to say how the decision maker will reason in Newcomb's problem by observing the decision maker and by detecting which kind of character the decision maker has. And Mackie (1977, p. 219) claims that under this description of Newcomb's problem it is idle to ask which decision is rational for the decision maker. For the decision maker will follow his characteristic style of reasoning and will do what the predictor has predicted. Thus the decision maker's decision isn't free, it is already determined by the decision maker's character.

But if one regards taking both boxes as the decision maker's rational decision in this case, one is treating that decision as a still free one, so Mackie (1977, p. 219).[129] Moreover, under this description of Newcomb's problem the decision maker's decision is already determined by the decision maker's character; likewise, the content of B2 is determined by the predictor's prediction. Thus it would be arbitrary to rely on the causally determined fixity of the content of B2 and ignore the causally determined fixity of the decision maker's decision and vice versa, so Mackie (1977, p. 220). Therefore under this description of Newcomb's problem the decision maker isn't free to make a decision, so that Newcomb's problem is no decision problem at all.

There are several responses to this particular case: First, the predictor just has to be a little bit better than chance for Newcomb's problem to arise. Lewis (1979a) rightly claims that there is a conflict between an evidential conception of rationality and a causal conception of rationality if and only if the conditional utility of taking both boxes is less than the conditional utility of taking B2, which is the case if and only if the average estimated reliability of the predictor is greater than $(1+r)/2$, which is 0.5005 if

[129]Mackie (1977) doesn't make clear how taking both boxes is connected to making a free decision and how the decision maker having a characteristic style of reasoning is connected to making a free decision. Yet he implicitly suggests the following: If the decision maker has the characteristic style of reasoning of a B2-taker, and if he were free from his reasoning style, he would maximise his gain by taking both boxes; and if the decision maker has the characteristic style of reasoning of a both boxes-taker, and if he were free from his reasoning style, he would maximise his gain by taking both boxes. But the decision maker isn't free. For his decision is already determined by his character.

utility is proportional to money.[130] Thus the predictor just has to be a little bit better than chance which is easy to accomplish. One doesn't need a very good psychologist for that. A friend who knows the decision maker well is enough.

Second, one could argue like Dummett (1993, p. 372) that the decision maker's thought "My decision isn't really open, because the predictor already knows my decision", or because my decision is already determined by my character, will not help the decision maker to decide and will not dispense the decision maker from making a decision. Thus the only kind of freedom which is relevant for Newcomb's problem is that the decision maker is free to take B2 or both boxes in the following sense, so Dummett (1993, p. 375): The decision maker can rule out the possibility that he attempts to carry out his decision, but finds himself unable to do so. Thus Mackie's argument in the case of the very good psychologist that Newcomb's problem in its intended interpretation cannot occur depends on how you explicate freedom of will.

All in all Mackie's argument for his most reasonable case that Newcomb's problem is no decision problem is not convincing. For the present purposes it is not necessary to look at his other cases. For I just wanted to show that Newcomb's problem can, in principle, be conceived as a decision problem. To achieve this aim one just has to show that Newcomb's problem can be realised as a decision problem at least in one case. Thus Jeffrey's, Schlesinger's, and Mackie's attempts to exclude Newcomb's problem from the range of genuine decision problems have failed.

With regard to Jeffrey's (1996) decision kinematics the following overall conclusion can be drawn: Jeffrey's decision kinematics proposes the wrong solution to Newcomb's problem. For as we have already seen in chapter 1.3 and in chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 1, the decision maker should take both boxes in Newcomb's problem.

## 2.6 Eells' Common Cause-Solution

Like Nozick (1969) Eells (1982) considers Newcomb's problem as a conflict between the principle of strong dominance and the principle of maximising conditional utility. In opposition to Nozick (1969) Eells (1982) thinks that the conflict can and should be resolved in favour of the principle of maximising conditional utility. This should be so,

[130] For a proof see Sobel (1994, pp. 37-38).

because the principle of strong dominance in contrast to the principle of maximising conditional utility is limited in its applicability (Eells 1985).

## *The Causal Structure of Newcomb's Problem*

Eells (1982) believes that Newcomb's problem is structurally essentially identical to two Newcomblike problems, namely the eggs-benedict-for-breakfast-problem developed by Skyrms (1980, pp. 128-129) and Solomon's problem developed by Gibbard and Harper (1978, pp. 135-136). These Newcomblike problems are on first sight counterexamples to the principle of maximising conditional utility, for the principle doesn't take into account certain causal beliefs the decision maker might have, and are examples for the principle of strong dominance, so Eells (1982). In order to understand this claim I will present these Newcomblike problems and Eells' argumentation with regard to these problems. The eggs-benedict-for-breakfast-problem by Skyrms (1980, pp. 128-129) is:

> "Suppose that the connection between hardening of the arteries and cholesterol intake turned out to be like this: hardening of the arteries is not caused by cholesterol intake ...; rather it is caused by a lesion in the artery wall. In an advanced state these lesions will catch cholesterol from the blood, ... . Moreover, imagine that once someone develops the lesion he tends to increase his cholesterol intake. We do not know what mechanism accounts for this effect of the lesion. We do, however, know that the increased cholesterol intake is beneficial; it somehow slows the development of the lesion. Cholesterol intake among those who do not have the lesion appears to have no effect on vascular health. ..., what would a rational man who believed the account do when made an offer of Eggs Benedict for breakfast?"

While the principle of maximising conditional utility recommends to refuse the eggs, the principle of strong dominance recommends to eat them. For in the former case the decision maker reasons (Eells 1982, p. 89):

> "For if he believes that almost all cases of high cholesterol intake are caused by the lesion and that the lesion is very efficacious in producing high cholesterol intake, then it is plausible that his subjective probabilities are such that the probability of having the lesion is higher conditional on high cholesterol intake than conditional on low to medium cholesterol intake. Thus, since the probability of hardening of the arteries conditional

> on having the lesion is very high and since the desirability of hardening of the arteries is very low, it seems that *PMCEU* recommends not eating the eggs benedict."[131]

In the latter case the decision maker argues as follows: Eating the eggs leads to higher utilities of the possible outcomes in comparison with refusing the eggs for every possible state of the world, namely having the lesion vs. not having the lesion in combination with developing arteriosclerosis vs. not developing arteriosclerosis, so that the principle of strong dominance recommends to eat the eggs.

Eells (1982) claims that eating the eggs is the rational possible action in the eggs-benedict-for-breakfast-problem. For even though the credence of the lesion and thus of arteriosclerosis given high cholesterol intake is higher than the credence of the lesion and thus of arteriosclerosis given low to medium cholesterol intake, eating the eggs doesn't cause the lesion or arteriosclerosis. Eells (1982) agrees with Skyrms (1980) that if the decision maker refuses the eggs in order to maximise his conditional utility, he tries to influence the cause by suppressing its symptoms.

Because Solomon's problem (Gibbard and Harper 1978, pp. 135-136) was already introduced in chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 4, I will just shortly present Eells' argumentation for this problem. According to Eells (1982) the decision maker should reason along the same lines as in the previous case, so that sending for another man's wife is marked out as the rational possible action. Because Eells (1982) believes that eating the eggs and sending for another man's wife are the rational possible actions in these Newcomblike problems, and because he wants to defend the principle of maximising conditional utility, he provides another analysis of these Newcomblike problems in the following, so that these Newcomblike problems are on second sight no counterexamples of the principle of maximising conditional utility anymore, and the principle of maximising conditional utility agrees with the principle of strong dominance in its recommendations.

These Newcomblike problems are on second sight no counterexamples of the principle of maximising conditional utility, if you analyse these Newcomblike problems in the following way, so Eells (1985):

(1) Eells (1985) claims that both problems have the following common causal structure: A common cause $\pm CC$ (lesion in the artery wall in the eggs-benedict-for-

---

[131] "*PMCEU*" denotes the principle of maximising conditional expected utility.

breakfast-problem, and charisma in Solomon's problem) causes on the one hand a possible outcome ±*O* (arteriosclerosis and successful revolt) and causes on the other hand by means of causing some element of *R* a possible action ±*A* (high cholesterol intake and unjust behaviour). According to Eells (1985) *R* is the set of propositions describing the possible sets of beliefs and wants the decision maker might have in the respective decision situation. The following figure summarises the causal structure of both Newcomblike problems:

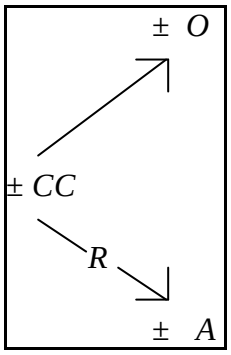


Figure 21. In the eggs-benedict-for-breakfast-problem and in Solomon's problem a common cause ±*CC* causes on the one hand a possible outcome ±*O* and causes on the other hand by means of causing some element of *R* a possible action ±*A*.

(2) Eells (1985) distinguishes between type-*A* beliefs and type-*B* beliefs. While type-*A* beliefs are beliefs of the decision maker about a randomly chosen rational decision maker, namely *r*, type-*B* beliefs are beliefs of the decision maker about himself, namely *i*. Eells claims that the decision maker should use type-*B* beliefs and not type-*A* beliefs for calculating his conditional utility. As a consequence the principle of maximising conditional utility and the principle of strong dominance give the same recommendations in the eggs-benedict-for-breakfast-problem and in Solomon's problem, so that these Newcomblike problems are no counterexamples of the principle of maximising conditional utility. In the next few paragraphs I will try to show how this result is obtained:

Eells maintains that in the eggs-benedict-for-breakfast-problem and in Solomon's problem the decision maker has the following credences, if he holds type-*A* beliefs:

$c(CC(r)|A(r)) > c(CC(r)|\neg A(r))$,

$c(A(r)|CC(r)) > c(A(r)|\neg CC(r))$.

Eells claims that in the eggs-benedict-for-breakfast-problem and in Solomon's problem the decision maker has the following credences, if he holds type-*B* beliefs:

$c(CC(i)|A(i)) = c(CC(i)|\neg A(i))$,

$c(A(i)|CC(i)) = c(A(i)|\neg CC(i))$.

Eells provides an argument to show that these equalities hold for a decision maker with type-*B* beliefs: First, Eells' argument contains three assumptions which are reasonable idealisations, so that the decision maker conceives himself as an ideal rational decision maker:

(1) Suppose that $\phi$ is the set of propositions that the decision maker considers in the respective decision situation. Suppose that the decision maker attributes credences to the members of $\phi$. Suppose that "$R\phi(i)$" denotes the proposition which says what the decision maker's credences and utilities are at decision time. Then assumption 1 is: The decision maker is certain what his credences and utilities are at decision time, that is

$c(R\phi(i)) = 1$.

(2) Suppose that "$D_A(i)$" respectively "$D_{\neg A}(i)$" denotes the proposition that the decision maker determines $A(i)$ respectively $\neg A(i)$ as the rational possible action. Then assumption 2 is: The decision maker is certain that he determines $A(i)$ respectively $\neg A(i)$ as the rational possible action if and only if he performs $A(i)$ respectively $\neg A(i)$, that is

$c(D_A(i) \leftrightarrow A(i)) = 1$.

(3) Assumption 3 is: The decision maker attributes the same credence to the following two propositions: (i) The decision maker determines $A(i)$ as the rational possible action given the decision maker's credences and utilities at decision time and the presence of the common cause. (ii) The decision maker determines $A(i)$ as the rational possible action given the decision maker's credences and utilities at decision time and the absence of the common cause. To put it into a formula:

$c(D_A(i)|R\phi(i) \cap CC(i)) = c(D_A(i)|R\phi(i) \cap \neg CC(i))$ (= screening assumption).

Second, starting off from these assumptions Eells' (1985) argument runs as follows:

From (2) and (3) it follows that

$c(A(i)|R\phi(i) \cap CC(i)) = c(A(i)|R\phi(i) \cap \neg CC(i))$ (= screening assumption).

From this and (1) it follows that

$c(A(i)|CC(i)) = c(A(i)|\neg CC(i))$,

$c(CC(i)|A(i)) = c(CC(i)|\neg A(i))$,

q.e.d.

Because Newcomb's problem is structurally essentially identical to the eggs-benedict-for-breakfast-problem and to Solomon's problem, and because type-*B* beliefs and not type-*A* beliefs should figure in the calculation of the decision maker's conditional utility, so Eells (1982), the decision maker should use the same kind of type-*B* beliefs for calculating his conditional utility in Newcomb's problem as in these Newcomblike problems. As a result the principle of maximising conditional utility recommends to take both boxes in Newcomb's problem which is in opposition to Jeffrey (1965). In the following Eells (1982) argues that the causal structure of Newcomb's problem is almost identical to the causal structure of the eggs-benedict-for-breakfast-problem and of Solomon's problem and that the differences between Newcomb's problem and these Newcomblike problems are unimportant. On that basis Eells shows that type-*B* beliefs lead to a 2-boxes-solution in Newcomb's problem.

According to Eells (1982) Newcomb's problem has the following causal structure: A common cause $\pm CC$ ($cc_1$: the common cause of $p_1$ and $a_1$, and $cc_2$: the common cause of $p_2$ and $a_2$) causes on the one hand by means of causing a prediction $\pm P$ ($p_1$: prediction to take both boxes, and $p_2$: prediction to take B2) a possible outcome $\pm O$ ($s_1$: \$ 0 is in B2, and $s_2$: \$ 1,000,000 is in B2) and causes on the other hand by means of causing some element of *R* a possible action $\pm A$ ($a_1$: taking both boxes, and $a_2$: taking B2). Therefore Eells (1982) claims that the causal structure of Newcomb's problem differs in one point from the causal structure of the eggs-benedict-for-breakfast-problem and of Solomon's problem, namely in adding a certain member $\pm P$ in the middle of the causal chain in Newcomb's problem going from the common cause $\pm CC$ via $\pm P$ to the possible outcome $\pm O$. Figure 22 shows the causal structure of Newcomb's problem:

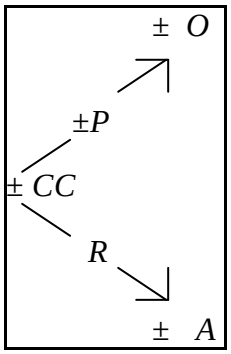


Figure 22. In Newcomb's problem a common cause $\pm CC$ causes on the one hand by means of causing a prediction $\pm P$ a possible outcome $\pm O$ and causes on the other hand by means of causing some element of *R* a possible action $\pm A$.

Eells (1982) maintains that there are two differences between Newcomb's problem on the one hand and the eggs-benedict-for-breakfast-problem and Solomon's problem on the other hand. First, in Newcomblike problems there is no predictor present having a high predictability. Second, in Newcomb's problem no common cause is mentioned. Eells (1982), however, believes that these differences are unimportant. For with regard to the first point you can reconstruct Solomon's problem in such a way that a predictor with a high predictability is present. This predictor causes successful revolts against kings for whom he predicts that they will perform unjust behaviour. His prediction is based on the kings having charisma or not. In this case the causal structure is identical to the causal structure of Newcomb's problem. And the rational possible action is still the unjust behaviour, so Eells. With regard to the second point Eells claims that if the decision maker believes in a high correlation between predictions and possible actions, then the decision maker must believe in a causal explanation for the predictor's success. Because the predictions don't cause the possible actions and vice versa, Eells believes that the only kind of causal explanation of the predictor's success which is compatible with Newcomb's problem is a common cause of the predictions and the possible actions.

Given this analysis of the situation in Newcomb's problem Eells (1982) tries to show that the decision maker's type-*B* beliefs lead to a 2-boxes-solution in Newcomb's problem: The following equalities hold for a decision maker with type-*B* beliefs:

$$c(cc_1(i)|a_1(i)) = c(cc_1(i)|a_2(i)),$$
$$c(cc_2(i)|a_1(i)) = c(cc_2(i)|a_2(i)).$$

These equalities are obtained by running the same kind of argument as in the eggs-benedict-for-breakfast-problem and in Solomon's problem. Eells (1982) continues by making a supposition: The decision maker believes that his possible actions don't interfere with the predictor's prediction process or with the predictor's filling of B2 according to his prediction, so that the following equalities should hold:

$$c(p_j(i)|cc_k(i)\cap a_1(i)) = c(p_j(i)|cc_k(i)\cap a_2(i)) \text{ for each } j, k = 1, 2,$$
$$c(s_j(i)|p_k(i)\cap cc_l(i)\cap a_1(i)) = c(s_j(i)|p_k(i)\cap cc_l(i)\cap a_2(i)) \text{ for each } j, k, l = 1, 2.$$

From this together with the following equalities

$$c(cc_1(i)|a_1(i)) = c(cc_1(i)|a_2(i)),$$
$$c(cc_2(i)|a_1(i)) = c(cc_2(i)|a_2(i)),$$

it follows that

$$c(s_1(i)|a_1(i)) = c(s_1(i)|a_2(i)),$$
$$c(s_2(i)|a_1(i)) = c(s_2(i)|a_2(i)).$$

Thus the calculation of the conditional utility results in a 2-boxes-solution for Newcomb's problem, and the principle of maximising conditional utility agrees with the principle of strong dominance in Newcomb's problem.

## *A Critique of Eells' Position*

Lewis (1981a, p. 10) rightly criticises Eells' (1981) position[132] which he calls the "Tickle Defence" for being limited in its applicability; for Eells' (1981) theory just applies to ideal rational decision makers. Lewis (1981a) maintains that rational decision theory shouldn't be limited to ideal rational decision makers, if rationality includes self-knowledge as Eells (1981) claims. For if rationality includes self-knowledge, there are also decision makers who are only almost sure before decision time how they will decide. Moreover, there are also dithery and self-deceptive decision makers, for whom decisions result in much more self-knowledge than thought experiments, so that the tickle doesn't work in their case. And why shouldn't we ask what decision would be rational for partly rational decision makers and whether their partly rational decision-making methods will lead them to decide for the rational possible action, Lewis (1981a) questions. Furthermore, Eells' (1981) theory doesn't give correct answers for partly rational decision makers in Newcomblike problems (Lewis 1981a).

But Lewis (1981a, p. 10) goes even one step further with his criticism. He doubts that rational decision theory applies to ideal rational decision makers. For how can ideal rational decision makers be uncertain what to decide for, so Lewis (1981a), and hence why should ideal rational decision makers deliberate. Moreover, Lewis (1981a) rightly questions whether ideal rational decision makers actually are very rational. For not only self-knowledge is an aspect of rationality, but also willing to learn from experience. And if the ideal rational decision maker's introspective data make him absolutely certain of his credences and utilities, then no amount of evidence can persuade him that those data are untrustworthy (Lewis 1981a).

Price (1986, p. 204) points out that Eells' Tickle Defence is dangerously self-referential. Eells' ideal rational decision maker not only has to know what his relevant beliefs and wants are, he also has to know whether they are the sort of beliefs and wants which will lead him to take both boxes. Price (1986, p. 205) continues by maintaining that Eells' Tickle Defence even has to deal with a more serious problem. For according to Price it is more natural to suppose that the common cause in

[132]Eells' 1981-position doesn't differ in substance from his positions in 1982 and 1985, so that Lewis' criticism of Eells' 1981-proposal carries over to Eells' 1982- and 1985-proposals.

Newcomb's problem is correlated with deciding to take both boxes than to suppose that it is correlated with actually taking both boxes. Price concludes that Jeffrey's (1983) ratificationism does a better job in this respect. He acknowledges that Jeffrey's ratificationism has other defects, though. With this Price opens up another area of investigation, namely which relation obtains between Eells' (1981, 1982, 1985) proposal of the common cause and Jeffrey's (1983) ratificationism.

In my opinion Eells' (1981, 1982, 1985) proposal of the common cause can be compared with Jeffrey's (1983) ratificationism. Both proposals rely on the fact that the decision maker perceives a gap between making a decision and performing the corresponding possible action. Yet one can improve both proposals by claiming that the causal structure in Newcomb's problem is even more detailed: In opposition to Price (cf. last paragraph) and in opposition to Eells (1981, 1982, 1985) and Jeffrey (1983) I think it is more natural to suppose that the common cause causes a certain set of beliefs and wants, which causes a certain decision, which again causes a certain possible action. Yet as we will see later on in this section (cf. Sober 1988) there may be objections against a common cause explanation of the predictor's prediction and the decision maker's decision in Newcomb's problem.

Eells' (1981, 1982, 1985) proposal of the common cause can also be compared with Price's (1986) proposal of decision maker probabilities. Price (1986, pp. 196-197) distinguishes between statistical generalisations and decision maker probabilities which is similar to Eells' distinction between type-*A* beliefs and type-*B* beliefs. The similarity can be described in the following way: While type-*A* beliefs are beliefs of the decision maker about a randomly chosen rational decision maker, statistical generalisations are probabilities of the decision maker about general reasons for action; and while type-*B* beliefs are beliefs of the decision maker about himself, decision maker probabilities are probabilities of the decision maker about his reasons for action. Price claims that statistical generalisations, like that the cancer gene occurs in 20% of smokers and in 2% of non-smokers, doesn't always licence an inference to corresponding probabilities in particular cases. For according to Price (1986, p. 199) when making a probabilistic judgement, one should take into account all the relevant available evidence (principle of total evidence). Furthermore, Price assumes that the decision maker's possible actions are immediately caused by the decision maker's reasons for his possible actions.

With regard to Newcomb's problem Price (1986, p. 208) concludes that in some cases the decision maker could have reasons to take both boxes, while in some other cases the decision maker could have reasons to take B2. With regard to the former

Price explains that if the prediction is a cause of the decision maker's possible action or the effect of some cause which the prediction and the decision maker's possible action have in common, then the decision maker is unable to take his possible actions to be probabilistically relevant to the actions of the predictor. With regard to the latter Price states that if the prediction isn't a cause of the decision maker's possible action or the effect of a common cause, then a judgement concerning the relevance of the decision maker's possible actions to the contents of B2 will not depend on an analysis of the causes of the decision maker's possible actions. Hence the decision maker can rationally act on the basis that by deciding to take only B2, he ensures that it contains $ 1,000,000. Thus in Newcomb's problem one doesn't infer in all cases from the statistical generalisation that the predictor is very reliable to the particular case that the decision maker's decision is very likely predicted correctly (for more details cf. Price 1983, p. 208).

Because Price appeals to the principle of total evidence, the question arises what is the relevant available evidence in Newcomb's problem? As we have already seen in chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 1 and ad 2, the predictor's high reliability is irrelevant for providing a solution to Newcomb's problem. Thus the predictor's high reliability can't be a reason for the decision maker's possible action.

Because the decision maker just wants to be rational and just wants to get as much money as he can, he cannot take the following reasons for his possible actions into account: The decision maker wants to demonstrate his greediness/non-greediness, his risk proneness/risk averseness, or any other property to which the decision maker ascribes a symbolic utility (cf. Nozick 1993 and chapter 4.3). Thus all these properties can't function as a reason for the decision maker's possible action.

Because the decision maker only knows that the predictor's prediction is causally independent of the decision maker's decision and besides that doesn't know anything about the causal relationship between the predictor's prediction and the decision maker's decision, only the former and not the latter can function as a reason for the decision maker's possible action. One can object to that by pointing out that the decision maker can assume a common cause for explaining the high correlation between the predictor's prediction and the decision maker's decision. Yet although this amounts to inference to the best explanation, the decision maker doesn't know that there is a common cause operating. Thus in my opinion it isn't justified to take a common cause as relevant available evidence in Newcomb's problem.

All in all the only factor which can be taken as relevant available evidence in Newcomb's problem is the causal independence of the predictor's prediction from the decision maker's decision. Thus the causal independence functions as a reason for the decision maker's possible action, and the decision maker's decision to take both boxes results. Therefore Price's conclusion that in some cases the decision maker has reasons to take both boxes, while in some other cases the decision maker has reasons to take B2, is false.

With regard to Newcomb's problem and Eells' proposal of the common cause the question remains what is the common cause of the decision maker's possible action and the predictor's prediction. One reasonable possibility is that the decision maker's character functions as a common cause for the decision maker's possible action and the predictor's prediction. This possibility is also in agreement with Nozick's description of how the predictor works (Nozick 1969, pp. 134-135). For the decision maker's character leads to a certain diagnosis of the decision maker's character by the predictor, which then leads to a certain prediction of the decision maker's decision, which finally leads to the predictor putting the respective money in B2. Thus the decision maker's character could be a common cause in Newcomb's problem.

Yet even if we have a plausible candidate for a common cause in Newcomb's problem, one could question whether it is reasonable to postulate a common cause in Newcomb's problem. For one could argue like Sober (1988, p. 217) that in some cases observed correlations shouldn't be explained by postulating a common cause. And perhaps one could go one step further and claim that Newcomb's problem is such a case.[133] Sober (1988, p. 215) gives the following example to demonstrate that in some cases observed correlations shouldn't be explained by postulating a common cause: The sea level in Venice and the cost of bread in Britain have both monotonically increased in the past two centuries, so that they are strongly positively correlated with each other. In this case and given the rest of our beliefs, so Sober (1988, p. 215), it is more plausible to believe that separate causes are responsible for the monotonous rise of the sea level in Venice and the monotonous rise of the cost of bread in Britain than to believe that a common cause is responsible for both phenomena.

Yet then the question arises: When should one prefer a common cause explanation to a separate cause explanation? Sober (1988) gives the answer in form of

[133]Sober (1988) doesn't discuss Newcomb's problem in his article.

an inequality: A common cause explanation is more probable than a separate cause explanation if and only if

$$P(E|CC)P(CC) > P(E|SC)P(SC),$$

where "*E*" denotes the evidence, "*CC*" denotes the common cause, and "*SC*" denotes the separate cause. This inequality is obtained by calculating the posterior probabilities of the common cause hypothesis and of the separate cause hypothesis given the evidence which both hypotheses have to explain, by taking the right hand parts of the following equations, and by cancelling out the equal denominators in these equations:

$$P(CC|E) = P(E|CC)P(CC)/P(E),$$

$$P(SC|E) = P(E|SC)P(SC)/P(E).$$

Furthermore, Sober (1988) claims that correlations are evidence for a common cause only in the context of a background theory and exemplifies this by means of the following variation of the Venice/Britain example (Sober 1988, p. 216): While the increase in Venetian sea levels and in British bread prices doesn't have a common cause that is not also a cause of French industrialisation, the increase in Venetian sea levels and in British bread prices does have a common cause that is not also a cause of Samoan migrations.[134]

Sober (1988, p. 224) makes some valuable remarks by stating that the principle of the common cause is a parsimony principle which is supposed to constrain causal explanation. Yet there is no principle which claims that one cause is always better than two, so Sober (1988). Furthermore, there is no principle which states that in a context, in which the investigator doesn't know anything about the subject under investigation, one cause has to be preferred to two causes, Sober (1988) claims.

There is another area worth of interest with regard to Eells' theory: Can the decision maker assign credences to his possible actions? Eells' answer would be yes. For when Eells (1982) claims that the following equalities hold for a decision maker with type-*B* beliefs in Newcomb's problem

$$c(cc_1(i)|a_1(i)) = c(cc_1(i)|a_2(i)),$$

$$c(cc_2(i)|a_1(i)) = c(cc_2(i)|a_2(i)),$$

he assumes that the decision maker can assign credences to his possible actions, which - as we will see in chapter 3.6 - the decision maker shouldn't do.

---

[134] A background theory for this variation of the Venice/Britain example could be a theory about climate conditions in Europe, while there is no background theory available which also includes Samoan migrations. I wish to thank Andreas Blank for this example of a background theory.

With regard to Eells' (1981, 1982, 1985) proposal of the common cause I conclude: Eells (1981, 1982, 1985) proposes the right solution to Newcomb's problem, namely to take both boxes. Yet Eells' theory is limited in its applicability. For it only applies to ideal rational decision makers. Furthermore, Eells' proposal is dangerously self-referential. For the decision maker has to know his relevant wants and beliefs and has to know whether they will lead him to take both boxes. Moreover, Eells' approach can be compared with Jeffrey's (1983) ratificationism and with Price's (1986) proposal of decision maker probabilities. For both Eells' theory and Jeffrey's theory rely on the gap between making a decision and performing the corresponding possible action; and both Eells and Price distinguish between general and particular cases, and both claim that the decision maker should consider his decision as belonging to a particular case. Although Eells (1981, 1982, 1985) proposes a common cause for Newcomb's problem, he doesn't specify any common cause for Newcomb's problem. Finally, Eells assumes that the decision maker can assign credences to his possible actions, which the decision maker shouldn't do.

## 2.7 Summary

Evidential decision theories do not provide adequate solutions to Newcomb's problem. For either they propose wrong solutions to Newcomb's problem, like Jeffrey (1965) and Jeffrey (1996), or they propose right solutions to Newcomb's problem, but their reasons for coming to the 2-boxes-solution are not adequate, like Jeffrey (1983), Jeffrey (1988), and Eells (1981, 1982, 1985).

Jeffrey's (1983) ratificationism, Jeffrey's (1988) probabilism, Jeffrey's (1996) decision kinematics, and Eells' (1981, 1982, 1985) proposal of the common cause fail to provide an adequate solution to Newcomb's problem, partly because they rely on Jeffrey's (1965) logic of decision, which has some serious defects, and partly because they have their own serious defects. Furthermore, Jeffrey's later proposals (Jeffrey 1983, 1988, 1996) are rather complicated, so that they lose the simplicity of Jeffrey's (1965) logic of decision and are on a par with causal decision theories with regard to their complexity.

# Chapter 3

# Causal Decision Theories

## 3.1 Introduction

### *Causation Theories as a Basis for Causal Decision Theories*

After having established in chapter 2 that causation should figure as a primitive term in rational decision theory, let's take a closer look at causal decision theories, and let's try to find out whether they provide adequate solutions to Newcomb's problem. In particular let's try to find out which causal decision theory provides an adequate solution to Newcomb's problem. For at least the following theories about causation are available (cf. Dowe forthcoming), and it can be doubted that all of them are suitable for a basis of causal decision theories:

(1) The regularity theory of causation (Hume 1978),

(2) the probabilistic theory of causation (Suppes 1970; Spohn 1983),

(3) the counterfactual theory of causation (cf. Lewis 1986),

(4) the manipulability theory of causation (Price 1996),

(5) the transference theory of causation (Aronson 1971; Fair 1979),

(6) the process theory of causation (Salmon 1984; Dowe 1992).

**Ad 1**: The regularity theory of causation states that a possible action is the cause of a possible outcome, if such possible actions always lead to such possible outcomes (Hume 1978, section 14, p. 172).

**Ad 2**: There is a contemporary version of the regularity theory of causation, namely the probabilistic theory of causation, which says that a possible action is the cause of a possible outcome, if it raises the probability of that outcome (for example Suppes 1970; Spohn 1983). Yet if one takes a probabilistic theory of causation as a basis for a causal decision theory, one has to make sure that causal independence cannot be reduced to probabilistic independence. For then Newcomb's problem becomes a fictitious problem which is of no interest anymore.

**Ad 3**: Another theory of causation is the counterfactual theory which claims that a possible action is the cause of a possible outcome, if it is the case that were that possible action performed, that possible outcome would follow (cf. Lewis 1986). Lewis' (1986) own theory of causation is a probabilistic counterfactual theory, though.

**Ad 4**: The manipulability theory of causation says that a possible action is the cause of a possible outcome, if we can use it as a means for bringing about that possible outcome (Price 1996). According to Mellor (1995, p. 80) a causal decision theory which is based on the manipulability theory of causation is circular, though.

**Ad 5**: The transference theory of causation states that a possible action is the cause of a possible outcome, if there is a transference of energy or momentum from that possible action to that possible outcome (Aronson 1971; Fair 1979).

**Ad 6**: The process theory of causation says that a possible action is the cause of a possible outcome, if there is a set of causal processes and interactions linking the two, where causal processes are delineated from other kinds of processes by their ability to transmit a mark (Salmon 1984) or by their possession of a conserved quantity (Dowe 1992).

While Gibbard and Harper (1978) use a counterfactual theory of causation in their causal decision theory, Sobel (1986) and Lewis (1981a) use a probabilistic counterfactual theory of causation in their causal decision theories. Skyrms (1982)[135] and Spohn (1978) defend a probabilistic theory of causation. Thus let's see whether these theories of causation in these causal decision theories provide an adequate basis for solving Newcomb's problem.

## *Order of Presentation*

The most prominent causal decision theories are[136]:

(1) Gibbard and Harper's (1978) *U*-utility,

(2) Skyrms' (1980, 1984) *K*-utility,

(3) Sobel's (1986) advancement of Jeffrey's logic of decision,

(4) Lewis' (1981a) unification of causal decision theories,

(5) Spohn's (1977, 1978) principle.

I will use this order of presentation for the following reason: Although the chronological order demands that Spohn's proposal (1977, 1978) should be presented and criticised first, this is inadvisable because of its contents. For Spohn (1978) already

---

[135]Skyrms (1982, pp. 696-697) formulates his causal decision theory by means of causal propensities.

[136]Joyce (in press, p. 48) claims that Savage's (1954/1972) rational decision theory is consistent with either a causal or a non-causal interpretation. Thus Savage's rational decision theory could be the first causal decision theory. According to Rabinowicz (1982, p. 321) the term causal decision theory stems from Lewis (1981a). This partly explains why it is in some cases difficult to classify rational decision theories in purely evidential or purely causal ones.

criticises Gibbard and Harper's (1978) approach. Moreover, this criticism can be extended to Sobel's (1986) and Lewis' (1981a) proposals. Furthermore, independently of Spohn's criticism of Gibbard and Harper's proposal Spohn's (1977, 1978) principle isn't connected in its substance to any of the other theories. Therefore I found it best to put Spohn's approach at the end of the list of the causal decision theories. Lewis' (1981a) unification of causal decision theories is set at the fourth position, because Lewis deals with Gibbard and Harper's (1978) proposal, Skyrms' (1980) approach, and Sobel's (unpublished) proposal which is a predecessor of Sobel's 1986-approach. Gibbard and Harper (1978), Skyrms (1980, 1984), and Sobel (1986) are set forth in chronological order, for nothing speaks against a chronological order in their cases.

## *Similarities and Differences between Causal Decision Theories*

The common characteristics of these causal decision theories are:

(1) They have causation as a primitive term in their theory.

(2) They propose the 2-boxes-solution in Newcomb's problem.

What distinguishes all five proposals from each other is:

(1) They propose different definitions of the utility of a possible action and therefore also propose different maximising principles.

(2) They add certain rationality constraints on the decision maker's credences.

(3) They motivate their solutions to Newcomb's problem by drawing analogies to different Newcomblike problems.

**Ad 1**: Gibbard and Harper (1978) propose the following definition of the utility of a possible action:

$$U(a_i) = \sum_{j=1}^{m} c(a_i \mathrel{\square\!\!\rightarrow} o_j)u(o_j),$$

where "$a_i \mathrel{\square\!\!\rightarrow} o_j$" means "if I were to do $a_i$, then $o_j$ would be the case". Within Jeffrey's (1965) terminology Gibbard and Harper's (1978) $U(a_i)$ takes the form:

$$U(a_i) = \sum_{j=1}^{m} c(a_i \mathrel{\square\!\!\rightarrow} s_j)u(o_{ij}).$$

According to Gibbard and Harper (1978) the decision maker should adopt the principle of maximising *U*-utility.

Skyrms (1980) claims that the decision maker should calculate the utility of a possible action along the following lines:

$$K(a_i) = \sum_{k=1}^{l}\sum_{j=1}^{m} c(K_k)c(C_j|K_k \cap a_i)u(K_k \cap C_j \cap a_i),$$

where $C_j$ are factors which are at decision time within the causal influence of the decision maker, and where $K_k$ are factors which are at decision time outside the causal influence of the decision maker and which are causally relevant for the possible outcomes of the decision maker. Skyrms' (1980) view is that the decision maker should use the principle of maximising *K*-utility. In 1984 Skyrms modifies his causal decision theory of 1980 by formulating it in terms of credence and chance. Yet this modification remains compatible with his proposal of 1980.[137]

Sobel (1986) proposes the following definition of the utility of a possible action: For any possible action $a_i$, so that $c(\circledcirc a_i) > 0$, and for any possible world $w$

$$U(a_i) = \sum_w \{ch(w|a_i)\}u(w),$$

where "$\circledcirc a_i$" means "$a_i$ is possibly open", "$\{ch(w|a)\}$" is called a ramified chance, and $\{ch(w|a)\} = \sum_y c[(a_i \diamond\!\!\rightarrow_y w)|\circledcirc a_i]y$. Furthermore, "$a_i \diamond\!\!\rightarrow_y w$" is a practical chance conditional, which says "if $a_i$ were the case, then $w$ might be the case with a chance of $y$". Moreover, $y$ ranges from 0 to 1, and $\sum_y c(a_i \diamond\!\!\rightarrow_y w) = 1$. According to Sobel (1986) the decision maker should adopt the

> **principle of maximising utility**: In a given decision situation *D* the decision maker *X* should decide for a possible action $a_i$ with maximal utility.

Lewis (1981a) maintains that the decision maker should calculate the *U*-utility of a possible action in the following way:

$$U(a_i) = \sum_{k=1}^{1} c(K_k)u(a_i \cap K_k),$$

where $K_k$ are dependency hypotheses. According to Lewis (1981a) the decision maker should use the

> **principle of maximising *U*-utility**: In a given decision situation *D* the decision maker *X* should decide for an option $a_i$ with maximal *U*-utility.

Lewis (1981a) states that Gibbard and Harper (1978), Skyrms (1980), and Sobel (unpublished) implicitly share this maximising principle with him. Furthermore, Lewis (1981a) claims that the dependency hypotheses consist of subjunctive conditionals with

---

[137] In 1985 Skyrms clarifies his causal decision theory. For he distinguishes between proximate and ultimate possible outcomes and claims that proximate possible outcomes are sufficient for calculating *K*-utility. Furthermore, Skyrms investigates the dynamics of rational deliberation (Skyrms 1984), the argument of the long run (Skyrms 1984), that is success in the long run, self-recommending rational decision theories (Skyrms 1982), which are also investigated by Nozick (1993, pp. 47-48), the problem of representation theorems for causal decision theories (Skyrms 1984), and the economics of collecting additional information (Skyrms 1982).

chance propositions as possible outcomes. Lewis (1981a) doesn't claim that this latter thesis is shared by Gibbard and Harper (1978), Skyrms (1980), and Sobel (unpublished). Thus Lewis (1981a) finally proposes the following definition of the utility of a possible action:

$$U(a_i) = \sum_{j=1}^{m} \sum_{q} c(a_i \mathbin{\square\!\!\rightarrow} [P(s_j) = q])q u(a_i \cap s_j),$$

where for any partition member $s_j$ and any number q ranging from 0 to 1 $[P(s_j) = q]$ is the proposition that holds only at those worlds where the chance of $s_j$ at the decision maker's decision time equals q. Moreover, Lewis (1981a) claims that the decision maker should maximise this utility.

Spohn (1978) doesn't propose a new definition of the utility of a possible action and therefore also doesn't propose a new maximising principle. Spohn's (1978) solution to Newcomb's problem is to take both boxes, for the decision maker should apply the principle of strong dominance in Newcomb's problem. The decision maker shouldn't maximise his conditional utility in Newcomb's problem because of Spohn's principle (cf. Ad 2).

**Ad 2**: Spohn (1977, 1978) adds a rationality constraint on the decision maker's credences: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts.*"[138] (Spohn 1977, p. 114) I will call this principle, which is implicitly used by Savage (1954/1972) and Fishburn (1964), Spohn's principle.

**Ad 3**: Gibbard and Harper (1978) motivate their solution to Newcomb's problem by drawing an analogy to Solomon's problem (Gibbard and Harper 1978, pp. 135-136). Skyrms motivates his solution to Newcomb's problem by drawing analogies to other Newcomblike problems: Fisher's problem (Skyrms 1984, pp. 65-66), the prisoner's dilemma (Skyrms 1984, pp. 67-68), the eggs-benedict-for-breakfast-problem (Skyrms 1980, p. 129), the Calvinists' problem (Skyrms 1980, p. 130), and uncle Albert's problem (Skyrms 1982, p. 700).

## 3.2 Gibbard and Harper's *U*-Utility

Gibbard and Harper (1978) view Newcomb's problem as a conflict between the principle of maximising *U*-utility and the principle of maximising *V*-utility. In opposition

[138] Trivial conditional probabilities, like $c(A|A) = 1$ for a possible action $A$ or $c(A'|A) = 0$ for two disjunctive possible actions $A$ and $A'$, are not considered.

to them Nozick (1969) and Eells (1982) consider Newcomb's problem as a conflict between the principle of strong dominance and the principle of maximising conditional utility. Gibbard and Harper (1978) resolve the conflict in favour of *U*-utility. Gibbard and Harper's (1978) *V*-utility is identical to Jeffrey's (1965) conditional utility; their *U*-utility of a possible action measures the expected efficacy of that possible action in causally bringing about possible outcomes which the decision maker wants to have. Gibbard and Harper interpret the *V*-utility of a possible action in such a way that it measures the welcomeness of the news that the decision maker will perform that possible action. They claim that such news may be welcome for two different reasons: First, the news of a possible action may be welcome, because that possible action leads to a possible outcome which the decision maker wants to have. Second, the news of a possible action may be welcome, because that possible action is evidence for a possible state of the world which the decision maker wants to obtain. Gibbard and Harper motivate their solution to Newcomb's problem by claiming that it has the same structure as Solomon's problem (Gibbard and Harper 1978, pp. 135-136) which is solved by applying the principle of maximising *U*-utility.

How is *U*-utility measured? According to Gibbard and Harper (1978) rational decision-making involves subjunctive conditionals of the form "if I were to do $a_i$, then $o_j$ would be the case". This will be denoted by "$a_i \boxrightarrow o_j$". The decision maker attributes credences to the subjunctive conditionals he considers for decision-making. For normally the decision maker doesn't know for sure what would be the case, if he performed a certain possible action. The decision maker ascribes utilities to his possible outcomes, so that the *U*-utility of a possible action is calculated as follows, so Gibbard and Harper (1978):

$$U(a_i) = \sum_{j=1}^{m} c(a_i \boxrightarrow o_j)u(o_j).$$

According to Gibbard and Harper (1978) the decision maker should use the

**principle of maximising *U*-utility**: In a given decision situation *D* the decision maker *X* should decide for a possible action $a_i$ with maximal *U*-utility.

What are the truth conditions of subjunctive conditionals? According to Gibbard and Harper (1978) the antecedent of a subjunctive conditional can be false, that is the decision maker can believe subjunctive conditionals of the form $a_i \boxrightarrow s_j$ for possible actions he will perform and for possible actions he will not perform. The subjunctive conditional $a_i \boxrightarrow s_j$ is true, if $a_i$'s holding causally brought about $s_j$'s holding, but also if

only $s_j$ held. In general Gibbard and Harper (1978) give in that they don't have a full theory of subjunctive conditionals and that they appeal to the readers' intuitions about subjunctive conditionals instead.

Gibbard and Harper (1978) develop Stalnaker's (1968, 1972/1981) and Stalnaker and Thomason's (1970) semantics of possible worlds: Suppose that $a_i$ is a possible action which the decision maker might decide for at time *t*. Suppose that an $a_i$-world is a possible world, which is like the actual world before *t*, in which the decision maker decides for $a_i$ at *t*, in which he performs $a_i$, and which obeys physical laws from time *t* on. Suppose $w_{a_i}$ is an $a_i$-world at *t*, which is most like the actual world at time *t*. Therefore $w_{a_i}$ is a possible world, which unfolds after *t* by obeying physical laws, and whose initial conditions at time *t* are minimally different from conditions in the actual world at *t*, so that $a_i$ is true in $w_{a_i}$. Then $a_i \boxrightarrow s_j$ is true in $w_{a_i}$.

According to Gibbard and Harper (1978) two axioms hold on this theory, namely

**axiom 1.** $(a_i \cap (a_i \boxrightarrow s_j) \rightarrow s_j$,

**axiom 2.** $(a_i \boxrightarrow \neg s_j) \leftrightarrow \neg(a_i \boxrightarrow s_j)$.

A consequence of these axioms is the following, so Gibbard and Harper (1978):

**Consequence 1.** $a_i \rightarrow ((a_i \boxrightarrow s_j) \leftrightarrow s_j)$.

According to Gibbard and Harper (1978) a sufficient condition for the *U*-utility and the *V*-utility of a possible action to be the same is:

**Condition 1 (stochastic independence).** $P(a_i \boxrightarrow o_j | a_i) = P(a_i \boxrightarrow o_j)$.

Condition 1 means that the subjunctive conditional $a_i \boxrightarrow o_j$ is stochastically independent of the possible action $a_i$. Because Gibbard and Harper (1978) claim that stochastic independence is the same as epistemic independence, which is in opposition to Kyburg's (1980, 1988) explicit distinction between stochastic vs. properly epistemic independence, condition 1 may be reformulated for epistemic independence:

**Condition 1 (epistemic independence).** $c(a_i \boxrightarrow o_j | a_i) = c(a_i \boxrightarrow o_j)$.

Gibbard and Harper's (1978) reason for the identity claim of stochastic and epistemic independence is: $c(a_i \boxrightarrow o_j | a_i)$ is the credence the decision maker rationally attributes to the subjunctive conditional $a_i \boxrightarrow o_j$, when he learns that he performs $a_i$.

Within Jeffrey's (1965) terminology Gibbard and Harper's (1978) $U(a_i)$ and $V(a_i)$ are:

$$U(a_i) = \sum_{j=1}^{m} c(a_i \boxrightarrow s_j) u(o_{ij}),$$

$$V(a_i) = \sum_{j=1}^{m} c(s_j | a_i) u(o_{ij}).$$

And the two forms of condition 1 are in Jeffrey's (1965) terminology:

**Condition 2 (stochastic independence).** $P(a_i \boxrightarrow s_j|a_i) = P(a_i \boxrightarrow s_j)$.

**Condition 2 (epistemic independence).** $c(a_i \boxrightarrow s_j|a_i) = c(a_i \boxrightarrow s_j)$.

How does all of this apply to Solomon's problem and Newcomb's problem? With regard to Solomon's problem Gibbard and Harper claim the following: If Solomon believes that a common cause causally influences his possible actions and his possible outcomes, then condition 2 fails for $R$ and $\neg R$, and $U$-utility differs from $V$-utility. Suppose "$R$" denotes "there will be a successful revolt", "$\neg R$" denotes "there will not be a successful revolt", '$B$" denotes "I take Bathseba", and "$\neg B$" denotes "I don't take Bathseba". Then the subjunctive conditional $B \boxrightarrow R$ isn't epistemically independent of $B$. For it is the case that

$$c(B \boxrightarrow R|B) > c(B \boxrightarrow R).$$

To see that this inequality holds, Gibbard and Harper (1978) present the following argument: Because Solomon knows that $B$'s holding wouldn't causally bring about $R$'s holding, he ascribes the same credence to $B \boxrightarrow R$ as to $R$, so that $c(B \boxrightarrow R) = c(R)$ and $c(B \boxrightarrow R|B) = c(R|B)$. If, however, Solomon learnt that $B$, he would have reason to believe that he was uncharismatic and therefore prone to a successful revolt, so that $c(R|B) > c(R)$. From all of this follows that

$$[c(B \boxrightarrow R|B) = c(R|B)] > [c(R) = c(B \boxrightarrow R)],$$

so that the subjunctive conditional $B \boxrightarrow R$ isn't epistemically independent of $B$.

To see that $U$-utility differs from $V$-utility in Solomon's problem Gibbard and Harper (1978) argue in the following way:

First, one has to calculate $U(B)$ and $U(\neg B)$:

$$U(B) = c(B \boxrightarrow \neg R)u(B\cap\neg R)+c(B \boxrightarrow R)u(B\cap R),$$

$$U(\neg B) = c(\neg B \boxrightarrow \neg R)u(\neg B\cap\neg R)+c(\neg B \boxrightarrow R)u(\neg B\cap R).$$

Because $c(B \boxrightarrow R) = c(R)$ and similarly $c(\neg B \boxrightarrow R) = c(R)$, $c(B \boxrightarrow R) = c(\neg B \boxrightarrow R)$. Likewise $c(B \boxrightarrow \neg R) = c(\neg B \boxrightarrow \neg R)$. Furthermore, Solomon believes that $u(B\cap \neg R) > u(\neg B\cap\neg R)$ and that $u(B\cap R) > u(\neg B\cap R)$, because $u(B\cap\neg R) = 10$, $u(\neg B\cap\neg R) = 9$, $u(B\cap R) = 1$, and $u(\neg B\cap R) = 0$. Therefore $U(B) > U(\neg B)$.

Second, I will consider $V$-utility: If Solomon learnt that $B$, he would have reason to believe that he was uncharismatic and therefore prone to a successful revolt, whereas if Solomon learnt that $\neg B$, he would have reason to believe that he was charismatic and therefore not prone to a successful revolt. Thus $c(R|B) > c(R|\neg B)$. Suppose the difference between these credences is $\varepsilon > 1/9$, and suppose $c(R|\neg B) = \alpha$ and $c(R|B) = \alpha+\varepsilon$, then $V(B)$ and $V(\neg B)$ are calculated:

$V(B) = c(\neg R|B)u(B\cap\neg R)+c(R|B)u(B\cap R)$,
$= 10(1-\alpha-\varepsilon)+1(\alpha+\varepsilon)$,
$= 10-9\alpha-9\varepsilon$,
$V(\neg B) = c(\neg R|\neg B)u(\neg B\cap\neg R)+c(R|\neg B)u(\neg B\cap R)$,
$= 9(1-\alpha)+0$,
$= 9-9\alpha$.

Because $\varepsilon > 1/9$, $V(\neg B) > V(B)$. And $U$-utility differs from $V$-utility in Solomon's problem, because $U(B) > U(\neg B)$, while $V(\neg B) > V(B)$.

Gibbard and Harper (1978) argue that the principle of maximising $V$-utility yields the irrational recommendation in Solomon's problem. For not taking Bathseba would be evidence for a wanted possible outcome, namely no successful revolt, without causally bringing about the wanted possible outcome itself.

## *Gibbard and Harper on Newcomb's Problem*

With regard to Newcomb's problem Gibbard and Harper (1978) argue as follows: While the principle of maximising $V$-utility recommends to take B2, for it is identical with Jeffrey's (1965) principle of maximising conditional utility, the principle of maximising $U$-utility recommends to take both boxes as can be seen by the following: Suppose that the credence of $s_1$ is $\mu$ and that the credence of $s_2$ is $1-\mu$. Because $s_1$ and $s_2$ are causally independent of the decision maker's possible actions $a_1$ and $a_2$, the credences of the decision maker are:

$c(a_1 \boxdot\!\!\rightarrow s_1) = \mu$,
$c(a_2 \boxdot\!\!\rightarrow s_1) = \mu$,
$c(a_1 \boxdot\!\!\rightarrow s_2) = 1-\mu$,
$c(a_2 \boxdot\!\!\rightarrow s_2) = 1-\mu$.

Thus the calculation of the $U$-utilities of the possible actions yields:

$U(a_1) = c(a_1 \boxdot\!\!\rightarrow s_1)u(\$\ 1{,}000)+c(a_1 \boxdot\!\!\rightarrow s_2)u(\$\ 1{,}001{,}000)$,
$= 1000\mu+(1-\mu)1{,}001{,}000$,
$= 1{,}001{,}000-1{,}000{,}000\mu$.
$U(a_2) = c(a_2 \boxdot\!\!\rightarrow s_1)u(\$\ 0)+c(a_2 \boxdot\!\!\rightarrow s_2)u(\$\ 1{,}000{,}000)$,
$= \mu(0)+(1-\mu)1{,}000{,}000$,
$= 1{,}000{,}000-1{,}000{,}000\mu$.

Because $U(a_1) > U(a_2)$ for every $\mu$, the principle of maximising $U$-utility recommends to take both boxes.

## *Gibbard and Harper on "Why Ain'cha Rich?"*

According to Gibbard and Harper (1978) the argument, which says "If you're so smart, why ain't you rich?", and which is stated but not defended by Lewis (1981b), is the following:

**Premise 1**: Decision makers who follow the principle of maximising *V*-utility and decision makers who follow the principle of maximising *U*-utility want to be millionaires.

**Premise 2**: While decision makers who follow the principle of maximising *V*-utility tend to leave Newcomb's problem as millionaires, decision makers who follow the principle of maximising *U*-utility don't.

**Conclusion**: Therefore decision makers who follow the principle of maximising *V*-utility make the rational decision.

In opposition to this Gibbard and Harper (1978) claim that this argument has the following moral: If a predictor with a high predictability rewards predicted irrationality, then irrationality will be rewarded.

Gibbard and Harper (1978) argue for this moral by presenting the following variation of Newcomb's problem: First, the decision maker takes the content of B2. Then he decides whether he additionally takes the content of B1 or not. The predictor is very reliable. Everything else stays the same as in the original version of Newcomb's problem. 1% of the 1,000,000 decision makers, who were confronted with this variation of Newcomb's problem, have found $ 1,000,000 in B2, and 1% of these have taken the content of B1, too. Decision makers who just took $ 1,000,000 argue in the following way, when they are asked, why they didn't take the content of B1, too: "If I were to take the additional $ 1,000, I wouldn't be a millionaire." According to Gibbard and Harper (1978) the decision to just take $ 1,000,000 is irrational on grounds of the principle of maximising *U*-utility and on grounds of the principle of maximising *V*-utility. For these decision makers know that they have $ 1,000,000, so that $V(a_1)$ = 1,001,000 and $V(a_2)$ = 1,000,000. Because $V(a_1) > V(a_2)$, the principle of maximising *V*-utility recommends to take both boxes. Therefore even if you take the view of decision makers who follow the principle of maximising *V*-utility, this variation of Newcomb's problem will make irrational decision makers millionaires (Gibbard and Harper 1978).

## *Gibbard and Harper on the Infallible Predictor*

Gibbard and Harper (1978) consider a variation of Newcomb's problem in which the predictor is known to be infallible. While the principle of maximising *U*-utility recommends to take both boxes in this case, the decision makers' intuitions go in the

other direction, namely to take B2. The latter derives from the decision maker's certainty that he will be a millionaire if and only if he takes only B2.

Gibbard and Harper (1978) provide three possible explanations for the decision makers' intuitions: (1) The decision makers have a tendency to causally bring about evidence for a wanted possible state of the world, even if it is known that the possible action which causally brings about that evidence doesn't causally bring about the wanted possible state of the world itself. (2) The decision makers may feel the force of the argument "If you're so smart, why ain't you rich?". Because this argument leads to an irrational recommendation in the above mentioned variation of Newcomb's problem, so Gibbard and Harper (1978), there is something wrong with it. (3) The decision makers make the fallacious inference from: "Either I will take B2 and be a millionaire, or I will take both boxes and be a non-millionaire." to "If I were to take B2, I would be a millionaire, and if I were to take both boxes, I would be a non-millionaire.". This inference is fallacious, so Gibbard and Harper (1978), because of the following reasoning: If the decision maker knows that he will take B2, then he knows that $ 1,000,000 is in B2. Thus he knows that he will be a millionaire. But because he knows that $ 1,000,000 is already in B2, he knows that even if he were to take both boxes, he would be a millionaire. If the decision maker knows that he will take both boxes, then he knows that $ 0 is in B2. Thus he knows that he will be a non-millionaire. But because he knows that $ 0 is in B2, he knows that even if he were to take B2, he would be a non-millionaire. If the decision maker doesn't know what he will do, he can conclude neither that B2 contains $ 1,000,000 nor that B2 contains $ 0. Therefore neither (i) "If I took B2, I would be a millionaire." nor (ii) "If I took both boxes, I would be a non-millionaire." follow from the decision maker's knowledge.

## *A Critique of Gibbard and Harper's Position*

With regard to Gibbard and Harper's (1978) formulation and moral of the "why ain't you rich?"-argument (cf. Lewis 1981b and chapter 3.5) the following can be said: In my opinion premise 2 should get a different formulation. For as it stands it suggests that all decision makers who follow the principle of maximising *V*-utility become millionaires in Newcomb's problem, while all decision makers who follow the principle of maximising *U*-utility don't become millionaires in Newcomb's problem. And this is clearly false, if the original version of Newcomb's problem is our starting-point. For in the original version only most of the decision makers who follow the principle of maximising *V*-utility become millionaires, and only most of the decision makers who

follow the principle of maximising *U*-utility don't become millionaires. Thus premise 2 should be reformulated in the following way:

> **Premise 2**: Most of the decision makers who follow the principle of maximising *V*-utility leave Newcomb's problem as millionaires, and most of the decision makers who follow the principle of maximising *U*-utility don't leave Newcomb's problem as millionaires.

Together with premise 1, that is

> decision makers who follow the principle of maximising *V*-utility and decision makers who follow the principle of maximising *U*-utility want to be millionaires,

the conclusion of the "why ain't you rich?"-argument, that is

> therefore decision makers who follow the principle of maximising *V*-utility make the rational decision,

doesn't follow.[139] Thus in opposition to Gibbard and Harper my moral of the "why ain't you rich?"-argument is that the argument isn't sound. Furthermore, in order to come to an overall conclusion in this "why ain't you rich?"-argument one has to provide reasons why most of the decision makers who follow the principle of maximising *V*-utility leave Newcomb's problem as millionaires, why most of the decision makers who follow the principle of maximising *U*-utility don't leave Newcomb's problem as millionaires, and whether these reasons also apply to the decision maker under consideration.[140] Therefore as long as this latter task isn't fulfilled, no overall conclusion can be drawn with regard to the "why ain't you rich?"-argument.

Gibbard and Harper (1978) claim that it is rational to take both boxes, if the predictor in Newcomb's problem is known to be infallible. Unfortunately Gibbard and Harper (1978) don't explain what they mean by infallibility. First, beforehand it should

---

[139]One can object to that conclusion by claiming that premise 1 can be changed, too, so that it states: Decision makers who follow the principle of maximising *V*-utility and decision makers who follow the principle of maximising *U*-utility want to be millionaires with the highest probability. And together with the changed premise 2 the conclusion of the "why ain't you rich?"-argument does follow.

[140]One can, for example, imagine a case in which most of the decision makers who follow the principle of maximising *V*-utility have a feature in common, which most of the decision makers who follow the principle of maximising *U*-utility don't have; furthermore, it is unclear whether the decision maker under consideration has this feature. Moreover, this feature could be responsible for a correct prediction of the predictor. Thus if the decision maker under consideration doesn't have this feature, the conclusion of the changed "why ain't you rich?"-argument doesn't follow. Yet if the decision maker under consideration does have this feature, the conclusion of the changed "why ain't you rich"-argument does follow.

be clear that the predictor's infallibility corresponds to a 100% predictability. Second, there are several possibilities to explicate infallibility (cf. Ledwig 1997). So let's see whether it is rational to take both boxes under different explications of infallibility. One way to explicate the infallibility of the predictor is:

(1) The predictor is infallible if and only if he predicts correctly in the actual world.

Sobel (1988a) says analogously that the predictor actually never makes incorrect predictions. Another way to explicate the infallibility of the predictor is as follows:

(2) The predictor is infallible if and only if he predicts correctly in all possible worlds.

Sobel (1988a) says analogously that the predictor is incapable of erring. With regard to both explications one could ask whether they are equivalent to necessarily correct predictions of the predictor. Whereas the equivalence holds in the second case, because it couldn't happen contingently that the predictor predicts correctly in all possible worlds, it doesn't hold in the first case, because it could happen contingently that the predictor predicts correctly in the actual world.

We associate with the infallibility of the predictor that it is impossible that he predicts incorrectly, that is he necessarily makes correct predictions. Now we can talk, for example, about analytical necessity, causal necessity, or logical necessity and can examine which mode of necessity is met by the proposition, the state of affairs, etc. that the infallible predictor predicts correctly.

So let's start with an explication of the predictor's infallibility in terms of analytical necessity:

(A) The predictor is infallible if and only if it is analytically necessary that he makes correct predictions.

Furthermore, following Kant's definition in his introduction (A7-A8, B10-B11) to *Kritik der reinen Vernunft* (1990) a judgement is analytical, if the term of the predicate is already contained in the term of the subject, so that nothing new is added in this way. In his *Prolegomena* (1976) Kant adds the criterion that the negation of an analytical judgement leads to a self-contradiction, so that analytical judgements are necessarily true. Thus a proposition *P* is analytically necessary if and only if the term of the predicate is already contained in the term of the subject and if the negation of an analytical proposition leads to a self-contradiction.

Now the question arises whether the proposition that the infallible predictor predicts correctly is analytically necessary? If we apply explication (A) and Kant's

explication of analytically necessary to the proposition that the infallible predictor makes correct predictions, we have to judge it as analytically necessary. Besides if one looks at the proposition that all bachelors are unmarried men, and if one judges it as analytical, how can one say that the proposition that all infallible predictors make correct predictions isn't analytical? Thus the proposition that the infallible predictor makes correct predictions is analytically necessary.

But what is the justification for the predictor's infallibility, or how does it come that the predictor is infallible? The answer seems to suggest itself: There must be a cause of the predictor's infallibility. We therefore look at another explication of the predictor's infallibility:

> (B) The predictor is infallible if and only if it is causally necessary that he makes correct predictions.

Furthermore, a state of affairs *P* is causally necessary in the actual world if and only if *P* follows from the combination of the antecedence conditions in the actual world with the natural laws in the actual world.

Unfortunately Newcomb's problem is insofar underdetermined that it doesn't specify all of the relevant antecedence conditions and all of the relevant natural laws. However, one can imagine the following scenarios in which the infallibility of the predictor is causally necessary: (a) The predictor predicts and determines the decision maker's decision. This case, however, isn't without problems, because the decision maker cannot decide freely anymore. (b) There is a common cause which unequivocally determines the prediction of the predictor and the decision of the decision maker, so that no other factor can intervene in the course of events, and there is a maximum correlation of 1 between prediction and decision. Yet if the decision maker's decision is completely determined by the common cause, one can also doubt that the decision maker decides freely. An exception is the possibility that the decision maker's will is the common cause of the decision maker's decision and the predictor's prediction. Thus if the state of affairs that the infallible predictor makes correct predictions is causally necessary, the problem arises that the decision maker cannot decide freely anymore. Because to judge something as a decision requires that the decision maker is free in making a decision, the causal necessity of the predictor's infallibility is irrelevant for decision-making.[141]

---

[141] One could object to that conclusion by pointing out that there are different explications of a free decision possible, so that the causal necessity of the predictor's infallibility might not be irrelevant for decision-making after all (cf. chapter 4.4).

Nevertheless one usually judges two events - like the prediction and the predicted event - as being logically independent of each other. To examine this let's have a look at another explication of the infallibility of the predictor:

(C) The predictor is infallible if and only if it is logically necessary that he makes correct predictions.

Moreover, a proposition *P* is logically necessary if and only if *P* is true and the negation of *P* is false in all possible worlds.

Thus is the proposition that the infallible predictor predicts correctly logically necessary? If we apply explication (C) and my explication of logically necessary to the proposition that the infallible predictor predicts correctly, we get the following result: The proposition that the infallible predictor makes correct predictions is true in all possible worlds, and the negation of the proposition that the infallible predictor makes correct predictions is false in all possible worlds. This sounds correct to me, so that the proposition that the infallible predictor predicts correctly is logically necessary.

Yet the following question has to be affirmed before accepting this conclusion: Is an infallible predictor logically possible? In personal communication Wolfgang Spohn denied this for the following reason: One can always imagine possible worlds in which an infallible predictor predicts incorrectly, that is from a logical point of view one cannot believe that there is an infallible predictor. Furthermore, in my opinion the actual world could be one of those worlds in which an infallible predictor predicts incorrectly. Thus one cannot rightfully claim the logical necessity of the proposition that the infallible predictor predicts correctly. For an infallible predictor is logically impossible.

Therefore even though the proposition that the infallible predictor makes correct predictions is analytically necessary, an infallible predictor isn't logically possible. Furthermore, the state of affairs that the infallible predictor makes correct predictions is causally necessary cannot arise in Newcomb's problem, if we require of decisions in a certain sense that they can be freely made. Thus Gibbard and Harper's claim that it is rational to take both boxes, if the predictor is infallible, can be criticised on the following grounds: First, there are different explications of infallibility possible, so that it is questionable to which explication Gibbard and Harper's claim applies. Second, an infallible predictor is logically impossible, so that Gibbard and Harper's recommendation to take both boxes, if the predictor is infallible, is in vain.

Spohn (1978, pp. 183-184) rightly criticises Gibbard and Harper (1978) for not providing an acceptable logic of subjunctive conditionals; moreover, he criticises that Gibbard and Harper don't provide a non-standard probability theory for subjunctive

conditionals. Finally, Spohn demands that Gibbard and Harper should rewrite their causal decision theory after having satisfied the first two points.[142] With regard to the second point I would like to object. For as Howard Sobel pointed out to me we don't need a non-standard probability theory for subjunctive conditionals, if we conceive subjunctive conditionals as a kind of proposition, and if there is no problem - and I don't see any - with applying probability theory to propositions.[143] One may object against this that the logical problems of subjunctive conditionals are not solved by indicating that subjunctive conditionals are propositions. Yet in my opinion a probability theory for subjunctive conditionals doesn't have to solve their logical problems.

Skyrms (1980, p. 132, footnote 6) points out that subjunctive conditionals which are used in the calculation of the *U*-utility of a possible action must be given a non-backtracking interpretation in Gibbard and Harper's theory. Thus the following kind of reasoning is forbidden, so Eells (1985, p. 193): It is highly probable that if Solomon were to send for the woman, he would be the kind of person who would do that, that is he would be uncharismatic and revolt prone; it is less probable that if Solomon were to abstain from the woman, he would be the kind of person who would send for her, that is he would be uncharismatic and revolt prone. Therefore $c(B \mathrel{\square\!\!\rightarrow} R) > c(\neg B \mathrel{\square\!\!\rightarrow} R)$, Eells (1985) concludes. An appropriate non-backtracking interpretation is the following, so Eells: If a possible action *B* has no causal efficacy in bringing about either a possible state of the world *R* or a possible state of the world $\neg R$, then $c(B \mathrel{\square\!\!\rightarrow} R)$ must equal $c(\neg B \mathrel{\square\!\!\rightarrow} R)$.

Eells (1985) rightly criticises that Gibbard and Harper (1978) have not constructed a theory of subjunctive conditionals with a non-backtracking effect; they rather appeal to their readers' intuitions, which has its problems, so Eells (1982, p. 105).[144] For there may be decision makers who have weak intuitions about subjunctive

---

[142]Spohn's (1978) criticism with regard to subjunctive conditionals doesn't only apply to Gibbard and Harper's (1978) causal decision theory, but to all rational decision theories which use subjunctive conditionals. Yet in 1978, when Spohn wrote his dissertation, there was no other rational decision theory around - except Gibbard and Harper's proposal - which used subjunctive conditionals.

[143]In personal communication Howard Sobel wrote to me (29th of August, 1998): "Subjunctive conditionals are a kind of proposition. There is nothing special about their probabilities. For example, for incompatible propositions, p and q, P(p v q) = P(p) + P(q). p and q are any propositions. They can (one or both) be subject-predicate propositions, generalisations, material conditionals, subjunctive conditionals. It doesn't matter."

[144]Although Gibbard and Harper (1978) have not constructed a theory of subjunctive conditionals with a non-backtracking effect, they explicitly state that they only consider worlds

conditionals, or there may be decision makers who have strong backtracking intuitions. I for one belong to the decision makers who have weak intuitions about subjunctive conditionals. Thus Gibbard and Harper's causal decision theory is limited in its applicability.

While Gibbard and Harper (1978) claim that stochastic independence and epistemic independence are the same, Kyburg (1980, 1988, and chapter 4.4) distinguishes between stochastic independence and properly epistemic independence. Thus the question arises whether Gibbard and Harper on the one hand and Kyburg on the other hand have the same concepts in mind, when they talk about stochastic independence and epistemic independence/properly epistemic independence. While Gibbard and Harper (1978) claim that stochastic independence can be expressed by the following equation

$$P(a_i \boxrightarrow o_j|a_i) = P(a_i \boxrightarrow o_j),$$

which means that the subjunctive conditional $a_i \boxrightarrow o_j$ is stochastically independent of the possible action $a_i$, Kyburg (1980, p. 150) claims that stochastic probability refers to the probability with which a toss of a coin will land heads. Thus both Gibbard and Harper's stochastic probability and Kyburg's stochastic probability refer to a general probability.

With regard to epistemic independence Gibbard and Harper (1978) claim that it can be expressed by the following equation

$$c(a_i \boxrightarrow o_j|a_i) = c(a_i \boxrightarrow o_j),$$

which means that the subjunctive conditional $a_i \boxrightarrow o_j$ is epistemically independent of the possible action $a_i$, whereas Kyburg (1980, p. 150) states that properly epistemic probability refers to the probability with which this coin lands heads. Furthermore, Gibbard and Harper's $c(a_i \boxrightarrow o_j|a_i)$ is the credence the decision maker rationally attributes to the subjunctive conditional $a_i \boxrightarrow o_j$, when he learns that he performs $a_i$. Thus both Gibbard and Harper's epistemic probability and Kyburg's properly epistemic probability refer to a particular probability.

Therefore although Gibbard and Harper (1978) on the one hand and Kyburg (1980) on the other hand have similar concepts in mind, when they talk about stochastic independence and epistemic independence/properly epistemic independence, they make different claims about the relationship between the two. Yet there are two ways to distinguish between Gibbard and Harper's position and Kyburg's position. For Gibbard

---

in which the past is like the actual past, for the decision maker cannot alter the past at decision time.

and Harper use causality as a primitive term in their proposal and also in their formulation of stochastic and epistemic independence, whereas Kyburg doesn't use causality as a primitive term in his proposal and therefore also doesn't use causality in his formulation of stochastic and properly epistemic probability. But does this distinction explain why Gibbard and Harper make an identity claim between stochastic independence and epistemic independence and why Kyburg doesn't? Actually not. Furthermore, in contrast to Kyburg (1980, p. 149) Gibbard and Harper (1978) believe in conditioning. Yet this distinction also doesn't explain the identity claim of Gibbard and Harper and the non-identity claim of Kyburg.

But there is another way to account for the difference between Gibbard and Harper's and Kyburg's position. For one could claim that Gibbard and Harper's epistemic probability doesn't refer to a particular probability. For Gibbard and Harper's $c(a_i \boxdot\!\!\rightarrow o_j|a_i)$ could either be the credence which a decision maker rationally attributes to the subjunctive conditional $a_i \boxdot\!\!\rightarrow o_j$, when he learns that he performs $a_i$ or be the credence which this decision maker rationally attributes to the subjunctive conditional $a_i \boxdot\!\!\rightarrow o_j$, when he learns that he performs $a_i$. Thus Gibbard and Harper's epistemic probability is an epistemic one, because a or this decision maker learns that he performs $a_i$, while Kyburg's properly epistemic probability is properly epistemic, because the decision maker learns the probability with which a particular coin lands heads. Yet in Kyburg's (1980) terminology stochastic probabilities and properly epistemic probabilities are both epistemic probabilities, because they both refer to a body of knowledge. In Gibbard and Harper's terminology, however, stochastic probabilities aren't epistemic probabilities. Thus Gibbard and Harper on the one hand and Kyburg on the other hand have different concepts in mind, when they talk about stochastic independence and epistemic independence respectively properly epistemic independence.

Skyrms (1985) claims that Gibbard and Harper's (1978) causal decision theory is formulated in terms of ultimate possible outcomes, which specify everything that the decision maker cares about in the decision problem, while Jeffrey's (1965) logic of decision is partition invariant and therefore doesn't need ultimate possible outcomes, proximate possible outcomes, or any possible outcomes for his evidential decision theory to work. In opposition to Gibbard and Harper (1978) and to Jeffrey (1965) Skyrms' (1984) own causal decision theory doesn't require ultimate possible outcomes. Furthermore, Skyrms (1985) shows that one can propose a causal decision theory which is formulated solely in terms of proximate possible outcomes.

Skyrms (1985) presents the following variation of Newcomb's problem to illustrate what a proximate possible outcome is and what an ultimate possible outcome is: The decision maker's task is to decide for a philosophical advisor who decides for the decision maker to take both boxes or to take B2. There are two advisors present, namely professor A and professor B. While the decision maker believes that it is probable that professor A will take B2, if the decision maker decides for him, the decision maker believes that it is probable that professor B will take both boxes, if the decision maker decides for him. Furthermore, the predictor is an expert on the psychology of Newcomblike problems and is the psychiatrist of the decision maker, of professor A, and of professor B. Moreover, the predictor has put $ 1,000,000 in B2 if and only if he had predicted that B2 will be taken. While the proximate possible outcome is the decision for B2 or for both boxes, the ultimate possible outcome is the monetary payoff. Everything else stays the same as in the original version of Newcomb's problem.

According to Skyrms (1985) in some versions of this variation of Newcomb's problem a causal decision theorist will decide for professor B and an evidential decision theorist will decide for professor A; furthermore, a decision maker who applies Gibbard and Harper's causal decision theory fallaciously to proximate possible outcomes gets uncausal results, in the sense that the results are in accordance with evidential decision theory. Thus Gibbard and Harper were right to formulate their causal decision theory in terms of ultimate possible outcomes, so Skyrms (1985). Yet Skyrms (1985) shows that a causal decision theory can also be formulated in terms of proximate possible outcomes. And by stating that Gibbard and Harper's causal decision theory in contrast to Skyrms' own theory requires ultimate possible outcomes to work, Skyrms (1985) implicitly evaluates his own causal decision theory as superior to Gibbard and Harper's causal decision theory. Unfortunately Skyrms (1985) doesn't give a reason why a causal decision theory which requires ultimate possible outcomes should be preferred to a causal decision theory which doesn't require ultimate possible outcomes. In my opinion one reason suggests itself: Proximate possible outcomes require less knowledge from the decision maker than ultimate possible outcomes. Thus less is demanded from the decision maker, if a causal decision theory is formulated in terms of proximate possible outcomes, which is an advantage.

According to Rabinowicz (1988, p. 409) Sobel's (1986) causal decision theory in opposition to Gibbard and Harper's (1978) causal decision theory doesn't presuppose the validity of conditional excluded middle, which in my opinion is an advantage of

Sobel's theory and a disadvantage of Gibbard and Harper's theory. For as Lewis (1981a) shows, the principle of conditional excluded middle is open to two objections which can't be overcome completely (cf. chapter 3.5).

Gibbard and Harper's (1978) causal decision theory proposes the right solution to Newcomb's problem, namely to take both boxes. But Gibbard and Harper's formulation of the "why ain't you rich?"-argument can be criticised, so that this argument turns out to be unsound. Furthermore, with regard to Gibbard and Harper's claim that it is rational to take both boxes, if the predictor is infallible, several explications of infallibility can be distinguished from each other. It is unclear to me which explication Gibbard and Harper have in mind, when they talk about the predictor's infallibility. Furthermore, one can argue that an infallible predictor is logically impossible, so that Gibbard and Harper's claim is in vain. In general the following can be criticised with regard to Gibbard and Harper's causal decision theory: Their theory, which is formulated in terms of subjunctive conditionals, doesn't provide a logic of subjunctive conditionals, but relies on the readers' intuitions instead which may be diverging. Furthermore, Gibbard and Harper still have to provide a theory of subjunctive conditionals with a non-backtracking effect. Moreover, Gibbard and Harper's causal decision theory is limited in its applicability, because it doesn't apply to decision makers who have weak intuitions about subjunctive conditionals and/or who have strong backtracking intuitions. Gibbard and Harper's causal decision theory is formulated in terms of ultimate possible outcomes and not in terms of proximate possible outcomes which is a disadvantage of their theory, for proximate possible outcomes demand less knowledge from the decision maker than ultimate possible outcomes. Finally, Gibbard and Harper's theory presupposes the validity of conditional excluded middle which is open to two objections which can't be completely overcome.

## 3.3 Skyrms' *K*-Utility

According to Skyrms (1984) Aristotle was the first to claim that causal, modal, or counterfactual distinctions are central for rational decision theory. For in the Nichomachean Ethics, so Skyrms (1984), Aristotle (1926) states that decision makers deliberate about factors which they can causally affect by their possible actions and not about factors which are already fixed:

> "As for deliberation, do people deliberate about everything - are all things possible objects of deliberation -, or are there some things about which

> deliberation is impossible? The term 'object of deliberation' presumably must not be taken to include things about which a fool or a madman might deliberate, but to mean what a sensible person might deliberate about.
>
> Well then, nobody deliberates about things eternal, such as the order of the universe, or the incommensurability of the diagonal and the side of a square. Nor yet about things that change but follow a regular process, whether from necessity or by nature or through some other cause: such phenomenae for instance as the solstices and the sunrise. Nor about irregular occurrences, such as droughts and rains. Nor about the results of chance, such as finding a hidden treasure. The reason that we do not deliberate about these things is that none of them can be effected by our agency." (Skyrms 1984, p. 132)

Likewise Skyrms (1980) makes the distinction between factors $C_j$ which the decision maker can causally influence by his possible actions at decision time and factors $K_k$ which are outside the decision maker's causal influence at decision time, but which are causally relevant to the possible outcomes of the decision maker's possible actions. Therefore Skyrms (1980) calculates the *K*-utility of a possible action in the following way:

$$K(a_i) = \sum_{k=1}^{l}\sum_{j=1}^{m} c(K_k)c(C_j|K_k\cap a_i)u(K_k\cap C_j\cap a_i).$$

Skyrms (1980) claims that the decision maker should adopt the

> **principle of maximising *K*-utility**: In a given decision situation $D$ the decision maker $X$ should decide for a possible action $a_i$ with maximal *K*-utility.

For the case in which the decision maker doesn't know which factors are outside the decision maker's causal influence at decision time Skyrms (1980, p. 136) makes clear that the decision maker can always construct hypotheses about which factors are outside of his causal influence, "such that the truth values of these hypotheses are factors" which are outside of his causal influence. Unfortunately Skyrms doesn't explain why the latter has to be the case. Skyrms continues in the following way: Suppose the factors, which are outside the decision maker's causal influence according to the decision maker's hypothesis $H_O$, are the $K_{Ok}$s, and the factors, which are inside the decision maker's causal influence according to the decision maker's hypothesis $H_O$, are the $C_{Oj}$s. Then the new factors which are outside the

decision maker's causal influence are the $H_o \cap K_{ok}$s, and the *K*-utility of a possible action is calculated in the following way:

$$K(a_i) = \sum_{k=1}^{l} \sum_{j=1}^{m} \sum_{o=1}^{p} c(H_o \cap K_{ok}) c(C_{oj} | H_o \cap K_{ok} \cap a_i) u(H_o \cap K_{ok} \cap C_{oj} \cap a_i).$$

Skyrms (1980, 1982, 1984) points out that Newcomblike problems are counterexamples to the principle of maximising conditional utility and are examples for the principle of maximising *K*-utility. Skyrms justifies his claim by the following reasoning: Probabilistic dependence in Newcomblike problems isn't causal dependence (Skyrms 1984); furthermore, probabilistic dependence in Newcomblike problems is generated by various causal and non-causal beliefs of the decision maker, for example, the belief that the possible actions of the decision maker cause possible states of the world, or the belief that possible states of the world cause possible actions of the decision maker, or the belief that there is an analogy between the two. Skyrms (1982, 1984) considers these probabilistic dependencies in Newcomblike problems as cases of spurious correlation in which the decision maker's possible actions and possible outcomes are correlated because of being effects of a common cause. Skyrms (1984) mentions Fisher's problem, Newcomb's problem, and the prisoner's dilemma as Newcomblike problems; Skyrms (1980) introduces the eggs-benedict-for-breakfast-problem, Fisher's problem, the Calvinists' problem, and Newcomb's problem as Newcomblike problems. In 1982 Skyrms adds uncle Albert's problem to the Newcomblike problems.

To get an impression of Skyrms' argumentation in Newcomblike problems, let's have a look at Fisher's problem, the Calvinists' problem, and uncle Albert's problem. First, consider Skyrms' argumentation in Fisher's problem (Skyrms 1984, p. 65):

> "Suppose, contrary to our present best knowledge, that the correlation between smoking and cancer was due not to smoking's being a causative factor in the aetiology of lung cancer; instead, that tendencies both to smoke and to develop lung cancer were effects of a common genetic cause. ... Suppose that you *know* that the hypothesis is true, and that the evidential relevance of your smoking to your getting lung cancer is only by virtue of your smoking being evidence that you have the gene. Your smoking and your having cancer are probabilistically independent for you *conditional* on having the gene and *conditional* on not having it, although your smoking is, over all, positively relevant for you to decide not to smoke in order to lower the probability of your having cancer. Would it not then

> be foolish for you to decide not to smoke in order to lower the probability of your having cancer?"

Second, look how Skyrms argues in the Calvinists' problem (Skyrms 1980, p. 130):

> "A crude caricature of Calvinism holds that it is already (in time) decided who are the elect and who are the damned. A sign of being one of the elect is leading a virtuous life, dissolute living being a mark of the damned. It is supposed to be, according to this story, an inducement to virtue that it raises the probability of being one of the elect. It is an odd utilitarian who would buy such an inducement, and who would not at least prefer rewards for virtue under an arrangement with a last judgement."

Third, consider Skyrms' argumentation in uncle Albert's problem (Skyrms 1982, p. 700):

> "Uncle Albert believes that, for him, going to the doctor is a symptom of being genuinely ill. Indeed, he believes that, whatever his other symptoms, if he finds himself in the doctor's office in addition, he is more likely to be ill than if not. This need not be an irrational belief. Now, in certain cases, conditional expected utility will recommend staying home because of the diminished prospects associated with going in to be examined. But Uncle Albert goes anyway. He's no fool. He knows that staying home won't help him if he's really ill."

Skyrms (1984) claims that his causal decision theory is more an advancement of Savage's (1954/1972) rational decision theory than an advancement of Stalnaker's (1972/1981) counterfactual proposal, although it is compatible with the latter because of Skyrms' (1984) Bayesian theory of conditionals. Skyrms (1984) modifies his causal decision theory of 1980 by claiming that his causal decision theory can be formulated in terms of credence and chance. Skyrms (1984) also substitutes his factors $C_j$ of 1980 by $C_j$ which are the decision maker's possible outcomes and substitutes his factors $K_k$ of 1980 by $K_k$ which are the possible states of the world. According to Skyrms (1984) the decision maker cannot decide for a possible state of the world, but he can decide for a possible action. Skyrms (1984) then defines the *K*-utility of a possible action in two steps. Relatively to a possible state of the world $K_k$ the possible action $a_i$ has the following objective utility:

$$U_{K_k}(a_i) = \sum_{j=1}^{m} ch_{K_k,a_i}(C_j)u_{K_k}(C_j \cap a_i).$$

If the decision maker knows which possible state of the world obtains, he should decide for the possible action with the highest objective utility, so Skyrms (1984).

If the decision maker doesn't know which possible state of the world obtains, he doesn't know which possible action has the highest objective utility, so that he should decide for the possible action with the highest subjective utility. According to Skyrms (1984) the subjective utility of a possible action is calculated in the following way:

$$U(a_i) = \sum_{k=1}^{l} c(K_k) \sum_{j=1}^{m} ch_{K_k,a_i}(C_j) u_{K_k}(C_j \cap a_i).$$

Skyrms (1984) uses unconditional credences in the latter formula, because by definition the decision maker's possible actions don't causally influence the possible states of the world.

By assuming that there are a finite number of possible states of the world, of possible actions, and of possible outcomes, and that each conjunction of possible state of the world and possible action has a positive prior credence, Skyrms (1984, p. 70) can rewrite chance as credence given the factors which determine chance according to the decision maker, so that the following equation holds:

$$ch_{K_k,a_i}(C_j) = c(C_j | K_k \cap a_i).$$

From this equation Skyrms (1984) obtains his *K*-utility of 1980:

$$K(a_i) = \sum_{k=1}^{l} \sum_{j=1}^{m} c(K_k) c(C_j | K_k \cap a_i) u(K_k \cap C_j \cap a_i).$$

## *Skyrms on Newcomb's Problem*

Skyrms (1984) analyses Newcomb's problem in the following way: The possible states of the world are: ($K_1$) \$ 0 is in B2, and \$ 1,000 is in B1; ($K_2$) \$ 1,000,000 is in B2, and \$ 1,000 is in B1. The possible actions of the decision maker are: ($a_1$) take both boxes; ($a_2$) take B2. The decision maker's possible outcomes are: ($C_{11}$) \$ 1,000; ($C_{12}$) \$ 0; ($C_{21}$) \$ 1,001,000; ($C_{22}$) \$ 1,000,000. Skyrms (1984) claims that in Newcomb's problem the chances are either 0 or 1. For the chance of getting \$ 1,000,000 is 1, if the decision maker takes B2 and there is \$ 1,000,000 in it. Relative to a possible state of the world the objective utility of taking both boxes is higher than the objective utility of taking B2 for each possible state of the world. For $U_{K_1}(a_1)$ = \$ 1,000, while $U_{K_1}(a_2)$ = \$ 0; and $U_{K_2}(a_1)$ = \$ 1,001,000, while $U_{K_2}(a_2)$ = \$ 1,000,000. As a result of the chances being either 0 or 1, the subjective utility of taking both boxes is also higher than the subjective utility of taking B2 for each possible state of the world. Therefore Skyrms (1984) recommends the decision maker to take both boxes.

Fisher's problem and the prisoner's dilemma, as Skyrms (1984) points out, mainly differ in one point from Newcomb's problem, namely the chances are neither 0

nor 1 in Fisher's problem and the prisoner's dilemma; nevertheless Skyrms' (1984) causal decision theory recommends to smoke in Fisher's problem and to confess in the prisoner's dilemma in contrast to the recommendations of Jeffrey's (1965) logic of decision.

## *A Critique of Skyrms' Position*

While Jeffrey (1988) takes credences as estimates of chances and Sobel (1986) defines the utility of a possible action in terms of chances, Skyrms (1984) defines his subjective utility in terms of credence and chance, so that the question arises whether it is an advantage or a disadvantage to use chances in a rational decision theory. On the one hand one could argue that it is a disadvantage for a rational decision theory to be formulated in terms of chances. For one could doubt to have epistemic access to chances, if chances were really something objective. Yet as Spohn (1988, p. 105) points out: While credence is "overtly epistemological", chance is "covertly epistemological".[145] On the other hand if one can specify the relationship between credence and chance clearly, for example, by means of Lewis' (1980) principal principle, then it is an advantage for a rational decision theory to be formulated in terms of chances. For it would be a gain in objectivity to have chances in comparison to credences.

With regard to Skyrms' (1984) causal decision theory it has to be made clear that Skyrms has a subjectivist understanding of chance. For in Skyrms' theory chances are to be construed in terms of credences given the factors which determine chance according to the decision maker, that is $ch_{K_k,a_i}(C_j) = c(C_j|K_k \cap a_i)$. Thus Skyrms' causal decision theory is in the ultimate analysis completely subjective. On first sight Sobel (1986) also has a subjectivist understanding of chance. For ramified chances are to be construed in terms of credences, that is $\{ch(w|a)\} = \sum_y c[(a_i \diamond_y \rightarrow w)|\circledcirc a_i]y$.

---

[145]In personal communication from the 1st of July, 1999, Wolfgang Spohn explained to me the difference between overtly epistemological and covertly epistemological (the translation is mine): "Overtly epistemological is everything which we describe, when we describe doxastic states and their changes and just the whole epistemological range of phenomena. Covertly epistemological is everything which is not overtly epistemological, but which nevertheless is to be described relatively to doxastic states and so ultimately only in relation to overtly epistemological things. The problem with this distinction only consists in the fact that it is usually philosophically highly controversial whether something is covertly epistemological or not. The dispute arises in the case of causation and of course in the case of objective probabilities. If one understands the latter, however, like Jeffrey as objectified credence, it would be in my sense covertly epistemological."

Thus Sobel's causal decision theory seems to be completely subjective in the ultimate analysis. Yet on second thoughts one has to object to that conclusion. For "$a_i \diamond_y \rightarrow w$" is a practical chance conditional, which says "if $a_i$ were the case, then $w$ might be the case with a chance of $y$". Thus ramified chances are to be construed in terms of credences which again are to be construed in terms of chances, so that Sobel's causal decision theory is in the ultimate analysis objective.

In chapter 2.4, in the section on a critique of Jeffrey's probabilism, I concluded that Jeffrey's probabilism is to be preferred to Skyrms' causal decision theory with regard to what the theory epistemically demands from the decision maker. For it looks on first sight as if Skyrms' theory is for an idealised decision maker who has epistemic access to chances, whereas Jeffrey's probabilism isn't restricted in this sense. Yet if we take into account that Skyrms construes chances in terms of credences, then Skyrms' theory isn't restricted to idealised decision makers who have epistemic access to chances, so that my conclusion that Jeffrey's probabilism is to be preferred to Skyrms' causal decision theory with regard to what the theory epistemically demands from the decision maker doesn't hold anymore. Rather it is the case that both theories are on a par with regard to epistemic demands from the decision maker. Yet Sobel's (1986) causal decision theory which is objective in the ultimate analysis is for an idealised decision maker who has epistemic access to chances and therefore is limited in its applicability. Thus Jeffrey's probabilism and Skyrms' causal decision theory are to be preferred to Sobel's causal decision theory with regard to what they epistemically demand from the decision maker.

Skyrms (1984) uses unconditional credences for calculating the subjective utility of a possible action, because he believes that the decision maker's possible actions don't causally influence the possible states of the world. By doing so Skyrms in fact doesn't violate Spohn's principle (cf. chapter 3.6), although he never explicitly endorses Spohn's principle. As we will see in chapter 3.6 Spohn's principle is valid, so that it is a virtue of Skyrms' causal decision theory to be in agreement with Spohn's principle.

Furthermore, Skyrms' (1984) analysis of Newcomb's problem seems to be correct, for he partitions the possible states of the world in the right way (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory). Moreover, Skyrms (1984) excludes partitions of the possible states of the world like $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$. For these partitions don't meet Skyrms' (1984) requirement that the possible states of the world are causally independent of the decision maker's possible

actions. Thus Skyrms' (1984) theory can be recommended with regard to its partition of the possible states of the world in Newcomb's problem.

It is a virtue of Skyrms' theory that it doesn't use subjunctive conditionals. For an adequate logic of subjunctive conditionals still has to be provided.[146]

In opposition to Gibbard and Harper's (1978) causal decision theory Skyrms' (1984) causal decision theory doesn't require to be formulated in terms of ultimate possible outcomes (cf. Skyrms 1985) which is an advantage of his theory. For ultimate possible outcomes demand more knowledge from the decision maker than proximate possible outcomes (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position).

With regard to Newcomb's problem Skyrms' (1980, 1982, 1984) causal decision theory recommends the right solution, namely to take both boxes. Furthermore, Skyrms' theory is in the ultimate analysis completely subjective, so that it doesn't demand objectivity from the decision maker. Moreover, Skyrms uses unconditional credences for calculating the utility of a possible action, thereby obeying Spohn's principle and thereby assuming that the decision maker has full control over his possible actions. The latter seems reasonable to assume. For only exceptional circumstances can prevent the decision maker from taking B2 or both boxes. Additionally, Skyrms partitions Newcomb's problem correctly (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory). Furthermore, Skyrms' causal decision theory doesn't use subjunctive conditionals, so that Skyrms' theory can also be applied to decision makers who have weak intuitions about subjunctive conditionals and/or who have strong backtracking intuitions. Decision makers with strong backtracking intuitions are no problem for Skyrms' theory, because Skyrms demands that the possible states of the world are causally independent of the decision maker's possible actions. Furthermore, Newcomb's problem can very easily generate backtracking intuitions in the decision maker, so that Skyrms' theory which is able to deal with them is very valuable. Finally, Skyrms' causal decision theory doesn't require ultimate possible outcomes for it to work, so that less knowledge is demanded from the decision maker. For example, if one applies Skyrms' causal decision theory to some variations of Newcomb's problem (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position), the decision maker only has to know his decision for B2 or for both boxes, which is the proximate possible outcome, whereas he doesn't have to know his

[146] An overview of problems associated with subjunctive conditionals can be found in Edgington (1995).

monetary payoff, which is the ultimate possible outcome. Thus Skyrms' theory is very valuable with regard to what it epistemically demands from the decision maker. So far Skyrms' causal decision theory seems to be the most adequate rational decision theory.

## 3.4 Sobel's Advancement of Jeffrey's Logic of Decision

Sobel (1990, p. 243) claims that a rational decision theory should be embeddable in a rational hope theory (cf. Sobel 1985b, p. 199, footnote 7) like Jeffrey's (1965, 1983) evidential decision theory, that is "the theory of rational acts should be a part of a general theory of rational desires for facts" (Sobel 1990, p. 244). For if one rationally decides for something, it is rational to hope for it, but not vice versa.

With his causal decision theory Sobel (1986) tries to improve Jeffrey's (1965) logic of decision. For on the one hand Sobel (1986) believes that Jeffrey's (1965) logic of decision gives wrong recommendations in Newcomblike problems, because causality doesn't figure in Jeffrey's theory. And on the other hand Sobel (1986) regards Jeffrey's theory as attractive, because it works for all partitions of the possible states of the world. Therefore Sobel (1986) develops a causal decision theory which is derived from Jeffrey's (1965) logic of decision by means of four definitions and which works for all natural partitions of the possible states of the world.[147] Sobel (1986) claims that his causal decision theory is equivalent to Lewis' (1981a) causal decision theory.[148] Furthermore, Sobel (1986) claims that his theory is to be preferred to Lewis' theory, for

[147]With regard to my questions what a natural partition is and whether Sobel like Skyrms claims that causal partitions are the natural ones Sobel gave me the following answer (personal communication from the 21st of August, 1998): "I mean by 'natural partitions' the partitions of circumstances that come naturally to mind (!) when one is conducting a Jeffrey-style (ch. 1 of *Logic of Decision*) Bayesian analysis of a decision. For example, {he confesses, he does not confess} in a prisoners's dilemma, and {there is $M in Box 2, there is no money in Box 2} as well as {Predictor predicted correctly, Predictor predicted incorrectly}. No, I do not consider Skyrms's partition into complete causal hypotheses to be 'natural,' as I do not consider ultimate partitions into possible worlds to be 'natural'."

[148]Rabinowicz (1982), however, claims that Lewis' (1981a) causal decision theory and Sobel's (unpublished) causal decision theory aren't equivalent to each other. For according to Rabinowicz Lewis tries to establish this equivalence by introducing an extra constraint on Sobel's theory which leads to counterintuitive consequences. Yet Lewis (1981a) admits that he imposes an extra constraint on Sobel's theory - a constraint which he deems plausible in ordinary cases. Furthermore, Sobel (1986) gives in that there are differences between his theory and Lewis' theory, but points out that these differences are not great. Thus Rabinowicz, Sobel, and Lewis seem to agree that the equivalence between Sobel's causal decision theory and Lewis' causal decision theory is restricted, yet can be established. They disagree whether it is reasonable to accept the extra constraint.

first, Sobel's definition of utility is nearer to Jeffrey's (1965) definition of utility and thereby makes possible quick access to Jeffrey's logic of decision. Second, Sobel's definition of utility doesn't presuppose another definition of desirability. And third, Sobel's definition of utility makes it possible that his causal decision theory is a natural extension of standard modal logic.

The derivation of Sobel's (1986) causal decision theory from Jeffrey's (1965) evidential decision theory proceeds in three steps and by means of four definitions. As a preliminary Sobel (1986) introduces practical conditionals of the form "p $\Box\!\!\rightarrow$ q" which measure causal potentials in opposition to conditional credences which measure evidential potentials. For Sobel (1986) believes that probabilities which figure in the utility calculation should be able to measure causal potentials and not evidential potentials. The practical conditional of the form "p $\Box\!\!\rightarrow$ q", which takes the form of a subjunctive conditional, is true if and only if either (1) q is the case and would remain the case even if p were the case or (2) although q is not the case, were p the case, q would be the case. According to Sobel (1986) the practical conditional of the form "p $\Box\!\!\rightarrow$ q" expresses a causal conditional, but not a purely causal conditional.[149] For if in the prisoner's dilemma the decision maker is certain that the other decision maker will confess, the decision maker is certain that (the decision maker doesn't confess $\Box\!\!\rightarrow$ the other decision maker will confess). The decision maker should be certain of this practical conditional even if the other decision maker's confession isn't causally determined, but determined in some other way (Sobel 1986).

First, Sobel (1986) proposes the following modification to make Jeffrey's (1965) logic of decision causal: Sobel (1986) substitutes Jeffrey's conditional credences with credences of practical conditionals, so that he gets

> **definition 1.** For any possible action $a_i$ and any possible world $w$ the utility of a possible action $a_i$ is defined in the following way:
> $U(a_i) = \sum_w c(a_i \Box\!\!\rightarrow w)u(w)$.

According to Sobel (1986) this definition resembles the definition of Stalnaker (1972/1981) and the definition of Gibbard and Harper's (1978) *U*-utility.

---

[149] When I asked Sobel what a not purely causal conditional is, he responded (personal communication from the 21st of August, 1998): "Can I make this business of practical, not 'purely' causal conditionals plainer? Probably not! The problem I think I see is that subjunctive conditionals figure in ordinary rational deliberations, and that we are not in our ordinary rational deliberations committed to causal determinism for people's actions."

Sobel (1986) rejects definition 1, for it is limited in two ways: Definition 1 is only apt for decision makers and possible actions which fulfil the following two constraints: (i) The decision makers don't believe in chances, and they believe that for every possible action there is a unique possible world which would be realised, if this possible action were realised. (ii) The decision maker is certain that the possible actions are open to him. In the following Sobel (1986) tries to get rid of these limitations by proposing other definitions. First, Sobel (1986) deals with the first limitation, so that he gets definition 2, second, he deals with the second limitation, so that he gets definition 3. Third, he deals with both limitations, so that he gets definition 4.

In order to improve his theory with regard to limitation 1 Sobel (1986) considers the following case: Suppose a decision maker who believes in an indeterministic world is certain that it is not the case that if $a_i$ were realised, it would be realised in $w$. Then Sobel (1986) claims that the following belief of this decision maker is consistent with the above mentioned supposition: If $a_i$ were realised, it might be realised in some world $w$ with a certain chance. And Sobel (1986) thinks that the utility of a possible action should be able to accommodate for such beliefs in chances. Therefore Sobel (1986) explicates the weights in the definition of the utility of a possible action in terms of chances: Let "$a_i \Diamond_x \rightarrow s_j$" be a practical chance conditional, which says "if $a_i$ were the case, then $s_j$ might be the case with a chance of $x$". Sobel (1986) then substitutes the credence of the causal conditional in definition 1 by the chance of $w$ given $a_i$ in

> **definition 2.** For any possible action $a_i$ and any possible world $w$ the utility of a possible action $a_i$ is defined in the following way:
> $U(a_i) = \sum_w ch(w|a_i)u(w)$,
> where $ch(w|a) = \sum_y c(a_i \Diamond_y \rightarrow w)y$, and where y ranges from 0 to 1 and
> $\sum_y c(a_i \Diamond_y \rightarrow w) = 1$.

According to Sobel (1986) definition 2 contains practical chance conditionals and not conditionals with chance outcomes à la Lewis (1981a).[150]

---

[150]With regard to my question whether he has a justification for putting the probability in the connective and not in the consequent like Lewis (1981a) Sobel gave me the following answer (personal communication from the 27th of August, 1998): "David Lewis, Wlodek Rabinowicz, and I engaged in a difficult correspondence on this question in 1981. My sense is that mine is a more unified and in ordinary terms natural theory of 'woulds', 'mights', and 'chances'. David replies in his *Philosophical Papers* Volume II (OUP) briefly to Wlodek's discussion ('Two Causal Decision Theories: Lewis vs. Sobel', T. Pauli (ed.), *Philosophical Essays Dedicated to Lennart Aqvist on his Fiftieth Birthday*, Uppsala: Filosofiska Studier, 1982). David comments on a case in which a 'coin is tossed fairly, with equal chance of head and tails' that 'will in fact fall heads'. A is the proposition that the coin is tossed. He says that according to his theory, at the

In order to improve his theory with regard to limitation 2 Sobel (1986) makes clear why definition 1 is limited to possible actions (1) which are causally possible and (2) which are open to the decision maker and why the former (1) are comprehended by the latter (2).

Definition 1 is limited to possible actions which are causally possible (Sobel 1986). For if a possible action $a_i$ isn't causally possible, then either (1) the proposition $a_i \mathrel{\square\!\!\rightarrow} s_j$ fails because a presupposition is missing and the causal conditional $a_i \mathrel{\square\!\!\rightarrow} s_j$ is neither true nor false or (2) although the causal conditional $a_i \mathrel{\square\!\!\rightarrow} s_j$ is true, it is true by stipulation to secure completeness and simplicity of theory.

Definition 1 is limited to possible actions which are open to the decision maker, that is definition 1 is limited to options (Sobel 1986, p. 420). For suppose the decision maker is almost certain that the future of a particular business firm is very good regardless of the number of investors, that many people share this view and will invest in this firm, and that it is therefore not open to the decision maker to be one of the few investors of this firm. Furthermore, the decision maker is almost certain that being an investor of this firm is open to the decision maker only if the future of this business firm is bad. Then the decision maker is almost certain that if he were one of the few investors of this firm, its future would be very good and that being one the few investors of this firm wouldn't have a causal influence on its future. Therefore the utility of the decision maker being one of the few investors of this firm is very great according to definition 1. In this case, however, the decision maker is almost certain that being one of the few investors of this firm isn't open to him, and that if it were open to him, the future of this firm would be bad. Therefore the utility of the decision maker being one of the few investors of this firm shouldn't be great in opposition to definition 1. Sobel

---

time of the toss, 'if it were that A, then there would be some chance that the coin would fall tails.... There would be some chance of it; but it would not happen.' That is, for David, at the moment of the toss, if it were the case that A, then it would be that there is a chance that H, and if it were the case that A, then it would be that it is not the case that H. These -- which spell out David's 'There would be some chance of it; but it would not happen' -- are awkward and at best odd sounding words. I think that, in ordinary terms, that if it were the case that A, then there is a 'chance' that H if only if it is not the case that it 'would' be that it is not the case that H. I think that the words 'there would be some chance of it; but it would not happen' understood naturally express an impossibility. I say of the case, that at the moment of the toss, if it were the case that A, then it might be the case that H, though it is not the case that if it were the case that A, then it would be the case that H. While 'wordy' there is nothing odd or puzzling about that. 'Though it might, it is not that it definitely would' is a way of saying there is a chance, but only a chance. Ordinarily 'chances' run between 0 and 1. David gets more specific by qualifying 'chance' -- for example, 'there would be a .5-chance that'. I get more specific by qualifying 'might' -- for example, 'it might, with a chance of .5, be that'."

(1986) concludes that the utility of a possible action of whose performability the decision maker isn't certain shouldn't depend on the decision maker's credence that if it were to happen, then it would have a certain effect, but on his credence that if he were to perform it, then it would have a certain effect.

In order to eliminate limitation 2 of definition 1, Sobel (1986) moves to

**definition 3.** For any possible action $a_i$, so that $c(\bigcirc a_i) > 0$, and for any possible world $w$ the utility of a possible action $a_i$ is defined in the following way:

$$U(a_i) = \sum_w c(a_i \mathbin{\Box\!\!\rightarrow} w)|\bigcirc a_i)u(w),$$

where "$\bigcirc a_i$" means "$a_i$ is open".

According to Sobel (1986, p. 421, 1994, p. 156) "... an action is to be *open* for an agent if and only if either it will take place or this agent can definitely do it or bring it about ...".[151] Sobel (1986) claims that being open is a stronger condition than being causally possible. For in the prisoner's dilemma the decision maker's confessing alone is causally possible, but it is not open to the decision maker, if the other decision maker confesses.

Sobel (1986, p. 421, 1994, p. 156) claims that the following principles for openness are valid: For every proposition $Y$

$$Y \rightarrow \bigcirc Y.$$

For any propositions $Y$ and $Z$, so that $Y$ entails $Z$

$$\bigcirc Y \rightarrow \bigcirc Z.$$

Furthermore, Sobel (1986, p. 421, 1994, pp. 156-157) assumes that if $a_i$ is an option in a decision problem, then $c(\bigcirc a_i) = 1$, and $a_i$ would be realised at every probable possible world if and only if the decision maker performed it or he decided for it, that is the

---

[151] I asked Sobel what kind of action concept he has in mind, when he defines openness of a possible action in such a way. Sobel answered (personal communication from the 21st of August, 1998): "I notice that I define 'open' and 'option' on pp. 156-7 in 'Notes' somewhat differently than I define 'open', 'completely under control', and 'option' on p. 176 of 'Partitions'. Also, my definitions of 'open' on pp. 156 and 176 are somewhat different. I think you would say that my definitions in 'Partitions' make actions 'active,' whereas those in 'Notes' leave room for 'passive' actions. I like the later definitions better. My definition of 'option' on pp. 156-7 does not, I think, agree with my gloss on options in the first sentence on p. 157." ['Notes' refers to Sobel (1986). 'Partitions' refers to Sobel (1989). Sobel's page specifications refer to Sobel (1994), which is a collection of Sobel's essays on decision-making.] In 1994 on p. 157 Sobel writes: "Options in a decision problem are to be things the agent is sure are completely under his control." And in 1994 on p. 176 Sobel writes: "Intuitively, for the agent in a decision problem, a thing is to be *completely under control* if and only if he can, by choice, do it or not, and it can happen only by his doing or choice. For such an agent, a thing is to be *open* if and only if he can bring it about by choice; ... . A thing is an *option* in a decision problem if and only if the agent is sure that it is completely under his control."

decision maker believes that options in a decision problem are completely under his control. Therefore Sobel (1986, p. 421, 1994, p. 157) assumes for any option $a_i$ in a decision problem and any proposition *Y* that the following principle is valid:

$(a_i \Box\!\!\rightarrow Y) \rightarrow \bigcirc Y$.

According to Sobel (1986) definition 4 encompasses the improvements of definition 2 and definition 3:

> **Definition 4.** For any possible action $a_i$, so that $c(\circledcirc a_i) > 0$, and for any possible world *w* the utility of a possible action $a_i$ is defined in the following way:
> $U(a_i) = \sum_w \{ch(w|a_i)\}u(w)$,
>
> where "$\circledcirc a_i$" means "$a_i$ is possibly open", where "$\{ch(w|a)\}$" is called a ramified chance, and where $\{ch(w|a)\} = \sum_y c[(a_i \Diamond_y \rightarrow w)|\circledcirc a_i]y$.

According to Sobel (1986, p. 422, 1994, p. 158) a possible action is possibly open for a decision maker if and only if there is a chance that it will happen or that this decision maker can perform it. Sobel (1986) believes that the condition of possible openness is related to but weaker than openness. He claims that there are valid principles for the connection between possible openness and openness and for the difference between them (cf. Sobel 1986, p. 423, 1994, pp. 158-159).

Furthermore, Sobel (1986, p. 422, 1994, p. 158) claims that the following principles for possible openness are valid: For every possible action $a_i$, proposition *Y*, and *x*

$\Diamond_x a_i \rightarrow \circledcirc a_i$;

and if $a_i$ entails *Y*,

$\circledcirc a_i \rightarrow \circledcirc Y$.

For every option $a_i$, proposition *Y*, at every probable possible world

$a_i \Diamond_x \rightarrow Y \rightarrow \circledcirc Y$.

According to Sobel (1986) the two conditions of openness and of possible openness are equivalent, if the following holds: The decision maker doesn't believe in conditional chances, and these conditional chances differ from objective conditional certainties and impossibilities. In such a case the decision maker is certain that a possible action is open to him if and only if he is certain that it is definitely open to him. But the two conditions don't have to be equivalent, if the decision maker believes in chances.

## *Sobel on Partitions of the Possible States of the World*

With regard to partitions of the possible states of the world Sobel (1986) claims that the analogue of Jeffrey's (1965) unrestricted partition theorem for the theory of definition 4 would be the following: For any possible action $a_i$, so that $c(\circledcirc a_i) > 0$, and for a partition $S$

$$U(a_i) = \Sigma_{s_j: s_j \in S \& c[\circledcirc(a_i \cap s_j)] > 0} \{ch(s_j|a_i\} U(a_i \cap s_j).$$

Furthermore, Sobel (1986) claims that it is attractive, but not plausible that Jeffrey's (1965) logic of decision works for every partition of the possible states of the world. Therefore Sobel (1986) provides two theoretically adequate partition theorems[152]: The first restricts the analogue to sufficiently fine partitions, whereas the second restricts the analogue to sufficiently exclusive partitions. According to Sobel (1986) a partition $S$ is sufficiently fine relative to a possible action $a_i$ if and only if $S$ analyses $a_i$ into parts or versions[153] whose utilities equal those of their probably possibly open parts or versions; and a partition $S$ is sufficiently exclusive relative to a possible action $a_i$ if and only if $S$ analyses $a_i$ into parts or versions, so that the decision maker is certain that not both of every pair are possibly open.

According to Sobel (1986) sufficiently fine partitions make all utility-relevant distinctions, and sufficiently exclusive partitions need not be sufficiently fine. Furthermore, natural partitions are always sufficiently exclusive, so Sobel (1986). For in the prisoner's dilemma each decision maker will be certain that not both confessing together with the other decision maker and confessing alone are possibly open.

Sobel (1986) claims that all natural partitions are theoretically adequate, but that not all theoretically adequate partitions are natural. For the chance partition[154] {I take both boxes, I take B2} would be an unnatural partition of the possible states of the world in Newcomb's problem. This partition would be theoretically adequate, though, for it is sufficiently exclusive. The partition is sufficiently exclusive, because the

---

[152]For further partition theorems see Sobel (1989).

[153]With regard to my question what kind of primitive terms Sobel uses in his causal decision theory Sobel gave me the following answer (personal communication from the 21st of August, 1998): "I work with partitions of options (possibilities for choices) and partitions of circumstances (circumstances in which choices might take place) and use 'c' to suggest 'circumstances'. Conjunctions of options and circumstances perhaps approximate 'consequences of options'. I think I usually refer to them, however, as 'versions of options'."

[154]Sobel (1986, p. 431, 1994, p. 166) gives this definition of a chance partition: For any proposition $a_i$ and any partition $S$, $S$ is a chance partition relative to $a_i$ if and only if the decision maker is certain of the following: If $a_i$ is causally possible, exactly one member of $S$ is true at each nearest $a_i$-world.

decision maker is certain that of (I take both boxes∩I take both boxes) and (I take both boxes∩ I take B2) only the first is open.

Sobel gives two partition theorems for sufficiently fine chance partitions and for sufficiently exclusive chance partitions (for the proofs cf. Sobel 1986, pp. 431-432 and pp. 435-436):

(1) **Partition theorem for sufficiently fine chance partitions**: For any option $a_i$ in a decision problem and any partition $S$: If $c(\circledcirc a_i) > 0$ and $S$ is a sufficiently fine chance partition relative to $a_i$, then the following holds:

$$U(a_i) = \sum_{s_j : s_j \in S \& c[\circledcirc(a_i \cap s_j] > 0} \{ch(s_j|a_i\} U(a_i \cap s_j).$$

(2) **Partition theorem for sufficiently exclusive chance partitions**: For any option $a_i$ in a decision problem and any partition $S$: If $c(\circledcirc a_i) > 0$ and $S$ is a sufficiently exclusive chance partition relative to $a_i$, then the following holds:

$$U(a_i) = \sum_{s_j : s_j \in S \& c[\circledcirc(a_i \cap s_j] > 0} \{ch(s_j|a_i\} U(a_i \cap s_j).$$

## *Sobel on Newcomb's Problem*

According to Sobel (1990) Newcomb's problem belongs to the class of Newcomblike problems which he presents as coherent challenges to evidential decision theories. (For Sobel's (1990) analysis of Newcomblike problems cf. chapter 1.7.) In opposition to Nozick (1969) and in agreement with Gibbard and Harper (1978) Sobel (1988c) argues that Newcomb's problem isn't a conflict between the principle of maximising conditional utility and the principle of strong dominance, but is a conflict between two maximising principles, namely an evidential maximising principle and a causal maximising principle. Furthermore, each maximising principle has a dominance principle as a corollary with which it doesn't conflict, so Sobel (1988c) (cf. chapter 1.2, in the section on the maximising principles and the principles of dominance with probabilistic and causal independence). In the following I will try to show how Sobel questions Newcomb's problem as a conflict between the principle of maximising conditional utility and the principle of dominance and how he establishes Newcomb's problem as a conflict between an evidential and a causal maximising principle.[155]

[155]Like Gibbard and Harper (1978) Sobel (1988c) distinguishes between the principle of maximising *V*-utility and the principle of maximising *U*-utility. While Sobel's *V*-utility corresponds to Jeffrey's (1965) conditional utility, Sobel's *U*-utility is in accordance with his causal decision theory of 1986.

First of all, Sobel (1988c) claims that taking both boxes is the right recommendation for the following two natural partitions of the possible states of the world in Newcomb's problem:

(1) $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$.

(2) $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$.

Yet from an empirical point of view, so Sobel (1988c), decision makers who take both boxes tend to favour the first partition, which is a causal partition, and decision makers who take B2 tend to favour the second partition, which is an evidential partition.

Thus both Jeffrey (1965) and Sobel (1988c) go against the empirical point of view in Newcomb's problem, when Jeffrey (1965) tries to defend a 1-box-solution for all partitions of the possible states of the world, and when Sobel (1988c) tries to defend a 2-boxes-solution for all natural partitions of the possible states of the world. Furthermore, on first sight it seems counterintuitive to get a 2-boxes-solution, if the decision maker partitions the possible states of the world evidentially; yet on second sight and with the aid of his practical conditionals Sobel is able to get this solution, which we will see later on. Figure 23 summarises the decision situation of Newcomb's problem for the first partition of the possible states of the world, and figure 24 summarises the decision situation of Newcomb's problem for the second partition.

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | \$ 1,000 | \$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | \$ 0 | \$ 1,000,000 |

Figure 23. Decision matrix of the possible outcomes which result from combining the possible actions $a_1, a_2$ with the first partition of the possible states of the world $s_1, s_2$.

| | $s_1$: The predictor has predicted correctly at $t_1$. | $s_2$: The predictor hasn't predicted correctly at $t_1$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | \$ 1,000 | \$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | \$ 1,000,000 | \$ 0 |

Figure 24. Decision matrix of the possible outcomes which result from combining the possible actions $a_1$, $a_2$ with the second partition of the possible states of the world $s_1$, $s_2$.

With regard to the first partition the decision maker believes that *c*(I take both boxes □→ there is \$ 1,000,000 in B2) = *c*(there is \$ 1,000,000 in B2), that *c*(I take B2 □→ there is \$ 1,000,000 in B2) = *c*(there is \$ 1,000,000 in B2), that *c*(I take both boxes □→ there is \$ 0 in B2) = *c*(there is \$ 0 in B2), and that *c*(I take B2 □→ there is \$ 0 in B2) = *c*(there is \$ 0 in B2), so Sobel (1988c). For the decision maker is certain that the content of B2 is causally independent of the decision maker's possible actions to take both boxes and to take B2. Sobel (1988c) sets *c*(there is \$ 1,000,000 in B2) = p, so that *c*(there is \$ 0 in B2) = (1-p). Then the following figure summarises the probabilities of practical conditionals for the combination of the possible actions with the first partition of the possible states of the world.

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | p | 1-p |
| $a_2$: I take the content of B2 at $t_3$. | p | 1-p |

Figure 25. Matrix of the probabilities of practical conditionals for the combination of the possible actions $a_1$, $a_2$ with the first partition of the possible states of the world $s_1$, $s_2$.

According to Sobel (1988c) practical independence is sufficient for the dominance principle of causal decision theory to apply. Sobel (1988c) defines practical independence in the following way:

$s_j$ is practically independent of $a_i$ if and only if $c(a_i \Box\!\!\rightarrow s_j) = c(s_j)$.

Sobel (1988c) claims that the condition of practical independence is satisfied in this case, for the rows in the matrix of the probabilities of practical conditionals are the same. Therefore the application of the dominance principle of causal decision theory leads to taking both boxes.

Furthermore, Sobel (1988c) defines certain causal independence in the following way:

$s_j$ is certainly causally independent of $a_i$ if and only if $c[(a_i \Box\!\!\rightarrow s_j)$ iff $s_j] = 1$.

Thus $s_j$ is practically independent of $a_i$, if $s_j$ is certainly causally independent of $a_i$.

Sobel (1988c) claims that in this case the recommendation of the subsidiary dominance principle of causal decision theory agrees with the recommendation of the fundamental principle of maximising utility of causal decision theory. Thus with regard to the first partition Sobel's (1988c) calculation of the utilities in accordance with causal decision theory yields the following:

$U(a_1)$ = p(1,001,000)+(1-p)1,000,
= p(1,000,000)+1,000,
$U(a_2)$ = p(1,000,000)+(1-p)0,
= p(1,000,000).

Therefore in this case Sobel (1988c) recommends to take both boxes.

In opposition to that evidential decision theory (Jeffrey 1965) yields the recommendation to take B2, so Sobel (1988c). For evidential decision theory doesn't use probabilities of practical conditionals, but uses conditional credences, that is credences of states of the world given the possible actions. The following figure summarises the conditional credences for the combination of the possible actions with the first partition of the possible states of the world.

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | [0] | [1] |
| $a_2$: I take the content of B2 at $t_3$. | [1] | [0] |

Figure 26. Matrix of the conditional credences for the combination of the possible actions $a_1$, $a_2$ with the first partition of the possible states of the world $s_1$, $s_2$. ("[ ]" means "nearly".)

According to Sobel (1988c) evidential independence is sufficient for the dominance principle of evidential decision theory to apply. Sobel (1988c) defines evidential independence in the following way:

$s_j$ is evidentially independent of $a_i$ if and only if $c(s_j|a_i) = c(s_j)$.

Sobel (1988c) claims that this case doesn't meet the condition of evidential independence, because the rows in the matrix of the conditional credences aren't the same. Therefore the subsidiary dominance principle of evidential decision theory cannot be applied. But the fundamental principle of maximising utility of evidential decision theory can be applied. Thus with regard to the first partition Sobel's (1988c) calculation of the utilities in accordance with evidential decision theory yields:

$V(a_1)$ = [0]1,001,000+[1]1,000,
= [1,000],
$V(a_2)$ = [1]1,000,000+[0]0,
= [1,000,000].

Therefore in this case evidential decision theory recommends to take B2, so Sobel (1988c).

With regard to the second partition Sobel (1988c) sets $c$(the predictor has predicted B2) = p, which equals or nearly equals $c$(there is \$ 1,000,000 in B2), so that p is used in the same way or almost in the same way as in the first partition. Then the following figure summarises the probabilities of practical conditionals for the

combination of the possible actions with the second partition of the possible states of the world.

| | $s_1$: The predictor has predicted correctly at $t_1$. | $s_2$: The predictor hasn't predicted correctly at $t_1$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | 1-p | p |
| $a_2$: I take the content of B2 at $t_3$. | p | 1-p |

Figure 27. Matrix of the probabilities of practical conditionals for the combination of the possible actions $a_1$, $a_2$ with the second partition of the possible states of the world $s_1$, $s_2$.

Sobel (1988c) claims that the dominance principle doesn't apply in this case, because there isn't a dominant possible action. Furthermore, the condition of practical independence isn't fulfilled in this case. For the rows in the matrix of the probabilities of practical conditionals aren't the same, so Sobel (1988c). Yet even in this case the principle of maximising utility of causal decision theory yields taking both boxes as in the first case. For with regard to the second partition Sobel's (1988c) calculation of the utilities in accordance with causal decision theory results in the following:

$U(a_1)$ = (1-p)1,000+ p(1,001,000),
= 1,000+p(1,000,000),
$U(a_2)$ = p(1,000,000)+(1-p)0,
= p(1,000,000).

Therefore in this case Sobel's (1988c) causal decision theory recommends to take both boxes.

In opposition to that evidential decision theory (Jeffrey 1965) again yields the recommendation to take B2, so Sobel (1988c). The following figure summarises the conditional credences for the combination of the possible actions with the second partition of the possible states of the world:

| | $s_1$: The predictor has predicted correctly at $t_1$. | $s_2$: The predictor hasn't predicted correctly at $t_1$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | [1] | [0] |
| $a_2$: I take the content of B2 at $t_3$. | [1] | [0] |

Figure 28. Matrix of the conditional credences for the combination of the possible actions $a_1, a_2$ with the second partition of the possible states of the world $s_1, s_2$.

Sobel (1988c) claims that this case meets the condition of evidential independence, because the rows in the matrix of the conditional credences are the same. But the dominance principle of evidential decision theory cannot be applied, for there isn't a dominant possible action. The fundamental principle of maximising utility of evidential decision theory can be applied, though. Thus with regard to the second partition Sobel's calculation of the utilities in accordance with evidential decision theory yields:

$V(a_1)$ = [1]1,000+[0]1,001,000,
= [1,000],
$V(a_2)$ = [1]1,000,000+[0]0,
= [1,000,000].

Therefore in this case evidential decision theory recommends to take B2, so Sobel (1988c). Furthermore, Sobel has established that Newcomb's principle isn't a conflict between the principle of maximising conditional utility and the principle of dominance, but is a conflict between an evidential and a causal maximising principle. Because Sobel defends a causal decision theory (Sobel 1986), he solves the conflict by giving up the evidential maximising principle and by maintaining the causal maximising principle.

## *A Critique of Sobel's Position*

Sobel's (1986) causal decision theory can be compared with Eells' (1981, 1982, 1985) evidential decision theory, because both theories belong to the most elaborated rational decision theories, and because both theorists stick to their 2-boxes-solution in Newcomb's problem over the years. Yet Sobel's theory is much more complicated; in fact it is so complicated that from the viewpoint of economical theory building it

shouldn't be recommended as a rational decision theory. The following features make Sobel's theory so complicated:

(1) Practical conditionals which provide the basis for Sobel's practical chance conditionals;

(2) conditional chances in the calculation of the utility of a possible action;

(3) the distinction between causally possible, open, and possibly open;

(4) the distinction between natural partitions, sufficiently fine partitions, and sufficiently exclusive partitions.

Furthermore, these features can be criticised or praised on the following grounds:

**Ad 1**: Even though Sobel explains why he considers practical conditionals as causal conditionals, but not as purely causal conditionals, and even though he gives truth conditions for practical conditionals, my intuitions with regard to practical conditionals are rather vague. Furthermore, if one of the primitive terms of a theory is vague, the theory itself becomes vague. Moreover, when Sobel explains why he considers practical conditionals as causal conditionals, but not as purely causal conditionals, he doesn't make it clear in his example of the prisoner's dilemma in which way the other decision maker's confession is determined, if it isn't causally determined. Thus Sobel's theory has a weak foundation which he himself even admits.

Sobel uses practical chance conditionals, but doesn't provide a logic for these conditionals; he just states that the vagueness of these conditionals should be resolved appropriately to contexts of decision which allows for a lot of arbitrariness. For Sobel doesn't state conditions how an appropriate resolution to contexts of decision looks like. Thus Sobel's theory still has to be supplemented by conditions for an appropriate resolution to contexts of decision with regard to practical chance conditionals.

**Ad 2**: Sobel uses conditional chances in his calculation of the utility of a possible action which is problematical for the following reasons:

(1) First, Sobel defines his ramified chances, $\{ch(w|a)\}$, by credences, that is $\{ch(w|a)\} = \sum_y c[(a_i \Diamond_y \rightarrow w)|◎a_i]y$. Then these credences are about practical chance conditionals, so that Sobel's theory is in the ultimate analysis objective, and therefore is limited in its applicability to decision makers who believe in chances which is a problem. For there may be decision makers who don't believe in chances, because they don't have epistemic access to chances, at best they can approximate chances by the law of large numbers or some other limit theorem. But even then, what happens

with short and medium runs? Yet this also depends on what kind of chance concept Sobel favours. For if Sobel is an objectivist, his chances hold for all decision makers, while if Sobel is a subjectivist, his chances emerge from the decision maker's credences, so that chances are relative to the decision maker's credences and are covertly epistemological (cf. Spohn 1988, p. 105). Because Sobel doesn't define his chances in terms of credences and because he doesn't refer to the law of large numbers, to some other limit theorem, or to Lewis' (1980) principal principle, Sobel seems to have an objectivist understanding of chance.

Furthermore, if one has a subjectivist understanding of chance, like Skyrms (1984) or Jeffrey (1988)[156], then the decision maker's epistemic access to chances is by means of his credences. Moreover, with a subjectivist understanding of chance the decision maker's epistemic access to chances is only a problem, if chances are approximated by means of the law of large numbers or by means of some other limit theorem. For if the decision maker isn't able to go to the limit at decision time, but is stuck at short or medium runs, then he - at best - has a rough approximation of his chances. This can happen to the decision maker in Jeffrey's (1988) probabilism (cf. chapter 2.4, in the section on a critique of Jeffrey's probabilism), but not to the decision maker in Skyrms' (1984) causal decision theory. For with regard to the latter there is no limit theorem involved in Skyrms' rewriting of chances in terms of credences.

But if one has an objectivist understanding of chance like Sobel (1986), then the question how the decision maker can have epistemic access to chances becomes vital. For the decision maker cannot refer to his credences for an approximation of his chances, so that it is unclear to me where the decision maker gets his chances from. Furthermore, Sobel doesn't provide us with any clue how the decision maker obtains chances. Moreover, if there are decision makers who don't believe in chances, they cannot even make a recourse to their credences for calculating the utility of a possible action in Sobel's theory. Therefore Sobel's theory which relies on chances fails to provide an account of how the decision maker obtains chances, and what to do when the decision maker fails to obtain chances; these two defects still have to be overcome for Sobel's theory to work properly in this respect.

(2) The decision maker is supposed to condition his practical chance conditionals by possible actions which are possibly open for him. Furthermore, the

---

[156]In my opinion one can understand Jeffrey's (1988) final credences as objectified credences, so that Jeffrey actually employs a chance concept in his calculation of the utility of a possible action.

decision maker is supposed to assign a credence to this combined term, so that Spohn's principle is violated (cf. chapter 3.6), which states that "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts.*" (Spohn 1977, p. 114). Because Spohn's principle seems to be valid (cf. chapter 3.6), Sobel's formula for calculating the utility of a possible action is false. One could object to that conclusion by claiming that Sobel's possible actions which are possibly open for the decision maker are probabilistic acts à la Jeffrey (1965, 1983) (cf. chapter 2.2), so that Spohn's principle doesn't apply to them. For Spohn's principle only applies to acts and not to probabilistic acts.

Thus the question has to be answered whether Sobel's possible actions which are possibly open for the decision maker are probabilistic acts. While a possible action is possibly open for a decision maker if and only if there is a chance that it will happen or that this decision maker can perform it (Sobel 1986, p. 422, 1994, p. 158), in the case of probabilistic acts the decision maker believes that a possible action isn't in his power to make true; he can only try to make it true, if he wants to (cf. chapter 2.2). If one looks at Sobel's definition there can be three general reasons why there is only a chance that the decision maker can perform a particular possible action. (1) There are inside factors which may prevent the decision maker from performing a particular possible action. (2) There are outside factors which may prevent the decision maker from performing a particular possible action. (3) There are inside and outside factors which may prevent the decision maker from performing a particular possible action. Yet all these reasons amount to the following: The decision maker doesn't have full control over this particular possible action. Thus Sobel's possible actions which are possibly open for the decision maker are probabilistic acts à la Jeffrey, and Spohn's principle isn't violated.[157]

**Ad 3**: I think that Sobel's distinction between "causally possible", "open", and "possibly open" is very valuable. Unfortunately Sobel doesn't make clear how "causally possible" and "possibly open" are related to each other, although he states that being causally possible is comprehended by being open.

**Ad 4**: With regard to Sobel's (1986) distinction between natural partitions, sufficiently fine partitions, and sufficiently exclusive partitions one can say that Sobel has to be praised for making it explicit that rational decision theory has to determine which partitions of the possible states of the world should figure in analysing decision

[157] Yet in chapter 3.6 we will see that Spohn's principle even applies to probabilistic acts, so that Spohn's principle is violated in Sobel's theory after all.

situations. Yet Sobel's account can be criticised, because Sobel's causal decision theory works for natural partitions of the possible states of the world, which isn't equivalent with rational partitions of the possible states of the world. Thus the aim of Sobel's theory is different from what I demanded from a rational decision theory to accomplish, namely first, to find out which possible actions and which possible states of the world should figure in decision situations, second, to find out which possible action should the decision maker decide for. Sobel's aim differs from my aim with regard to the first point. While Sobel wants to determine which partitions of the possible states of the world are natural for analysing a decision situation, I would like to determine which partitions of the possible states of the world are rational for a decision situation. Thus the question arises which aim is the better one for a rational decision theory.

To answer this question we first have to know what natural partitions of the possible states of the world are. According to Sobel natural partitions of the possible states of the world are partitions that come naturally to mind, when one is conducting a Jeffrey-style (1965, 1983) Bayesian analysis of a decision. With regard to Newcomb's problem this includes the following two partitions:

(1) $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$.

(2) $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$.

Yet I'm sure there are decision makers who would deny that both partitions come naturally to their mind, when they are conducting a Jeffrey-style Bayesian analysis of a decision in Newcomb's problem. For Jeffrey's (1965) logic of decision is an evidential decision theory, so that the evidential partition of the possible states of the world seems to be more natural than the causal partition. Furthermore, if one considers a rational decision theory which is partition invariant, like Jeffrey's (1965) logic of decision, it is difficult to say which partitions come naturally to mind. Because Sobel doesn't make it more precise what it means to come naturally to mind, when one is conducting a Jeffrey-style Bayesian analysis of a decision, the term natural partitions remains vague and allows for a lot of arbitrariness.

Thus although the term natural partitions of the possible states of the world is vague, the question still is which rational decision theory is the better one - a rational

decision theory which works for natural partitions or a rational decision theory which works for rational partitions.

To answer this question it will help to look at Rabinowicz' (1982) persuasive claim that Lewis' (1981a) point of view to apply rational decision theory not only to fully rational decision makers, but also to partly rational decision makers must have some limits. For we cannot allow the decision maker to be as irrational as he wants to be. Rabinowicz supports his point of view by the following example: Rational decision theory would have difficulties with a decision maker whose beliefs cannot be represented by a credence function, because he believes in impossibilities, and/or whose wants cannot be consistently represented by a utility function. Thus rational decision theory has to set some limits to the irrationality of the decision maker.

Let's consider the following cases to find some limits to the irrationality of the decision maker:

(1) The decision maker's belief in clairvoyance or in revelation;

(2) the decision maker's belief in backwards causation;

(3) the decision maker's belief in chances;

(4) the decision maker's belief in natural partitions of the possible states of the world.

(1) The decision maker's belief in clairvoyance or in revelation seem to me clear cut cases for non-permissible irrationalities of the decision maker. For there is no scientifically accepted theory which speaks in favour of clairvoyance or revelation. Furthermore, common sense tells us that there is no clairvoyance or revelation in this world. Thus the decision maker's belief in clairvoyance or in revelation are clear cut cases for non-permissible irrationalities of the decision maker.

(2) The decision maker's belief in backwards causation seems to me a borderline case with regard to non-permissible irrationalities of the decision maker. For according to Dowe, Oakley, and Rosier (forthcoming) backwards causation is involved in at least four different hypotheses in twentieth century theoretical physics which are the following: (1) Tachyons, that is particles postulated to travel faster than the speed of light, involve backwards causation (cf. Horwich 1987). (2) The Gödel loop (Gödel 1949), that is there are solutions to the field equations of Einstein's general theory of relativity which allow for the possibility of closed causal loops. (3) The Feynman electron (Feynman 1949), that is Feynman's hypothesis that the positron is an electron which travels backwards in time. (4) The backwards in time interpretation of quantum mechanics (cf. for example Cramer 1986), that is the correlations in Bell experiments

(Bell 1964) are explained, if one supposes that a later measurement can affect an earlier state backwards in time. Thus backwards causation seems to be at least in science a hypothesis worth for discussion. Yet there are a lot of philosophers arguing against backwards causation. Mellor (1981, chapter 10, 1991b, 1995, pp. 224 ff.) even holds that all backwards causation entails causal loops and that there is an a priori argument against causal loops, so that backwards causation can be ruled out a priori, too. Furthermore, common sense tells us that there is no backwards causation in our world. All in all the decision maker's belief in backwards causation isn't a clear cut case for non-permissible irrationalities of the decision maker.

(3) The decision maker's belief in chances is no clear cut case for a non-permissible irrationality of the decision maker. For even if we don't have any epistemic access to chances, as long as some decision theorists deem it rational to use chances for calculating the utility of a possible action, it would be too much to demand from the decision maker to question the adequacy of chances for calculating the utility of a possible action. Furthermore, common sense doesn't speak against chances. Moreover, there is no a priori argument which shows that chances are downright impossible. Thus the decision maker's belief in chances is no clear cut case for a non-permissible irrationality of the decision maker.

(4) I don't think that the decision maker's belief in natural partitions is a clear cut case for a non-permissible irrationality of the decision maker. For as long as decision theorists still discuss which partitions of the possible states of the world are adequate for analysing decision situations, it would be too much to demand from the decision maker to know which partitions are the adequate ones. Thus the decision maker's belief in natural partitions doesn't seem to be a non-permissible irrationality.

Yet there are some partitions of the possible states of the world in Newcomb's problem which everybody would consider as non-permissible irrationalities of the decision maker. For example, the following two partitions of the possible states of the world are such cases: (1) $s_1$: I take the content of both boxes at $t_3$. vs. $s_2$: I take the content of B2 at $t_3$. (2) $s_1$: I take the boxes with my left hand. vs. $s_2$: I take the boxes not with my left hand (cf. chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 2). For with regard to the first case nothing should be doubly represented in the decision matrix, so that the possible actions of the decision maker shouldn't figure as the possible states of the world in Newcomb's problem. Furthermore, possible outcomes should ensue from the combination of the possible actions of the decision maker with the possible states of

the world which seems to be impossible, if the possible actions are doubly represented in the decision matrix. With regard to the second case nothing of irrelevance should figure in the decision matrix, so that the hand with which the decision maker takes the boxes shouldn't figure in the decision matrix. Moreover, Sobel wouldn't consider these two partitions as natural ones. Thus although some partitions of the possible states of the world in Newcomb's problems are clear cut cases for non-permissible irrationalities of the decision maker, Sobel's natural partitions don't seem to belong to these cases.

Therefore although Skyrms' (1980, 1982, 1984) causal decision theory seems to be adequate as a rational decision theory for fully rational decision makers who use rational partitions of the possible states of the world (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory, and chapter 3.3), Sobel's (1986) causal decision theory even seems to be adequate as a rational decision theory for partly rational decision makers who don't commit the blunder of being non-permissible irrational by using every partition of the possible states of the world, but use natural partitions of the possible states of the world instead.

According to Rabinowicz (1988, p. 409) Sobel's (1986) causal decision theory - in opposition to Gibbard and Harper's (1978) causal decision theory - doesn't presuppose the validity of conditional excluded middle, which in my opinion is an advantage of Sobel's theory. For as Lewis (1981a) shows the principle of conditional excluded middle is open to two objections which can't be overcome completely (cf. chapter 3.5). Thus Sobel's causal decision theory can be recommended for not presupposing the validity of conditional excluded middle.

With regard to Newcomb's problem Sobel's (1986) causal decision theory proposes the right solution, namely to take both boxes. Yet Sobel's causal decision theory is much too complicated from an economical point of view of theory building - one exception is that Sobel doesn't presuppose the validity of conditional excluded middle. Furthermore, Sobel's practical conditionals which provide the basis for his practical chance conditionals are vague. Moreover, Sobel doesn't provide a logic for his practical chance conditionals.

Because Sobel's causal decision theory relies in its ultimate analysis on practical chance conditionals with an objectivist understanding of chance, it is limited to decision makers who believe in chances. Yet with regard to Newcomb's problem there may be decision makers who don't believe in chances, so that they cannot come to a solution, if they apply Sobel's theory. Furthermore, it is unclear what the chances are in the practical chance conditional "if it were the case that $a_i$, then it might - with a chance of

*x* - be the case that $s_j$" in Newcomb's problem. For the decision maker only knows that the predictor's reliability in making correct predictions is very high. Yet the interesting thing is - regardless which chances figure in the calculation of the utility of a possible action - the decision to take both boxes ensues, so that the decision maker doesn't have to know the chances in Newcomb's problem. He just can place any number ranging from 0 to 1 in the formula for calculating the utility of a possible action, and the 2-boxes-solution results.

Sobel's distinction between "causally possible", "open", and "possibly open" can be applied to Newcomb's problem. For the decision maker's possible actions to take B2/both boxes are causally possible; furthermore, they are open to the decision maker and even possibly open, so that the utility of a possible action can be calculated by means of Sobel's four definitions. Yet if the decision maker believes that his possible actions are not even possibly open for him, one cannot calculate the utility of a possible action by means of definition 4 in Sobel's theory.

Sobel's causal decision theory works with natural partitions of the possible states of the world, so that two partitions of the possible states of the world, a causal and an evidential one, are marked out as natural in Newcomb's problem. Yet natural partitions of the possible states of the world aren't equivalent with rational partitions of the possible states of the world which fully rational decision makers are supposed to use. Because natural partitions of the possible states of the world include rational partitions of the possible states of the world and not vice versa, and because decision makers who use natural partitions don't commit the blunder of being non-permissible irrational, Sobel's theory could work for partly rational decision makers as well as for fully rational decision makers.

## 3.5 Lewis' Unification of Causal Decision Theories

Lewis' (1981a) causal decision theory is based on two theses: First, the thesis that the decision maker should maximise *U*-utility calculated by means of dependency hypotheses; according to Lewis (1981a) this thesis is implicitly accepted by Gibbard and Harper (1978), Skyrms (1980), and Sobel (unpublished).[158] Second, the thesis that the dependency hypotheses are conjunctions of probabilistic full patterns.

[158]In the following I will not deal with Lewis' (1981a) proofs for this, because a solution to Newcomb's problem isn't affected by the correctness or incorrectness of these proofs.

Lewis' (1981a) causal decision theory can be seen as an advancement of Gibbard and Harper's (1978) causal decision theory and is built on Lewis' (1973) semantics of possible worlds. The decision maker's beliefs and wants are idealistically represented by the decision maker's credence and utility functions which are defined over single possible worlds. Each world $w$ has a credence $c(w)$ which measures the decision maker's credence that $w$ is the actual world. The credences range from 0 to 1 and sum to 1. And each world $w$ has a utility $u(w)$ which measures how satisfactory it seems for the decision maker that $w$ is the actual world. The utilities range over a linear scale with arbitrary 0 and unit. Lewis (1981a) also defines credence for sets of worlds and calls these sets propositions. According to Lewis (1981a) a proposition holds only at those worlds which are members of the proposition.

Assuming that the numbers of worlds are finite Lewis (1981a) defines unconditional credence: For any proposition $Y$,

$$c(Y) = \sum_{w \in Y} c(w).$$

Conditional credence is defined in the following standard way, so Lewis (1981a): If $c(Z) > 0$, then

$$c(Y|Z) = c(Y \cap Z)/c(Z).$$

Lewis (1981a) defines $V$-utility:

$$V(Y) = \sum_{w} c(w|Y)u(w),$$
$$= \sum_{w \in Y} c(w)u(w)/c(Y).$$

According to Lewis (1981a, p. 6) a partition is a set of propositions of which exactly one holds at any world.[159] Suppose the variable $T$ ranges over any partition. Then Lewis' (1981a) definitions yield the following two rules of additivity for credence on the one hand and for the product of credence and $V$-utility on the other hand:

$$c(Y) = \sum_{Z} c(Y \cap T),$$
$$c(Y)V(Y) = \sum_{T} c(Y \cap T)u(Y \cap T).$$

Then the averaging rule for $V$-utility holds:

$$V(Y) = \sum_{T} c(T|Y)u(Y \cap T).$$

From this Lewis (1981a) derives an alternative definition of $V$-utility. Suppose, for any number $v$, $[V = v]$ is a value-level proposition that holds at just those worlds $w$ for which $V(w) = v$. Because the value-level propositions build a partition, the following alternative definition of $V$-utility holds:

---

[159] Lewis (1981a) doesn't go into more detail here.

$$V(Y) = \sum_{v} c([V = v]|Y)v.$$

With regard to the decision maker's possible actions Lewis (1981a) claims: Suppose that one has a partition of propositions which distinguishes worlds on grounds of the decision maker acting differently, that the decision maker can act in such a way that he makes any one of these propositions hold, and that he cannot act in such a way that he makes any proposition obtain which implies, but is not implied by a proposition in the partition. Furthermore, suppose the partition yields the most detailed specifications of the decision maker's possible actions over which he has control. Then according to Lewis (1981a) this is the partition of the decision maker's alternative options. Lewis (1981a) uses the variable $a_i$ to range over these options.

## *Lewis on Newcomblike Problems*

Lewis (1981a) claims that Newcomblike problems show that causal distinctions are needed in rational decision theory. For Jeffrey's (1965) logic of decision gives the wrong recommendation in Newcomblike problems. It recommends an option which is good news for a certain state of the world over which the decision maker has no control.

According to Lewis (1981a) Newcomblike problems have the following structure: Suppose someone offers you a small good, so that you can take it or not. And suppose you may suffer a great evil, but you believe this evil thing is completely outside your control. Furthermore, suppose you believe some prior state of the world, which might obtain or not, and which is completely outside your control, would be useful for deciding to take the good and for suffering the evil. Therefore taking the good would be evidence that the prior state of the world obtains, and that your chance of suffering the evil is higher than you believed, which is bad news, but not a reason for declining the good. For if the prior state of the world obtains, you cannot change it by declining the good, you only shield yourself from the bad news, which is useless. Therefore you should take the good. Because Jeffrey's (1965) logic of decision recommends the irrational option, namely declining the good, Lewis (1981a) rejects Jeffrey's theory.

Lewis (1981a) puts Newcomblike problems in the following form: $G$ and $\neg G$ be the propositions that you take the good and that you decline it. $E$ and $\neg E$ be the propositions that you suffer the evil and that you don't suffer it. Furthermore, let the good add $+g$ to the utility of a world, and let the evil add $-e$ to the utility of a world; suppose $+g$ and $-e$ are additive; and there is an arbitrary 0 when $+g$ and $-e$ are absent. Then the $V$-utilities of $G$ and $\neg G$ are:

$$V(G) = c(E|G)u(E\cap G)+c(\neg E|G)u(\neg E\cap G),$$
$$=-ec(E|G)+g,$$
$$V(\neg G) = c(E|\neg G)u(E\cap\neg G)+c(\neg E|\neg G)u(\neg E\cap\neg G),$$
$$=-ec(E|\neg G).$$

This means that $\neg G$ is *V*-maximising if and only if the difference $(c(E|G)-c(E|\neg G)) > g/e$.

According to Lewis (1981a) Newcomb's problem differs from the other Newcomblike problems in that $c(E|G)$ is close to 1 and $c(E|\neg G)$ is close to 0, so that declining the good, that is taking B2 in Newcomb's problem, is *V*-maximising by a large amount. Lewis (1981a) claims that in comparison with Newcomb's problem the eggs-benedict-for-breakfast-problem (cf. Skyrms 1980, p. 129), Fisher's problem (cf. Skyrms 1984, p. 65), the loafing-weak-heart-problem (cf. Lewis 1981a, pp. 9-10), and the prisoner's dilemma (cf. Lewis 1979a) are more moderate Newcomblike problems. For with regard to them the difference between $c(E|G)$ and $c(E|\neg G)$ is greater than $g/e$, but is much smaller than in Newcomb's problem, so Lewis (1981a). Furthermore, Lewis (1981a) thinks that these Newcomblike problems are more realistic than Newcomb's problem.

## *Lewis on Dependency Hypotheses*

Lewis (1981a) supposes that the decision maker knows how the things he is interested in do and do not causally depend on his present possible actions. Lewis (1981a) calls the proposition which the decision maker knows a dependency hypothesis for that decision maker at that time, that is a maximally specific proposition which contains information about how the things the decision maker is interested in do and do not causally depend on his present possible actions. Since each proposition has a truth value, and since the dependency hypotheses are maximally specific and cannot differ without conflicting, they build a partition.

Lewis (1981a) uses the terminology of Gibbard and Harper (1978) by distinguishing between *V*-utility and *U*-utility. Lewis claims that within a single dependency hypothesis *V*-maximising is correct. Failures of *V*-maximising appear only if the decision maker spreads his credence over several dependency hypotheses, and only if the decision maker's possible actions might be evidence for some dependency hypotheses and against other dependency hypotheses.

According to Lewis (1981a) the decision maker should do the following, if he spreads his credence over several dependency hypotheses (Lewis uses $K_k$ to range over dependency hypotheses): First, Lewis defines the *U*-utility of an option:

$$U(a_i) = \sum_{k=1}^{l} c(K_k)u(a_i \cap K_k).$$

Second, Lewis (1981a) claims that the decision maker should adopt the

> **principle of maximising *U*-utility**: In a given decision situation *D* the decision maker *X* should decide for an option $a_i$ with maximal *U*-utility.

In a further step Lewis (1981a) defines the *V*-utility of an option:

$$V(a_i) = \sum_{k=1}^{l} c(K_k|a_i)u(a_i \cap K_k).$$

And the principle of maximising *V*-utility is (Lewis 1981a):

> **The principle of maximising *V*-utility**: In a given decision situation *D* the decision maker *X* should decide for an option $a_i$ with maximal *V*-utility.

According to Lewis (1981a) Jeffrey's (1965) logic of decision is an adequate rational decision theory, whenever the dependency hypotheses are probabilistically independent of the options, so that all of the $c(K_k|a_i)$'s equal the corresponding $c(K_k)$'s.

Lewis (1981a) thinks that his rational decision theory is causal in two different respects. For on the one hand the dependency hypotheses are causal in their content. And on the other hand the dependency hypotheses themselves are causally independent of the decision maker's possible actions. According to Lewis (1981a) the correct partition is the partition of the dependency hypotheses emphasising their causal content and not their causal independence.

Lewis (1981a) shortly considers the case in which any of the credences $c(a_i \cap K_k) = 0$. Then $u(a_i \cap K_k)$ becomes an undefined sum of quotients with denominator 0, so that $U(a_i)$ is undefined and cannot be compared with the utilities of the other options. According to Lewis (1981a), however, this case will never arise. For $c(a_i \cap K_k) = 0$, that is absolute certainty, amounts to a firm resolve never to change your mind, no matter what happens, and that is objectionable, so Lewis (1981a).

In the following Lewis (1981a) expands his causal decision theory by applying the averaging rule to $u(a_i \cap K_k)$ in Lewis' definition of the *U*-utility of an option (suppose *Z* ranges over any partition):

$$U(a_i) = \sum_{k=1}^{l} \sum_{Z} c(K_k)c(Z|a_i \cap K_k)u(a_i \cap K_k \cap Z).$$

According to Lewis (1981a) a partition is rich if and only if for every member $s_j$ of that partition and for every option $a_i$ and dependency hypotheses $K_k$, $u(a_i \cap K_k \cap s_j) = u(a_i \cap s_j)$. In the following Lewis (1981a) uses $s_j$ to range over rich partitions.

Suppose the partition is rich, then the utility terms in the previous equation can be partly factored out, so that Lewis (1981a) obtains:

$$U(a_i) = \sum_{j=1}^{m} (\sum_{k=1}^{l} c(K_k)c(s_j|a_i \cap K_k))u(a_i \cap s_j).$$

## *Lewis on Gibbard and Harper*

Because Lewis' (1981a) causal decision theory is an advancement of Gibbard and Harper's (1978) causal decision theory, Lewis' account of Gibbard and Harper's theory and the derivation of Lewis' theory from Gibbard and Harper's theory are shortly presented.

First, I start with Lewis' (1981a) account of Gibbard and Harper's (1978) causal decision theory: In ordinary language dependency hypotheses are usually expressed by counterfactuals. For suppose the decision maker wants Bruce, the cat, to purr. The decision maker's options are brushing, stroking, and leaving alone. The alternatives that Bruce purrs loudly, softly, or not at all build a rich partition. Then much of the decision maker's credence goes to the following dependency hypothesis which is expressed by three counterfactuals:

I brush Bruce $\Box\!\!\rightarrow$ he purrs loudly;

I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly;

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr.

This dependency hypothesis states that soft and loud purring are causally dependent on the decision maker's options. Furthermore, it specifies the extent of the decision maker's influence, namely full control, and it specifies the direction of the decision maker's influence, namely what he must do to get what. But the decision maker also gives some credence to other dependency hypotheses, for example, to the following:

I brush Bruce $\Box\!\!\rightarrow$ he doesn't purr;

I stroke Bruce $\Box\!\!\rightarrow$ he doesn't purr;

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr.

This dependency hypothesis states that the lack of purring is causally independent of the decision maker's options. Comparing both dependency hypotheses with each other the pattern of counterfactuals and not one counterfactual alone expresses causal dependence or causal independence.

According to Lewis (1981a) Gibbard and Harper (1978) put the following constraints on their counterfactuals: The antecedent and the consequent of their

counterfactuals must specify entirely distinct occurrences which can cause in case of the antecedent and which can be caused in case of the consequent. Furthermore, backtracking counterfactuals are excluded. Lewis (1981a) states that Gibbard and Harper (1978) define causal counterfactuals in the following way: Causal counterfactuals can belong to patterns of causal dependence or causal independence. According to Lewis (1981a) Gibbard and Harper (1978) consider causal counterfactuals of the form $a_i \boxrightarrow s_j$, and he states that Gibbard and Harper define a full pattern as a set consisting of one counterfactual of the form $a_i \boxrightarrow s_j$ for each option. Furthermore, Lewis (1981a) claims that the conjunction of the counterfactuals in a full pattern yield a dependency hypothesis in Gibbard and Harper's (1978) causal decision theory. Finally, Gibbard and Harper (1978) assume that there is a full pattern for each world (Lewis 1981a). According to Lewis (1981a) this completes Gibbard and Harper's (1978) causal decision theory.

Second, I present Lewis' (1981a) reasons for improving Gibbard and Harper's (1978) causal decision theory and the derivation of Lewis' (1981a) theory from Gibbard and Harper's (1978) theory.

According to Lewis (1981a) Gibbard and Harper's (1978) assumption that there is a full pattern for each world is a consequence of Stalnaker's (1968) principle of conditional excluded middle. The latter is open to two objections, so Lewis (1981a).

The first objection, which Lewis (1981a) calls the arbitrariness objection, is: The principle of conditional excluded middle makes arbitrary decisions. For if you follow the principle, the way things would be on a false, but possible supposition $X$ is no less specific than the things actually are. But according to Lewis (1981a) some questions about how things would be, if $X$ were the case, have no non-arbitrary answers like the question: If you had a sister, would she like blintzes? Lewis (1981a) claims that the arbitrariness objection is no problem for Gibbard and Harper's (1978) causal decision theory. For the less specific the supposition $X$, the less it settles, and the more far-fetched the supposition $X$, the less can be settled by actuality; and the less is settled otherwise, the more must be settled arbitrarily or not at all. But the supposition that the decision maker realises one of his options is neither unspecific nor far-fetched. Therefore the arbitrariness objection may work against the principle of conditional excluded middle, but not against Gibbard and Harper's (1978) assumption that there is a full pattern for each world (Lewis 1981a).

The second objection is the chance objection (Lewis 1981a): Suppose the decision maker believes that the actual world may be indeterministic, so that many

things are settled by chance processes. Then the decision maker may give little credence to worlds where full patterns hold or to counterfactuals of the form $a_i \Box\!\!\rightarrow s_j$ that make up these full patterns. Therefore Gibbard and Harper's (1978) conception of dependency hypotheses as conjunctions of full patterns is too narrow, so Lewis (1981a). In the following Lewis (1981a) tries to adjust Gibbard and Harper's (1978) causal decision theory to the case, in which decision makers give credence to indeterministic worlds, by distributing the decision maker's credence over contingent propositions about single-case objective chances.

## *Lewis' Causal Decision Theory*

Suppose the decision maker decides for a rich partition which meets Gibbard and Harper's (1978) demand for distinct occurrences. Suppose the variable p ranges over probability distributions for this rich partition, so that the decision maker assigns a number p($s_j$) to each $s_j$ in the partition ranging from 0 to 1, and so that the p($s_j$)'s sum to 1. Suppose [P = p] is the proposition that only holds at those worlds where the chances of the $s_j$'s are correctly given by the function p. In Lewis' (1981a) terminology [P = p] is a chance proposition, and the chance propositions build a partition. $a_i \Box\!\!\rightarrow$ [P = p] is a causal counterfactual which goes from the decision maker's options to the chance propositions. Lewis (1981a) then defines a probabilistic full pattern as a set which contains exactly one causal counterfactual of the form $a_i \Box\!\!\rightarrow$ [P = p] for each option. Lewis (1981a) claims that the conjunction of the causal counterfactuals in any probabilistic full pattern is a causal dependency hypothesis which specifies causal dependence or independence of the chances of the $s_j$'s on the $a_i$'s.

Lewis (1981a) gives the following three chance propositions [P = $p_1$], [P = $p_2$], and [P = $p_3$] as examples:

[P = $p_1$]: The chance that Bruce purrs loudly is 50%; the chance that he purrs softly is 40%; and the chance that he purrs not at all is 10%.

[P = $p_2$]: The chance that Bruce purrs loudly is 30%; the chance that he purrs softly is 50%; and the chance that he purrs not at all is 20%.

[P = $p_3$]: The chance that Bruce purrs loudly is 10%; the chance that he purrs softly is 10%; and the chance that he purrs not at all is 80%.

Lewis (1981a) gives the following example of a causal dependency hypothesis which is the conjunction of causal counterfactuals in a probabilistic full pattern:

I brush Bruce $\Box\!\!\rightarrow$ [P = $p_1$] holds;

I stroke Bruce $\Box\!\!\rightarrow$ [P = $p_2$] holds;

I leave Bruce alone $\Box\!\!\rightarrow$ [P = $p_3$] holds.

According to Lewis (1981a) this causal dependency hypothesis expresses the decision maker's influence on loud purring and soft purring. Furthermore, it expresses the extent and the direction of the decision maker's influence.

Lewis (1981a) postulates that Gibbard and Harper's (1978) dependency hypotheses, namely the conjunctions of full patterns, can be subsumed under Lewis' (1981a) dependency hypotheses, namely the conjunctions of probabilistic full patterns. For a chance proposition can state that one of the $s_j$'s has a chance of 1, whereas all of the others have a chance of 0, so that Gibbard and Harper's (1978) dependency hypotheses are an extreme case of Lewis' (1981a) dependency hypotheses. Furthermore, the conjunctions of mixed full patterns, which consist partly of $a_i \Box\!\!\rightarrow s_j$ and partly of $a_i \Box\!\!\rightarrow$ [P = p], can also be subsumed under the conjunctions of probabilistic full patterns.

Lewis (1981a) calculates the utility of an option in the following way: Lewis considers a particular option $a_i$ and a particular partition member $s_j$. If a dependency hypothesis $K_k$ is the conjunction of a probabilistic full pattern, then $K_k$ implies $a_i \Box\!\!\rightarrow$ [P = p] for some p. Furthermore, $a_i \cap K_k$ implies [P = p]; and $c(s_j|a_i \cap K_k) = p(s_j)$. The $K_k$'s which are conjunctions of probabilistic full patterns build a partition of $a_i \Box\!\!\rightarrow$ [P = p] for any p. Hence the following equation holds, so Lewis (1981a):

$$\sum_{p} c(a_i \Box\!\!\rightarrow [P = p])p(s_j) = \sum_{k=1}^{l} c(K_k)c(s_j|a_i \cap K_k).$$

Then Lewis (1981a) substitutes the left part of this equation into the following formula:

$$U(a_i) = \sum_{j=1}^{m}(\sum_{k=1}^{l} c(K_k)c(s_j|a_i \cap K_k))u(a_i \cap s_j),$$

and gets a formula for utility in terms of counterfactuals with chance propositions as consequents:

$$U(a_i) = \sum_{j=1}^{m}\sum_{p} c(a_i \Box\!\!\rightarrow [P = p])p(s_j)u(a_i \cap s_j).$$

Furthermore, Lewis (1981a) gives the following formula for utility in terms of counterfactuals with chance propositions of single partition members as consequents (Lewis sets that, for any partition member $s_j$ and any number q ranging from 0 to 1, [P($s_j$) = q] is the proposition that holds only at those worlds where the chance of $s_j$ at the decision maker's option realisation time equals q.):

$$U(a_i) = \sum_{j=1}^{m}\sum_{q} c(a_i \Box\!\!\rightarrow [P(s_j) = q])qu(a_i \cap s_j).$$

## *Lewis on "Why Ain'cha Rich?"*

In his (1981b) Lewis considers the dispute between the *V*-maximisers and the *U*-maximisers with regard to Newcomb's problem from a different point of view. For Lewis (1981b) claims that *V*-maximisers believe in indicative conditionals in Newcomb's problem, whereas *U*-maximisers believe in subjunctive conditionals in Newcomb's problem. The indicative conditionals are: If I take B2, I will be a millionaire; if I take both boxes, I will not be a millionaire. The subjunctive conditional is: If I took only B2, I would be poorer by $ 1,000 than I will be after taking both boxes. Lewis (1981b) furthermore specifies that the *U*-maximisers don't believe in backtracking conditionals, only in normal subjunctive conditionals like the one above. The belief in the indicative conditionals vs. the belief in the subjunctive conditionals is backed up by the following beliefs: Whereas the *V*-maximisers think that they have to make a decision between $ 1,000,000 and $ 1,000, the *U*-maximisers think that they have no choice between taking or not taking $ 1,000,000.

Lewis (1981b) continues by stating the "Why Ain'cha Rich?"-argument of the *V*-maximisers: If the *U*-maximisers are so rational, why aren't they rich? While the *V*-maximisers end up being millionaires, the *U*-maximisers end up being not-millionaires. According to the *V*-maximisers the *U*-maximisers aren't rich, because they have decided not to be millionaires.

According to Lewis (1981b) the *U*-maximisers including himself reply to the argument in the following way: We never had a choice to become millionaires. For when we made our decisions, there weren't $ 1,000,000 to be had, that is $ 1,000,000 were reserved for the irrational *V*-maximisers. Lewis (1981b) furthermore specifies the beliefs of the *U*-maximisers by stating that in their view irrationality is richly prerewarded by $ 1,000,000 being in B2, so that according to their opinion irrationality doesn't cause $ 1,000,000 to be in B2.

With regard to the status of this moral Lewis (1981b) claims that it is only one further *U*-maximiser-doctrine which the *V*-maximisers can consistently deny. For the *U*-maximisers can consistently keep believing that the *V*-maximiser's reward of $ 1,000,000 and the *U*-maximiser's reward of $ 0 proves nothing. And for the *V*-maximisers can consistently keep believing otherwise, because from their point of view it is impossible to be sure at decision time that the irrational decision and not the rational decision will be richly prerewarded. The *V*-maximiser's expectation that taking B2 will be richly prerewarded is enough to regard that decision as rational.

## *A Critique of Lewis' Position*

Like Spohn (1977, 1978) and Sobel (1986) Lewis (1981a) makes something already implicitly accepted explicit: Lewis highlights the fact that in causal decision theories the decision maker should maximise *U*-utility calculated by means of dependency hypotheses. Yet while Gibbard and Harper's (1978) causal decision theory is meant to work for decision makers who believe in deterministic worlds, Lewis' (1981a) causal decision theory and Sobel's (1986) causal decision theory are meant to work even for decision makers who believe in indeterministic worlds. Thus Lewis' theory and Sobel's theory have a wider range of applicability than Gibbard and Harper's theory in this respect.

There is one more parallel between Lewis' causal decision theory and Sobel's causal decision theory. For according to Rabinowicz (1985, p. 178) Lewis' (1981a) rich partitions of the possible states of the world are similar to Sobel's (1985b) sufficiently fine partitions of the possible states of the world. Yet Rabinowicz (1982) also claims that Lewis' theory and Sobel's theory aren't equivalent with each other (cf. chapter 3.4). Furthermore, because of the difference in meaning between "would" and "might" it would be quite astonishing to find out that Lewis' would-subjunctive conditionals have the same truth conditions as Sobel's might-subjunctive conditionals. But as long as Sobel doesn't provide a logic for his might-subjunctive conditionals, as Lewis (1973) has done for his would-subjunctive conditionals, Sobel's and Lewis' account of subjunctive conditionals cannot be compared with each other.

Rabinowicz (1982) rightly criticises Lewis' (1981a) claim that the decision maker's ascription of a credence of 0 to $a_i \cap K_k$ will never arise. For according to Rabinowicz there are cases, when the assignment of a credence of 0 to $a_i \cap K_k$ is necessary. Rabinowicz (1982) tries to prove the latter in the following way: If one ascribes a credence of 0 to $a_i \cap K_k$, then $u(a_i \cap K_k)$ becomes an undefined sum of quotients with denominator 0, and $U(a_i)$ is undefined and cannot be compared with the utilities of the other options. According to Lewis (1981a), however, this case will never arise, because $c(a_i \cap K_k) = 0$, that is absolute certainty, amounts to a firm resolve never to change your mind, and that is objectionable. Yet Rabinowicz (1982) points out that the assignment of a credence of 0 to $a_i \cap K_k$ is necessary, whenever $a_i \cap K_k$ is empty. Thus to maintain his point of view Lewis must presuppose that $a_i \cap K_k$ is non-empty, which can be criticised.

In order to do that Rabinowicz (1982) first defines a certain group of propositions, which correspond to Lewis' dependency hypotheses, in terms of the

tendency function $T$, which takes as its argument pairs $(w, X)$, and which assigns to every such pair a probability distribution $T_{w, X}$ on worlds: Let two worlds, $w$ and $w'$, have the same tendencies on the decision maker's options if and only if for any option $a_i$ $T_{w, a_i} = T_{w', a_i}$. Let a non-empty equivalence class with regard to the relation of having the same tendencies on the decision maker's options be called a tendency proposition. Furthermore, let the variable $S$ range over such propositions. Moreover, if a tendency proposition holds at a world $w$, we shall refer to this proposition as $S_w$.

On that basis Rabinowicz (1982) criticises Lewis' presupposition that $a_i \cap K_k$ or equivalently $a_i \cap S_w$ is non-empty: If we suppose that for some merely possible world $w$ the decision maker cannot perform his option $a_i$ in $w$, the tendencies that characterise $w$ exclude $a_i$. Furthermore, this doesn't show that $a_i$ cannot be performed in the actual world, so that $a_i$ may be an option for the decision maker even though for some merely possible world $w$ $a_i \cap S_w$ is empty, Rabinowicz (1982) claims. Therefore there are cases in which the decision maker has to ascribe a credence of 0 to $a_i \cap K_k$.

With regard to Lewis' (1973) semantics of possible worlds Horgan (1981) rightly claims that the comparative overall similarity among possible worlds is inherently vague. Horgan (1981) makes this vagueness explicit by means of the 1-box-argument and the 2-boxes-argument in Newcomb's problem. Yet Lewis doesn't deny the vagueness of the comparative overall similarity among possible worlds. For according to Lewis (1979c) there is the standard resolution of vagueness which is appropriate in most contexts and also in Newcomb's problem, and there is the non-standard resolution of vagueness, which Horgan calls the backtracking resolution and which is appropriate in certain unusual contexts. According to Horgan (1981), though, the backtracking resolution of vagueness is pragmatically appropriate for Newcomb's problem. For although the backtracking resolution of vagueness is appropriate for the 1-box-argument in Newcomb's problem and the standard resolution of vagueness is appropriate for the 2-boxes-argument in Newcomb's problem, only the 1-box-argument is pragmatically appropriate for Newcomb's problem (Horgan 1981).

In order to evaluate Horgan's point of view, look at (1) Lewis' semantics of possible worlds, (2) Lewis' distinction between the standard resolution of vagueness and the non-standard resolution of vagueness, and consider (3) Horgan's claim that the backtracking resolution of vagueness is appropriate for the 1-box-argument and that the standard resolution of vagueness is appropriate for the 2-boxes-argument.

(1) According to Horgan (1981) Lewis (1973) states the following truth conditions for counterfactuals: A counterfactual $P \mathrel{\square\!\!\rightarrow} Q$ is true at a possible world $w$ if

and only if either (1) there are no possible worlds at which *P* is true or (2) some *P*-world at which *Q* is true is more similar over all to *w* than is any *P*-world at which *Q* is not true. Furthermore, Horgan takes over Stalnaker's (1968) assumption, which Lewis doesn't use, that for any world *w* and any proposition *P* if there are any possible *P*-worlds, then there is a unique *P*-world closest to *w*. Thus Horgan (1981) states the following truth conditions for counterfactuals: A counterfactual $P \mathrel{\square\!\!\rightarrow} Q$ is (non-vacuously) true if and only if *Q* is true at the closest *P*-world. Furthermore, Horgan assumes that no counterfactual, which he discusses, is vacuous.

(2) Lewis discusses the case that past events are dependent on present events and comes to the following conclusions (Lewis 1979c, pp. 456-457):

> "We know that present conditions have their past causes. We can persuade ourselves ... that if the present were different then these past causes would have to be different, else they would have caused the present to be as it actually is. Given such an argument - ... a *back-tracking argument* - we willingly grant that if the present were different, the past would be different too. ... But the persuasion does not last. We very easily slip back into our usual sort of counterfactual reasoning, and implicitly assume ... that facts about earlier times are counterfactually independent of facts about later times. ... What is going on ... can best be explained as follows. (1) Counterfactuals are infected with vagueness ... . Different ways of (partly) resolving the vagueness are appropriate in different contexts. ... (2) We ordinarily resolve the vagueness ... in such a way that counterfactual dependence is asymmetric (except perhaps in cases of time travel or the like). Under this standard resolution, back-tracking arguments are mistaken: ... . (3) Some special contexts favour a different resolution of vagueness, one under which the past depends counterfactually on the present and some back-tracking arguments are correct. ... (4) A counterfactual saying that the past would be different if the present were somehow different may come out true under the special resolution of its vagueness, but false under the standard resolution."

Furthermore, Lewis (1979c, p. 472) claims that the standard resolution of vagueness obeys the following system of weights or priorities:

> "(1) It is of the first importance to avoid big, complicated, varied, widespread violations of law.

(2) It is of the second importance to maximise the spatiotemporal region throughout which perfect match of particular fact prevails.

(3) It is of the third importance to avoid even small, simple, localised violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly."

On the basis of these remarks Horgan (1981) comes to the following conclusion with regard to the standard resolution of vagueness: If the antecedent of a counterfactual $a_i \Box\!\!\rightarrow s_j$ is a statement describing a particular possible action, then the $a_i$-world $w$ most similar to actuality will have these features: (i) There is perfect match of particular fact between $w$ and our actual world until a moment before the possible action $a_i$ in $w$; (ii) a small, simple, localised violation of actual-world law occurs in $w$ very shortly before $a_i$, so that $a_i$ is brought about; (iii) no other violations of actual-world law occur.

(3) With regard to the 1-box-argument in Newcomb's problem the backtracking resolution of vagueness is appropriate for the following two premises, so Horgan (1981):

**Premise 1**. If I were to decide for both boxes, then the predictor would have predicted this.

**Premise 2**. If I were to decide for B2, then the predictor would have predicted this.

Horgan (1981) makes this claim more precise by explaining that the backtracking resolution differs from the standard resolution insofar as it gives the most weight to the predictive correctness in the nearest possible world where the decision maker takes both boxes and in the nearest possible world where the decision maker takes B2. Under this backtracking resolution premise 1 and premise 2 are both true. For the closest world in which the decision maker takes both boxes is one in which the predictor correctly predicted this and put $ 0 in B2, and the closest world in which the decision maker takes B2 is one in which the predictor correctly predicted this and put $ 1,000,000 in B2.

Under the standard resolution of vagueness premise 1 and premise 2 cannot both be true, Horgan (1981) maintains. For the predictor has already made his prediction, when the decision maker decides, so that there is already $ 1,000,000 in B2 or not. Thus the actual-world prediction and the actual-world state of B2 remain intact in the closest world in which the decision maker takes both boxes and in the closest

world in which the decision maker takes B2 only. Therefore either premise 1 or premise 2 must be false (Horgan 1981).

In general Horgan (1981) concludes: If we compare the backtracking resolution with the standard resolution, then under the backtracking resolution the predictor's predictive correctness is a more important parameter of similarity than is maximisation of the spatiotemporal region through which perfect match of particular fact prevails, while under the standard resolution the maximisation of spatiotemporal region through which perfect match of particular fact prevails is a more important parameter of similarity than the predictor's predictive correctness.

With regard to the 2-boxes-argument in Newcomb's problem the standard resolution of vagueness is appropriate for the following two premises (Horgan 1981):

**Premise 3**. If B2 contains $ 1,000,000, then I would get $ 1,001,000 if I decided for both boxes.

**Premise 4**. If B2 contains $ 0, then I would get $ 0 if I decided for B2.

Under the standard resolution premise 3 and premise 4 are both true (Horgan 1981). Premise 3, for example, is true if and only if either there is not $ 1,000,000 in B2 or the decision maker gets $ 1,001,000 in the closest world *w* to actuality in which the decision maker takes both boxes, Horgan claims. For suppose that there is $ 1,000,000 in B2. Then because perfect match of particular fact prevails between our actual world and *w* until shortly before the decision maker's decision in *w* (since the predictor predicts in *w* that the decision maker will decide for B2 and therefore puts $ 1,000,000 in B2), the decision maker gets $ 1,001,000 in *w*. Furthermore, under the standard resolution it doesn't matter that the predictor predicts incorrectly in *w*. Thus premise 3 is true. An analogous argument holds for premise 4.

Under the backtracking resolution of vagueness premise 3 and premise 4 cannot both be true, Horgan (1981) maintains. For suppose there is $ 1,000,000 in B2. Then because the predictor predicts the decision maker's decision correctly in our actual world, and because preservation of this predictive correctness has greater weight of similarity than does preservation of his actual prediction, the decision maker only gets $ 1,000 in the closest world where he takes both boxes and not $ 1,001,000. Thus premise 3 is false, if there is $ 1,000,000 in B2, so Horgan. On account of analogous reasons premise 4 is false, if there is $ 0 in B2.

Because Horgan's claim that the backtracking resolution of vagueness is appropriate for the 1-box-argument in Newcomb's problem can already be criticised, we don't have to deal with Horgan's metaargument that the backtracking resolution of

vagueness is pragmatically appropriate for the 1-box-argument, while the standard resolution of vagueness isn't pragmatically appropriate for the 2-boxes-argument. My criticism of Horgan's point of view amounts to the following: Because premise 1 and premise 2 of the 1-box-argument can be criticised for containing not well-specified possible states of the world (cf. Schramm unpublished), the backtracking resolution of vagueness cannot be appropriate for the 1-box-argument in Newcomb's problem. Premise 1 and premise 2 contain not well-specified possible states of the world, because the predictor's prediction refers to two possible states of the world. It refers to the earlier possible state of the world, when the prediction was made, and it refers to the later possible state of the world, when the predicted possible action takes place. While the former possible state of the world already obtains, when the decision maker decides, it is logically possible that the latter possible state of the world doesn't turn out as predicted, so that the possible state of the world isn't well-specified. To be more precise the possible state of the world is over-specified. For it already determines a future possible state of the world as true. Thus premise 1 and premise 2 are both false.

If one considers the limit case in which the predictor is infallible in the sense that it is logically necessary that he makes correct predictions (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position), and if one asks whether such a predictor is logically possible, it becomes even more apparent that premise 1 and premise 2 contain not well-specified possible states of the world. For even though the proposition that the infallible predictor predicts correctly is logically necessary (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position), such an infallible predictor is logically impossible. The latter can be seen by Spohn's argument (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position): One can always imagine possible worlds in which an infallible predictor predicts incorrectly, that is from a logical point of view one cannot believe that there is an infallible predictor, and the actual world could be one of these possible worlds. Thus if one assumes that premise 1 and premise 2 are about such an infallible predictor, they would be false. For there could be always possible worlds in which the predictor predicts incorrectly, and the actual world could be one of these possible worlds.

Therefore because premise 1 and premise 2 of the 1-box-argument can already be criticised for containing not well-specified possible states of the world, Horgan's (1981) backtracking resolution of vagueness cannot be appropriate for the 1-box-argument in Newcomb's problem either.

Furthermore, to demand of the possible states of the world to be well-specified leads to an exclusion of an evidential partition of the possible states of the world in Newcomb's problem. For an evidential partition of the possible states of the world, like $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$., is over-specified. As in the case of the 1-box-argument the evidential partition refers to the earlier possible state of the world, when the prediction was made, and it refers to the later possible state of the world, when the predicted possible action takes place. While the former possible state of the world already obtains, when the decision maker decides, it is logically possible that the latter possible state of the world doesn't turn out as predicted, so that the possible state of the world isn't well-specified. To be more precise the possible state of the world is over-specified. For it already determines a future possible state of the world as true. Thus an evidential partition of the possible states of the world shouldn't be used, when making a decision in Newcomb's problem.

In my opinion Lewis neglects the possible states of the world, when he specifies causal dependence or independence by means of dependency hypotheses. This can be illustrated by means of Lewis' account of Gibbard and Harper's causal decision theory, but carries over to Lewis' own causal decision theory. For if one, for example, considers the dependency hypothesis

I brush Bruce $\Box\!\!\rightarrow$ he purrs loudly,

I stroke Bruce $\Box\!\!\rightarrow$ he purrs softly,

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr,

which according to Lewis (1981a) states that soft and loud purring are causally dependent on the decision maker's options, and compares it with the dependency hypothesis

I brush Bruce $\Box\!\!\rightarrow$ he doesn't purr,

I stroke Bruce $\Box\!\!\rightarrow$ he doesn't purr,

I leave Bruce alone $\Box\!\!\rightarrow$ he doesn't purr,

which according to Lewis (1981a) states that the lack of purring is causally independent of the decision maker's options, then the following can be said: With regard to the latter dependency hypothesis Lewis implicitly assumes that Bruce is in the possible state of the world of non-purring before the decision maker acts. For if one assumes that Bruce is in the possible state of the world of purring before the decision maker acts, the dependency hypothesis states that the lack of purring is causally dependent on the decision maker's possible actions. Thus Lewis' claim that the pattern of counterfactuals

and not one counterfactual alone expresses causal dependence or causal independence has to be modified: If the possible state of the world doesn't differ from the consequent with regard to all counterfactuals of a dependency hypothesis, then we have causal independence; otherwise we have causal dependence.[160]

Now we come to the structure of Newcomblike problems: Lewis (1981a) claims that if the prior state of the world obtains, for example, the cancer gene in Fisher's problem (Skyrms 1984, p. 65), you cannot change it by declining the good, that is smoking, you only shield yourself from the bad news, that is your chance of getting lung cancer is higher than you believed, which is useless. Yet in my opinion it may not be useless to shield yourself from the bad news. For the bad news that your chance of getting lung cancer is higher than you believed may have utterly devastating effects in your life[161]: You may be afraid of getting lung cancer all the time; you may stop enjoying your life, because there is always the threat present of getting lung cancer; you may be afraid of going to the physician, because you are afraid that he discovers lung cancer, etc. That is the bad news is so bad that it changes your whole life into a miserable life, so that it is actually better to shield yourself from the bad news by not-smoking. The reason why it is not useless to shield yourself from the bad news is: It takes the cancer gene longer to cause lung cancer than to cause the decision maker to smoke. Thus there is a time span between your decision to smoke respectively not to smoke and your getting lung cancer respectively not-getting lung cancer, in which it would be best to have the illusion not to get lung cancer, so that you should decide not to smoke.[162]

---

[160]With regard to this objection to his causal decision theory Lewis responded in the following way (personal communication from the 6th of August, 1999): "Counterfactual conditionals normally have factual background implicitly built in -- for example, factual background pertaining to the previous state of the world. That's how counterfactuals differ from strict conditionals. That's why it would be pointless to build factual background in all over again by adding it to the antecedent."

[161]One may object that if the bad news has such devastating effects in your life, then it is more than just bad news. Furthermore, Lewis (1981a) just wants to discuss cases in which no other significant payoffs are at stake, so that my criticism of Lewis' structure of Newcomblike problems doesn't apply (cf. also footnote 162).

[162]With regard to this objection to his causal decision theory Lewis answered (personal communication from the 6th of August, 1999): "Yes, of course it can be pleasant not to get bad news. Please remember that I was discussing a case in which 'no other significant payoffs are at stake' (section 4, first paragraph). The pleasure of avoiding bad news is therefore ex hypothesi an insignificant payoff; shielding yourself from bad news gains you no significant payoff and is therefore useless, as I said it was, and contrary to the opinion of the noncausal decision theorists. If you would rather discuss some different case in which the pleasure of avoiding bad

In Newcomb's problem this amounts to the following: If the prior state of the world obtains, for example, your greedy character in Newcomb's problem, you cannot change it by declining the good, that is taking both boxes, you only shield yourself from the bad news, that is your chance of $ 0 in B2 is higher than you believed, which is useless. I think in this case it is actually useless to shield yourself from the bad news. For one can assume that it takes the greedy character not longer to cause the predictor's diagnosis which leads to the predictor's prediction, so that there is $ 0 in B2, than to cause the decision maker's decision to take both boxes. Thus one can assume that there is no time span between your decision to take both boxes respectively to take B2 and your getting $ 1,000 respectively $ 1,000,000, in which it would be best to have the illusion not to get $ 1,000. Therefore you should decide to take both boxes.[163]

With regard to Newcomb's problem Lewis' (1981a) causal decision theory proposes the right solution, namely to take both boxes. Furthermore, Lewis' causal decision theory - like Sobel's (1986) causal decision theory and unlike Gibbard and Harper's (1978) causal decision theory - is even meant to work for decision makers who believe in indeterministic worlds. Thus even decision makers who hold the following causal dependency hypothesis in Newcomb's problem are taken account of in Lewis' causal decision theory:

I take both boxes $\Box\!\!\rightarrow$ [P = $p_1$] holds,

I take B2 $\Box\!\!\rightarrow$ [P = $p_2$] holds,

where [P = $p_1$] and [P = $p_2$] are the following two chance propositions:

[P = $p_1$]: The chance that there is $ 1,000,000 in B2 is 50%; the chance that there is $ 0 in B2 is 50%.

[P = $p_2$]: The chance that there is $ 1,000,000 in B2 is 30%; the chance that there is $ 0 in B2 is 70%.

Yet with regard to Lewis' (1973) semantics of possible worlds Horgan (1981) rightly claims that the comparative overall similarity among possible worlds is inherently vague, so that the decision maker could either argue that if I were to take both boxes, then the predictor would have predicted this or argue that if I were to take both boxes, then I would get $ 1,001,000. Yet because premise 1 and premise 2 of the 1-box-argument can already be criticised for containing not well-specified possible states of the world,

---

news is a significant payoff, of course you may; but you should not expect what I said to apply to that different case."

[163]In chapter 3.2, in the section on a critique of Gibbard and Harper's position, I have already dealt with the "Why Ain'cha Rich?"-argument of the *V*-maximisers, and I have nothing new to add to it.

Horgan's (1981) backtracking resolution of vagueness cannot be appropriate for the 1-box-argument in Newcomb's problem either. Thus Lewis' standard resolution of vagueness for the 2-boxes-argument in Newcomb's problem wins after all. Furthermore, Lewis neglects the possible states of the world, when he specifies causal dependence or causal independence by means of dependency hypotheses, so that in Newcomb's problem it makes a difference that either the possible state of the world is there is $ 1,000,000 in B2 and the dependency hypothesis is the following:

I take both boxes $\Box\!\!\rightarrow$ there is $ 1,000,000 in B2,

I take B2 $\Box\!\!\rightarrow$ there is $ 1,000,000 in B2,

or the possible state of the world is there is $ 0 in B2 and the dependency hypothesis is the following:

I take both boxes $\Box\!\!\rightarrow$ there is $ 1,000,000 in B2,

I take B2 $\Box\!\!\rightarrow$ there is $ 1,000,000 in B2.

For in the former case we have causal independence, while in the latter case we have causal dependence. Yet subjunctive conditionals already have factual background implicitly built in, so Lewis, so that my objection is no objection after all. With regard to the structure of Newcomblike problems Lewis neglects the fact that the shielding off from the bad news may not be useless, if the shielding off takes a long time, which is the case in Fisher's problem, but not the case in Newcomb's problem. Yet Lewis pointed out to me that he only considers cases in which no other significant payoffs are at stake, so that this objection loses its force, too.

## 3.6 Spohn's Principle[164]

Spohn's (1977, 1978) causal decision theory, which is an advancement of Fishburn's theory (1964), is valuable for making explicit a principle which is used by Savage (1954/1972) and Fishburn (1964). The principle is the following: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts.*"[165] (Spohn 1977, p. 114) This principle, which I henceforth call Spohn's principle, isn't used in the rational decision theories of Jeffrey (1965) and of Luce and Krantz (1971), Spohn (1978) claims. According to Spohn (1977) his principle

[164]For a discussion of Spohn's principle see Ledwig (1999b); this discussion differs in some ways from the treatment of Spohn's principle in this chapter, though. I prefer the latter.

[165]Trivial conditional credences, like $c(a_1|a_1) = 1$ for a possible action $a_1$ or $c(a_2|a_1) = 0$ for two disjunctive possible actions $a_1$ and $a_2$, are not considered.

is important, because it has implications for the concept of action, Newcomb's problem, the theory of causality, and freedom of will.[166]

In a recent paper Spohn (in press) makes his claim with regard to Newcomb's problem more precise. In this paper Spohn points out that the solution to Newcomb's problem depends on the number of trials the decision maker has to go through. Whereas in a 1-shot Newcomb's problem Spohn's (1978) causal decision theory by means of Spohn's principle and the dominance principle results in taking both boxes (even in the case of an infallible predictor), in a finitely iterated Newcomb's problem with the same decision maker the application of Spohn's (in press) generalised conception of strategic rationality[167] results in taking B2. According to Spohn (in press) an infinitely iterated Newcomb's problem isn't possible for decision makers, who have a finite life, so no recommendation results.[168] I will not deal with Spohn's solution to a finitely iterated Newcomb's problem with the same decision maker, because a finitely iterated Newcomb's problem with the same decision maker seems to involve quite different arguments (cf. chapter 1.6), like the backwards induction argument (cf. Spohn 1999, section 5), than the 1-shot Newcomb's problem, so that we actually seem to deal with two different problems whose solutions need not coincide.

What kind of implications has Spohn's principle for a 1-shot Newcomb's problem? If Spohn's principle is valid, the evidential decision theories (for example Jeffrey 1965) cannot give a solution to Newcomb's problem, for they do not use his principle. The causal decision theories which use subjunctive conditionals (for example Gibbard and Harper 1978) are problematical, because they still have to provide a logic of subjunctive conditionals, a probability theory for subjunctive conditionals, and a corresponding rational decision theory (Spohn 1978, pp. 183-184). Kyburg (1980) with

---

[166]In 1999 (pp. 44-45) Spohn modifies his claim with regard to his principle. For he postulates that in the case of strategic thinking, that is in the case of sequential decision-making, the decision maker can ascribe subjective probabilities to his possible actions. Unfortunately he doesn't give a justification for this claim. In opposition to a finitely iterated Newcomb's problem there is no sequential decision-making in a 1-shot Newcomb's problem, so that his modified claim doesn't apply to a 1-shot Newcomb's problem. I will not deal with Spohn's modified claim, for I just want to deal with a 1-shot Newcomb's problem.

[167]Spohn (1999) argues that a strategy should make future possible actions contingent on the internal future decision situation of the decision maker, and not on the external future decision situation of the decision maker, the latter being the standard view. According to Spohn (1999) the standard view is a special case of his view.

[168]Even if we construe Newcomb's problem as a cyclic problem with different decision makers, we still obtain a finitely iterated Newcomb's problem in my opinion, because the number of decision makers are still finite (cf. Selten and Holtz-Wooders 1995).

his proposal of the distinction between epistemic vs. stochastic independence can provide a solution to Newcomb's problem, though. For he doesn't violate Spohn's principle. According to Spohn (1978) Nozick's (1969) solution to Newcomb's problem can be criticised (cf. chapter 1.8), so that Nozick fails to provide an adequate solution to Newcomb's problem. Nozick's (1993) proposal of combining various decision principles violates Spohn's principle by using evidential decision principles. Furthermore, its usage of causal decision principles is open to the above mentioned criticism, namely that causal decision theories which use subjunctive conditionals still have to provide a logic of subjunctive conditionals, a probability theory for subjunctive conditionals, and a corresponding rational decision theory. Meek and Glymour (1994) try to mediate between evidential and causal decision theories not providing a unique solution to Newcomb's problem: If the decision maker views his possible actions as interventions in the system, he doesn't condition his credences by his possible actions (Meek and Glymour 1994). Thus the decision maker is in harmony with causal decision theory and doesn't violate Spohn's principle. But if the decision maker views his possible actions as non-interventions in the system, he conditions his credences by his possible actions, Meek and Glymour (1994) claim. Thus the decision maker is in harmony with evidential decision theory, but violates Spohn's principle. Because Meek and Glymour (1994) permit that the decision maker conditions his credences by his possible actions, Meek and Glymour's (1994) proposal violates Spohn's principle.

Therefore embracing Spohn's principle only Spohn's (1978) solution, Skyrms' (1980, 1982, 1984, and chapter 3.3) solution, and Kyburg's (1980, 1988, and chapter 4.4) solution to Newcomb's problem are left. Spohn's solution is valuable for its simplicity in contrast to the causal decision theories with subjunctive conditionals. Spohn's (1978) solution is: Because the predictor's prediction is earlier than the decision maker's decision, it is false to use credences of the possible states of the world given the possible actions of the decision maker. Taking this into account you rather have to suppose that these credences have a fixed value independently of the decision maker's decision, so that absolute credences are used for the predictor's predictions. Furthermore, because taking both boxes dominates taking B2, the decision to take both boxes results (Spohn 1978). And this even holds for the case of an infallible predictor. But Spohn's (1978) aim with regard to Newcomb's problem is to find a coherent rational decision theory for each tendency to decide. Spohn (1978) acknowledges that Gibbard and Harper (1978) have presented such a proposal, but complains that Gibbard

and Harper's (1978) principle of maximising *V*-utility violates Spohn's principle. Spohn (1978) himself, though, doesn't come up with a different solution to fulfil his aim.

Because Spohn's principle is important for solving Newcomb's problem, Spohn (1977, 1978) makes his principle more precise:

> (1) Spohn's principle refers to future possible actions of the decision maker.
>
> (2) Credences for possible actions don't manifest themselves in the willingness to bet on these possible actions.
>
> (3) Spohn's principle requires that possible actions are things which are under the decision maker's full control relatively to the decision model describing him.
>
> (4) A theoretical reason for Spohn's principle is that credences for possible actions cannot manifest themselves in these possible actions.
>
> (5) An immediate consequence of Spohn's principle is that unconditional credences for events which probabilistically depend on possible actions are forbidden.

In the following I will explain each of these five points in detail.

## *Spohn's Principle Refers to Future Possible Actions of the Decision Maker*

Spohn's principle doesn't refer to past actions and possible actions of other persons, but only to possible actions which are open to the decision maker in his decision model, that is to future possible actions of the decision maker. Spohn (1977, p. 115) concedes, however, that decision makers frequently have and utter beliefs about their future possible actions like the following:

> (1) "I believe it is improbable that I will wear shorts during the next winter."

Yet Spohn (1977, p. 115) wants this utterance to be understood in such a way that it doesn't express a credence for a possible action, but a credence for a decision situation:

> (2) "I believe it is improbable that I will get into a decision situation during the next winter in which it would be best to wear shorts."

Conceding this opinion to Spohn for a while consider what happens with the following utterance:

> (3) "I believe it is improbable that I will run 100 meters in 9 seconds during the next year."

Spohn's reformulation would be the following:

(4) "I believe it is improbable that I will get into a decision situation during the next year in which it would be best to run 100 meters in 9 seconds."

This reformulation seems non-sensical, for it wouldn't matter how much I tried I never would be able to run 100 meters in 9 seconds. If the Olympic Games were to happen next year and I were in this decision situation, it would be best for me to run 100 meters in 9 seconds, but nevertheless I wouldn't be able to do it because of my bodily constitution. I could at most try to run 100 meters in 9 seconds. With this we have reached what Jeffrey (1965, 1968, 1983) terms probabilistic acts and tries; Jeffrey (1983, p. 177) postulates probabilistic acts, which are possible actions with probabilistic possible outcomes, in analogy to probabilistic observations.

Thus on first sight it seems plausible to suppose that Jeffrey and Spohn are talking about two different things, when they speak of probabilities for possible actions. While Jeffrey considers utterances like (3), Spohn looks at utterances like (1). A difference can be pointed out between both utterances: While "wearing shorts" is not quantified, "running 100 meters" is quantified by 9 seconds. Cases which are quantified seem to be exactly the actions with probabilistic possible outcomes à la Jeffrey. This can also be shown by pointing out that the fifth utterance in comparison to the sixth utterance sounds strange:

(5) "I will try to wear shorts during the next winter."

(6) "I will try to run 100 meters in 9 seconds during the next year."

One can object to that by showing that there are also actions with probabilistic possible outcomes which are not quantified such as in the following utterance:

(7) "It is improbable that I will climb Mount Everest during the next winter."

The following reformulation à la Jeffrey sounds appropriate, too:

(8) "I will try to climb Mount Everest during the next winter."

The last example suggests that the utterances (1), (3), and (7) nevertheless do not differ from each other. Jeffrey (1965, chapter 5.8, 1968, 1983, chapter 5.8) takes exactly this position. Jeffrey believes that all possible actions - acts and probabilistic acts - are probabilistic: That is my wearing shorts during the next winter is probably successful. For it is an act which is in my power to make true, if I want to. This act is only probably successful, because as Jeffrey (1983, p. 83) states "... one can always take a strict point of view from which the agent can only try to perform the act indicated ...", so that failures of carrying out the act are taken into account in Jeffrey's theory. And my

running 100 meters in 9 seconds during the next year improbably leads to success. For it is a probabilistic act which isn't in my power to make true, if I want to. I can only try to make it true. In opposition to Jeffrey (1983) Spohn doesn't explain how utterances of type (3) should be understood. And if we look at Spohn's principle, he doesn't have to explain them, for Spohn's principle just refers to acts and not to probabilistic acts.

From Spohn's and Jeffrey's detailed explanations one can analyse the utterances (1) and (3) in the following way: In the first example it is on the one hand improbable that the decision maker will get into a decision situation in which it would be best to wear shorts during the next winter, on the other hand trying to perform this possible action leads probably to success. In the third example, however, it is on the one hand probable that the decision maker will get into a decision situation in which it would be best to run 100 meters in 9 seconds during the next year, on the other hand trying to perform this possible action improbably leads to success. Utterances like "It is improbable that ..." can therefore refer to different objects like decision situations, possible actions, events, etc. I propose to speak of probabilities for possible actions just in case the probability statement refers to the possible action, for this corresponds to our natural usage of language.

Against this one could object that the decision maker can only ascribe probabilities to possible actions with reference to the given decision situation. For utterances like

(9) "It is improbable that I - regardless of the given decision situation - will fly to the next galaxy."

are false, because you can always imagine decision situations in which it would be possible and best to fly to the next galaxy. For example, somewhere in the future it could be possible that we would be able to build very fast space shuttles, so that we could reach the next galaxy within a few years and that we would need a substance which we could only get from a planet of the next galaxy, so that it would best for us to fly to the next galaxy. Thus I propose that the decision maker ascribes probabilities to possible actions conditional on the given decision situation.

But there is one way to rescue Spohn's principle. For Spohn could argue that his principle is satisfied, if the decision maker puts the probabilities of accomplishment in the possible states of the world and not in the possible actions.[169] To see how this works, let's consider the case in which the decision maker's possible action is to walk

[169]I would like to thank James Joyce for drawing my attention to this possibility.

across the room: This possible action is often easy to accomplish, but there may be circumstances, that is possible states of the world, which may prevent the decision maker from accomplishing it. In such a case one replaces the decision maker's possible action "I walk across the room." by "I try to walk across the room." and partitions the possible states of the world in the following way: $s_1$ & I will walk across the room, if I try. vs. $s_1$ & I will not walk across the room, if I try. Furthermore, this procedure can be generalised, so that it doesn't only apply to acts, but also to probabilistic acts. For in Jeffrey's (1983, p. 178) "trying to hit the bullseye", which he regards as a probabilistic act, the same procedure is possible: The decision maker's possible action "I hit the bullseye." is replaced by "I try to hit the bullseye.", and the possible states of the world are partitioned in the following way: $s_1$ & I will hit the bullseye, if I try. vs. $s_1$ & I will not hit the bullseye, if I try. Therefore Spohn's principle cannot only be rescued for acts, but can also be extended to probabilistic acts.

This whole procedure, namely putting the probabilities of accomplishment in the possible states of the world and not in the possible actions, can also be applied to Jeffrey's (1983) ratificationism and Jeffrey's (1988) probabilism. For in Jeffrey's (1983) ratificationism the decision maker conditions by the final decision to perform the possible actions, and in Jeffrey's (1988) probabilism the decision maker conditions by the performance of the possible actions like in Jeffrey's (1965) logic of decision. Thus with regard to Jeffrey's ratificationism if one assumes that "deciding" is a possible action, one can replace the decision maker's possible action "I decide to walk across the room." by "I try to decide to walk across the room." and partition the possible states of the world in the following way: $s_1$ & I will decide to walk across the room, if I try. vs. $s_1$ & I will not decide to walk across the room, if I try. And with regard to Jeffrey's probabilism one can replace the decision maker's possible action "I walk across the room." by "I try to walk across the room." or "I hit the bullseye." by "I try to hit the bullseye." and partition the possible states of the world like in the paragraph above. Therefore by means of Spohn's principle not only Jeffrey's logic of decision, but also Jeffrey's ratificationism, and Jeffrey's probabilism become falsified. Because Eells' (1981, 1982, 1985) proposal of the common cause and Jeffrey's (1996) decision kinematics don't differ from Jeffrey's logic of decision with regard to probabilities for possible actions, the same conclusion holds for them, too.

Up to now Spohn has never considered this whole procedure, so that it is possible that he is against it. In case of the latter he could argue that his principle demands that any adequate quantitative decision model must not even implicitly contain

any probabilities for acts. And by shifting the probabilities from the possible actions to the possible states of the world Spohn could claim that the decision model implicitly contains probabilities for acts. Unfortunately Spohn doesn't tell us what implicit containment with respect to Spohn's principle means. Furthermore, Spohn could argue that in an adequate quantitative decision model one cannot shift probabilities from possible actions to possible states of the world. But Spohn doesn't specify what an adequate quantitative decision model is. Thus I will leave these objections unanswered.

There is another way to support Spohn's principle: According to Segerberg (1993) we have to distinguish between a first-person perspective and a third-person perspective in rational decision theory. From a third-person perspective we can estimate the probability of a coin landing heads, likewise we can estimate the probability which possible action a decision maker will decide for, Segerberg (1993) maintains. Thus from a third-person perspective the decision maker puts the probabilities of accomplishment in the possible actions, so that Spohn's principle is violated. From a first-person perspective the decision maker cannot estimate the probability which possible action he will decide for, Segerberg (1993) points out. For otherwise the decision maker could equally reason in the following two ways, if he tries to decide whether to do $a_1$, which Segerberg (1993, p. 276) considers as strange: (1) The probability of my doing $a_1$ is 0.95, therefore I will do $a_1$. (2) The probability of my doing $a_1$ is 0.95, therefore I will not do $a_1$. Segerberg (1993) claims that the oddity stems from the decision maker taking the third-person perspective. Thus from a first-person perspective the decision maker doesn't put the probabilities of accomplishment in the possible actions, so that Spohn's principle isn't violated. Because a decision maker deliberates from a first-person perspective, Spohn's rationality constraint is obeyed.

## *Credences for Possible Actions Don't Manifest Themselves in the Willingness to Bet on these Possible Actions*

According to Spohn (1978) it is commonly accepted that credences for particular events manifest themselves in the willingness to bet on these events with appropriate betting odds and that this doesn't apply to credences for possible actions. For the willingness of the agent to bet on his own possible action just depends on his gain, that is on the stake of his betting partner, so that his own stake is quite irrelevant. He will accept the bet only if his gain is so high that his possible action together with his gain is to be preferred to all alternatives. Spohn (1978, p. 73) illustrates this by the following example:

> "If someone ... offered me DM 30 to watch a particular movie, I would presumably accept, while I wouldn't accept it, if someone offered me only DM 5 - both being independent of what the other person would get from me, if I didn't watch the movie."[170]

Considering Spohn's example I would like to ask whether it is a bet at all; Spohn (1978, p. 73, footnote 22) concedes that it is only a bet-like agreement. If we suppose that it is a bet, one has to ask whether it is really the case that I would join a bet at DM 30 and wouldn't join a bet at DM 5 if my own stake was > DM 30, ≥ DM 5 and ≤ DM 30, or < DM 5. Let's consider the case that my stake is DM 10,000 (negative change from reference point[171]). If money were important to me, and if I owned just a little bit more than DM 10,000 (reference point), then I would neither at DM 30 nor at DM 5 agree (positive change from reference point), for there could always happen something, which could hinder me from watching the movie, and the risk to lose DM 10,000 would be too high for me even if I were highly risk prone[172]. By considering this case we see that the willingness of the agent to bet on his own action doesn't depend solely on the stake of his betting partner, but also on his own stake. Furthermore, such factors as risk aversion, reference point, and the amount of change from the reference point seem to influence the willingness of the agent to bet.

In my opinion the willingness of the agent to bet on his own possible action given a real bet depends in the following way on his own stake and on the stake of his betting partner: Suppose that I bet with X that I will run 100 meters in 9 seconds during the next year. If the stake of X were DM 1,000, we would nevertheless argue without considering the reference point and the amount of change from the reference point that my stake given a moderate risk aversion would depend in the following way on my credence for reaching the goal: The first maxim is to keep my own stake as low as possible, for I want to keep my own possible loss as low as possible. But the more improbable I think it is that I will reach my goal, the lesser my stake should be in comparison to your stake; and the more probable it is that I will reach my goal, the higher my stake could be in comparison to your stake; if, however, the credence that I will reach my goal is $c = 0.5$, my stake could equal the stake of my betting partner.

---

[170] The translation is mine.

[171] Kahneman and Tversky (1988) say that you should consider value as a function with two arguments: The asset position that serves as reference point and the magnitude of the change (positive or negative) from that reference point.

[172] Arrow (1971) and Pratt (1964) have done pioneer work on the topic of risk aversion.

There is one way to rescue Spohn's claim. For one could argue that the following justification for his claim is more plausible[173]: Credences for possible actions don't manifest themselves in the willingness to bet on these possible actions, because the decision maker has to factor in the pleasure or displeasure associated with doing $a_i$ in the willingness to bet on $a_i$. Thus the more the decision maker likes to do $a_i$, the more the decision maker is willing to bet on his doing $a_i$; and the less the decision maker likes to do $a_i$, the less the decision maker is willing to bet on his doing $a_i$. Therefore the decision maker's willingness to bet on his possible actions isn't only influenced by his credences for his possible actions, but also by the decision maker's degree of pleasure associated with doing his possible actions.

## *Spohn's Principle Requires that Possible Actions Are under the Decision Maker's Full Control*

According to Spohn (1977) his principle requires that possible actions are things which are under the decision maker's full control not in an absolute sense, but relatively to the decision model describing him. Spohn (1977, p. 114) illustrates this by the following example: When the decision maker wants to go from here to there, this possible action isn't under his full control in an absolute sense, because there could be obstacles. But for reasons of simplicity a decision model can assume that possible actions are under the decision maker's full control. Thus Spohn (1977) ideally assumes that possible actions are under the decision maker's full control.

The requirement of full control over possible actions has some problematic consequences. If the decision maker's possible actions are not under the decision maker's full control, this includes that the decision maker must not be willing to bet his life on these possible actions and that the decision maker must assign a credence $< 1$ to these possible actions; but if the decision maker's possible actions are under the decision maker's full control relatively to the decision model describing him, this includes that the decision maker must be willing to bet his life on these possible actions and that the decision maker must assign a credence of 1 to these possible actions.[174]

---

[173]I wish to thank James Joyce for pointing out this possibility to me.

[174]One can object that this inclusion doesn't follow, if one understands $c = 1$ as an idealisation for reasons of simplicity. Yet it seems to me that decision theorists actually make this inclusion. For when I asked Howard Sobel in personal communication of the 28th of July, 1998: "If you say that options are things of which the decision maker is certain that they are completely under his control, so that the decision maker would assign a subjective probability of one to that option, would you also say that the decision maker would be willing to bet his life on that option?", he

Two questions arise with regard to these consequences: First, is a decision maker rational who must be willing to bet his life on his possible actions? Second, is Spohn's causal decision theory limited in its applicability? With regard to these questions one can answer the following: Because Spohn's causal decision theory is for an idealised decision maker who is always sure what he can do, it must be rational for this decision maker to be willing to bet his life on his possible actions.[175] At the same time this means that Spohn's causal decision theory is limited in its applicability. For Spohn's requirement demands that his causal decision theory is for an idealised decision maker. Thus Lewis' (1981a) criticism of Eells' (1981) position can be applied to Spohn's causal decision theory, too. For as Lewis (1981a) claims with regard to Eells' (1981) position, rational decision theory shouldn't be limited to ideal rational decision makers, if rationality includes self-knowledge. Because the decision maker has to assign a credence of 1 to his possible actions in Spohn's causal decision theory, Spohn's conception of rationality seems to include self-knowledge, too. Thus Lewis' criticism of Eells' position also holds for Spohn's causal decision theory, that is if rationality includes self-knowledge, there are also decision makers who are not sure what they can do. Furthermore, we not only want to know what decision would be rational for idealised decision makers, that is for fully rational decision makers, we also want to know what decision would be rational for partly rational decision makers and whether their partly rational decision-making methods will lead them to decide for the rational possible action. In Newcomblike problems Spohn's causal decision theory doesn't give an answer for partly rational decision makers who are not sure what they can do.

Yet one can defend Spohn's position by claiming that this limited applicability is adequate for a rational decision theory. For Lewis (1981a, p. 14, footnote 11) writes: "... what is open to reconsideration does not have a credence of zero or one ... ." And one could argue that possible actions in Spohn's theory are not open to reconsideration, so that the decision maker can ascribe a credence of 1 to his possible actions. Possible actions are not open to reconsideration for the following reason: Even if Spohn's conception of rationality includes self-knowledge, this self-knowledge is limited. For

---

answered on the 21st of August, 1998: "Yes! The theory is for a simple case of an idealised decision maker who is always sure what he can do. Perhaps you and I are never sure of anything."

[175]Yet one can also maintain the view that $c = 1$ is an idealisation of a normal decision maker or of the decision theorists for a normal decision maker. Spohn doesn't make his claim more precise with regard to this point.

according to Segerberg (1993) the decision maker deliberates from a first-person perspective, so that he cannot assign probabilities to his possible actions.

Yet one can question whether Spohn's principle requires the decision maker's full control over his actions. For if Spohn wants utterances like

> (1) "I believe it is improbable that I will wear shorts during the next winter."

to be understood in such a way that they express probabilities for decision situations, this doesn't entail anything about control over actions yet. But for the validity of Spohn's principle he could also presuppose the decision maker's full control over his actions. For if he reformulates utterances like (1) in such a way that they say the following "I believe it is improbable that I will get into a decision situation during the next winter in which it would best to wear shorts.", he presupposes that you can perform such actions as "wearing shorts" without having any problems. But isn't it rather unrealistic to assume that we have full control over our actions? Couldn't it be a much more realistic point of view to presuppose just a partial control over our actions? If one just considers how many men and women try to lose weight day by day, how many smokers try unsuccessfully to quit smoking every day, etc., doubts about full control over our actions come up. Yet it is always possible to put the probabilities of accomplishment from the possible actions into the possible states of the world, so that Spohn's requirement that the decision maker has full control over his possible actions can be satisfied after all.

## *Credences for Possible Actions Cannot Manifest Themselves in these*

Spohn (1978) mentions a theoretical reason for Spohn's principle, namely that credences for possible actions cannot manifest themselves in these possible actions. And Spohn (1977) gives the following argument to support his claim: Credences for possible actions play no part in rational decision-making. For in rational decision-making it is only important how much the decision maker likes his possible actions, what he believes to result from his possible actions, and how much he likes his possible outcomes. In the end the decision maker decides for the possible action which he likes most regardless of its credence. According to Spohn (1977) decision models should only capture the decision maker's cognitive and motivational dispositions which manifest themselves in his decisions and preferences. Credences for possible actions are not of this sort. For one could tell neither from the decision maker's decisions nor from his

preferences what his credences are. Therefore credences for possible actions shouldn't be contained in decision models.

Nevertheless, Spohn (1978) admits, decision makers sometimes do ascribe credences to their future possible actions and credences to events which are earlier than their possible actions conditional on their possible actions. But these kind of ascriptions are a form of theorising about the decision maker's own possible actions which cannot manifest itself in the decision maker's own possible actions, but which manifests itself for the most part in talking about the decision maker's own possible actions. Hence these kind of ascriptions are out of place in decision models which just want to capture theoretically the decision maker's possible actions (Spohn 1978).

But even if one cannot tell from the decision maker's decisions or preferences what the decision maker's credences are, one could always ask the decision maker what his credences are, so that they could figure in decision models.

## *Unconditional Credences for Events which Probabilistically Depend on Possible Actions Are Forbidden*

Spohn (1977) claims that if his principle holds, unconditional credences for events which are probabilistically dependent on possible actions are forbidden. For if a particular event $E$ is probabilistically dependent on a possible action, that is $c(E|a_1) \neq c(E|\neg a_1)$ for some possible action $a_1$ and its complement $\neg a_1$, then $c(a_1)$ can be inferred from $c(E)$ by means of $c(a_1) = [c(E)\text{-}c(E|\neg a_1)]/[c(E|a_1)\text{-}c(E|\neg a_1)]$. Therefore if a decision maker attributes an unconditional credence to an event, he must assume that this event is probabilistically independent of his possible actions.

## *A Critique of Spohn's Position*

By means of Spohn's principle the decision maker cannot take his possible actions as evidence of the possible states of the world.[176] For the decision maker's credence function cannot be modified by the evidence of the possible actions, if Spohn's principle demands that the decision maker shouldn't assign any credences to his possible actions. Other rational decision theories also contain that the decision maker cannot take his possible actions as evidence of the possible states of the world.[177] For in Jeffrey's ratificationism (1983) the decision maker takes his decisions, but not his

[176]I would like to thank Andreas Blank for drawing my attention to this fact.

[177]Wolfgang Spohn pointed out to me, though, that he was the first one to make the decision maker's possible actions evidentially irrelevant.

possible actions as evidence of the possible states of the world, in Eells' (1981, 1982, 1985) proposal of the common cause the decision maker's beliefs and wants and not his possible actions are evidence of the possible state of the world, and in Kyburg's (1980, 1988) proposal of maximising properly epistemic utility the decision maker cannot take his free possible actions as evidence of the possible states of the world when he is deciding.

Spohn's solution to Newcomb's problem is very simple. For if Spohn views Newcomb's problem as a conflict between Jeffrey's (1965) principle of maximising conditional utility and the principle of dominance, and if he makes Jeffrey's principle of maximising conditional utility invalid by Spohn's principle, the principle of dominance wins out easily. Yet what is the connection between the principle of dominance and causal decision theory? Spohn doesn't tell us that. But as we have already seen in chapter 1.2 the principle of dominance with probabilistic independence and the principle of dominance with causal independence are corollaries of respective maximising principles. Furthermore, we have shown in chapter 1.2 that the principle of dominance with probabilistic independence belongs to evidential decision theory, whereas the principle of dominance with causal independence belongs to causal decision theory. Moreover, we have seen in chapter 1.4 that the principle of strong dominance with causal independence can be applied to Newcomb's problem, because the decision maker believes that the possible states of the world are causally independent of the possible actions of the decision maker. Thus the respective causal decision theory can be applied to this decision problem. Therefore Spohn's solution to Newcomb's problem could be extended by providing an adequate causal decision theory.

Even if this is the correct solution to Newcomb's problem, what happens to all the decision problems in which the principle of dominance cannot be applied? The principle of dominance is restricted in its range. It can only be applied, when the decision maker believes that the possible actions of the decision maker don't causally influence the possible states of the world, or the possible actions of any other decision maker (cf. chapter 1.2). Thus even if Spohn's solution to Newcomb's problem is correct, we want to have a rational decision theory which is applicable in all cases. Therefore although Spohn's principle is valid, Spohn's solution to Newcomb's problem isn't satisfactory because of the limited applicability of the principle of dominance.

## 3.7 Summary

Causal decision theories do provide adequate solutions to Newcomb's problem. Yet their reasons for coming to the 2-boxes-solution are not always adequate as can be seen by the following main results of this chapter:

(1) Gibbard and Harper's (1978) and Lewis' (1981b) formulation of the "why ain't you rich?"-argument of the *V*-maximisers, which they state, but don't defend, is unsound. Thus my moral of this argument differs from Gibbard and Harper's moral, that is if a predictor with a high predictability rewards predicted irrationality, then irrationality will be rewarded, and Lewis' moral, that is irrationality is richly prerewarded by $ 1,000,000 being in B2.

(2) Gibbard and Harper's (1978) claim that it is rational to take both boxes, if the predictor is infallible, has to be made more precise by distinguishing between different kinds of infallibility, namely an infallibility which is analytically necessary, causally necessary, or logically necessary. Yet one can argue that an infallible predictor is logically impossible.

(3) While Gibbard and Harper's (1978) causal decision theory and Lewis' (1981a) causal decision theory are based on would-subjunctive conditionals, Sobel's (1986) causal decision theory is based on might-subjunctive conditionals. Whereas Gibbard and Harper don't provide a logic of would-subjunctive conditionals, but rely on the decision maker's intuitions instead, so that their theory is limited to decision makers who have clear intuitions with regard to would-subjunctive conditionals, Lewis (1973) provides a logic of would-subjunctive conditionals, as already Spohn (1978, p. 183) has pointed out. Sobel doesn't provide a logic of might-subjunctive conditionals, so that his theory is limited to decision makers who have clear intuitions with regard to might-subjunctive conditionals. Therefore Lewis' causal decision theory is to be preferred to Gibbard and Harper's causal decision theory and Sobel's causal decision theory with regard to this point.

(4) While Gibbard and Harper's (1978) causal decision theory is formulated in terms of ultimate possible outcomes, Skyrms (1985) proposes a causal decision theory which works for proximate possible outcomes. Because proximate possible outcomes demand less knowledge from the decision maker than ultimate possible outcomes, Skyrms' causal decision theory is to be preferred to Gibbard and Harper's causal decision theory in this respect.

(5) Gibbard and Harper's (1978) causal decision theory presupposes the validity of conditional excluded middle which is open to two objections which can't be completely overcome. In opposition to this Sobel (1986) and Lewis (1981a) don't presuppose the validity of conditional excluded middle. Because from an economical point of view of theory building it is better to presuppose viewer things, and because the objections can't be completely overcome, Sobel's and Lewis' theories are to be preferred to Gibbard and Harper's theory with regard to this point.

(6) Whereas Gibbard and Harper's (1978) causal decision theory is limited to decision makers who believe in deterministic worlds, Skyrms', Sobel's, Lewis', and Spohn's causal decision theories are even meant to work for decision makers who believe in indeterministic worlds. Thus Skyrms', Sobel's, Lewis' and Spohn's causal decision theories are to be preferred to Gibbard and Harper's causal decision theory in this respect.

(7) Gibbard and Harper's (1978) causal decision theory is limited to decision makers with non-backtracking intuitions, although Newcomb's problem arouses backtracking intuitions. Because it is difficult to determine whether backtracking intuitions are rational, and because rational decision theory is even supposed to apply to partly rational decision makers, this limitation is a disadvantage of Gibbard and Harper's theory. In opposition to Gibbard and Harper (1978) Skyrms (1984) can account for backtracking intuitions of the decision maker. For Skyrms demands that the possible states of the world are causally independent of the decision maker's possible actions. Thus Skyrms' theory is to be preferred to Gibbard and Harper's theory with regard to this point.

(8) Skyrms' (1980, 1982, 1984) causal decision theory is the rational decision theory which provides so far the best solution to Newcomb's problem.

(9) Skyrms' (1984) causal decision theory is in the ultimate analysis completely subjective, so that it isn't limited to decision makers who believe in chances. In opposition to this Sobel's (1986) causal decision theory relies in the ultimate analysis on practical chance conditionals with an objectivist understanding of chance, so that it is limited to decision makers who believe in chances. Gibbard and Harper's (1978) causal decision theory is completely subjective, and Lewis' (1981a) causal decision theory is in the ultimate analysis completely subjective, so that their causal decision theories aren't

limited to decision makers who believe in chances.[178] Thus Gibbard and Harper's, Skyrms' and Lewis' theories are to be preferred to Sobel's theory in this respect.

(10) In Skyrms' (1984) causal decision theory unconditional credences are used for calculating the utility of a possible action. Thus Spohn's (1977, 1978) principle, which is valid and which applies to acts and even to probabilistic acts, is not violated. Because Gibbard and Harper (1978), Sobel (1986), and Lewis (1981a) use subjunctive conditionals in their causal decision theories, they don't violate Spohn's principle either. Thus all causal decision theories are on a par with regard to this point.

(11) In Skyrms' (1984) causal decision theory the possible states of the world are partitioned rationally, so that Skyrms' causal decision theory works for in this respect fully rational decision makers, whereas in Sobel's (1986) causal decision theory the possible states of the world are partitioned naturally. Because all rational partitions are natural, but not all natural partitions are rational, Sobel's causal decision theory is even meant to work for in this respect partly rational decision makers. Thus with regard to this point Sobel's causal decision theory has a wider range of applicability than Skyrms' causal decision theory. Yet with regard to the tasks I wanted a rational decision theory to accomplish (cf. chapter 1.2, in the section on the matrix formulation of Newcomb's problem), Skyrms' theory is to be preferred to Sobel's theory. For Skyrms' theory determines the rational partitions of the possible states of the world.

(12) Skyrms' (1984) causal decision theory isn't formulated in terms of subjunctive conditionals - although his theory is compatible to the latter because of his (Skyrms 1984) Bayesian theory of conditionals -, so that his theory isn't limited in its applicability in this respect. In opposition to this Gibbard and Harper's (1978), Sobel's (1986), and Lewis' (1981a) causal decision theory are formulated in terms of subjunctive conditionals, so that decision makers who have weak intuitions with regard to subjunctive conditionals are not accounted for.

(13) Sobel's (1986) causal decision theory is much too complicated from an economical point of view of theory building, so that it shouldn't be recommended as a rational decision theory. The following features make Sobel's theory so complicated: (i) Practical conditionals which provide the basis for Sobel's practical chance conditionals; (ii) conditional chances in the calculation of the utility of a possible action; (iii) the distinction between causally possible, open, and possibly open; (iv) the distinction

[178] According to Spohn (1988, p. 131) Lewis' (1980) chance concept is covertly epistemological.

between natural partitions, sufficiently fine partitions, and sufficiently exclusive partitions.

(14) Sobel's (1986) practical conditionals which provide the basis for his practical chance conditionals are vague. Furthermore, if one of the primitive terms of a theory is vague, the theory itself becomes vague. Moreover, Sobel uses practical chance conditionals, but doesn't provide a logic for these conditionals; he just states that the vagueness of these conditionals should be resolved appropriately to contexts of decision which allows for a lot of arbitrariness.

(15) Sobel's (1986) distinction between possible actions which are causally possible, open, and possibly open can be applied to Newcomb's problem. Yet if the decision maker believes that his possible actions are not even possibly open for him, one cannot calculate the utility of a possible action by means of definition 4 in Sobel's theory, as Sobel requires for his causal decision theory to work.

(16) Rabinowicz (1982) claims that Lewis' (1981a) point of view to apply rational decision theory not only to fully rational decision makers, but also to partly rational decision makers must have some limits. For we cannot allow the decision maker to be as irrational as he wants to be.

(17) Horgan (1981) claims that the comparative overall similarity among possible worlds in Lewis' (1981a) causal decision theory is inherently vague. Yet Horgan's backtracking resolution of vagueness isn't appropriate for the 1-box-argument in Newcomb's problem either, so that Lewis' standard resolution of vagueness for the 2-boxes-argument wins after all.

(18) Lewis (1981a) doesn't neglect the possible states of the world, when he specifies causal dependence or causal independence by means of dependency hypotheses. For subjunctive conditionals normally have factual background implicitly built in.

(19) With regard to the structure of Newcomblike problems Lewis (1981a) doesn't neglect the fact that the shielding off from the bad news may not be useless, if the shielding off takes a long time. For he just considers cases in which no other significant payoffs are at stake.

(20) A rational decision theory has to demand from the possible states of the world to be well-specified, which leads to an exclusion of the evidential partition of the possible states of the world in Newcomb's problem (cf. chapter 3.5, in the section on a critique of Lewis' position).

(21) Spohn (1978) applies the principle of dominance to Newcomb's problem, although the principle of dominance is restricted in its range, so that it cannot be applied to all decision problems.

## *Outlook*

In the following chapter other proposals will be looked at which stress the decision maker's perspective. There are several motivations for looking at these other proposals: First, if one assumes like Lewis (1981a) that rational decision theory has to apply not only to fully rational decision makers, but also to partly rational decision makers - as long as the decision maker's wants and beliefs are not non-permissible irrational -, so that the decision maker's perspective is taken account of, rational decision theory could take a completely different form and provide other solutions to Newcomb's problem. Second, although the concept of probability is relatively unproblematical in rational decision theory, for the decision maker assigns probabilities subjectively, the concept of causation could be viewed as relatively problematical in rational decision theory, although causal decision theorists take for granted that the concept of causation is without problems.[179] Third, the decision maker's perspective could make a difference to what factors are viewed as causes in Newcomb's problem and therefore could lead to different solutions to Newcomb's problem. Thus let's have a look at these other proposals, and let's try to find out whether they provide better solutions to Newcomb's problem than the traditional causal decision theories.

[179]Dowe (forthcoming), for example, points out that causality in connection with identity poses problems for causal decision theory. For in the conserved quantity theory of causation (Dowe 1992) a causal process is defined in terms of identity through time, so that if we are to provide a satisfactory account of the causal links involved, then we need to show that those links are epistemically accessible, that is we have to know that the identity criterion is met. Yet the latter point doesn't play a role for the rational decision theories of chapter 3 and chapter 4. For these theories don't have the conserved quantity theory of causation as a basis.

# Chapter 4

# Other Proposals

## 4.1 Introduction

### *Order of Presentation*

The most prominent other proposals which stress the decision maker's perspective are:

(1) Meek and Glymour's (1994) distinction between conditioning and intervening,

(2) Nozick's (1993) combination of various decision principles,

(3) Kyburg's (1980, 1988) distinction between epistemic vs. stochastic independence on the basis of his inductive logic (Kyburg 1974).

Furthermore, as a fourth proposal I will try to defend the view that a 1-shot Newcomb's problem is a game against nature.

I will present and criticise these proposals in this order for the following reason: I'm aware that the chronological order demands that Kyburg's (1980, 1988) proposal should be presented first followed by Nozick's (1993) approach and Meek and Glymour's (1994) proposal. But Meek and Glymour's and Nozick's proposals are much nearer in substance to the previous two chapters than Kyburg's proposal. While Meek and Glymour's approach can be conceived as a mediation between evidential and causal decision theories and Nozick's approach as a compromise between evidential and causal decision theories, Kyburg's proposal cannot be classified in any of these terms. Kyburg's approach doesn't seem to be related to evidential and causal decision theories at all. Therefore I will set Kyburg's approach at the third place of the list. Nozick's proposal is set at second place, because besides giving evidential decision theory and causal decision theory its due Nozick proposes the so called symbolic utility of a possible action which is neither related to evidential decision theory nor to causal decision theory. In opposition to Nozick and to Kyburg Meek and Glymour deal only with evidential and causal decision theories. Thus to maintain thematic continuity Meek and Glymour's proposal is presented first followed by Nozick's proposal and Kyburg's

approach.[180] My own proposal comes last. For I needed to evaluate every solution to Newcomb's problem first before giving my own solution. Furthermore, although my solution provides a foundation for Skyrms' (1980, 1982, 1984) causal decision theory, my proposal shouldn't be placed in the chapter on causal decision theories, because first, it is unrelated to causal decision theories with regard to its primitive terms, and second, it shares the common characteristic of the other proposals.

## *Similarities and Differences between these Other Proposals*

The common characteristic of these other proposals is that all four proposals stress the decision maker's perspective. While Meek and Glymour (1994) claim that a solution to Newcomb's problem only depends upon what the decision maker believes about the causal relationships within a decision situation, Nozick (1993) postulates that a solution to Newcomb's problem only depends upon how much the decision maker believes in certain decision principles. Kyburg (1988) claims for Newcomblike problems that the decision maker's decision is evidence for an observer, but no evidence for the decision maker, unless the decision maker looks back on one of his possible actions. My own proposal entails that the decision maker should view Newcomb's problem as a game against nature.

What distinguishes all four proposals from each other is:

---

[180] Yet one objection remains. For one could argue that Meek and Glymour's (1994) proposal should be set forth in the chapter on causal decision theories. For Meek and Glymour propose a method for calculating the effect of an intervention. But Meek and Glymour (1994, pp. 1014-1015) explicitly state the following:

> "Our analysis of the dispute between causal and 'evidential' decision theory does not put us neatly on either side, ... . From our perspective, whether causal or 'evidential' recommendations should be followed depends only on what one believes about the causal character of a context of decision. We agree with 'evidential' decision theorists that nothing but an ordinary calculation of the maximum expected utility is required; we agree with causal decision theorists that sometimes the relevant probabilities in the calculation are not the obvious conditional probabilities; ... . The method that we suggest is to use the combined causal graph, the knowledge of the direct effects of a manipulation, and the Manipulation theorem. To this extent we are causal decision theorists. But if one believes that a decision will not produce an intervention in the system, our calculation will agree with 'evidential' decision theorists; we only disagree with those critics of causal decision theory who believe that an action *is* an intervention in the sense we have described."

Furthermore, in footnote 22 on p. 1015 Meek and Glymour (1994) speak of a "happy reconciliation" between evidential and causal decision theories. And there are even more text passages going in that direction in Meek and Glymour's article. Thus I think it is unfair to the authors' intention to place their proposal in the chapter on causal decision theories.

(1) They propose different definitions of the utility of a possible action and therefore also propose different maximising principles.

(2) They add structure to Newcomb's problem.

(3) They loosen certain rationality constraints on probabilities.

(4) They propose different solutions to Newcomb's problem.

**Ad 1**: Nozick (1993) proposes the following definition of the combined utility of a possible action:

$CoU(a_i)$ = $CU(a_i)c$(principle of maximising conditional utility)+$CDTU(a_i)c$(principle of maximising utility of causal decision theory)+$SU(a_i)c$(principle of maximising symbolic utility)+ ...,

where $CU(a_i)$ is Jeffrey's (1965) conditional utility of a possible action $a_i$, $CDTU(a_i)$ is the utility of a possible action $a_i$ of one of the versions of causal decision theory, $SU(a_i)$ is the symbolic utility of a possible action $a_i$. Unfortunately Nozick (1993) doesn't provide a formula for calculating the symbolic utility of a possible action $a_i$. Furthermore, the respective utilities are weighed by the decision maker's confidence in the corresponding maximising principles (Nozick 1993); the decision maker's confidence might be represented by degrees of confidence between 0 and 1 inclusive that sum to 1, or that don't sum to 1, or by confidence weightings that are not degrees between 0 and 1, Nozick (1993) claims. Moreover, the decision maker can add every other plausible utility with its respective weight into the formula of the combined utility. According to Nozick (1993) the decision maker should adopt the principle of maximising combined utility.

Kyburg (1980) claims that the decision maker should calculate the utility in the following way:

$$U(\ulcorner D \in a_i \urcorner, K) = \sum_{j=1}^{m} P(K, \ulcorner D \in a_i \urcorner, \ulcorner D \in o_j \urcorner) u(\ulcorner D \in o_j \cap a_i \urcorner),$$

where $\ulcorner\ \urcorner$ are Quine's corners, $D$ is the decision situation, $a_i$ is a type of a possible action, $K$ is the decision maker's body of knowledge, and $o_j$ is a type of a possible outcome. Furthermore, Kyburg (1980) distinguishes between properly epistemic utility and stochastic utility: If $D$ takes the form $\ulcorner \alpha P \urcorner$, which is an instance of a type of a decision situation, then $U(\ulcorner D \in a_i \urcorner, K)$ is stochastic; if $D$ doesn't take the form $\ulcorner \alpha P \urcorner$, then $U(\ulcorner D \in a_i \urcorner, K)$ is properly epistemic. Kyburg (1980) claims that the decision maker should use the principle of maximising properly epistemic utility.

Meek and Glymour (1994) don't propose a new definition of the utility of a possible action and therefore also don't propose a new maximising principle. I don't

propose a new definition of the utility of a possible action and therefore also don't propose a new maximising principle. For in my opinion Skyrms' (1984) definition and maximising principle is sufficient in this respect.

**Ad 2**: Meek and Glymour (1994) add the following distinction to the analysis of Newcomblike problems: The decision maker can consider possible actions as interventions or as no interventions. Kyburg (1980) adds a knowledge structure to the analysis of Newcomb's problem. For Kyburg's properly epistemic probability and his stochastic probability are relativised to the decision maker's body of knowledge *K*. In my view the decision maker should consider Newcomb's problem as a game against nature. Yet he could also conceive Newcomb's problem as a game with two decision makers. For although the former is the rational thing to do, the latter is not non-permissible irrational.

**Ad 3**: In Kyburg's (1980) theory probabilities are conceptualised as intervals[181] and not as point probabilities. As a consequence the utilities of the possible actions also take the shape of intervals and not the shape of point utilities. With regard to this modification the question arises why aren't utilities conceptualised as intervals in Kyburg's theory, too? And is this modification necessary? Both questions will be answered later on. Furthermore, Kyburg (1980) doesn't speak about credences or chances, but about epistemic probabilities. Epistemic probabilities are relativised to a body of knowledge, so Kyburg (1988, p. 77), but are objective, because given a body of knowledge all probabilities relative to that body of knowledge are determined by logic alone. Kyburg (1980, p. 150) claims that epistemic probability refers to the probability with which this coin lands heads or to the probability that a toss of a coin will land heads; while Kyburg (1980) calls the former probability "properly epistemic", Kyburg calls the latter probability "stochastic".

**Ad 4**: Meek and Glymour (1994) don't mark the 1-box-solution or the 2-boxes-solution as the correct solution to Newcomb's problem from the very beginning. But if the decision maker believes that his decision is no intervention, then the 1-box-solution recommends itself. And if the decision maker believes that his decision is an intervention, then the 2-boxes-solution is the rational one. For in the former case the decision maker has to condition by a different event than in the latter case. Nozick's (1993) principle of maximising combined utility doesn't mark the 1-box-solution or the 2-

[181] Kyburg (1974, pp. 264-267) claims that in certain special cases interval probabilities which are defined in a logical and syntactical way fulfil within his framework the axioms of a generalised mathematical probability calculus.

boxes-solution as the correct solution to Newcomb's problem from the very beginning, too. For the decision maker's solution to Newcomb's problem just depends upon which decision principles the decision maker favours and how much confidence he has in the respective decision principles, so Nozick (1993). Kyburg (1980) and I propose the 2-boxes-solution for Newcomb's problem.

## 4.2 Meek and Glymour's Distinction between Conditioning and Intervening

Meek and Glymour (1994) argue that the discussion between evidential decision theorists and causal decision theorists over Newcomblike problems is misplaced, because both the evidential decision theorists and the causal decision theorists don't distinguish between conditioning on an event E and conditioning on an event I which is a possible action, that is in Meek and Glymour's terminology an "intervention", to causally bring about E. By contrast, Meek and Glymour (1994) emphasise the importance of this distinction and place it in the framework of a theory of causation and directed graphs, which was developed by Spirtes, Glymour, and Scheines (1993). Meek and Glymour (1994) don't deal explicitly with Newcomb's problem; they just analyse Fisher's problem (Skyrms 1984, p. 65), which they consider as a Newcomblike problem.

According to Meek and Glymour (1994) philosophers agree that the causal relations in Fisher's problem are represented by a common cause schema:

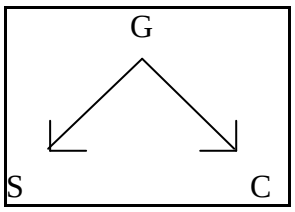


Figure 29. The causal relations in Fisher's problem with "G" denoting the genetic factor, "S" denoting smoking, and "C" denoting cancer.

Furthermore, most philosophers agree that these causal relations imply the following factorisation of the probabilities (Meek and Glymour 1994):

$$P(S \cap C \cap G) = P(S|G)P(C|G)P(G),$$

where $P(S \cap C \cap G)$ is the probability distribution over S, C, and G. Philosophers agree that if smoking and cancer have no causal effect on each other and only have the

genetic factor as their common cause, then smoking and cancer are probabilistically independent of each other given the genetic factor, that is:

$$S \perp\!\!\!\perp C|G,$$

which is a direct consequence of the factorisation formula above, so Meek and Glymour (1994).

Meek and Glymour (1994) claim that Reichenbach (1956) formulates some general principles connecting causal hypotheses and probability constraints. These principles are generalised by Kiiveri and Speed (1982) to the "Markov condition". Meek and Glymour (1994) introduce the Markov condition in two versions:

> **The frequency version of the causal Markov condition**: Suppose *G* is a directed acyclic graph which describes the causal relations among a set **V** of variables, where every common cause of a variation of two or more variables in **V** is itself in **V**, suppose **P** is a population, whose members share the same causal relations *G*, and suppose *P* is the frequency distribution of variables in **V**. Then every variable **X** in **V** is probabilistically independent of its nondescendants, that is variables it doesn't causally affect, given its parents.
>
> **The subjective version of the causal Markov condition**: Suppose *G* is a directed acyclic graph which describes the causal relations among a set **V** of variables, where every common cause of a variation of two or more variables in **V** is itself in **V**, and suppose *P* is the credence distribution over **V** given *G*. Then *G* and *P* should be related in such a way that every variable **X** in **V** is probabilistically independent of its nondescendants given its parents.

Meek and Glymour (1994) claim that in Fisher's problem the probabilistic independence relation follows from treating smoking and cancer as binary variables and from the fact that for sets of random variables **X**, **Y**, and **Z** the following holds: If **X** is probabilistically independent of **Y** given **Z**, then each member of **X** is probabilistically independent of each member of **Y** given **Z**. Meek and Glymour point out that the causal structure constrains the probabilistic structure and not the other way around, if the causal Markov condition obtains.

According to Meek and Glymour the causal Markov condition entails the following version of the common cause principle: If in a system of variables, which satisfies the causal Markov condition, **X** and **Y** don't causally influence each other, but are probabilistically dependent on each other, then there exists a set **Z** of variables,

which doesn't contain **X** and **Y**, but which causes **X** and **Y**, and **X** and **Y** are probabilistically independent of each other given **Z**.

Meek and Glymour (1994) continue to analyse Fisher's problem by introducing a special case with an additional causal factor: If the will intervenes, the will alone determines whether the decision maker smokes or not, and if the will doesn't intervene, the genetic factor alone determines whether the decision maker smokes or not. The following figure summarises the causal situation for this special case ("$G_{Comb}$" denotes the expanded graph; in this case the graph is expanded by W, which is the will):

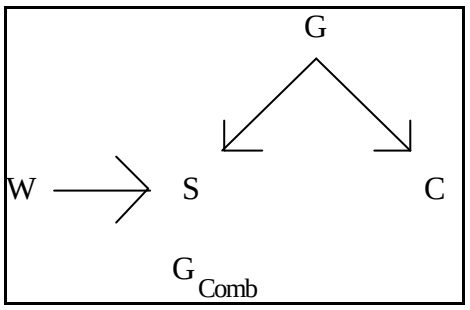


Figure 30. The causal situation in Fisher's problem in a special case: If the will intervenes, the will alone determines whether the decision maker smokes or not, and if the will doesn't intervene, the genetic factor alone determines whether the decision maker smokes or not.

Meek and Glymour (1994) claim that according to the causal Markov condition the probability distribution must satisfy the following equation in this case ($P_{Comb}$ is the probability distribution of the expanded graph $G_{Comb}$, which satisfies the causal Markov condition for $G_{Comb}$):

$P_{Comb}(S \cap C \cap G \cap W) = P(S|W \cap G)P(C|G)P(G)P(W)$.

Furthermore, Meek and Glymour (1994) suppose that the following three states of the will are relevant in this case:

(i) intervention for smoking,

(ii) intervention against smoking, and

(iii) no intervention.

These three states of the will result in the following three different conditional probability distributions over S, C, and G, Meek and Glymour (1994) state:

(1) $P_{Comb}(S \cap C \cap G|W = i) = P(S|W = i \cap G)P(C|G)P(G)$,

(2) $P_{Comb}(S \cap C \cap G|W = ii) = P(S|W = ii \cap G)P(C|G)P(G)$,

(3) $P_{Comb}(S \cap C \cap G|W = iii) = P(S|W = iii \cap G)P(C|G)P(G)$.

If S = 1 stands for smoking, and if S = 0 stands for not-smoking, then the following equations hold, so Meek and Glymour (1994):

(1) $P(S = 1|W = i \cap G) = P(S = 1|W = i) = 1$,

(2) $P(S = 1|W = ii \cap G) = P(S = 1|W = ii) = 0$, and

(3) $P(S|W = iii \cap G) = P(S|G)$.

Meek and Glymour (1994) claim that these conditional probability distributions yield the following three different probability distributions of C given S:

(1) $P_{Comb}(C|S = 1 \cap W = i) = P(C|W = i) = P(C)$,

(2) $P_{Comb}(C|S = 1 \cap W = ii) =$ undefined,

(3) $P_{Comb}(C|S = 1 \cap W = iii) = P(C|S = 1) = [\Sigma_G P(C|G)P(S = 1|G)P(G)]/P(S = 1)$.

With regard to these three probability distributions Meek and Glymour (1994) point out that while the first probability distribution is used by causal decision theorists, the third probability distribution is used by evidential decision theorists.[182] Whereas causal decision theorists condition on an event that is an intervention (e. g. willing to smoke), evidential decision theorists condition on an event that is not an intervention (e. g. smoking). These different conditioning practices stem from different beliefs about the causal structure in decision-making: While causal decision theorists believe that an intervention occurs, when the decision maker decides to smoke or decides not to smoke, evidential decision theorists believe that no intervention occurs, when the decision maker decides to smoke or decides not to smoke.

Meek and Glymour (1994) claim that the calculations in this example illustrate a general theorem, that is the manipulation theorem, which gives the reason for interpreting the directed graphs as causal hypotheses. According to Meek and Glymour causal claims entail claims about interventions, that is manipulations.

Meek and Glymour (1994) derive the manipulation theorem in the following way: Let $G_{Unman}$ be the directed graph which describes the causal relations in a system or population of systems. In Fisher's problem $G_{Unman}$ is as follows:

---

[182]To be more precise than Meek and Glymour (1994) the first probability distribution is used by Savage (1954/1972), who may count as a causal decision theorist, while the third probability distribution is used by Jeffrey (1965), who is an evidential decision theorist.

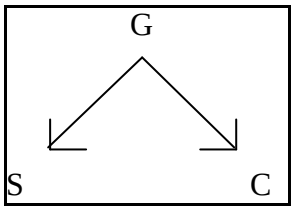


Figure 31. $G_{Unman}$ in Fisher's problem.

For manipulating $G_{Unman}$ a state of affairs M (in Fisher's problem it is W) has to be changed, so that values are forced on some of the variables in $G_{Unman}$. Therefore suppose $G_{Unman}$ is expanded by an additional variable M, that is the manipulation, so that directed edges from the variable M to whatever variables in $G_{Unman}$ are directly controlled by M. Furthermore, Meek and Glymour make the following suppositions:

(1) M takes on the value 0, if there is no manipulation.

(2) The expanded graph which is $G_{Unman}$ plus M and the additional edges is called $G_{Comb}$.

(3) $P_{Unman}$ is the probability distribution of the causal structure $G_{Unman}$, which satisfies the causal Markov condition for $G_{Unman}$.

(4) $P_{Comb}$ is the probability distribution of the expanded causal structure $G_{Comb}$, which satisfies the causal Markov condition for $G_{Comb}$.

In this case $P_{Comb}( \,|M = 0) = P_{Unman}$.

Moreover, suppose $P_{Man=i}$ is the probability distribution of the original variables in $G_{Unman}$ given a manipulation M = i. Then it follows from the causal Markov condition that, for any value i of M, $P_{Comb}( \,|M = i)$ can be computed from $G_{Unman}$, $P_{Unman}$, and the probabilities of the variables in $G_{Unman}$ given M = i which are directly causally influenced by M.

Following from the causal Markov condition this is Meek and Glymour's (1994) **manipulation theorem**:

> One obtains the manipulated distribution $P_{Man=i}$, if one replaces the conditional probabilities for the variables X, which are directly causally influenced by M, in the factorisation of $P_{Unman}$ by $P_{Comb}(X|M = i)$.

With the help of this manipulation theorem the differences between evidential and causal decision theory are explained by the fact that different events - non-interventions in the case of evidential decision theory, interventions in the case of causal decision theory - are conditioned on.

## *Consequences for Newcomb's Problem*

I will apply Meek and Glymour's distinction between conditioning and intervening to a special case of Newcomb's problem: If the will intervenes, the will alone determines whether the decision maker takes both boxes or takes B2, and if the will doesn't intervene, the common cause alone determines whether the decision maker takes both boxes or takes B2.[183] The following figure summarises the causal situation for this special case ("$G_{Comb}$" denotes the expanded graph; in this case the graph is expanded by W, which is the will; "CC" denotes the common cause, "BB" denotes taking both boxes, "PBB" denotes the predictor's prediction to take both boxes, "W" denotes the will):

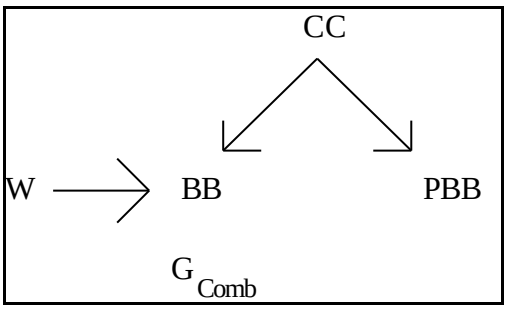


Figure 32. The causal situation in Newcomb's problem for a special case: If the will intervenes, the will alone determines whether the decision maker takes both boxes or takes B2, and if the will doesn't intervene, the common cause alone determines whether the decision maker takes both boxes or takes B2.

According to the causal Markov condition the probability distribution must satisfy the following equation in this case ($P_{Comb}$ is the probability distribution of the expanded graph $G_{Comb}$, which satisfies the causal Markov condition for $G_{Comb}$):

$$P_{Comb}(BB \cap PBB \cap CC \cap W) = P(BB|W \cap CC)P(PBB|CC)P(CC)P(W).$$

Furthermore, the following three states of the will are relevant in this case:

(i) intervention for taking both boxes,

(ii) intervention for taking B2, and

(iii) no intervention.

[183] One can object against this special case of Newcomb's problem and also against Meek and Glymour's special case of Fisher's problem that is isn't a decision problem anymore. For if a different cause than the will causes the possible action of the decision maker, one could claim that whatever the decision maker does isn't an action, but merely a kind of behaviour.

These three states of the will result in the following three different conditional probability distributions over BB, PBB, and CC:

(1) $P_{Comb}(BB \cap PBB \cap CC|W = i) = P(BB|W = i \cap CC)P(PBB|CC)P(CC)$,

(2) $P_{Comb}(BB \cap PBB \cap CC|W = ii) = P(BB|W = ii \cap CC)P(PBB|CC)P(CC)$,

(3) $P_{Comb}(BB \cap PBB \cap CC|W = iii) = P(BB|W = iii \cap CC)P(PBB|CC)P(CC)$.

If BB = 1 stands for taking both boxes, and if BB = 0 stands for taking B2, then the following equations hold:

(1) $P(BB = 1|W = i \cap CC) = P(BB = 1|W = i) = 1$,

(2) $P(BB = 1|W = ii \cap CC) = P(BB = 1|W = ii) = 0$, and

(3) $P(BB|W = iii \cap CC) = P(BB|CC)$.

These conditional probability distributions yield the following three different probability distributions of PBB given BB:

(1) $P_{Comb}(PBB|BB = 1 \cap W = i) = P(PBB|W = i) = P(PBB)$,

(2) $P_{Comb}(PBB|BB = 1 \cap W = ii) =$ undefined,

(3) $P_{Comb}(PBB|BB = 1 \cap W = iii) = P(PBB|BB = 1) =$

$[\Sigma_{CC}P(PBB|CC)P(BB = 1|CC)P(CC)]/P(BB = 1)$.

While the first probability distribution is used by causal decision theorists, the third probability distribution is used by evidential decision theorists. Whereas causal decision theorists condition on the event of an intervention, that is willing to take both boxes, evidential decision theorists condition on an event, that is taking both boxes, that is not an intervention. These different conditioning practices stem from different beliefs about the causal structure in decision-making: While causal decision theorists believe that an intervention occurs, when the decision maker decides to take both boxes or decides to take B2, evidential decision theorists believe that no intervention occurs, when the decision maker decides to take both boxes or decides to take B2.

## *A Critique of Meek and Glymour's Position*

Meek and Glymour's (1994) proposal to distinguish between conditioning and intervening can be criticised on the following grounds:

Although it is very valuable that Meek and Glymour (1994) provide an explanation why evidential and causal decision theorists differ in their recommendations in Newcomblike problems, it would have been more valuable to have a criterion for deciding, when the decision maker should view his possible actions as interventions, and when the decision maker should view his possible actions as non-interventions. Thus

after all Meek and Glymour's proposal leaves us undecided what to decide for in Newcomblike problems.

Meek and Glymour (1994) don't demand from the decision maker that he should always take one point of view in decision-making, so that in one case the decision maker can view his possible actions as interventions and in another case he can view his possible actions as non-interventions, which leads to an arbitrariness in decision-making. This arbitrariness, which bears similarities to the arbitrariness in Nozick's (1993) proposal (cf. chapter 4.3, in the section on a critique of Nozick's position in 1993), can in part be explained away by the fact that rational decision makers should be able to learn from experience, so that if the decision maker learns that his possible actions should be viewed as interventions in the system, his decisions should change, too. Yet Meek and Glymour don't state how learning from experience should be accounted for in their framework.

Meek and Glymour's (1994) proposal is limited in its applicability. For Meek and Glymour don't consider the case in which the decision maker views his possible actions partly as interventions and partly as non-interventions[184], so they cannot say what is rational to decide for in this case. Furthermore, Meek and Glymour don't give a recommendation for the case in which the decision maker doesn't know whether he should view his possible actions as interventions or not. This last case is important, because it highlights the fact that Meek and Glymour's proposal demands a certain self-knowledge from the decision maker. Thus if Meek and Glymour demand this self-knowledge from the decision maker in order for him to be rational, and if he lacks this self-knowledge, Lewis' (1981a) criticism of Eells' (1981, 1982, 1985) proposal of the common cause can be transferred to Meek and Glymour's proposal: Why shouldn't we ask what decision would be rational for partly rational decision makers and whether their partly rational decision-making methods will lead them to decide for the rational possible action. Moreover, it doesn't seem to me a non-permissible irrationality of the decision maker to lack this kind of self-knowledge, so that it seems reasonable to me to ask what is rational for such partly rational decision makers. Because Meek and Glymour don't give an answer to this question, their proposal is inadequate in this respect. Thus Meek and Glymour's proposal is limited to decision makers who have so

[184]With regard to the case in which the decision maker views his possible actions partly as interventions and partly as non-interventions it is of no importance whether such a case is possible at all. It is only important that the decision maker believes that such a case is possible.

much self-knowledge as to view their possible actions either as interventions or as non-interventions.

Another interesting case for Meek and Glymour's proposal is the finitely iterated Newcomb's problem with the same decision maker in opposition to the 1-shot Newcomb's problem: For if we follow Sorensen's (1985) analysis of the finitely iterated Newcomb's problem with the same decision maker, the decision maker should view his possible actions as reputation building behaviour. Furthermore, if the decision maker views his possible actions as reputation building behaviour in the finitely iterated Newcomb's problem, he should view his possible actions as interventions in the system. For in the finitely iterated Newcomb's problem the decision maker wants to influence the predictor's prediction by his possible actions. Thus the decision to take both boxes ensues. Yet Sorensen (1985) comes to the opposite recommendation (cf. chapter 1.6).

In the 1-shot Newcomb's problem the decision maker shouldn't view his possible actions as reputation building behaviour, so Sorensen (1985). For in the 1-shot Newcomb's problem the decision maker cannot build up a reputation. Moreover, if the decision maker doesn't view his possible actions as reputation building behaviour in the 1-shot Newcomb's problem, it is unclear to me whether he should view his possible actions as interventions in the system. For besides reputation building behaviour there may be other reasons why the decision maker should view his possible actions as interventions or as no interventions in the system. Thus no overall conclusion can be drawn for the 1-shot Newcomb's problem.

A further question for Meek and Glymour's (1994) proposal is whether different perspectives, namely (1) first-person perspective and (2) third-person perspective both deliberating about present decisions, and (3) first-person perspective deliberating about past decisions, make a difference for recommendations in decision-making or for evaluations of decisions.

Although Meek and Glymour don't take a stance with regard to this question, the following position seems plausible in my view: As Jones and Nisbett (1971) have shown there is a difference in attributing behaviour, if one compares first-person perspective with third-person perspective deliberating about present behaviour. From the first-person perspective deliberating about present behaviour the agent's behaviour seems to be for the most part determined by the agent's situation, while from the third-person perspective deliberating about present behaviour the agent's behaviour seems to be for the most part determined by the agent. Furthermore, Storms (1973) has shown

that from the first-person perspective deliberating about past behaviour the agent's behaviour seems to be for the most part determined by the agent.

Thus if the decision maker deliberates from the first-person perspective about his present behaviour in which his behaviour is for the most part determined by his situation, it is unlikely that he views his possible actions as interventions in the system. Yet if one deliberates from the third-person perspective about the decision maker's present behaviour in which the decision maker's behaviour is for the most part determined by the decision maker, it is likely that one views his possible actions as interventions in the system. Furthermore, if the decision maker deliberates from the first-person perspective about his past behaviour in which his behaviour is for the most part determined by himself, it is likely that he views his possible actions as interventions in the system.

Applying this to Newcomb's problem one gets the following results: If the decision maker deliberates from the first-person perspective about his present behaviour, it is likely that a decision to take B2 ensues. Whereas if one deliberates from the third-person perspective about the decision maker's present behaviour, or if the decision maker deliberates from the first-person perspective about his past behaviour, it is likely that the deliberating person recommends or evaluates the 2-boxes-solution as rational.

Summing up, the main problem with Meek and Glymour's proposal consists in the fact that it is unclear when, why, and in which cases the decision maker should view his possible actions as interventions or not. Thus Meek and Glymour's proposal is in this sense unsatisfying.

Yet there may be a way to find out whether the decision maker should view his possible actions as interventions or not. For first, there may be a reason why it is rational to consider the will as a cause in Newcomb's problem. Second, one can ask whether it is a non-permissible irrationality of the decision maker not to consider the will as a cause in Newcomb's problem. With regard to the first point the following can be said: Leaving the problem of freedom of will aside it seems to me reasonable to assume that decisions should be at least partly determined by the will. Would anyone call a decision a decision, if the will were not involved as a cause? I doubt that. With regard to the second point I don't think that the decision maker is non-permissible irrational, if he doesn't consider the will as a cause in Newcomb's problem. For the decision maker might simply neglect the fact that the will is a causal factor. Furthermore, no logical impossibility is involved, if the decision maker doesn't view the will as a cause, and

there is no a priori argument to show that the will has to be viewed as a causal factor. Finally, there is up to now no generally accepted scientific theory about the psychological process of decision-making available. Thus although it is rational to consider the will as a cause in Newcomb's problem, it isn't non-permissible irrational of the decision maker not to consider the will as a cause in Newcomb's problem.

There may be reasons why the decision maker shouldn't view his possible actions as no interventions. For if the decision maker views his possible actions as no interventions, he conditions by his possible actions in the formula for calculating the utility of a possible action, thereby violating Spohn's principle. But if the decision maker views his possible actions as interventions, he doesn't condition by his possible actions in the formula for calculating the utility of a possible action, thereby obeying Spohn's principle. Thus the decision maker shouldn't view his possible actions as non-interventions.

With regard to Newcomb's problem Meek and Glymour's (1994) proposal to distinguish between conditioning and intervening may propose the right solution, namely to take both boxes. But it may also propose the wrong solution, namely to take B2. For within Meek and Glymour's proposal the solution to Newcomb's problem depends entirely on the beliefs of the decision maker about the causal structure in decision-making: While decision makers who take both boxes consider the will as a cause in Newcomb's problem, decision makers who take B2 don't consider the will as a cause. Because decision makers can change their point of view in decision-making within Meek and Glymour's proposal, so that in one case the decision maker can view his possible actions as interventions and in another case he can view his possible actions as no interventions, a certain arbitrariness in decision-making results. This arbitrariness can also show itself in the decision maker's attitude towards the possible actions in Newcomb's problem. Furthermore, Meek and Glymour's proposal is limited to decision makers who have so much self-knowledge as to view their possible actions either as interventions or as no interventions, which holds for Newcomb's problem, too.

If one applies Meek and Glymour's proposal to the finitely iterated Newcomb's problem with the same decision maker, and if one follows Sorensen's (1985) analysis of the finitely iterated Newcomb's problem, so that reputation building plays a role in decision-making, then in opposition to Sorensen's recommendation the decision to take both boxes is rational. Furthermore, within Meek and Glymour's proposal different perspectives, namely (1) first-person perspective and (2) third-person perspective both deliberating about present decisions, and (3) first-person perspective deliberating about

past decisions, make a difference for recommendations in decision-making or for evaluations of decisions, therefore also in Newcomb's problem.

Although it is rational to consider the will as a cause in Newcomb's problem, it isn't non-permissible irrational of the decision maker not to consider the will as a cause in Newcomb's problem. Finally, in Newcomb's problem the decision maker shouldn't view his possible actions as no interventions. For if the decision maker views his possible actions as no interventions, he conditions by his possible actions in the formula for calculating the utility of a possible action, thereby violating Spohn's principle.

Although Meek and Glymour (1994) provide a good explanation why evidential and causal decision theorists differ in their recommendations in Newcomblike problems, they don't provide a criterion for deciding which position is the rational one.

## 4.3 Nozick's Proposal of the Combination of Various Decision Principles

Nozick (1993) wants to provide an adequate solution to Newcomb's problem, which he wants to apply to the prisoner's dilemma; in order to fulfil this aim Nozick (1993) revises his former solution (Nozick 1969) to Newcomb's problem. Nozick (1993) claims that the standard normative rational decision theory has to be expanded, so that it also takes into account the symbolic meaning of possible actions. According to Nozick (1993) no solution of Newcomb's problem has been convincing. Furthermore, human beings are fallible, so Nozick (1993). Therefore it is irrational to put absolute confidence in one decision principle, Nozick (1993) claims. To support this conclusion Nozick (1993) gives the following example: In Newcomb's problem the decision maker's confidence in their favoured argument and therefore also their decision can be shaken, if the amount of money in B1 is varied. Decision makers who initially decided to take both boxes are unwilling to follow the principle of strong dominance, if the amount of money in B1 is lowered to $ 1; decision makers who initially decided to take B2 are unwilling to follow the principle of maximising conditional utility, if the amount of money in B1 is raised to $ 900,000. Therefore no decision maker has absolute confidence in the decision principle which he uses in the original version of Newcomb's problem, so Nozick (1993).

This result leads Nozick (1993) to develop a rational decision theory with the principle of maximising combined utility, where the combined utility of a possible action $a_i$ results from taking the utility of evidential decision theory for this possible action $a_i$

weighing it subjectively by the decision maker's confidence in the principle of maximising conditional utility and adding the utility of causal decision theory for this possible action $a_i$ weighing it subjectively by the decision maker's confidence in the principle of maximising utility of causal decision theory. Furthermore, Nozick (1993) claims that the weights aren't measures of uncertainty, but are measures of the legitimate force of each principle, so that we obtain a normative rational decision theory.

Nozick (1993) brings forward the following thought experiment to question the confidence of authors on rational decision theory with regard to their own views and to support his own view: Suppose there are pills which can transform the decision makers into consistent followers of each principle. In this case, Nozick (1993) claims, it is unclear whether each decision principle will recommend itself. For the recommendations of the decision principles depend on how the world is like, so Nozick (1993). If the world consists of many decision situations like Newcomb's problem, then taking the pill of evidential decision theory will have better causal consequences, so that the principle of causal decision theory will recommend to take the pill of evidential decision theory. If the world consists of many decision situations like Gibbard and Harper's (1978) Solomon case, then taking the pill of evidential decision theory will forgo significant benefits, so that the principle of evidential decision theory will recommend to take the pill of causal decision theory; furthermore, in this case the principle of causal decision theory will recommend to take the pill of causal decision theory, too.

Nozick (1993) continues by claiming that further decision principles, which are plausible from the decision maker's point of view and which have their own weights, can be included in the formula of the combined utility of a possible action $a_i$. In particular Nozick (1993) proposes that the following factor is added to the formula of the combined utility of a possible action $a_i$: The symbolic utility of a possible action $a_i$ with its own weight. Nozick (1993) points out that symbolic utility doesn't differ in kind from the utility of evidential decision theory and from the utility of causal decision theory; yet symbolic utility differs from the utility of evidential decision theory and from the utility of causal decision theory in that the connection is symbolic rather than evidential or causal.[185] The symbolic utility of a possible action $a_i$ is determined by the

[185]I would like to thank Brian Skyrms for drawing my attention to the fact - and I agree with him - that symbolic utility does differ in kind from the utility of evidential decision theory and the utility of causal decision theory. For Nozick (1993) doesn't provide a formula for calculating the

fact that the possible actions $a_i$ have symbolic connections to the possible outcomes or to other possible actions, Nozick (1993) claims.

## *A Critique of Nozick's Position in 1993*

Nozick's (1993) proposal of the combination of various decision principles can be criticised on the following grounds:

First, Nozick's (1993) proposal is uneconomical. For in his theory figure several decision principles and not one which is the case in all other rational decision theories. Thus Nozick's proposal of the combination of various decision principles shouldn't be chosen as a rational decision theory, if simplicity of the theory were a crucial factor.

Second, another critical point is that if the decision maker follows Nozick's (1993) proposal, the decision maker can make every possible action rationally justifiable. To obtain this result the decision maker just has to introduce a new decision principle or has to change the weights of the respective decision principles. Furthermore, Nozick doesn't give any rational criteria for selecting new decision principles/for maintaining old decision principles or for changing the decision maker's confidence in the respective decision principles/for determining the decision maker's confidence in the respective decision principles. Thus Nozick's proposal of the combination of various decision principles leaves many questions open and permits a lot of arbitrariness on the side of the decision maker.

Third, moreover, the symbolic utility of a possible action can lead to another kind of arbitrariness on the side of the decision maker. For possible actions can have different symbolic meanings for different decision makers and/or for the same decision maker at different decision times. This can be seen by the following: Taking both boxes in Newcomb's problem can symbolise for one decision maker "being a greedy person", but can symbolise for another decision maker "being a rational person"; for a third decision maker taking both boxes can also symbolise "being an irrational person". Or taking both boxes in Newcomb's problem can symbolise for a decision maker at $t_0$ "being a greedy person", but can symbolise for the same decision maker at $t_1$ "being a rational person"; taking both boxes can also symbolise "being an irrational person" for the same decision maker at $t_2$. Thus while the evidential utility of a possible action and the causal utility of a possible action each leads to one particular recommendation in Newcomb's problem over different decision makers and over the same decision maker

---

symbolic utility of a possible action, whereas causal and evidential decision theory provide such formulas.

at different decision times, the symbolic utility of a possible action can lead to different recommendations in Newcomb's problem for different decision makers and for the same decision maker at different decision times. Thus if the decision maker only follows the principle of maximising symbolic utility, he can make every possible action rationally justifiable.

Fourth, in my opinion being rational seems at least to demand that decision makers are coherent within their beliefs and wants and at least to some extent consistent over time. Thus the question arises whether Nozick's proposal fulfils these two criteria. My answer is no, for the following reasons: With regard to the coherence factor one can doubt whether it is coherent to believe at the same time (although in different degrees) in the principle of maximising causal utility, in the principle of maximising evidential utility, etc. For these different maximising principles exclude each other. Either causation should be a primitive term in rational decision theory or not. Furthermore, if causation should just in some cases or for some decision makers be a primitive term in rational decision theory and in some other cases or for some other decision makers not, Spohn's (1978, p. 182) criticism of Nozick's 1969-proposal can be transferred to Nozick's 1993-proposal. For then Nozick's (1993) theory doesn't provide a unifying solution to Newcomb's problem which is undesirable. Moreover, although one can argue like Lewis (1981a) that rational decision theory has to work for fully rational as well as for partly rational decision makers, this doesn't mean that the rational decision theory itself should be built in such a way that incoherence is part of the theory. For if a rational decision theory includes maximising principles which exclude each other, one can argue that such a theory is incoherent. Thus Nozick's (1993) proposal of the combination of various decision principles shouldn't be recommended as a rational decision theory, for it allows for incoherence.

With regard to the problem of consistency over time some prima facie arbitrariness on the side of the decision maker can be traced back to the learning factor. For being rational seems to demand that decision makers learn from experience. Thus decision makers can be inconsistent over time, but still rational, if the decision makers' beliefs and wants change adequately, that is rationally, by experience. Yet Nozick (1993) doesn't tell us how beliefs and wants change rationally by experience, so that Nozick's theory is inadequate in that respect. Moreover, consistency over time isn't demanded from the decision maker in Nozick's proposal. For the decision maker can just introduce a new decision principle, can weigh the decision principles differently, or can assign different symbolic meanings to his possible actions to justify his decision as

rational, so that in the most extreme case the decision maker can even justify opposite possible actions in the same decision situation at different times as rational. Therefore Nozick's (1993) proposal of the combination of various decision principles shouldn't be recommended as a rational decision theory, for it allows for inconsistency over time.

Fifth, in my opinion Nozick's (1993) thought experiment in which he asks whether the respective decision principles, the causal and the evidential one, recommend themselves is very valuable. Furthermore, Nozick isn't the only one to address the issue of self-recommending rational decision theories. For Skyrms (1982) also discusses this issue. Indeed this topic is important, because a rational decision theory which doesn't recommend itself by its own standards would be incoherent.

Yet Nozick's thought experiment can be criticised on the following grounds: First, Nozick's thought experiment actually consists of the finitely iterated Newcomb's problem and the finitely iterated Solomon's problem with the same decision maker and/or with different decision makers and not of the 1-shot Newcomb's problem and the 1-shot Solomon's problem. Sorensen (1985) has already shown that in the finitely iterated Newcomb's problem with the same decision maker reputation building plays a role, whereas in the 1-shot Newcomb's problem reputation building doesn't play a role (cf. chapter 1.6). Yet in the 1-shot Newcomb's problem the decision maker isn't able to build up a reputation and therefore cannot influence the predictor. Moreover, according to Sorensen (1985) the decision maker should not build up a reputation as a both boxes taker, but as a B2 taker in the finitely iterated Newcomb's problem with the same decision maker. For in the former case the predictor will be influenced to put $ 0 in B2, whereas in the latter case the predictor will be influenced to put $ 1,000,000 in B2. Thus the role of reputation building can make a difference to what is rational to decide for.

One can argue because of the reputation building factor that the finitely iterated Newcomb's problem is actually a game with two decision makers, whereas the 1-shot Newcomb's problem is actually a game against nature, so that it is reasonable to assume that the solution to the finitely iterated Newcomb's problem differs from the solution to the 1-shot Newcomb's problem. Thus Nozick's conclusion with regard to the different decision principles, namely that causal decision theory will recommend to take the pill of evidential decision theory in the finitely iterated Newcomb's problem, can be traced back to the reputation building factor. For to take the pill of evidential decision theory in the finitely iterated Newcomb's problem results in taking B2.

Nozick presupposes in his thought experiment that causal decision theory leads to different solutions in Newcomb's problem and in Solomon's problem (Gibbard and

Harper 1978), but actually causal decision theory leads to equivalent solutions in these cases as Gibbard and Harper (1978) have already shown (cf. chapter 2.2, and chapter 3.2). Or to say it in other words: It is unclear to me how Nozick can claim that in the finitely iterated Newcomb's problem the principle of causal decision theory will recommend to take the pill of evidential decision theory, while in the finitely iterated Solomon's problem the principle of causal decision theory will recommend to take the pill of causal decision theory. Therefore Nozick's thought experiment has to be made more precise - and I don't see how this could be done - to establish that the respective decision principles don't recommend themselves.

If the decision maker uses the principle of maximising evidential utility for decision-making which is permitted within Nozick's (1993) proposal, he thereby violates Spohn's (1977, 1978, and cf. chapter 3.6) principle which is valid. For the principle of maximising evidential utility demands from the decision maker to condition by his possible actions. Thus the decision maker has to ascribe credences to his possible actions which is forbidden by Spohn's principle. Therefore Nozick's proposal can be criticised for violating Spohn's principle.

With regard to Newcomb's problem Nozick's (1993) proposal of the combination of various decision principles may propose the right solution, namely to take both boxes. But it may also propose the wrong solution, namely to take B2. For within Nozick's proposal the solution to Newcomb's problem depends entirely on the decision maker, that is it depends on which decision principles the decision maker favours and how much he weighs them subjectively. Yet this dependence on the decision maker allows for a lot of arbitrariness. For first, Nozick doesn't demand that the decision maker is fully rational, he can also be partly rational or irrational. Second, the decision maker can make every possible action rationally justifiable by introducing new decision principles, by weighing them differently, and/or by giving different symbolic meanings to his possible actions. Furthermore, Nozick's proposal with regard to Newcomb's problem can be criticised on the following grounds: First, the decision maker has to make a decision which decision principles to use and how much to weigh them before making a decision in Newcomb's problem which is very uneconomical. Second, within Nozick's proposal the decision maker is incoherent, if he uses decision principles which exclude each other. Third, the decision maker can be inconsistent over time. For Nozick's approach doesn't demand consistency of the decision maker. Therefore Nozick's proposal doesn't provide a rational solution to Newcomb's problem. Finally, Nozick's

approach can be criticised for allowing the violation of Spohn's principle in general and therefore also in Newcomb's problem.

Although Nozick's (1993) proposal of the combination of various decision principles may be adequate as a descriptive rational decision theory, it is inadequate as a normative rational decision theory.

## 4.4 Kyburg's Distinction between Epistemic vs. Stochastic Independence

In opposition to Gibbard and Harper (1978) Kyburg (1980) isn't persuaded by the usefulness of subjunctive conditionals and of conditioning, but believes in the distinction between epistemic and stochastic independence which surprisingly leads to the same conclusions as Gibbard and Harper's (1978) approach with regard to Newcomb's problem. Kyburg (1980) claims that in deliberation epistemic probabilities play the role of Gibbard and Harper's (1978) probabilities of subjunctive conditionals, and that in deliberation stochastic conditional probabilities play the role of Gibbard and Harper's (1978) conditional probabilities.[186] Furthermore, unlike Gibbard and Harper (1978) Kyburg (1980) distinguishes between possible actions and types of possible actions and between possible outcomes and types of possible outcomes. Kyburg (1980) claims without giving any argument that types of possible actions and types of possible outcomes should be used in the calculation of the utility of a possible action. Another innovation of Kyburg (1974, 1980) is that he deals with interval probabilities. Kyburg's (1980) rational decision theory and therefore also his solution to Newcomb's problem is set within the framework of his inductive logic (Kyburg 1974).

---

[186]With regard to my question whether epistemic probabilities are the same as credences or degrees of belief Kyburg replied (personal communication from the 15th of September, 1998): "No. Credences and degrees of belief are properties of people. I'm not sure what credences are, but degrees of belief are generally intended to be real-valued, and I don't even think they exist. In any event, epistemic probabilities represent logical relations: P(statement|evidence) = [interval], whether anybody ever had that evidence or not. A question of logic, not psychology. Normative, not descriptive." Furthermore, I asked Kyburg whether stochastic probabilities are the same as statistical probabilities in Kyburg's terminology and what the relation of stochastic probabilities to chances and relative frequencies is. Kyburg gave me the following answer (personal communication from the 15th of September, 1998): "The short answer is yes; and identity. Of course this doesn't say much about the relations among the four kinds of things you mention. But I don't think those relations are terribly interesting/useful from the point of view of an agent. What the agent has to deal with are (I think) relative frequencies in (large) finite sets of events/objects. Sometimes these large finite sets may be 'hypothetical', so we inch toward chances ... ."

In the following I will present Kyburg's (1974, 1980) framework: Let $K$ be a rational corpus[187] which consists of a set of statements in a language $L$. Kyburg (1980) assumes that rational corpora are (1) deductively closed and (2) consistent, that is:

(1) $CnK \subset K$ ($CnK$ are the consequences of $K$),

(2) $\neg\ulcorner 0 = 1\urcorner \in K$.[188]

Furthermore, Kyburg (1980) defines the expansion of the rational corpus $K$ by a statement $S$, that is $K$ and $S$, (1) as the deductive closure of $K \cup \{S\}$, if $S$ is consistent with $K$, and (2) as the empty set, if $S$ is inconsistent with $K$, that is:

(1) $K$ and $S = Cn(K \cup \{S\})$, if $\neg\ulcorner 0 = 1\urcorner \in Cn(K \cup \{S\})$,

(2) $K$ and $S = 0$, if $\neg[\neg\ulcorner 0 = 1\urcorner \in Cn(K \cup \{S\})]$.

Within Kyburg's (1980) terminology a subjunctive conditional of the form $A \Box\!\!\rightarrow B$ is represented by "$B \in K$ and $A$", but according to Kyburg (1980) the language $L$ doesn't contain a counterfactual connective. Statistical statements are included in the language $L$ and will have the following form: $\ulcorner S(A, B, p, q)\urcorner$, which will be interpreted as "the frequency with which $A$'s are $B$'s lies between $p$ and $q$". Kyburg (1980) focuses on the strongest of such statements in $K$, so that he gives the following definition:

$\ulcorner S(A, B, p, q)\urcorner^* \in K$ if and only if $\ulcorner S(A, B, p, q)\urcorner \in K$;

and if $\ulcorner S(A, B, r, s)\urcorner \in K$, then $\ulcorner (p, q) \subset (r, s)\urcorner$ is a theorem.

With regard to probability Kyburg (1980, p. 150) makes the following distinctions: Because of statistical knowledge we say that the probability of this coin landing heads in the next toss is a half. Kyburg (1980) claims that epistemic probability refers to the probability with which this coin lands heads or to the probability that a toss of a coin will land heads; while he calls the former probability "properly epistemic", he calls the latter probability "stochastic". For with regard to stochastic probability statements, like "the probability that a toss of a coin will land heads", Kyburg (1980) claims that they differ from properly epistemic probability statements in the following way: The use of the indefinite article in the stochastic probability statements refers to a specific reference class, that is to tosses of coins in the example. Furthermore, Kyburg

---

[187] Kyburg (1974, p. 188) uses the term rational corpus instead of body of knowledge, because the thesis which he favours, namely that we know the items in our body of knowledge, is difficult to defend.

[188] Kyburg (1974, p. 157) gives the following example to clarify his usage of Quine's corners $\ulcorner\ \urcorner$: "... as a sign in the metalanguage '~' is ambiguous, functioning in some contexts as a one-place sentential connective (expressing negation), and in other contexts as a term, denoting a certain expression of the object language, namely '~'. These latter contexts will be easily identifiable by the fact that the signs in question occur in an expression enclosed in Quine's corners."

(1980) supposes that the indefinite article in "a toss" in a stochastic probability statement means "a described toss which is random relative to *K*". On the basis of this supposition Kyburg (1980) shows that the probability of "an *A* is a *B*" is the interval (*p*, *q*) relative to *K* if and only if *K* contains a statement to the effect that the measure of *B*'s among *A*'s lies between *p* and *q* and contains no stronger statement.

Within his framework Kyburg (1980) continues to provide the following terminology: "*RAN*(*a*, *B*, *C*, *K*)" means "*a* is a random member of *B* with respect to *C* relative to the rational corpus *K*".[189] Kyburg (1980) points out that this relation just obtains in case *B* is an appropriate reference class[190] for the assessment of the probability of "$a \in C$" given *K*.

In Kyburg's (1980) theory probability is defined for equivalence classes of statements:

*P*(*K*, *S*) = (*p*, *q*) if and only if there exist terms *a*, *B*, and *C* in the language *L*, so that the following holds:

(1) $\ulcorner a \in C \leftrightarrow S \urcorner \in K$,

(2) $\ulcorner S(B, C, p, q) \urcorner^* \in K$,

(3) *RAN*(*a*, *B*, *C*, *K*).

Therefore Kyburg (1980) proposes a logical and mainly syntactical definition of probability. Kyburg (1974, p. 285) claims that under some restrictions and relative to a strictly consistent rational corpus every statement of the language *L* has a probability, and that in some cases the classical probability calculus and a generalised probability calculus for interval probabilities obtain.

Kyburg (1980) introduces an operator $\alpha$ into the language *L* for the following purpose: $\alpha$ will be combined with appropriate terms for the position of *A* in $\ulcorner S(A, B, p, q) \urcorner$, that is terms which are taken to denote reference sets, to form terms which are taken as random members of those reference sets with regard to being a member of any set which is denoted by an appropriate term for the position of *B* in $\ulcorner S(A, B, p, q) \urcorner$. According to Kyburg (1980) this corresponds to one usage of the indefinite article

---

[189] According to Kyburg (1974, p. 222) randomness is a 4-termed relation in which *a* refers to an object, *B* to a class, *C* to a property, and *K* to a rational corpus.

[190] With regard to my questions when *B* is an appropriate reference class in *RAN*(*a*, *B*, *C*, *K*) and what an appropriate reference class is Kyburg gives the following answer (personal communication from the 15th of September, 1998): "A very deep question. Pollock seeks to answer it in terms of projectibility; I don't think that works. On my (1998) view a set of reference formulas and target formulas is simply characteristic of the language *L* in which we are expressing our knowledge. Then I'm obliged to explain why one language is better than another -- but I think I'm obliged to do that anyway."

in English which is exemplified in the sentence: "The probability that *a* coin toss will land heads is a half."

Kyburg (1980) poses two axioms for the operator α:

(1) If *B* and *C* are appropriate terms for ⌜*S*(*B*, *C*, *p*, *q*)⌝, then *RAN*(α*B*, *B*, *C*, *K*).

(2) If *B*, ⌜*A*∩*B*⌝, and *C* are appropriate terms, and if ⌜α*A*∈ *B*⌝∈ *K*, then *RAN*(α(⌜*A*∩*B*⌝),⌜*A*∩*B*⌝, *C*, *K*).

Kyburg (1980) defines conditional probability as follows: The probability of *S* given *T* relative to *K* is the probability of *S* relative to *K* and *T*, that is:

*P*(*K*, *T*, *S*) = *P*(*K* and *T*, *S*).

In Kyburg's (1980) terminology the distinction between stochastic and epistemic probabilities relative to *K* takes the following syntactical form: If *S* has the form ⌜α*A*∈ *B*⌝, then *P*(*K*, *S*) is stochastic; and if *S* doesn't have the form ⌜α*A*∈ *B*⌝, then *P*(*K*, *S*) is epistemic, so Kyburg (1980). If *S* has the form ⌜α*A*∈ *B*⌝ and *T* has the form ⌜α*A*∈ *C*⌝, then *P*(*K*, *S*, *T*) is a stochastic conditional probability; and if the operator α doesn't occur in either of *S* or *T*, then *P*(*K*, *S*, *T*) is an epistemic conditional probability, so Kyburg (1980); and if the operator α occurs in any other way, then *P*(*K*, *S*, *T*) is neither a stochastic nor an epistemic conditional probability.

Kyburg (1980) then introduces two theorems (for their proofs cf. Kyburg 1980, p. 152):

(1) *P*(*K*, ⌜α*A*∈ *B*⌝) = (p, q) if and only if ⌜*S*(*A*, *B*, *p*, *q*)⌝*∈ *K*.

(2) *P*(*K*, ⌜α*A*∈ *B*⌝, ⌜α*A*∈ *C*⌝) = (p, q) if and only if ⌜*S*(*A*∩*B*, *C*, *p*, *q*)⌝*∈ *K*.

Kyburg (1980) defines probabilistic independence: *S* is independent of *T* given *K*, if the probability of *S* remains the same by the addition of *T* to *K*, that is:

*SInd*(*K*, *T*) if and only if *P*(*K*, *S*) = *P*(*K*, *T*, *S*).

Furthermore, if *S* and *T* have the forms ⌜α*A*∈ *B*⌝ and ⌜α*A*∈ *C*⌝ respectively, so Kyburg (1980), then *SInd*(*K*, *T*) expresses a stochastic independence. According to Kyburg (1980) properly epistemic independence and stochastic independence which is symmetrical differ from causal independence.

With the help of these approaches Kyburg (1980) defines utility:

$$U(\ulcorner D\in a_i\urcorner, K) = \sum_{j=1}^{m} P(K, \ulcorner D\in a_i\urcorner, \ulcorner D\in o_j\urcorner)u(\ulcorner D\in o_j\cap a_i\urcorner),$$

where $a_i$ is a type of a possible action and $o_j$ is a type of a possible outcome. Kyburg (1980) distinguishes between properly epistemic utility and stochastic utility: If *D*, that is

the decision situation, takes the form $\ulcorner \alpha P \urcorner$, then $U_{a_i,o_j}(\ulcorner D \in a_i \urcorner, K)$ is stochastic; if $D$ doesn't take the form $\ulcorner \alpha P \urcorner$, then $U_{a_i,o_j}(\ulcorner D \in a_i \urcorner, K)$ is properly epistemic. Instances of types of decision situations are represented by terms of the form $\ulcorner \alpha P \urcorner$. According to Kyburg (1980) the decision maker should adopt the

> **principle of maximising properly epistemic utility**: In a given decision situation $D$ the decision maker $X$ should decide for a possible action $a_i$ with maximal properly epistemic utility.

## *Kyburg on Newcomb's Problem*

Kyburg's (1980) analysis of Newcomb's problem is as follows: Suppose $K_n$ is the decision maker's rational corpus, $s$ is the subject's decision situation, $N$ is the set of Newcomb situations[191], $M$ is the set in which there is $ 1,000,000 in B2, and $R$ is the set in which the decision maker decides to take B2. Furthermore, Kyburg (1980) claims that the decision maker should forget the predictor, for the predictor introduces a competitive element distorting our and the decision maker's intuitions. Kyburg (1980) just supposes that we and the decision maker know that almost all $R$'s are $M$'s and that almost none $R$'s are $M$'s.

Kyburg's (1980) calculation of the stochastic utilities yields:

---

[191]When I asked Kyburg which kind of cases fall in the set of Newcomb situations and whether Solomon's problem falls in the set of Newcomb situations, Kyburg responded (personal communication from the 15th of September, 1998): "I don't exactly remember Solomon's problem; but as I recall I thought that all the cases reflected the same reference class problem: what I decide to do is perfectly good evidence for someone else, but is not perfectly good evidence for me, since I am deciding. (Or, I guess one might say, I think I'm deciding, and so long as I think that I can't use that decision as evidence ... .)" In 1988 Kyburg makes his position a little bit more precise: For Kyburg (1988, p. 80) distinguishes between an "act" and a "specimen of behaviour". If the decision maker decides, he regards his decision as a free act and thus doesn't construe it as evidence that changes the decision maker's reference class. But if the decision maker looks back on his decision, he regards it as a bit of behaviour and thus construes it as evidence that changes the decision maker's reference class. The same holds for an observer. Harper (1988, p. xiii) claims that these ideas of Kyburg have their origin in Kant's treatment of the Freedom Antinomy. Yet one question remains. What is the connection between free acts and being no evidence? In my opinion the following answer suggests itself: When the decision maker decides he regards his decision as a free act, which leads to an interruption of the existent causal chains from the decision maker's point of view, so that he cannot regard his free action as evidence for the existent possible state of the world. An observer regards the decision maker's decision as a specimen of behaviour, which doesn't lead to an interruption of the existent causal chains from the observer's point of view, so that the observer can regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

$U(\ulcorner \alpha N \in R \urcorner, K_n)$ = $P(K_n$ and $\ulcorner \alpha N \in R \urcorner, \ulcorner \alpha N \in M \urcorner)\$ 1{,}000{,}000 + P(K_n$ and $\ulcorner \alpha N \in R \urcorner, \ulcorner \alpha N \in \neg M \urcorner)\$ 0,$

$\approx \$ 1{,}000{,}000,$

$U(\ulcorner \alpha N \in \neg R \urcorner, K_n)$ = $P(K_n$ and $\ulcorner \alpha N \in \neg R \urcorner, \ulcorner \alpha N \in M \urcorner)\$ 1{,}001{,}000 + P(K_n$ and $\ulcorner \alpha N \in \neg R \urcorner, \ulcorner \alpha N \in \neg M \urcorner)\$ 1{,}000,$

$\approx \$ 1{,}000.$

Under these conditions the decision maker should take B2.

According to Kyburg (1980), however, the decision maker isn't interested in the stochastic utilities, for the decision maker doesn't want to know the utility of a possible action to take both boxes and the utility of a possible action to take B2, but he wants to know the utility of his possible action to take both boxes and the utility of his possible action to take B2. Hence the decision maker is interested in the properly epistemic utilities which are calculated in the following way:

$U(\ulcorner s \in R \urcorner, K_n)$ = $P(K_n$ and $\ulcorner s \in R \urcorner, \ulcorner s \in M \urcorner)\$ 1{,}000{,}000 + P(K_n$ and $\ulcorner s \in R \urcorner, \ulcorner s \in \neg M \urcorner)\$ 0,$

$U(\ulcorner s \in \neg R \urcorner, K_n)$ = $P(K_n$ and $\ulcorner s \in \neg R \urcorner, \ulcorner s \in M \urcorner)\$ 1{,}001{,}000 + P(K_n$ and $\ulcorner s \in \neg R \urcorner, \ulcorner s \in \neg M \urcorner)\$ 1{,}000.$

Kyburg (1980) claims that it is in the power of the decision maker to make $\ulcorner s \in R \urcorner$ or $\ulcorner s \in \neg R \urcorner$ true[192], so that the decision maker has to believe that his decision situation $s$ falls in a subset $FN$ of $N$ in which $M$ is statistically independent of $R$; furthermore, the decision maker knows that the frequency with which \$ 1,000,000 is put in B2 lies between 0.1 and 0.5. Therefore the decision maker's rational corpus $K_n$ contains the statements:

$\ulcorner S(N, M, 0.1, 0.5) \urcorner,$

$\ulcorner S(FN \cap R, M, 0.1, 0.5) \urcorner,$

$\ulcorner S(FN \cap \neg R, M, 0.1, 0.5) \urcorner.$

Furthermore, Kyburg (1980) makes the following suppositions:

$RAN(s, \ulcorner FN \cap R \urcorner, M, K_n$ and $\ulcorner s \in R \urcorner),$

[192]I asked Kyburg whether this also means that the decision maker deliberates and therefore decides freely to take both boxes or just B2, so that his decision cannot be evidence for a prior possible state of the world, Kyburg answered (personal communication from the 15th of September, 1998): "I'm not sure that 'prior state' has much to do with anything. The whole idea of the examples is that there is no causal connection (whatever that may be!) between what the agent decides and what happens. But what the decision maker decides is perfectly good evidence for an observer, a third party."

that is $s$ is a random member of $\ulcorner FN \cap R \urcorner$ with respect to $M$ relative to the rational corpus $K_n$, which contains $\ulcorner s \in R \urcorner$, and

$RAN(s, \ulcorner FN \cap \neg R \urcorner, M, K_n$ and $\ulcorner s \in \neg R \urcorner)$,

that is $s$ is a random member of $\ulcorner FN \cap \neg R \urcorner$ with respect to $M$ relative to the rational corpus $K_n$, which contains $\ulcorner s \in \neg R \urcorner$,

so that

$P(K_n$ and $\ulcorner s \in R \urcorner, \ulcorner s \in M \urcorner)$

$= P(K_n$ and $\ulcorner s \in \neg R \urcorner, \ulcorner s \in M \urcorner)$

$= (0.1, 0.5)$.

Therefore the resultant properly epistemic utilities of the decision maker are:

$U(\ulcorner s \in R \urcorner, K_n) = (100{,}000, 500{,}000)$,

$U(\ulcorner s \in \neg R \urcorner, K_n) = (101{,}000, 501{,}000)$.

Hence the decision maker should take both boxes. Furthermore, the same recommendation results, so Kyburg (1980), (1) if the decision maker doesn't know the frequency with which $ 1,000,000 is put in B2, and (2) if the decision maker knows exactly the frequency with which $ 1,000,000 is put in B2. Therefore the recommendation to take both boxes results irrespective of the decision maker's knowledge of the frequency with which $ 1,000,000 is put in B2.

## *A Critique of Kyburg's Position*

Kyburg's (1980, 1988) proposal of maximising properly epistemic utility can be criticised for the following reasons:

In opposition to all other decision theorists Kyburg (1980) conceptualises probabilities as intervals. With regard to this modification two questions arise: First, is this modification necessary? Second, why aren't utilities conceptualised as intervals in Kyburg's theory, too?

With regard to the first question the following justification for using interval probabilities suggests itself, although Kyburg (1980) doesn't provide one: In rational decision theory one should use interval probabilities instead of point probabilities, because interval probabilities represent the decision maker's knowledge situation better than point probabilities. To give an example: The decision maker's knowledge situation that it is very likely that there is good weather tomorrow is better represented by the probability interval (0.7, 0.9) than by the point probability of P = 0.9. Furthermore, it is more realistic to claim that the probability of good weather tomorrow lies within the probability interval (0.7, 0.9) than to claim that the probability of good weather

tomorrow is P = 0.9. Therefore one should use interval probabilities instead of point probabilities in rational decision theory.

But as Kyburg (1974, pp. 264-267) already admits, in opposition to point probabilities interval probabilities only fulfil in certain special cases the axioms of a generalised mathematical probability calculus for intervals. This is a serious defect of interval probabilities. For as we have already seen in chapter 1.2, in the section on (subjective) utilities/objective utilities and credences/chances, it is a minimal requirement in rational decision theory that the decision maker's probabilities are coherent. This can be achieved by fulfilling the axioms of the mathematical probability calculus. Therefore Kyburg's interval probabilities are limited in their applicability to certain special cases.

Furthermore, in personal communication Brian Skyrms told me that his causal decision theory (Skyrms 1980, 1982, 1984) also works for interval probabilities, but that he uses point probabilities. Thus there must be some reasons why Skyrms doesn't use interval probabilities. Furthermore, all other decision theorists conceptualise their rational decision theories not in terms of interval probabilities, but in terms of point probabilities, so that they must have a justification for doing so, too.

Two reasons for point probabilities and against interval probabilities come to my mind: First, one should prefer a simpler rational decision theory to a more complicated one, if the former theory is as adequate as the latter for solving the problems in its field. Because a rational decision theory with point probabilities is simpler than a rational decision theory with interval probabilities, and because there is no reason why a rational decision theory with point probabilities shouldn't be able to deal with all the problems in its field, whereas a rational decision theory with interval probabilities is already limited in its applicability, a rational decision theory with point probabilities should be preferred to a rational decision theory with interval probabilities.

Second, if one uses interval probabilities one has to provide a justification for both borders, whereas if one uses point probabilities one just has to justify one point. The former can be seen by the following: If the decision maker claims that his knowledge situation that it is very likely that there is good weather tomorrow is best represented by the probability interval (0.7, 0.9), one can ask the decision maker why his knowledge situation isn't best represented by the probability interval (0.69, 0.91)? In the case of point probabilities, however, one can only ask the decision maker why he uses this point, for example, P = 0.9 in the weather example, and not any other one, like P = 0.91. Because in either case, in the case of interval probabilities and in the case of point probabilities, it is difficult to provide justifications, and because it is better to need

as few justifications as possible, point probabilities should be preferred to interval probabilities in rational decision theory.

With regard to the second question the following can be said: In 1988 Kyburg gives in that utilities like probabilities can be conceptualised as intervals, although no other decision theorist does that. The justification which I have given for interval probabilities can be transferred to interval utilities. Yet my two reasons for point probabilities and against interval probabilities can be transferred to reasons for point utilities and against interval utilities, too - with the exception of the violation of a generalised mathematical probability calculus for intervals in certain special cases -, so that in the end a rational decision theory with point utilities should be preferred to a rational decision theory with interval utilities. Therefore Kyburg (1980) was right in conceptualising utilities as points and not as intervals.

In opposition to all other decision theorists Kyburg (1980) uses types of possible actions, types of possible outcomes, and types of decision situations instead of using possible actions, possible outcomes, and decision situations, so that the question arises whether these modifications are necessary. Although Kyburg (1980) doesn't give a justification for using types of possible actions, types of possible outcomes, and types of decision situations instead of using possible actions, possible outcomes, and decision situations, the following justification suggests itself: One should use the former and not the latter, because by doing so the decision maker obtains recommendations for types of decision problems and not only for decision problems, which is very economical. Yet one has to provide a justification why some possible actions, possible outcomes, and possible decision situations belong to one type of possible actions, one type of possible outcomes, and one type of decision situations, whereas other possible actions, possible outcomes, decision situations don't belong to that particular type of possible actions, type of possible outcomes, and type of decision situations. In the case of possible actions, possible outcomes, and possible decision situations, however, no such justification is needed. Thus although these modifications are not necessary in rational decision theory, there is at least one point which speaks in favour of preferring types of possible actions, types of possible outcomes, and types of decision situations to possible actions, possible outcomes, and decision situations.

An advantage of Kyburg's (1980, 1988) theory is that he doesn't need possible states of the world or types of possible states of the world. For in Kyburg's theory types of decision situations seem to have taken their position. Thus in opposition to all causal decision theorists he doesn't have any problems with determining the correct partition or

the correct partitions of the possible states of the world, which is very nice. Furthermore, he doesn't have to provide partition theorems for the possible states of the world. Therefore it is an advantage of Kyburg's theory not to need possible states of the world or types of possible states of the world.

Another advantage of Kyburg's (1980) proposal is that it works without subjunctive conditionals.[193] For the decision maker's rational corpus just consists of a set of statements in the language $L$ which doesn't have a counterfactual connective. Thus in opposition to Gibbard and Harper (1978), Sobel (1986), and Lewis (1981a) Kyburg doesn't have to have a logic of subjunctive conditionals for his theory to work properly. Therefore Kyburg's theory is to be preferred to Gibbard and Harper's, Sobel's, and Lewis' theory in this respect.

Furthermore, Kyburg (1980) doesn't have causation as a primitive term in his theory; the only way that causation appears in his theory is by means of statistical knowledge in $K$. Thus Kyburg's theory is simpler than all causal decision theories in this respect. Kyburg even doesn't want causation to figure as a term in his theory at all. For in 1988 he states (p. 71) "Metaphysically speaking, causality is sheer superstition.". Therefore Kyburg achieves what he aims at.

Moreover, because Kyburg (1980) doesn't believe in conditioning, Spohn's principle isn't violated. In Newcomb's problem Kyburg (1980) claims that it is in the power of the decision maker to make $\ulcorner s \in R \urcorner$ or $\ulcorner s \in \neg R \urcorner$ true, so that the decision maker has to ascribe a probability of 1 to $\ulcorner s \in R \urcorner$ or to $\ulcorner s \in \neg R \urcorner$. Thus Spohn's principle isn't violated in this case either, which is an advantage of Kyburg's theory.

Kyburg (1988) claims that if the decision maker decides, he regards his decision as a free act and thus doesn't construe it as evidence that changes the decision maker's reference class. Furthermore, if the decision maker looks back on his decision, he regards it as a bit of behaviour and thus construes it as evidence that changes the decision maker's reference class. The same holds for an observer. Moreover, to make Kyburg's position coherent one has to the assume the following: When the decision maker decides, he regards his decision as a free act, which leads to an interruption of the existent causal chains from the decision maker's point of view, so that he cannot regard his free action as evidence for the existent possible state of the world. An observer regards the decision maker's decision as a specimen of behaviour, which doesn't lead to an interruption of the existent causal chains from the observer's point of

[193]For a valuable overview about the problems of subjunctive conditionals see Edgington (1995).

view, so that the observer can regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

Yet when Kyburg speaks of a free act, he implicitly assumes a certain meaning of freedom of will, so for Kyburg's theory to hold his position has to be made more precise in this respect. In order to understand my claim let's consider different meanings of freedom of will. According to Pfannkuche (unpublished) there are four different meanings of freedom of will:

(1) Absolute freedom of will formation,

(2) relative freedom of will formation,

(3) freedom of will,

(4) freedom of action.

**Ad 1**: Absolute freedom of will formation is given, if the agent can decide independently from all previous factors, what the agent's will is going to be, so that the agent's decision is only determined by his will. Pfannkuche ascribes this concept of freedom of will to Schopenhauer and Kant.

In my opinion Kyburg's position can be made coherent, if the decision maker regards his decision as a free act, in the sense that absolute freedom of will formation is given, and if the observer/the decision maker looking backwards in time regards the decision maker's decision not as a free act, in the sense that absolute freedom of will formation isn't given. For in the former case the decision maker can decide independently from all previous factors, what his will is going to be, so that there is no connection to the existent causal chains from the decision maker's point of view. Thus the decision maker cannot regard his decision as evidence for the existent possible state of the world. In the latter case the decision maker cannot decide independently from all previous factors, what his will is going to be, so that there is a connection to the existent causal chains from the observer's point of view. Thus the observer can regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

**Ad 2**: Relative freedom of will formation is the case, if the agent's will formation depends on characteristics of the agent, so that his will is determined by his beliefs, wishes, fears, and/or dispositions, etc. Moreover, an agent isn't free in this sense, if his will is influenced by hypnosis, and/or by means of neurosurgical operations, etc.

I think that Kyburg's theory cannot be made coherent, if the decision maker regards his decision as a free act, in the sense that relative freedom of will formation is given, and if the observer/the decision maker looking backwards in time regards the decision maker's decision not as a free act, in the sense that relative freedom of will formation isn't given. For in the former case the decision maker's will formation depends on characteristics of the decision maker, so that there is a connection to the existent causal chains from the decision maker's point of view. Thus the decision maker can regard his decision as evidence for the existent possible state of the world. In the latter case the decision maker's will formation doesn't depend on characteristics of the decision maker; the decision maker's will formation may depend on hypnosis, etc., though, so that there may be a connection to the existent causal chains from the observer's point of view. Thus the observer may regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

**Ad 3**: Freedom of will is given, if the agent isn't prevented by inner circumstances to do what he wants to do, so that one is dealing with the transposition of a given will into impulses of action or tryings to act and not with the formation of a will or of an intention. Moreover, an agent isn't free in this sense, if he wants to get rid of his drug addiction, but is forced by his inner emotions to inject himself drugs, if he wants to cross a square, but cannot do so because of his agoraphobia, etc.

In my opinion Kyburg's theory cannot be made coherent, if the decision maker regards his decision as a free act, in the sense that freedom of will is given, and if the observer/the decision maker looking backwards in time regards the decision maker's decision not as a free act, in the sense that freedom of will isn't given. For in the former case the decision maker isn't prevented by inner circumstances to do what he wants to do; yet he may be prevented by outer circumstances to do what he wants to do, so that there is a connection to the existent causal chains from the decision maker's point of view. Thus the decision maker can regard his decision as evidence for the existent possible state of the world. In the latter case the decision maker is prevented by inner circumstances to do what he wants to do, so that there is a connection to the existent causal chains from the observer's point of view. Thus the observer can regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

**Ad 4**: Freedom of action is given, if the agent isn't prevented by outer circumstances to do what he wants to do.

I think that Kyburg's theory cannot be made coherent, if the decision maker regards his decision as a free act, in the sense that freedom of action is given, and if the observer/the decision maker looking backwards in time regards the decision maker's decision not as a free act, in the sense that freedom of action isn't given. For in the former case the decision maker's decision isn't prevented by outer circumstances to do what he wants to do; yet he may be prevented by inner circumstances to do what he wants to do, so that there is a connection to the existent causal chains from the decision maker's point of view. Thus the decision maker can regard his decision as evidence for the existent possible state of the world. In the latter case the decision maker's decision is prevented by outer circumstances to do what he wants to do, so that there is a connection to the existent causal chains from the observer's point of view. Thus the observer can regard the decision maker's decision as evidence for the existent possible state of the world. The latter also holds for a decision maker looking backwards in time and evaluating his past decisions.

Therefore Kyburg's theory only holds, if he implicitly assumes a certain meaning of freedom of will, namely absolute freedom of will formation.

With regard to Newcomb's problem Kyburg's (1980) proposal to maximise properly epistemic utility proposes the right solution, namely to take both boxes. Yet Kyburg's modification to use interval probabilities instead of point probabilities in rational decision theory is an unnecessary modification which even leads to undesirable consequences and therefore should be dropped. Furthermore, in Newcomb's problem interval probabilities are of no relevance. For within Kyburg's theory the decision to take both boxes results irrespective of the decision maker's knowledge of the frequency with which $ 1,000,000 is put in B2. Interval utilities instead of point utilities could be of relevance for Newcomb's problem. For within Kyburg's theory the decision to take both boxes doesn't result irrespective of the decision maker's knowledge of the utilities. Yet there are arguments which speak against interval utilities and for point utilities, so that Kyburg was right in conceptualising utilities as points and not as intervals.

With regard to types of possible actions, types of possible outcomes, and types of decision situations instead of possible actions, possible outcomes, and decision situations the following can be said: By using the former and not by using the latter Kyburg's theory provides a decision-theoretical solution not only for Newcomb's problem, but also for Newcomblike problems, so that for economical reasons Kyburg's

theory has to be praised. Furthermore, within Kyburg's proposal there is no need of possible states of the world or types of possible states of the world, so that in opposition to all causal decision theories one doesn't have to determine the correct partition of the possible states of the world in Newcomb's problem or in any other decision problem, which is an advantage of his approach. Yet another question arises within Kyburg's theory, namely what is an appropriate reference class, which has to be answered for Newcomblike problems and therefore also for Newcomb's problem and which Kyburg hasn't answered properly.

In opposition to Gibbard and Harper's (1978), Sobel's (1986), and Lewis' (1981a) causal decision theories Kyburg's (1980) proposal works without subjunctive conditionals, so that Kyburg doesn't have to provide a logic of subjunctive conditionals for his theory to hold, which is an advantage of Kyburg's approach. For in Newcomb's problem it is difficult to determine whether the proposition "If I were to take both boxes, the predictor would have predicted this." or the proposition "If I were to take both boxes, there would be $ 1,000,000 in B2." is true, and whether the proposition "If I were to take B2, the predictor would have predicted this." or the proposition "If I were to take B2, there would be $ 0 in B2." is true.

In opposition to all causal decision theorists Kyburg (1980) doesn't have causation as a primitive term in his theory, so that Kyburg's proposal is simpler than all causal decision theories in this respect. Yet with regard to Newcomb's problem we have seen in chapter 1.4 that it consists of a conflict between probabilistic dependence and causal independence. Kyburg resolves the conflict by denying that there is causation in the world which is an adequate, although a very drastic way of doing so.

Kyburg's (1980) theory doesn't violate Spohn's (1977, 1978, and cf. chapter 3.6) principle and therefore also doesn't violate Spohn's principle with regard to Newcomb's problem, which is an advantage of Kyburg's theory.

In opposition to all other rational decision theories Kyburg's (1980) theory only holds, if he implicitly assumes absolute freedom of will formation. The latter is given, if the decision maker can decide independently from all previous factors, what his will is going to be. Applying this to Newcomb's problem the 2-boxes-solution only ensues, if the decision maker can decide independently from all previous factors, what his will is going to be. As a result there is an interruption of causal chains which has as a consequence that the decision maker cannot regard his decision as evidence for the existent possible state of the world. On the one hand it has to be praised that Kyburg makes the role of the will in rational decision theory explicit, on the other hand it is a

central weakness of Kyburg's theory that it only holds in case of absolute freedom of will formation.

## 4.5 Newcomb's Problem as a Game against Nature

Like Skyrms (1990a), who transfers the equilibrium concept from game theory to rational decision theory and who uses it for modelling the dynamics of rational deliberation, I would like to transfer a concept of game theory, namely the concept of a game against nature, to rational decision theory and wish to use it for providing a foundation for Skyrms' (1980, 1982, 1984) causal decision theory. Furthermore, like Brams (1983, pp. 41-65) I would like to use the concept of a game against nature for defending a certain solution to Newcomb's problem as rational. This concept which has to be distinguished from the concept of games with two or more decision makers explains why there are different recommendations (cf. Spohn in press) in a 1-shot Newcomb's problem vs. in a finitely iterated Newcomb's problem. (I will neither consider an infinitely iterated Newcomb's problem nor a cyclic Newcomb's problem here.) Furthermore, the concept of a game against nature explains why the predictor's high reliability and why the predictor's perspective (= observer) are irrelevant for providing a solution to Newcomb's problem. Moreover, if the decision maker conceives Newcomb's problem as a game against nature in contrast to a game with two or more decision makers, he takes a certain perspective, so that this approach of rational decision theory stresses the decision maker's perspective. Additionally, while the decision maker's belief in Newcomb's problem as a game against nature is rational, the decision maker's belief in Newcomb's problem as a game with two decision makers isn't non-permissible irrational.

In chapter 1.3, in the section on the classification of Newcomb's problem in game theory, Brams (1983, p. 181) defines games against nature in the following way: "A game against nature is a game in which one player is assumed to be 'nature', whose choices are neither conscious nor based on rational calculation but on chance instead." On this basis Brams classifies Newcomb's problem as a 1-person game against nature, but without giving any justification for doing so. Yet we have already seen in chapter 1.3 that it is problematic to view the predictor in Newcomb's problem as nature, whose decisions are based on chance. For the predictor's predictions are determined in some other way (cf. chapter 1.4). Thus this definition of a game against nature is inappropriate for Newcomb's problem. Therefore I propose a different definition: A

game against nature is a game in which one decision maker is assumed to be nature, whose decisions are already fixed and determinate.

By means of this definition the decision maker can view Newcomb's problem as a game against nature. For when the decision maker decides in Newcomb's problem, the predictor had already made his prediction and has put $ 1,000,000 in B2 or not, so that the predictor's prediction is already fixed and determinate. Furthermore, if the decision maker conceives Newcomb's problem as a game against nature and therefore takes a certain perspective with regard to the problem, the decision maker is confronted with two possible states of the world and not with two possible actions of the predictor besides his own possible actions.

I argued in chapter 1.3, in the section on the classification of Newcomb's problem in game theory, that it is rational to view a 1-shot Newcomb's problem as a game against nature, whereas it is rational to view a finitely iterated Newcomb's problem as a game with two decision makers. I will first deal with the latter and then with the former.

Sorensen (1985) claims that in a finitely iterated Newcomb's problem with the same decision maker the decision maker tries to build up a reputation as a B2 taker (cf. chapter 1.6). But why should the decision maker try to build up a reputation as a B2 taker, if he views the predictor as something belonging to the possible states of the world? The only way the decision maker can make sense of his reputation building behaviour is to view the predictor as playing an active role in determining the decision maker's possible outcomes. Moreover, the active role of the predictor is stressed by the following: Sorensen (1985) postulates that in a finitely iterated Newcomb's problem with the same decision maker the predictor's last prediction is affected by his opinion of the decision maker's decision-making tendencies. That is the predictor will put $ 1,000,000 in B2 in the last play, if he believes that the decision maker has the tendency to take B2; and the predictor will put $ 0 in B2 in the last play, if he believes that the decision maker has the tendency to take both boxes. Thus the predictor's future action is affected by the predictor's opinion of the decision maker's decision-making tendencies. Therefore a finitely iterated Newcomb's problem with the same decision maker seems to be a game with two decision makers, namely the decision maker and the predictor.

In opposition to that in a 1-shot Newcomb's problem the decision maker doesn't try to build up a reputation as a B2 taker, for one cannot build up a reputation in one single play. Furthermore, the predictor had already made his prediction in a 1-shot Newcomb's problem, when the decision maker decides, that is the predictor's prediction

had already become a part of the possible states of the world, when the decision maker decides (except in case backwards causation takes place in Newcomb's problem). Therefore a 1-shot Newcomb's problem seems to be a game against nature.

Conceiving Newcomb's problem as a game against nature also includes that the decision maker plays an active role in determining his possible outcomes. For according to Resnik (1987, p. 121) both in rational decision theory and in game theory the decision maker plays an active role in determining his possible outcomes. But what does playing an active role mean? In my opinion playing an active role at least means that the decision maker is free in the following sense (cf. Dummett 1993, p. 375): The decision maker can rule out the possibility that he attempts to carry out his decision, but finds himself unable to do so. For with regard to Newcomb's problem the decision maker's thought "My decision isn't really open, because the predictor already knows my decision" or my decision is already determined by my character will not help the decision maker to decide and will not dispense the decision maker from making a decision (cf. Dummett 1993, p. 372).

Why does the concept of a game against nature provide a foundation for Skyrms' (1980, 1982, 1984) causal decision theory? Although in Skyrms' (1984) causal decision theory by definition the decision maker's possible actions don't causally influence the possible states of the world, he doesn't give an argument why this should be the case. Furthermore, Newcomb's problem seems to be a counterexample of this definition. For it looks like as if the decision maker can influence the possible states of the world by his possible actions in Newcomb's problem. Yet if the decision maker considers Newcomb's problem as a game against nature, which is the rational thing to do (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory), it is clear that he cannot influence the possible states of the world by his possible actions. Moreover, I don't propose any further changes with regard to Skyrms' causal decision theory, so that there are no limitations with regard to freedom of will in contrast to Kyburg's (1980, 1988) theory and the decision maker should calculate the utility of a possible action as in Skyrms (1984) and use the corresponding maximising principle. As a result the decision maker should take both boxes in Newcomb's problem. Furthermore, by classifying Newcomb's problem as a game against nature Spohn's (1977, 1978) principle isn't violated. For the concept of a game against nature doesn't refer in any way to Spohn's principle.

Why does the concept of a game against nature explain that there are different recommendations in the case of a 1-shot Newcomb's problem vs. in the case of a

finitely iterated Newcomb's problem? Spohn (in press) demands that in the finitely iterated Newcomb's problem the decision maker should take B2, whereas in the 1-shot Newcomb's problem the decision maker should take both boxes. Yet by means of these recommendations Spohn can be criticised for not providing a unifying solution to Newcomb's problem. If one, however, conceives the finitely iterated Newcomb's problem as a game with two decision makers and the 1-shot Newcomb's problem as a game against nature, so that one actually has two different games at hand, the criticism against Spohn loses its substance. Furthermore, if the decision maker views a 1-shot Newcomb's problem as a game against nature, it doesn't make sense to build up a reputation as a B2 taker, whereas if the decision maker views a finitely iterated Newcomb's problem as a game with two decision makers, it makes sense to build up a reputation as a B2 taker. For in the former case he cannot influence the predictor by taking B2, while in the latter case he can at least try to influence the predictor by taking B2.

How does the concept of a game against nature make the predictor's high reliability and the predictor's perspective (= observer) irrelevant for providing a solution to Newcomb's problem? With regard to the predictor's high reliability the following can be said (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory): If a 1-shot Newcomb's problem is a game against nature, the following partition of the possible states of the world suggests itself as the correct partition: $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. Thus Newcomb's problem should be represented by the following decision matrix:

| | $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put \$ 0 in B2 at $t_2$. | $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put \$ 1,000,000 in B2 at $t_2$. |
|---|---|---|
| $a_1$: I take the content of both boxes at $t_3$. | $o_{11}$: \$ 1,000 | $o_{12}$: \$ 1,001,000 |
| $a_2$: I take the content of B2 at $t_3$. | $o_{21}$: \$ 0 | $o_{22}$: \$ 1,000,000 |

Figure 33. Decision matrix for Newcomb's problem of the possible outcomes $o_{11}, o_{12}, o_{21}, o_{22}$ which result from combining the possible actions $a_1, a_2$ with the possible states of the world $s_1, s_2$.

The reason why this is the correct partition is: If a 1-shot Newcomb's problem is a game against nature and nature is considered as something fixed and determinate, then the predictor's prediction is fixed and determinate, when the decision maker decides. Thus the predictor's prediction cannot be changed anymore to correspond to the predictor's high reliability and to the probabilistic dependence between the predictor's prediction and the decision maker's decision. In this way the predictor's high reliability and the probabilistic dependence become irrelevant to the solution of Newcomb's problem. And partitions of the possible states of the world which stress the predictor's correctness, like $s_1$: The predictor has predicted correctly at $t_1$ vs. $s_2$: The predictor hasn't predicted correctly at $t_1$, turn out to be wrong.

With regard to the predictor's perspective I would like to claim the following: If the decision maker conceives a 1-shot Newcomb's problem as a game against nature and therefore takes a certain perspective with regard to the problem, the decision maker is confronted with two possible states of the world and not with two possible actions of the predictor. In this way the predictor's perspective becomes irrelevant for providing a solution to Newcomb's problem, so that only the first-person perspective is relevant for decision-making in this case.

While the decision maker's belief that Newcomb's problem is a game against nature is rational (cf. chapter 1.3, in the section on the classification of Newcomb's problem in game theory), the decision maker's belief that Newcomb's problem is a game with two decision makers is no clear cut case for a non-permissible irrationality of

the decision maker. For Newcomb's problem is difficult to classify in game theory (cf. chapter 1.3). Moreover, there is no a priori argument which shows that Newcomb's problem as a game with two decision makers is downright impossible. Furthermore, as long as decision theorists don't question Newcomb's problem as a game with two decision makers, it would be too much to demand from the decision maker to do this. Thus the decision maker's belief in Newcomb's problem as a game with two decision makers is no clear cut case for a non-permissible irrationality of the decision maker.

## 4.6 Summary

Other proposals which stress the decision maker's perspective may provide adequate solutions to Newcomb's problem, namely to take both boxes (Meek and Glymour 1994; Nozick 1993; Kyburg 1980, 1988; my own proposal). Yet they may also provide inadequate solutions to Newcomb's problem, namely to take B2 (Meek and Glymour 1994; Nozick 1993).

The main results of this chapter are:

(1) Meek and Glymour's (1994) proposal to distinguish between conditioning and intervening provides a good explanation why evidential and causal decision theorists differ in their recommendations in Newcomblike problems, namely that evidential decision theorists don't view the will as a causal factor, while causal decision theorists view the will as a causal factor; yet Meek and Glymour's theory doesn't provide a criterion for deciding which position is the rational one, so that finally Meek and Glymour's theory is inadequate as a rational decision theory. Nozick's (1993) proposal of the combination of various decision principles is inadequate as a normative rational decision theory. It may be adequate as a descriptive rational decision theory. For in Nozick's theory it is only up to the decision maker to determine what rationality is. Kyburg's (1980, 1988) proposal to maximise properly epistemic utility is insofar inadequate as a rational decision theory as it holds only, if one assumes absolute freedom of will formation.

(2) While Meek and Glymour's (1994) theory doesn't make any assumptions with regard to different meanings of freedom of will, Kyburg's (1980, 1988) proposal to maximise properly epistemic utility only holds, if one assumes absolute freedom of will formation, which is Kant's conception of freedom of will and which is in accordance with Harper's (1988) claim that Kyburg's ideas have their origin in Kant's treatment of the Freedom Antinomy. Absolute freedom of will formation is given, if the decision

maker can decide independently from all previous factors, what his will is going to be. Thus in opposition to all other rational decision theories Meek and Glymour on the one hand and Kyburg on the other hand try to make the role of the will in decision-making explicit which is very valuable.

(3) While Meek and Glymour's (1994) proposal and Nozick's (1993) proposal of the combination of various decision principles allow for a lot of arbitrariness on the decision maker's side, Kyburg's (1980, 1988) proposal doesn't allow for any such arbitrariness. For within Meek and Glymour's proposal the decision maker can view his possible actions as interventions in one case and his possible actions as non-interventions in another case; within Nozick's proposal arbitrariness on the decision maker's side is possible, because the decision maker doesn't have to be fully rational; moreover, he can introduce new decision principles, can weigh them differently, and/or can assign different symbolic meanings to his possible actions.

(4) While Meek and Glymour's (1994) theory is limited to decision makers who have so much self-knowledge as to view their possible actions either as interventions or as non-interventions, Nozick's (1993) and Kyburg's (1980, 1988) theories aren't limited in this way.

(5) In opposition to Nozick's (1993) approach in Meek and Glymour's (1994) and Kyburg's (1980, 1988) theories different perspectives, namely first-person perspective and third-person perspective both deliberating about present decisions, and first-person perspective deliberating about past decisions, make a difference for recommendations in decision-making or for evaluations of decisions.

(6) In opposition to Kyburg's (1980, 1988) proposal to maximise properly epistemic utility Meek and Glymour's (1994) proposal to distinguish between conditioning and intervening and Nozick's (1993) proposal of the combination of various decision principles allow for the violation of Spohn's (1977, 1978, and cf. chapter 3.6) principle which is a disadvantage of their theories.

(7) In opposition to Gibbard and Harper's (1978), Sobel's (1986), and Lewis' (1981a) causal decision theories Meek and Glymour's (1994) and Kyburg's (1980, 1988) proposals work without subjunctive conditionals, whereas Nozick's (1993) proposal leaves it unspecified which causal decision theory the decision maker should favour - if he favours causal decision theories at all -, so that Meek and Glymour on the one side and Kyburg on the other side don't have to provide a logic of subjunctive conditionals, whereas in Nozick's case it is unclear whether he has to.

(8) Nozick's (1993) proposal of the combination of various decision principles and Kyburg's (1980, 1988) proposal of maximising properly epistemic utility are uneconomical in comparison to Meek and Glymour's (1994) distinction between conditioning and intervening. For in Nozick's proposal the decision maker can use various decision principles instead of one decision principle as is the case in all other rational decision theories, so that in Nozick's proposal the decision maker has to decide first which decision principles to use and how much to weigh them before making a decision, whereas in all other rational decision theories the decision maker can just apply the given decision principle. In Kyburg's proposal the decision maker uses interval probabilities instead of point probabilities as is the case in all other rational decision theories, so that the decision maker has to decide which boundaries are the correct ones in the case of interval probabilities, whereas the decision maker just has to determine one point in the case of point probabilities.

(9) Within Nozick's (1993) proposal the decision maker can be incoherent, if he uses decision principles which exclude each other, and inconsistent over time, if he uses different decision principles at different times and weighs them differently at different times, whereas in Kyburg's (1980, 1988) proposal the decision maker can neither be incoherent nor be inconsistent over time. In Meek and Glymour's (1994) proposal the decision maker cannot be incoherent, but he can be inconsistent over time. For with regard to the former he doesn't use decision principles which exclude each other, and with regard to the latter the decision maker can view his possible actions as interventions at one time and can view his possible actions as non-interventions at another time.

(10) In comparison with all other rational decision theories Nozick's (1993) proposal is the most unprecise of all. For while all other rational decision theories tell the decision maker how to calculate the utility of a possible action, Nozick's theory doesn't tell the decision maker how to calculate the symbolic utility of a possible action, if the decision maker wants to use the symbolic decision principle. Furthermore, Nozick doesn't specify which causal decision principle the decision maker should use, if the decision maker wants to use the causal decision principle.

(11) Although Kyburg's (1980, 1988) theory is uneconomical because of using interval probabilities, it is economical for using types of possible actions, types of possible outcomes, and types of decision situations instead of using possible actions, possible outcomes, and decision situations as is the case in all other rational decision theories. By using types of possible actions, etc. and not by using possible actions, etc.

Kyburg's theory provides decision-theoretical solutions for types of decision problems and not only for decision problems.

(12) Kyburg's theory is economical, because in opposition to all other rational decision theories it doesn't need possible states of the world or types of possible states of the world. As a result and in opposition to all causal decision theories Kyburg's theory doesn't have to determine the correct partition of the possible states of the world. Yet in opposition to all other rational decision theories Kyburg's theory has to answer what an appropriate reference class is.

(13) Kyburg's theory is economical, because in opposition to all causal decision theories his theory doesn't have causation as a primitive term. Kyburg's theory is even more economical than Meek and Glymour's (1994) and Nozick's (1993) proposals in this respect. For in Meek and Glymour's (1994) proposal causation figures as a primitive term, if the decision maker views his possible actions as interventions. In Nozick's (1993) theory causation figures as a primitive term, if the decision maker favours the causal decision principle.

(14) In comparison with all other rational decision theories Kyburg's (1980, 1988) proposal is the most innovative. For in opposition to all other decision theorists Kyburg uses interval probabilities, types of possible actions, types of possible outcomes, types of decision situations, and inductive logic (Kyburg 1974).

(15) Skyrms' (1980, 1982, 1984) causal decision theory is to be preferred to the other proposals of Meek and Glymour (1994), Nozick (1993), and Kyburg (1980, 1988). For Skyrms provides an adequate solution to Newcomb's problem, namely to take both boxes, whereas the other proposals may provide inadequate solutions to Newcomb's problem, namely to take B2 (Meek and Glymour 1994; Nozick 1993). Furthermore, it is a central weakness of Kyburg's (1980, 1988) proposal that it only holds, if one assumes absolute freedom of will formation, whereas Skyrms' causal decision theory isn't limited in this respect. Moreover, my own proposal to view Newcomb's problem as a game against nature provides a foundation for Skyrms' causal decision theory.

# Summary

All in all I come to the following conclusions with regard to Newcomb's problem:

(1) Newcomb's problem is a well-defined decision problem. For the predictor just has to be a little bit better than chance for Newcomb's problem to arise (cf. chapter 2.5, in the section on a critique of Jeffrey's decision kinematics).

(2) A 1-shot Newcomb's problem in contrast to a finitely iterated Newcomb's problem can be conceived as a 1-person game against nature, in which the correct partition of the possible states of the world is: $s_1$: The predictor has predicted at $t_1$ that I will take the content of both boxes, and he has put $ 0 in B2 at $t_2$. vs. $s_2$: The predictor has predicted at $t_1$ that I will take the content of B2, and he has put $ 1,000,000 in B2 at $t_2$. (cf. chapter 1.3).

(3) Newcomb's problem can be described as a conflict between the principle of maximising conditional utility and the principle of strong dominance with causal independence (cf. chapter 1.4). For on the one hand the principle of maximising conditional utility can be applied to Newcomb's problem. On the other hand the principle of strong dominance with causal independence can be applied to Newcomb's problem, because the decision maker believes that the possible states of the world are causally independent of the possible actions of the decision maker.

(4) (a) Evidential decision theories and (b) other proposals which stress the decision maker's perspective - with the exception of my own proposal - are inadequate rational decision theories (cf. chapter 2 and chapter 4). For (a) evidential decision theories do not provide adequate solutions to Newcomb's problem. Either they propose wrong solutions to Newcomb's problem, like Jeffrey (1965) and Jeffrey (1996), or they propose right solutions to Newcomb's problem, but their reasons for coming to the 2-boxes-solution are not adequate, like Jeffrey (1983), Jeffrey (1988), and Eells (1981, 1982, 1985). (b) Although other proposals which stress the decision maker's perspective may provide adequate solutions to Newcomb's problem, namely to take both boxes (Meek and Glymour 1994; Nozick 1993; Kyburg 1980, 1988; my own proposal), they may also provide inadequate solutions to Newcomb's problem, namely to take B2 (Meek and Glymour 1994; Nozick 1993), so that the latter two disqualify as adequate rational decision theories. Furthermore, it is a central weakness of Kyburg's (1980, 1988) proposal that it only holds, if one assumes absolute freedom of will formation.

(5) Causation should figure as a primitive term in rational decision theory. For the decision maker shouldn't seek for good news, but should seek for good results. With regard to Newcomb's problem this means that the decision maker should take both boxes (cf. chapter 2.2, in the section on the advantages and disadvantages of the central features of Jeffrey's logic of decision, ad 4).

(6) Skyrms' (1980, 1982, 1984) causal decision theory is presently the best rational decision theory. Its solution to Newcomb's problem consists in taking both boxes (cf. chapter 3.3). In my own proposal to view Newcomb's problem as a game against nature I give a foundation for Skyrms' causal decision theory. Skyrms' (1984) approach is in the ultimate analysis completely subjective, so that it isn't limited to decision makers who believe in chances. In Skyrms' (1984) proposal unconditional credences are used for calculating the utility of a possible action. Thus Spohn's (1977, 1978) principle isn't violated. Furthermore, in Skyrms' (1984) theory the possible states of the world are partitioned rationally, so that Skyrms' causal decision theory works for in this respect fully rational decision makers. Skyrms' (1984) proposal isn't formulated in terms of subjunctive conditionals - although his theory is compatible to the latter because of his (Skyrms 1984) Bayesian theory of conditionals -, so that his proposal isn't limited in its applicability in this respect. While Gibbard and Harper's (1978) causal decision theory is formulated in terms of ultimate possible outcomes, Skyrms (1985) proposes a causal decision theory which works for proximate possible outcomes. Because proximate possible outcomes demand less knowledge from the decision maker than ultimate possible outcomes, Skyrms' causal decision theory is to be preferred to Gibbard and Harper's causal decision theory in this respect. Whereas Gibbard and Harper's (1978) proposal is limited to decision makers who believe in deterministic worlds, Skyrms' approach is even meant to work for decision makers who believe in indeterministic worlds. Gibbard and Harper's (1978) theory is limited to decision makers with non-backtracking intuitions, although Newcomb's problem arouses backtracking intuitions. Skyrms (1984), however, can account for backtracking intuitions of the decision maker.

(7) Spohn's (1977, 1978) principle, which says: "*Any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts.*"[194] (Spohn 1977, p. 114), is valid for acts and for probabilistic acts (cf. chapter 3.6).

---

[194]Trivial conditional credences, like $c(a_1|a_1) = 1$ for a possible action $a_1$ or $c(a_2|a_1) = 0$ for two disjunctive possible actions $a_1$ and $a_2$, are not considered.

(8) The possible states of the world have to be well-specified in rational decision theory, which leads to an exclusion of the evidential partition of the possible states of the world in Newcomb's problem (cf. chapter 3.5, in the section on a critique of Lewis' position). For an evidential partition of the possible states of the world, like $s_1$: The predictor has predicted correctly at $t_1$. vs. $s_2$: The predictor hasn't predicted correctly at $t_1$., refers to the earlier possible state of the world, when the prediction was made, and it refers to the later possible state of the world, when the predicted possible action takes place. While the former possible state of the world already obtains, when the decision maker decides, it is logically possible that the latter possible state of the world doesn't turn out as predicted, so that the possible state of the world isn't well-specified. To be more precise the possible state of the world is over-specified. For it already determines a future possible state of the world as true.

(9) Rabinowicz (1982) rightly claims that Lewis' (1981a) suggestion to apply rational decision theory not only to fully rational decision makers, but also to partly rational decision makers must have some limits. For we cannot allow the decision maker to be as irrational as he wants to be (cf. chapter 3.4, in the section on a critique of Sobel's position).

(10) Gibbard and Harper's (1978) and Lewis' (1981b) formulation of the "why ain't you rich?"-argument is unsound (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position). For premise 2 should get a different formulation. As it stands it suggests that all decision makers who follow the principle of maximising *V*-utility become millionaires in Newcomb's problem, while all decision makers who follow the principle of maximising *U*-utility don't become millionaires in Newcomb's problem. And this is clearly false, if the original version of Newcomb's problem is our starting-point. Yet the correct formulation of premise 2 together with premise 1 doesn't yield the conclusion of the "why ain't you rich?"-argument.

(11) Gibbard and Harper's (1978) claim that it is rational to take both boxes, if the predictor is infallible, has to be made more precise by distinguishing between different kinds of infallibility, namely analytical necessity, causal necessity, and logical necessity (cf. chapter 3.2, in the section on a critique of Gibbard and Harper's position).

(12) Meek and Glymour (1994) and Kyburg (1980, 1988) rightly claim that the will plays a role in decision-making (cf. chapter 4.2 and chapter 4.4). Furthermore, Meek and Glymour (1994) and Kyburg (1980, 1988) rightly point out that different perspectives make a difference for recommendations in decision-making or for evaluations of decisions (cf. chapter 4.2 and chapter 4.4).

# Zusammenfassung

Alles in allem komme ich zu folgenden Schlußfolgerungen in bezug auf Newcombs Problem:

(1) Newcombs Problem ist ein wohldefiniertes Entscheidungsproblem. Denn der Vorhersager muß nur ein bißchen besser als der Zufall sein, damit Newcombs Problem entstehen kann. (s. Kapitel 2.5, im Abschnitt einer Kritik zu Jeffreys Entscheidungskinematik).

(2) Ein einmal durchgeführtes Newcombs Problem im Gegensatz zu einem endlich wiederholten Newcombs Problem kann als ein 1-Personen Spiel gegen die Natur aufgefaßt werden, in dem die korrekte Zerlegung der möglichen Weltzustände folgende ist: $s_1$: Der Vorhersager hat zu $t_1$ vorhergesagt, daß ich den Inhalt von beiden Boxen nehmen werde, und er hat zu $t_2$ $ 0 in B2 gelegt. vs. $s_2$: Der Vorhersager hat zu $t_1$ vorhergesagt, daß ich den Inhalt von B2 nehmen werde, und er hat zu $t_2$ $ 1,000,000 in B2 gelegt. (s. Kapitel 1.3).

(3) Newcombs Problem kann als ein Konflikt zwischen dem Prinzip der Maximierung des bedingten Nutzens und dem starken Dominanzprinzip mit kausaler Unabhängigkeit beschrieben werden (s. Kapitel 1.4). Denn auf der einen Seite kann man das Prinzip der Maximierung des bedingten Nutzens auf Newcombs Problem anwenden. Auf der anderen Seite kann das starke Dominanzprinzip mit kausaler Unabhängigkeit auf Newcombs Problem angewandt werden, und zwar weil der Entscheidende glaubt, daß die möglichen Weltzustände kausal unabhängig von den möglichen Handlungen des Entscheidenden sind.

(4) (a) Evidentielle Entscheidungstheorien und (b) Theorien, die die Perspektive des Entscheidenden betonen, - mein eigener Vorschlag ausgenommen - sind inadäquate rationale Entscheidungstheorien (s. Kapitel 2 und Kapitel 4). Denn (a) evidentielle Entscheidungstheorien liefern keine adäquaten Lösungen für Newcombs Problem. Entweder schlagen sie falsche Lösungen für Newcombs Problem vor, wie Jeffrey (1965) und Jeffrey (1996), oder sie schlagen richtige Lösungen für Newcombs Problem vor, aber ihre Gründe für eine 2-Boxen-Lösung sind nicht adäquat, wie Jeffrey (1983), Jeffrey (1988) und Eells (1981, 1982, 1985). (b) Obwohl die Theorien, die die Perspektive des Entscheidenden betonen, adäquate Lösungen für Newcombs Problem liefern können, nämlich beide Boxen zu nehmen (Meek und Glymour 1994; Nozick 1993; Kyburg 1980, 1988; mein eigener Vorschlag), können sie auch inadäquate

Lösungen für Newcombs Problem liefern, nämlich B2 zu nehmen (Meek und Glymour 1994; Nozick 1993), so daß sich die letzteren zwei als adäquate rationale Entscheidungstheorien disqualifizieren. Außerdem ist es eine zentrale Schwäche von Kyburgs (1980, 1988) Theorie, daß sie nur dann gilt, wenn man absolute Willensbildungsfreiheit annimmt.

(5) Kausalität sollte ein grundlegender Term in der rationalen Entscheidungstheorie sein. Denn der Entscheidende sollte nicht nach guten Neuigkeiten suchen, sondern nach guten Resultaten. In bezug auf Newcombs Problems bedeutet das, daß der Entscheidende beide Boxen nehmen sollte. (s. Kapitel 2.2, im Abschnitt zu den Vorteilen und Nachteilen der zentralen Merkmale von Jeffreys Entscheidungslogik, ad 4).

(6) Skyrms (1980, 1982, 1984) kausale Entscheidungstheorie ist gegenwärtig die beste rationale Entscheidungstheorie. Ihre Lösung für Newcombs Problem besteht im Nehmen beider Boxen (s. Kapitel 3.3). In meinem eigenem Vorschlag, Newcombs Problem als ein Spiel gegen die Natur zu betrachten, liefere ich, Skyrms kausaler Entscheidungstheorie eine Grundlage. Skyrms (1984) Theorie ist letztlich vollständig subjektiv, so daß sie nicht auf Entscheidende beschränkt ist, die an objektive Wahrscheinlichkeiten glauben. In Skyrms (1984) Vorschlag werden unbedingte subjektive Wahrscheinlichkeiten benutzt, um den Nutzen einer möglichen Handlung zu berechnen. Also wird Spohns (1977, 1978) Prinzip nicht verletzt. Außerdem werden in Skyrms (1984) Theorie die möglichen Weltzustände auf rationale Weise zerlegt, so daß Skyrms kausale Entscheidungstheorie für in bezug auf diesen Punkt vollständig rationale Entscheidende funktioniert. Skyrms (1984) Vorschlag ist nicht in Form von subjunktiven Konditionalen formuliert - obwohl seine Theorie wegen seiner (Skyrms 1984) Bayesschen Theorie der Konditionale kompatibel mit letzteren ist -, so daß sein Vorschlag in bezug auf diesen Punkt in seiner Anwendbarkeit nicht beschränkt ist. Während Gibbard und Harpers (1978) kausale Entscheidungstheorie in Form von letzten möglichen Folgen formuliert ist, schlägt Skyrms (1985) eine kausale Entscheidungstheorie vor, die mit nächsten möglichen Folgen arbeitet. Weil nächste mögliche Folgen weniger Wissen vom Entscheidenden verlangen als letzte mögliche Folgen, ist die kausale Entscheidungstheorie von Skyrms der Entscheidungstheorie von Gibbard und Harper in bezug auf diesen Punkt vorzuziehen. Während Gibbard und Harpers (1978) Vorschlag auf Entscheidende beschränkt ist, die an deterministische Welten glauben, soll Skyrms Vorschlag sogar für Entscheidende funktionieren, die an indeterministische Welten glauben. Gibbard und Harpers (1978) Theorie ist beschränkt

auf Entscheidende mit nicht-zurückverfolgenden Intuitionen, obwohl Newcombs Problem zurückverfolgende Intuitionen erzeugt. Skyrms (1984) kann jedoch mit zurückverfolgenden Intuitionen des Entscheidenden umgehen.

(7) Spohns (1977, 1978) Prinzip, welches besagt, daß adäquate quantitative Entscheidungsmodelle weder explizit noch implizit subjektive Wahrscheinlichkeiten für Handlungen enthalten dürfen[195], ist für Handlungen und für probabilistische Handlungen valide (s. Kapitel 3.6).

(8) In der rationalen Entscheidungstheorie müssen mögliche Weltzustände wohlspezifiziert sein, was zum Ausschluß der evidentiellen Zerlegung der möglichen Weltzustände in Newcombs Problem führt (s. Kapitel 3.5, im Abschnitt einer Kritik zur kausalen Entscheidungstheorie von Lewis). Denn eine evidentielle Zerlegung der möglichen Weltzustände, wie $s_1$: Der Vorhersager hat zu $t_1$ korrekt vorhergesagt. vs. $s_2$: Der Vorhersager hat zu $t_1$ nicht korrekt vorhergesagt., bezieht sich auf den früheren möglichen Weltzustand, wann die Vorhersage gemacht wurde, und sie bezieht sich auf den späteren möglichen Weltzustand, wann die vorhergesagte mögliche Handlung stattfinden wird. Während der erstere der beiden möglichen Weltzustände schon gegeben ist, wenn sich der Entscheidende entscheidet, ist es logisch möglich, daß der letztere der beiden möglichen Weltzustände nicht wie vorhergesagt eintritt, so daß dieser mögliche Weltzustand nicht wohlspezifiziert ist. Um präziser zu sein, dieser mögliche Weltzustand ist überspezifiziert. Denn er bestimmt schon einen zukünftigen möglichen Weltzustand als wahr.

(9) Rabinowicz (1982) behauptet richtigerweise, daß Lewis (1981a) Vorschlag eingeschränkt werden muß, rationale Entscheidungstheorie nicht nur auf vollständig rationale Entscheidende anzuwenden, sondern auch auf teilweise rationale Entscheidende. Denn wir können dem Entscheidenden nicht erlauben, so irrational zu sein, wie er möchte (s. Kapitel 3.4, im Abschnitt einer Kritik zur kausalen Entscheidungstheorie von Sobel).

(10) Gibbard und Harpers (1978) und Lewis (1981b) Formulierung des "Warum bist Du nicht reich?"-Argumentes ist nicht korrekt (s. Kapitel 3.2, im Abschnitt einer Kritik zur kausalen Entscheidungstheorie von Gibbard und Harper). Denn Prämisse 2 sollte anders formuliert werden. So wie sie formuliert ist, suggeriert sie, daß alle Entscheidenden, die dem Prinzip der Maximierung des *V*-Nutzens folgen, Millionäre in

---

[195]Triviale bedingte subjektive Wahrscheinlichkeiten, wie $c(a_1|a_1) = 1$ für eine mögliche Handlung $a_1$ oder $c(a_2|a_1) = 0$ für zwei disjunktive mögliche Handlungen $a_1$ und $a_2$, werden nicht betrachtet.

Newcombs Problem werden, während alle Entscheidenden, die dem Prinzip der Maximierung des *U*-Nutzens folgen, keine Millionäre in Newcombs Problem werden. Und das ist klarerweise falsch, wenn die ursprüngliche Version von Newcombs Problem unser Ausgangspunkt ist. Die korrekte Formulierung von Prämisse 2 zusammen mit Prämisse 1 liefert jedoch nicht die Konklusion des "Warum bist Du nicht reich?"-Argumentes.

(11) Gibbard und Harpers (1978) Behauptung, daß es rational ist, beide Boxen zu nehmen, wenn der Vorhersager unfehlbar ist, muß präzisiert werden, in dem man zwischen unterschiedlichen Arten der Unfehlbarkeit unterscheidet, nämlich analytischer Notwendigkeit, kausaler Notwendigkeit und logischer Notwendigkeit (s. Kapitel 3.2, im Abschnitt einer Kritik zur kausalen Entscheidungstheorie von Gibbard und Harper).

(12) Meek und Glymour (1994) und Kyburg (1980, 1988) behaupten richtigerweise, daß der Willen eine Rolle beim Entscheiden spielt (s. Kapitel 4.2 und Kapitel 4.4). Außerdem heben Meek und Glymour (1994) und Kyburg (1980, 1988) richtigerweise hervor, daß unterschiedliche Perspektiven einen Unterschied für Empfehlungen beim Entscheiden oder für die Bewertung von Entscheidungen machen (s. Kapitel 4.2 und Kapitel 4.4).

# References


Adams, R. M. (1995), "Possible Worlds", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pp. 633-634.

Albert, M. (1998), *The Logic of Risk and Uncertainty*, Habilitationsschrift, Universität Konstanz.

Anglin, W. S. (1981), "Backwards Causation", *Analysis* 41, pp. 86-91.

Aristotle (1926), *Nichomachean Ethics*, translated by H. Rackham, Heinemann, London.

Armendt, B. (1986), "A Foundation for Causal Decision Theory", *Topoi* 5, pp. 3-19.

Armstrong, D. M. (1997), *A World of States of Affairs*, Cambridge University Press, Cambridge, UK.

Arntzenius, F. (1990), "Physics and Common Causes", *Synthese* 82, pp. 77-96.

Aronson, J. (1971), "On the Grammar of 'Cause'", *Synthese* 22, pp. 414-430.

Arrow, K. J. (1971), *Essays in the Theory of Risk-Bearing*, Markham, Chicago.

Audi, R. (1986), "Intending Intentional Action, and Desire", in J. Marks (ed.), *The Ways of Desire*, Precedent, Chicago, pp. 17-38.

Aumann, R. J. (1976), "Agreeing to Disagree", *Annals of Statistics* 4, pp. 1236-1239.

Aumann, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55, pp. 1-18.

Axelrod, R., and Hamilton, W. D. (1981), "The Evolution of Cooperation", *Science* 211, pp. 1390-1396.

Bach, K. (1987), "Newcomb's Problem: The $ 1,000,000 Solution", *Canadian Journal of Philosophy* 17, pp. 409-426.

Bar-Hillel, M., and Margalit, A. (1972), "Newcomb's Paradox Revisited", *British Journal for the Philosophy of Science* 23, pp. 295-304.

Bar-Hillel, M., and Margalit, A. (1985), "Gideon's Paradox - a Paradox of Rationality", *Synthese* 63, pp. 139-155.

Barnes, R. E. (1997), "Rationality, Dispositions, and the Newcomb Paradox", *Philosophical Studies* 88, pp. 1-28.

Bell, J. (1964), "On the Einstein Podolsky Rosen Paradox", *Physics* 1, pp. 195-200.

Ben-Menahem, Y. (1986), "Newcomb's Paradox and Compatibilism", *Erkenntnis* 25, pp. 197-220.

Benditt, T. M., and Ross, D. J. (1976), "Newcomb's 'Paradox'", *British Journal for the Philosophy of Science* 27, pp. 161-164.

Binmore, K. (1992), *Fun and Games: A Text on Game Theory*, D. C. Heath and Company, Lexington, MA, and Toronto.

Bittner, R. (1992), "Was ist eine Entscheidung?", *Ethik und Sozialwissenschaften* 3, pp. 17-22.

Bolker, E. (1965), *Functions Resembling Quotients of Measures*, dissertation, Harvard University.

Bolker, E. (1966), "Functions Resembling Quotients of Measures", *Transactions of the American Mathematical Society* 124, pp. 292-312.

Bolker, E. (1967), "A Simultaneous Axiomatization of Utility and Subjective Probability", *Philosophy of Science* 34, pp. 333-340.

Brams, S. J. (1975), "Newcomb's Problem and Prisoners' Dilemma", *Journal of Conflict Resolution* 19, pp. 596-612.

Brams, S. J. (1983), *Superior Beings: If They Exist, How Would We Know?*, Springer, New York, Berlin, Heidelberg, and Tokyo.

Bratman, M. E. (1987), *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA, and London, UK.

Broome, J. (1989), "An Economic Newcomb Problem", *Analysis* 49, pp. 220-222.

Campbell, R. (1985), "Background for the Uninitiated", in R. Campbell, and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, Vancouver, pp. 3-41.

Cargile, J. (1975), "Newcomb's Paradox", *British Journal for the Philosophy of Science* 26, pp. 234-239.

Carnap, R. (1950/1962), *Logical Foundations of Probability*, University of Chicago Press, Chicago.

Craig, W. L. (1987), "Divine Foreknowledge and Newcomb's Paradox", *Philosophia* 17, pp. 331-350.

Cramer, J. (1986), "The Transactional Interpretation of Quantum Mechanics", *Reviews of Modern Physics* 58, pp. 647-687.

Davis, M. D. (1970), *Game Theory: A Nontechnical Introduction*, Basic Books, New York, and London.

de Finetti, B. (1937), "La prévision: ses lois logiques, ses sources subjectives", *Annales de l'Institut Henri Poincaré* 7, pp. 1-68.

Dowe, P. (1992), "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* 59, pp. 195-216.

Dowe, P. (forthcoming), "Causality and Identity", *Communication and Cognition*.

Dowe, P., Oakley, A., and Rosier, A. (forthcoming), "Backwards Causation", *Metascience*.

Dummett, M. (1964), "Bringing about the Past", *Philosophical Review* 73, pp. 338-359.

Dummett, M. (1993), *The Seas of Language*, Clarendon Press, Oxford.

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, MIT Press, Cambridge, MA, and London, UK.

Edgington, D. (1995), "On Conditionals", *Mind* 104, pp. 235-329.

Eells, E. (1981), "Causality, Utility, and Decision", *Synthese* 48, pp. 295-327.

Eells, E. (1982), *Rational Decision and Causality*, Cambridge University Press, Cambridge.

Eells, E. (1984), "Metatickles and the Dynamics of Deliberation", *Theory and Decision* 17, pp. 71-95.

Eells, E. (1985), "Causality, Decision, and Newcomb's Paradox", in R. Campbell, and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, Vancouver, pp. 183-213.

Eells, E., and Sober, E. (1983), "Probabilistic Causality and the Question of Transitivity", *Philosophy of Science* 50, pp. 35-57.

Eells, E., and Sober, E. (1986), "Common Causes and Decision Theory", *Philosophy of Science* 53, pp. 223-245.

Factor, R. L. (1978), "Newcomb's Paradox and Omniscience", *International Journal for Philosophy of Religion* 9, pp. 30-40.

Fair, D. (1979), "Causation and the Flow of Energy", *Erkenntnis* 14, pp. 219-250.

Feynman, R. (1949), "The Theory of Positrons", *Physical Review* 76, pp. 749-759.

Fischer, J. M. (1994), *The Metaphysics of Free Will: An Essay on Control*, Basil Blackwell, Cambridge, MA, and Oxford, UK.

Fishburn, P. C. (1964), *Decision and Value Theory*, Wiley, New York.

Fishburn, P. C. (1973), "A Mixture-Set Axiomatization of Conditional Subjective Expected Utility", *Econometrica* 41, pp. 1-25.

Frydman, R., O'Driscoll, G. Jr., and Schotter, A. (1982), "Rational Expectations of Government Policy: An Application of Newcomb's Problem", *Southern Economic Journal* 49, pp. 311-319.

Gärdenfors, P., and Sahlin, N.-E. (1988), "Introduction: Bayesian Decision Theory - Foundations and Problems", in P. Gärdenfors, and N.-E. Sahlin (eds.), *Decision, Probability, and Utility: Selected Readings*, Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, and Sydney, pp. 1-15.

Gärdenfors, P., and Sahlin, N.-E. (1988), "Introduction", in P. Gärdenfors, and N.-E. Sahlin (eds.), *Decision, Probability, and Utility: Selected Readings*, Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, and Sydney, pp. 335-339.

Gallois, A. (1979), "How Not to Make a Newcomb Choice", *Analysis* 39, pp. 49-53.

Gallois, A. (1981), "Locke on Causation, Compatibilism and Newcomb's Problem", *Analysis* 41, pp. 42-46.

Gardner, M. (1973), "Free Will Revisited, with a Mind-Bending Prediction Paradox by William Newcomb", *Scientific American* 229, pp. 104-109.

Gauthier, D. (1994), "Assure and Threaten", *Ethics* 104, pp. 690-721.

Gibbard, A., and Harper, W. L. (1978), "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, vol. 1, Reidel, Dordrecht, pp. 125-162.

Gödel, K. (1949), "A Remark about the Relationship between Relativity Theory and Idealistic Philosophy", in P. Schlipp (ed.), *Albert Einstein: Philosopher-Scientist*, Open Court, La Salle, IL, pp. 557-562.

Goldstein, L. (1993), "Inescapable Surprises and Acquirable Intentions", *Analysis* 53, pp. 93-99.

Hardin, R. (1995), "Game Theory", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pp. 292-293.

Hargreaves Heap, S. (1992), "Risk and Uncertainty", in Hargreaves Heap, S., Hollis, M., Lyons, B., Sugden, R., and Weale, A. (1992), *The Theory of Choice: A Critical Guide*, Basil Blackwell, Oxford, UK, and Cambridge, MA, pp. 349-350.

Harper, W. L. (1988), "Introduction", in W. L. Harper, and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, vol. 2, Kluwer, Dordrecht, pp. ix-xix.

Horgan, T. (1981), "Counterfactuals and Newcomb's Problem", *The Journal of Philosophy* 78, pp. 331-356.

Horgan, T. (1985), "Newcomb's Problem: A Stalemate", in R. Campbell, and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, Vancouver, pp. 223-234.

Horne, J. R. (1983), "Newcomb's Problem as a Theistic Problem", *International Journal for Philosophy of Religion* 14, pp. 217-223.

Horwich, P. (1987), *Asymmetries in Time*, MIT Press, Cambridge, MA.

Hubin, D., and Ross, G. (1985), "Newcomb's Perfect Predictor", *Noûs* 19, pp. 439-446.

Hudson, J. L. (1979), "Schlesinger on the Newcomb Problem", *Australasian Journal of Philosophy* 57, pp. 145-156.

Hume, D. (1978), *A Treatise of Human Nature*, L. A. Selby-Bigge (ed.), Clarendon, Oxford, second edition.

Hume, D. (1993), *Eine Untersuchung über den menschlichen Verstand*, translated by R. Richter, J. Kulenkampff (ed.), Meiner, Hamburg.

Hurley, S. L. (1991), "Newcomb's Problem, Prisoners' Dilemma, and Collective Action", *Synthese* 86, pp. 173-196.

Jeffrey, R. C. (1965), *The Logic of Decision*, McGraw-Hill, New York.

Jeffrey, R. C. (1977), "A Note on the Kinematics of Preference", *Erkenntnis* 11, pp. 135-141.

Jeffrey, R. C. (1978), "Axiomatizing the Logic of Decision", in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory*, vol. 1, Reidel, Dordrecht, pp. 227-231.

Jeffrey, R. C. (1981), "The Logic of Decision Defended", *Synthese* 48, pp. 473-492.

Jeffrey, R. C. (1983), *The Logic of Decision*, The University of Chicago Press, Chicago, and London, second edition.

Jeffrey, R. C. (1988), "How to Probabilize a Newcomb Problem", in J. H. Fetzer (ed.), *Probability and Causality*, Reidel, Dordrecht, pp. 241-251.

Jeffrey, R. C. (1996), "Decision Kinematics", in K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (eds.), *The Rational Foundations of Economic Behaviour*, Macmillan, Basingstoke, pp. 3-19.

Jones, E. E., and Nisbett, R. E. (1971), *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*, General Learning Press, New York.

Joyce, J. M. (in press), *The Foundations of Causal Decision Theory*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pre-print.

Kahneman, D. (1996), "New Challenges to the Rationality Assumption", in K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (eds.), *The Rational Foundations of Economic Behaviour*, Macmillan, Basingstoke, pp. 203-219.

Kahneman, D., and Tversky, A. (1988), "Prospect Theory: An Analysis of Decision under Risk", in P. Gärdenfors, and N.-E. Sahlin (eds.), *Decision, Probability, and Utility: Selected Readings*, Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, and Sydney, pp. 183-214, originally 1979 in: *Econometrica* 47, pp. 263-291.

Kamiyama, K. (forthcoming), "Finite Characterizations of Common Knowledge, Barcan Axiom and Reductionism", *Journal of the Japan Association for Philosophy of Science*, originally in Japanese.

Kant, I. (1976), *Prolegomena*, K. Vorländer (ed.), Meiner, Hamburg.

Kant, I. (1990), *Kritik der reinen Vernunft*, R. Schmidt (ed.), Meiner, Hamburg, third edition, reprint of the 1781- and 1787- (original-) editions.

Kaufmann, J. N. (1996), "The Belief-Desire Model of Decision Theory Needs a Third Component: Prospective Intentions", in M. Marion, and R. S. Cohen (eds.), *Québec Studies in the Philosophy of Science*, vol. 2, Kluwer, Dordrecht, pp. 215-227.

Kavka, G. S. (1983), "The Toxin Puzzle", *Analysis* 43, pp. 33-36.

Kennedy, R. (1995), "Dutch Book", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, p. 213.

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan, London.

Keynes, J. M. (1936), *The General Theory of Employment, Interest and Money*, Macmillan, London.

Kiiveri, H., and Speed, T. (1982), "Structural Analysis of Multivariate Data: A Review", in S. Leinhardt (ed.), *Sociological Methodology*, Jossey-Bass, San Francisco.

Knight, F. (1921), *Risk, Uncertainty, and Profit*, Houghton Mifflin, Boston.

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin.

Kusser, A. (1989), *Dimensionen der Kritik von Wünschen*, Athenäum, Frankfurt/M.

Kyburg, H. E. (1974), *The Logical Foundations of Statistical Inference*, Reidel, Dordrecht, and Boston.

Kyburg, H. E. (1980), "Acts and Conditional Probabilities", *Theory and Decision* 12, pp. 149-171.

Kyburg, H. E. (1988), "Powers", in W. L. Harper, and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, vol. 2, Kluwer, Dordrecht, pp. 71-82.

Lambert, K., and Brittan, G. G. (1991), *Eine Einführung in die Wissenschaftsphilosophie*, translated by J. Schulte, de Gruyter, Berlin, and New York.

Ledwig, M. (1994), *Überraschung und Handlung unter Berücksichtigung des Schemakonzeptes*, Diplomarbeit, Universität Bielefeld.

Ledwig, M. (1997), "Liefert Newcombs Problem einen Beweis menschlicher Willensfreiheit?", in P. Weingartner, G. Schurz, and G. Dorn (eds.), *Die Rolle der Pragmatik in der Gegenwartsphilosophie: Beiträge des 20. Internationalen Wittgenstein Symposiums 10. - 16. August, 1997, Kirchberg am Wechsel*, Bd. 2, Die Österreichische Ludwig Wittgenstein Gesellschaft, Kirchberg am Wechsel, pp. 520-526.

Ledwig, M. (1998), "Analytical Philosophy of Science in Quebec", *Metascience* 7, pp. 548-551.

Ledwig, M. (1999a), "Newcomb's Problem - a Gedankenexperiment of Probabilistic Dependence!?", in W. Löffler, and E. Runggaldier (eds.), *Vielfalt und Konvergenz der Philosophie. Vorträge des 5. Kongresses der Österreichischen Gesellschaft für Philosophie*, Bd. 1, Hölder-Pichler-Tempsky, Wien, pp. 94-97.

Ledwig, M. (1999b), "The Rationality of Probabilities for Actions in Decision Theory", *Proceedings of the Twentieth World Congress of Philosophy*, 10th-16th of August, 1998, Boston, MA, on-line at www.bu.edu/wcp/Papers/Acti/ActiLedw.htm.

Ledwig, M. (forthcoming), "Newcomb's Problem and Backwards Causation", in W. Spohn, M. Ledwig, and M. Esfeld (eds.), *Current Issues in Causation*, Mentis, Paderborn.

Lenzen, W. (1997), "Die Newcomb-Paradoxie - und ihre Lösung", in W. Lenzen (ed.), *Das weite Spektrum der analytischen Philosophie: Festschrift für Franz von Kutschera*, de Gruyter, Berlin, and New York, pp. 160-177.

Leslie, J. (1991), "Ensuring Two Bird Deaths with One Throw", *Mind* 100, pp. 73-86.

Levi, I. (1982), "A Note on Newcombmania", *The Journal of Philosophy* 79, pp. 337-342.

Lewis, D. (1969), *Convention: A Philosophical Study*, Harvard University Press, Cambridge, MA.

Lewis, D. (1970), "General Semantics", *Synthese* 22, pp. 18-67.

Lewis, D. (1973), *Counterfactuals*, Basil Blackwell, Oxford.

Lewis, D. (1979a), "Prisoners' Dilemma Is a Newcomb Problem", *Philosophy and Public Affairs* 8, pp. 235-240.

Lewis, D. (1979b), "Attitudes De Dicto and De Se", *Philosophical Review* 88, pp. 513-543.

Lewis, D. (1979c), "Counterfactual Dependence and Time's Arrow", *Noûs* 13, pp. 455-476.

Lewis, D. (1980), "A Subjectivist's Guide to Objective Chance", in R. C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*, vol. 2, University of California Press, Berkeley, and Los Angeles.

Lewis, D. (1981a), "Causal Decision Theory", *Australasian Journal of Philosophy* 59, pp. 5-30.

Lewis, D. (1981b), "`Why Ain'cha Rich?´", *Noûs* 15, pp. 377-380.

Lewis, D. (1986), "Causation", in Lewis, D. (1986), *Philosophical Papers*, vol. 2, Oxford University Press, New York, pp. 159-171, originally 1973 in: *Journal of Philosophy* 70, pp. 556-567.

Locke, D. (1978), "How to Make a Newcomb Choice", *Analysis* 38, pp. 17-23.

Locke, D. (1979), "Causation, Compatibilism and Newcomb's Problem", *Analysis* 39, pp. 210-211.

Luce, R. D., and Krantz, D. H. (1971), "Conditional Expected Utility", *Econometrica* 39, pp. 253-271.

Luce, R. D., and Raiffa, H. (1957), *Games and Decisions*, Wiley, New York, London, and Sydney.

MacCrimmon, K. R., and Larsson, S. (1979), "Utility Theory: Axioms versus 'Paradoxes'", in M. Allais, and O. Hagen (eds.), *Expected Utility Hypotheses and the Allais Paradox*, Reidel, Dordrecht, pp. 333-409.

Mackie, J. L. (1977), "Newcomb's Paradox and the Direction of Causation", *Canadian Journal of Philosophy* 7, pp. 213-225.

Matsuhisa, T., and Kamiyama, K. (1997), "Lattice Structure of Knowledge and Agreeing to Disagree", *Journal of Mathematical Economics* 27, pp. 389-410.

Meek, C., and Glymour, C. (1994), "Conditioning and Intervening", *British Journal for the Philosophy of Science* 45, pp. 1001-1021.

Meggle, G. (1985), "Grundbegriffe der rationalen Handlungstheorie", in G. Meggle (ed.), *Analytische Handlungstheorie, vol. 1: Handlungsbeschreibungen*, Suhrkamp, Frankfurt/M., pp. 415-428.

Mellor, D. H. (1981), *Real Time*, Cambridge University Press, Cambridge.

Mellor, D. H. (1991a), *Matters of Metaphysics*, Cambridge University Press, Cambridge.

Mellor, D. H. (1991b), "Causation and the Direction of Time", *Erkenntnis* 35, pp. 191-203.

Mellor, D. H. (1995), *The Facts of Causation*, Routledge, London.

Nozick, R. (1969), "Newcomb's Problem and Two Principles of Choice", in N. Rescher, D. Davidson, and C. G. Hempel (eds.), *Essays in Honor of Carl G. Hempel*, Reidel, Dordrecht, pp. 114-146.

Nozick, R. (1993), *The Nature of Rationality*, Princeton University Press, Princeton.

Olin, D. (1988), "Predictions, Intentions and the Prisoner's Dilemma", *The Philosophical Quarterly* 38, pp. 111-116.

Otte, R. (1985), "Probabilistic Causality and Simpson's Paradox", *Philosophy of Science* 52, pp. 110-125.

Parfit, D. (1984), *Reasons and Persons*, Clarendon Press, Oxford.

Perry, J. (1979), "The Problem of the Essential Indexical", *Noûs* 13, pp. 3-21.

Pettit, P. (1988), "The Prisoner's Dilemma Is an Unexploitable Newcomb Problem", *Synthese* 76, pp. 123-134.

Pfannkuche, W. (unpublished), *Wieviel Freiheit brauchen wir? Freiheit und Identität*, Habilitationsvortrag, TU-Berlin.

Pothast, U. (1980), *Die Unzulänglichkeit der Freiheitsbeweise: Zu einigen Lehrstücken aus der neueren Geschichte von Philosophie und Recht*, Suhrkamp, Frankfurt/M.

Pratt, J. W. (1964), "Risk Aversion in the Small and in the Large", *Econometrica* 32, pp. 122-136.

Price, H. (1986), "Against Causal Decision Theory", *Synthese* 67, pp. 195-212.

Price, H. (1996), *Time's Arrow and Archimedes' Point*, Oxford University Press, Oxford.

Rabinowicz, W. (1982), "Two Causal Decision Theories: Lewis vs. Sobel", in T. Pauli (ed.), *Philosophical Essays Dedicated to Lennart Aqvist*, Department of Philosophy, Uppsala, pp. 299-321.

Rabinowicz, W. (1985), "Ratificationism without Ratification: Jeffrey Meets Savage", *Theory and Decision* 19, pp. 171-200.

Rabinowicz, W. (1988), "Ratifiability and Stability", in P. Gärdenfors, and N.-E. Sahlin (eds.), *Decision, Probability, and Utility: Selected Readings*, Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, and Sydney, pp. 406-425.

Ramsey, F. P. (1931), "Truth and Probability", in R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Logical Essays*, Routledge, and Kegan Paul, London, pp. 156-198.

Ramsey, F. P. (1990), *Philosophical Papers*, D. H. Mellor (ed.), Cambridge University Press, Cambridge.

Rapoport, A. (1975), "Comment on Brams's Discussion of Newcomb's Paradox", *Journal of Conflict Resolution* 19, pp. 613-619.

Rapoport, A., and Chammah, A. M. (1965), *Prisoner's Dilemma - A Study in Conflict and Cooperation*, University of Michigan, Ann Arbor.

Reichenbach, H. (1956), *The Direction of Time*, University of California Press, Berkeley, Los Angeles, and London.

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

Salmon, W. C. (1978), "Why Ask 'Why'? - An Inquiry Concerning Scientific Explanation", *Proceedings and Addresses of the American Philosophical Association* 51, pp. 683-705.

Salmon, W. C. (1984), *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton.

Savage, L. J. (1954/1972), *The Foundations of Statistics*, Wiley, New York, Dover.

Schlesinger, G. (1974), "The Unpredictability of Free Choices", *British Journal for the Philosophy of Science* 25, pp. 209-221.

Schlesinger, G. (1976), "Unpredictability: A Reply to Cargile and to Benditt and Ross", *British Journal for the Philosophy of Science* 27, pp. 267-274.

Schlesinger, G. (1977a), "You Can Always Depend on the Perfect Judge", *Australasian Journal of Philosophy* 55, pp. 136-138.

Schlesinger, G. (1977b), *Religion and Scientific Method*, Reidel, Dordrecht.

Schmidt, C. (1996), "Comment", in K. J. Arrow, E. Colombatto, M. Perlman, and C. Schmidt (eds.), *The Rational Foundations of Economic Behaviour*, Macmillan, Basingstoke, pp. 20-23.

Schmidt, J. H. (1998), "Newcomb's Paradox Realized with Backward Causation", *British Journal for the Philosophy of Science* 49, pp. 67-87.

Schramm, A. (unpublished), *Ostrich, Newcomb, Weather Forecast*, private manuscript.

Segerberg, K. (1993), "Perspectives on Decisions", *Aristotelian Society Proceedings* 93, pp. 263-278.

Selten, R., and Holtz-Wooders, M. (1995), *Cyclic Games: An Introduction and Examples*, discussion paper B334, Universität Bonn, Sonderforschungsbereich 303.

Shafir, E., and Tversky, A. (1992), "Thinking through Uncertainty: Nonconsequential Reasoning and Choice", *Cognitive Psychology* 24, pp. 449-474.

Sklar, L. (1995), "Philosophy of Science", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pp. 611-615.

Skyrms, B. (1980), *Causal Necessity*, Yale University Press, New Haven, and London.

Skyrms, B. (1982), "Causal Decision Theory", *Journal of Philosophy* 79, pp. 695-711.

Skyrms, B. (1984), *Pragmatics and Empiricism*, Yale University Press, New Haven, and London.

Skyrms, B. (1985), "Discussion: Ultimate and Proximate Consequences in Causal Decision Theory", *Philosophy of Science* 52, pp. 608-611.

Skyrms, B. (1990a), *The Dynamics of Rational Deliberation*, Harvard University Press, Cambridge, MA, and London, UK.

Skyrms, B. (1990b), "Ratifiability and the Logic of Decision", in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.), *Midwest Studies in Philosophy, vol. XV: Philosophy of the Human Sciences*, University of Notre Dame Press, Notre Dame, IND, pp. 44-56.

Skyrms, B. (1992), "Theories of Probability", in J. Dancy, and E. Sosa (eds.), *A Companion to Epistemology*, Basil Blackwell, Cambridge, MA, pp. 374-378.

Slote, M. A. (1995), "Satisfice", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pp. 712-713.

Snow, P. (1985), "The Value of Information in Newcomb's Problem and the Prisoners' Dilemma", *Theory and Decision* 18, pp. 129-133.

Sobel, J. H. (unpublished), *Probability, Chance and Choice: A Theory of Rational Agency*, presented in part at a workshop on Pragmatics and Conditionals at the University of Western Ontario in May 1978.

Sobel, J. H. (1985a), "Not every Prisoner's Dilemma Is a Newcomb Problem", in R. Campbell, and L. Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, The University of British Columbia Press, Vancouver, pp. 263-274.

Sobel, J. H. (1985b), "Circumstances and Dominance in Causal Decision Theory", *Synthese* 63, pp. 167-202.

Sobel, J. H. (1986), "Notes on Decision Theory: Old Wine in New Bottles", *Australasian Journal of Philosophy* 64, pp. 407-437.

Sobel, J. H. (1988a), "Infallible Predictors", *The Philosophical Review* 97, pp. 3-24.

Sobel, J. H. (1988b), "Defenses and Conservative Revisions of Evidential Decision Theories: Metatickles and Ratificationism", *Synthese* 75, pp. 107-131.

Sobel, J. H. (1988c), "Metatickles, Ratificationism, and Newcomb-Like Problems without Dominance", in B. R. Munier (ed.), *Risk, Decision and Rationality*, Reidel, Dordrecht, pp. 483-501.

Sobel, J. H. (1989), "Partition Theorems for Causal Decision Theories", *Philosophy of Science* 56, pp. 71-93.

Sobel, J. H. (1990), "Newcomblike Problems", in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.), *Midwest Studies in Philosophy, vol. XV: Philosophy of the Human Sciences*, University of Notre Dame Press, Notre Dame, IND, pp. 224-255.

Sobel, J. H. (1991), "Some Versions of Newcomb's Problem are Prisoners' Dilemmas", *Synthese* 86, pp. 197-208.

Sobel, J. H. (1994), *Taking Chances: Essays on Rational Choice*, Cambridge University Press, Cambridge, UK, NewYork, NY, and Melbourne.

Sober, E. (1988), "The Principle of the Common Cause", in J. H. Fetzer (ed.), *Probability and Causality*, Reidel, Dordrecht, pp. 211-228.

Sorensen, R. A. (1985), "The Iterated Versions of Newcomb's Problem and the Prisoner's Dilemma", *Synthese* 63, pp. 157-166.

Sorensen, R. A. (1986), "A Strengthened Prediction Paradox", *The Philosophical Quarterly* 36, pp. 504-513.

Spirtes, P., Glymour, C., and Scheines, R. (1993), *Causation, Prediction and Search*, Springer, New York.

Spohn, W. (1977), "Where Luce and Krantz Do Really Generalize Savage's Decision Model", *Erkenntnis* 11, pp. 113-134.

Spohn, W. (1978), Grundlagen der Entscheidungstheorie, in *Monographien Wissenschaftstheorie und Grundlagenforschung*, No. 8, Scriptor, Kronberg/Ts.

Spohn, W. (1982, 1994), "How to Make Sense of Game Theory", in W. Stegmüller, W. Balzer, and W. Spohn (eds.), *Philosophy of Economics*, Springer, Berlin, pp. 239-270, German revision: "Wie läßt sich die Spieltheorie verstehen?", in J. Nida-Rümelin (ed.), *Praktische Rationalität: Grundlagenprobleme und ethische Anwendungen des rational choice-Paradigmas*, de Gruyter, Berlin, and New York, pp. 197-237.

Spohn, W. (1983), *Eine Theorie der Kausalität*, Habilitationsschrift, Ludwig-Maximilians-Universität München.

Spohn, W. (1986), "The Representation of Popper Measures", *Topoi* 5, pp. 69-74.

Spohn, W. (1988), "Ordinal Conditional Functions. A Dynamic Theory of Epistemic States", in W. L. Harper, and B. Skyrms (eds.), *Causation in Decision, Belief Change, and Statistics*, vol. 2, Kluwer, Dordrecht, pp. 105-134.

Spohn, W. (1993), "Wie kann die Theorie der Rationalität normativ und empirisch zugleich sein?", in L. H. Eckensberger, and U. Gähde (eds.), *Ethische Norm und empirische Hypothese*, Suhrkamp, Frankfurt/M., pp. 151-196.

Spohn, W. (1997), "Über die Gegenstände des Glaubens", in G. Meggle (ed.), *Analyomen 2: Proceedings of the 2nd Conference: Perspectives in Analytical Philosophy, vol. 1: Logic, Epistemology, Philosophy of Science*, de Gruyter, Berlin, and New York, pp. 291-321.

Spohn, W. (1999), *Strategic Rationality*, Forschungsbericht der DFG-Forschergruppe "Logik in der Philosophie", University of Constance.

Spohn, W. (in press), "A Rationalization of Cooperation in the Iterated Prisoner's Dilemma", in J. Nida-Rümelin, and W. Spohn (eds.), *Practical Rationality, Rules, and Structure*, Kluwer, Dordrecht.

Stalnaker, R. (1968), "A Theory of Conditionals", in *Studies in Logical Theory*, American Philosophical Quarterly Monograph Series, No. 2.

Stalnaker, R. (1972/1981), "A Letter to David Lewis", in W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*, Reidel, Dordrecht, pp. 151-152.

Stalnaker, R., and Thomason, R. (1970), "A Semantic Analysis of Conditional Logic", *Theoria* 36, pp. 23-42.

Storms, M. D. (1973), "Videotape and the Attribution Process: Reversing Actors' and Observers' Point of View, *Journal of Personality and Social Psychology* 27, pp. 165-175.

Suppe, F. (1989), *The Semantic Conception of Theories and Scientific Realism*, University of Illinois Press, Urbana.

Suppes, P. (1970), *A Probabilistic Theory of Causality*, North Holland Publishing Company, Amsterdam.

Suppes, P. (1974), *Probabilistic Metaphysics*, University of Uppsala Press, Uppsala.

Swain, C. G. (1988), "Cutting a Gordian Knot: The Solution to Newcomb's Problem", *Philosophical Studies* 53, pp. 391-409.

Taylor, M. (1976), *Anarchy and Cooperation*, Wiley, New York.

Voizard, A. (1996), "'If Cows Had Wings, We'd Carry Big Umbrellas.' An Almost Number-Free Note on Newcomb's Problem", in M. Marion, and R. S. Cohen (eds.), *Québec Studies in the Philosophy of Science*, vol. 2, Kluwer, Dordrecht, pp. 193-213.

Wagner, C. G. (1991), "Simpson's Paradox and the Fisher-Newcomb Problem", *Grazer Philosophische Studien* 40, pp. 185-194.

Wagner, S. J. (1995), "Proposition", in R. Audi (ed.), *The Cambridge Dictionary of Philosophy*, Cambridge University Press, Cambridge, UK, New York, NY, Melbourne, pp. 658-659.

Weintraub, R. (1995), "Psychological Determinism and Rationality", *Erkenntnis* 43, pp. 67-79.