Minds & Machines (2009) 19:61–92
DOI 10.1007/s11023-008-9126-2

Decision Theory, Intelligent Planning and Counterfactuals

Michael John Shaffer

Received: 9 February 2007 / Accepted: 4 November 2008 / Published online: 21 November 2008

Abstract The ontology of decision theory has been subject to considerable debate in the past, and discussion of just how we ought to view decision problems has revealed more than one interesting problem, as well as suggested some novel modifications of classical decision theory. In this paper it will be argued that Bayesian, or evidential, decision-theoretic characterizations of decision situations fail to adequately account for knowledge concerning the causal connections between acts, states, and outcomes in decision situations, and so they are incomplete. Second, it will be argues that when we attempt to incorporate the knowledge of such causal connections into Bayesian decision theory, a substantial technical problem arises for which there is no currently available solution that does not suffer from some damning objection or other. From a broader perspective, this then throws into question the use of decision theory as a model of human or machine planning.

Keywords Artificial intelligence · Bayesianism · Causality · Conditionals · Counterfactuals · Decision theory · Deliberation · Probabilities · Rationality · Utility · Planning

Introduction

The ontology of decision theory has been subject to considerable debate in the past, and discussion of just how we ought to view decision problems has revealed more than one interesting problem, as well as suggested some novel modifications of

M. J. Shaffer (✉)
Department of Philosophy (CH 365), St. Cloud State University, 720 4th Ave. South, St. Cloud, MN 56301, USA
e-mail: shaffermphil@hotmail.com

classical decision theory.[1] In this paper it will be argued that Bayesian, or evidential, decision-theoretic characterizations of decision situations fail to adequately account for knowledge concerning the causal connections between acts, states, and outcomes in decision situations, and so they are incomplete. Second, it will be argued that when we attempt to incorporate the knowledge of such causal connections into Bayesian decision theory, a substantial technical problem arises for which there is no currently available solution that does not suffer from some damning objection or other.[2]

From a broader perspective, this then throws into question the use of standard forms of decision theory as a model of human or machine planning. In the context of the model of human planning, decision theory serves both descriptive and normative functions. The philosophically motivated treatment in Resnik (1987), for example, makes this especially clear and the role of decision theory with respect to this context is to both model and guide human decision-making. In effect, decision theory so understood is intended to specify the rules by which human decision should proceed. What such theories purport to do is to formalize the process of deliberation that is supposed to occur priori to acting rationally. In the context of planning in machine intelligence, decision theory has been employed as a formal system for the construction of artificial agents that are guided by the normative rules constitutive of the theory of decisions and the overview provided in Todd (1999) illustrates this point quite well. The A.I. application of decision theory has had enormous influence in planning theory, the design of expert systems, the design of computer vision and, more generally, in the design of artificial agents. As a matter of fact, the decision-theoretic approach to A.I. is one of the most basic and fertile approaches to the host of theoretical problems encountered in the attempt to design artificially intelligent systems as, for example, both Russell and Norvig's (1995) and Korb and Nicholson's (2004) indicate. These sorts of attempts to model theoretical human planning and its application to acting also, of course, have considerable influence on more practical A.I. applications involved in getting A.I. agents to perform normatively correct actions in various specified environments. Russell and Norvig (1995, Chap. 13) in particular are especially clear that the theoretical foundation that decision theory is supposed to provide for this approach to A.I. is absolutely crucial for the resolving the practical issue of plan *execution*. But this should be no surprise to philosophers, for as far back as Aristotle's *Nicomachean Ethics* it has been assumed that practical rationality—acting rationally—presupposes theoretical rationality. In effect these classical approaches to A.I. are predicated on the idea that rational deliberation precedes action.

[1] See Gärdenfors and Sahlin (1988) for a good overview. Jeffrey (1965) contains the most well known revision of the classical view as presented in Savage (1954/1972). Levi (1992) examines the distinction between these two views of the ontology of decision theory.

[2] This issue should not be generally confused with the more specific problem concerning the effects that an agent's acts have on the outcomes of a decision that gave rise to what is called causal decision theory, although they are related issues. These more specific problems arise in virtue of issues like Newcomb's problem, and extensive discussion of this issue can be found in the articles reprinted in Gärdenfors and Sahlin (1988), part V, and, for example, in Eells (1981) and Levi (1975).

In any case, the first problem noted above, the problem concerning the lack of causal knowledge on the part of the agent, involves how we ought to view decision problems and what kinds of things a decision-maker must be taken to know if that agent can be said to be acting rationally. The suggestion is that a rational decision-maker must be taken to have some information concerning the causal laws that obtain in her world in order for her to act rationally. One view that challenges this assumption will be examined and rejected in a later section and so a critical conclusion will be drawn concerning Bayesian decision theory, as well as other fairly standard versions of decision theory.

In order for the desideratum concerning causal knowledge and the completeness of decision matricies to be satisfied, a decision-maker must be able to assign a well-defined probability to statements of the form 'If β were to do act a_n and the world is in state s_p, then it would result in outcome o_m' for each act, state, and possible outcome triple in a decision matrix. The need to be able to do this coherently has been recognized by Gibbard and Harper (1981), Stalnaker (1996), and Joyce (1999), among others, and so this issue is not especially breaking news. The theories that have been proposed as successors to Bayesian decision theory, versions of causal decision theory, attempt to do just this. Versions of causal decision theory explicitly incorporate these sorts of statements, what we might call *deliberative counterfactuals*, and they are simply future directed counterfactuals about what would happen if the agent in question were to do each act and the world turned out to be in a particular state. Such statements are the very essence of deliberative prediction about what we might do, and which claims of these sorts turn out to be true depends fundamentally on what causal laws operate in the world the agent inhabits.

As we shall see then, in accord with causal decision theory, a fully complete, maximally specified, decision matrix must incorporate knowledge of states of the world, including outcomes, but it must also incorporate knowledge of the causal dynamics of those states and their connections as they relate to the acts that the agent chooses to perform.[3] What we will then discover is that due to David Lewis' celebrated triviality results concerning the probabilities of conditionals, the most plausible suggestion for interpreting how such probabilities are to be defined leads to absurdity.[4] Moreover, the extant alternative proposals for doing so also all appear to fail to account for the probabilities of deliberative counterfactuals for one reason or another, and so it seems that there is no known, coherent, way to assign such probabilities to such statements. As a result a critical conclusion will be drawn concerning extant formulations of causal (i.e. non-Bayesian or evidential) decision theory. Hence, we must conclude that, as there is a failure in Bayesian decision theory with respect to the inclusion of causal knowledge and there is no extant coherent account of the probabilities of deliberative counterfactuals, the foundations of both Bayesian decision theory *and* causal decision theory rest on highly dubious bases.

[3] This issue is, in many respects, reminiscent of the notorious frame problem in artificial intelligence. See Pylyshyn, Shoham (1988) for perspectives on this problem. See Pollock (2002a) for a specific connection between the frame problem and deliberation.

[4] See Lewis (1976) and Lewis (1986).

This minimally suggests that currently both descriptive and normative decision theory have only questionable application to the issues of human and machine intelligence.[5] Of course this overall negative conclusion about the relevance of decision theory to substantive issues in human and machine intelligence will only hold *strongly* provided both that the argument concerning the incompleteness of classically described decision situations that will be elaborated below is unsolvable and that the problem of assigning probabilities to deliberative counterfactuals is really unsolvable. This may not, of course, turn out to be true. Nevertheless, in light of the failures of both evidential and causal decision theory to adequately model rational deliberation, the significance of decision-theoretic rationality will be addressed from the perspective of idealization in the context of issues in human and machine intelligence. What will ultimately be suggested here is that what the failures of both causal and evidential decision theory indicate is *not* that they are simply idealized models of rational behavior, but rather that they are simply false or inadequate models of deliberation and planning. As a result, practitioners in fields dealing with human and machine intelligence would do better to look at alternative non-decision-theoretic approaches to rational deliberation if they are to solve substantive problems in those fields.

Basic Decision Theory

Before addressing the two main concerns of this paper it will be useful to introduce the basic concepts that are the central fixtures of decision theory in its classical guise. Let us begin by laying out an example of a typical decision problem, and then it will be argued that, as classically construed, all such decision problems are descriptively incomplete. Presentation will follow Gärdenfors and Sahlin's (1988) treatment of classical decision theory as originally introduced by Leonard J. Savage in his (1954/1972).

The basic entity with which decision theory is concerned is a decision situation. A decision situation is composed of a set of acts or alternatives A where the individual acts will be denoted a_i with i ranging over the natural numbers, ℵ. In addition, in order to fully specify a classical decision matrix we must include the relevant states of the world S whose members will be denoted as s_j with j ranging over the natural numbers, ℵ. As Gärdenfors and Sahlin indicate, the incorporation of the states of the world is used to reflect our uncertainty about the consequences of our acts.[6] Associated with each act a_i is an outcome o_{ij} that depends on the state of the world s_j, where $o_{ij} \in O$, the set of all outcomes open in that situation. So, as Gärdenfors and Sahlin assert, "a decision situation is thus determined by sets of alternatives, states, and outcomes (1988, p. 2)." In other words, a decision situation is maximally specific according to classical decision theory if and only if it includes a complete specification of the acts, states and outcomes in that situation.

[5] See Horovitz et al. (1988) and Russell and Norvig (1995) for representative discussion of the role of decision theory in artificial intelligence.

[6] Gärdenfors and Sahlin (1988, pp. 2–6).

Schematically, a decision situation can be represented by a decision matrix. As a result, a decision matrix lays out the possible alternatives, states, and outcomes that constitute that decision situation in a convenient manner. So, in terms of classical decision theory, a decision matrix offers a convenient presentation of what classical decision theorists take to be a complete description of a decision situation; i.e. it contains all relevant factors needed to account for decision-making.

Consider the following decision situation as described in Savage (1954/1972):

> Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must either be used for the omelet or wasted altogether, lies unbroken beside the bowl. You must decide what to do with this unbroken egg. Perhaps it is not too great an oversimplification to say that you must decide among three acts only, namely, to break it into the bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection. Depending on the state of the egg, each of these three acts will have some consequences of concern to you… (pp. 13–14).

The decision matrix for this decision situation (with respect to some world w_j) can be represented quite easily as follows and denoted as d_1:

Act	State	
	Good (s_1)	Bad (s_2)
Break egg into bowl (a_1)	Six-egg omelet (o_{11})	No omelet; five good eggs destroyed (o_{21})
Break egg into saucer (a_2)	Six-egg omelet and a saucer to wash (o_{12})	Five-egg omelet and a saucer to wash (o_{22})
Throw egg away (a_3)	Five-egg omelet and a good egg destroyed (o_{13})	Five-egg omelet (o_{23})

This sort of schematic representation of a decision situation allows one to examine what outcomes will be caused by which acts if a given state of the world obtains, and so it can be used as a guide when an agent deliberates about this situation should she find herself in a relevantly similar real situation.

Given these fundamental concepts one can, of course, then begin to evaluate decision situations in terms of various concepts of rationality. Classical decision theory, again as presented by Gärdenfors and Sahlin, bases these sorts of evaluations on three basic assumptions:

A1. (*Values of outcomes*): The values of the outcomes in a decision situation are determined by a utility measure which assigns numerical values to the outcomes (1988, p. 3).

A2. (*Values of alternatives*): When determining the value of a decision alternative, the *only* information about the decision maker's wants and desires that is exploited is the utilities of the possible outcomes of the alternative (1988, p. 4).

A3. (*Information about states*): A decision maker's beliefs about the states of the world in a given situation can be represented by a unique probability measure defined over the states (1988, p. 4).

Most crucially, following A1, decision theorists introduce a measure of utility defined on O, such that in a given decision situation d_k each outcome o_{nm}, has a well-defined utility denoted as $u(o_{nm})$. This is a measure of the agent's preferences, her preference ordering and preference magnitudes, defined on the possible outcomes in a decision situation. Similarly, following A3, in order to reflect the decision-maker's ignorance about the states of the world, decision theorists introduce a unique probability measure on S, such that in a given decision situation d_k with states s_n the decision maker assigns a probability $P(s_j)$ for each $s_j \in s_n$ in accord with the probability calculus.[7] In incorporating A3 into the presentation of decision theory and interpreting the probability distribution over S as a measure of subjective—or creedal—probabilities we get what is typically referred to as Bayesian decision theory.

With this basic theory in hand one can then, of course, ask of any decision situation d_k what is the rational thing to do in that decision situation? As Gärdenfors and Sahlin explain, the fundamental answer to this question in terms of Bayesian decision theory is encapsulated in the principle referred to as *the principle of maximizing expected utility*. This principle is defined as follows:

P1 (*Maximizing expected utility*): In a given decision situation the decision maker should choose the alternative with the maximal expected utility (or one of the alternatives with maximal expected utility if there are more than one) (1988, p. 5).

The expected utility of a given act is, of course, defined as follows:

$$(\mathrm{EU})E(a_i) = P(s_1) \times u(o_{i1}) + P(s_2) \times u(o_{i2}) + \cdots + P(s_q) \times u(o_{iq}).$$

Notice that, when joined with A1–A3, what P1 tells us is that there is at least one act that is the act that the decision-maker in a decision situation should make. There may be more than one act that maximizes utility according to P1, but there will be at least one that is maximal in this sense. So, the addition of P1 to A1–A3 yields what we might call a normative extension of classical descriptive decision theory (A1–A3).

With respect to d_1 we might find that an agent has assigned utilities to the outcomes and probabilities to the states as follows: $u(o_{11}) = 6$, $u(o_{12}) = 4$, $u(o_{13}) = 2$, $u(o_{21}) = 1$, $u(o_{22}) = 3$, $u(o_{23}) = 5$, $P(s_1) = .5$, $P(s_2) = .5$. So, for example, the expected utility of a_1, i.e. of breaking the egg into the bowl, given these assignments is 3.5. The expected utilities of the other acts are as follows: $E(a_2) = 3.5$ and $E(a_3) = 3.5$. So in this case, our rule P1 would tell us that we are allowed to do any one of a_1, a_2, or a_3 as they are all utility maximizing acts.

Given this understanding of the Bayesian interpretation of classical decision theory in hand it will now be argued that no decision situation as defined by A1–A3

[7] The probabilities employed in decision theory are typically given a subjectivist interpretation. For considerations of other interpretations of the probability calculus see Good (1983) and Howson (1995).

is really descriptively complete such that A1 and A2 can be satisfied in a way that would allow a decision maker to obey P1 and still be considered to be rational. In other words, it will be argued that a decision-maker would need to know more than the facts, the acts open to the agent, the probabilities of the states of the world and the utilities of the outcomes in a decision situation if she were to be considered rational in the true sense of maximizing expected utility.

A Problem for Bayesian Decision Theory

Presumably, the main impetus behind the creation of decision theory was to provide a model of rational decisions in such a way that it could be employed by cognitive agents as a normative guide concerning how they ought to actually make decisions and potentially as a framework for the construction of artificial agents. What other purposes it could have? From the perspective of an actual decision situation faced by a real decision-maker what we hope is that the real situation is at least roughly isomorphic to a well-defined decision matrix and that by constructing a decision matrix the real decision-maker or the artificial agent can identify what would be the rational thing to do in that actual situation. Decision matrices are intended to serve as useful tools for deliberating, in an hypothetical way, about future actions that may subsequently be taken by these sorts of cognitive agents.

In order to make a rational decision about what to do the real decision-maker or programmer must, at least, be able to provide all the information that characterizes decision situations as defined above. This is especially true if we accept a normative version of Bayesian, or evidential, decision theory, as it seems obvious that epistemic oughts imply epistemic cans. If we cannot provide that information, or any other relevant information, needed for rational decision-making, then we cannot be required to act in accord with the prescriptions that depend on that information. In this section it will be argued that all such specifications are incomplete in a crucially important respect; i.e. there is information relevant to real decision situations that is absent in the classical and Bayesian accounts of decision situations. It will further be argued that the most obvious manner in which they could be made complete raises a deeply serious technical problem for Bayesian decision theorists for which there appear to be no extant solutions. So, we should conclude that rational agents could not be obligated to follow Bayesian decision theory's prescriptions.

The sense in which decision-theoretic descriptions of decision situations are incomplete involves the causal connection described by some law L or some singular causal statement l, between the acts, a_n, in a given decision situation, the relevant states of the world s_p, and the outcomes o_m, in that decision situation. In these sorts of situations decision theorists have simply assumed that the decision-maker or programmer is aware of the actual causal connections between the individual possible acts available to the agent and the particular possible outcomes that would result, depending on the relevant state of the world. Classical decision theorists and, in particular, Savage, tried to avoid this problem by *equating acts with their possible consequences*; more precisely as functions from states to consequences. In his 1954/1972 Savage claims that,

> If two different acts had the same consequences in every state of the world, there would from the present point of view be no point in considering them two different acts at all. An **act** may therefore be identified with its possible consequences. Or, more formally, an act is a function, attaching the consequence $f(s)$ to the state s (p. 14).

This attempt to identify acts with their consequences is, at best, a metaphysically dubious move that allowed classical decision theorists to ignore or perhaps just miss the important role that causal connections—and beliefs about causal connections—play in decision situations. In point of fact, if we were to accept Savage's suggestion concerning the identification of acts with functions from states outcomes and we recognize our general ignorance of causal laws and of the future, then, in accord with good Humean sense about causality, all acts would appear to be identical as any given act can be potentially identified with any outcome. Cracking open the egg and dropping it in the bowl is perfectly compatible with an infinite number of possible outcomes before the actual occurrence of such an outcome. Cracking open the egg may ruin all the eggs, it may turn the eggs into gold, the eggs might fly away, etc. The same will be true of every act ever available to an agent. What the agent must apparently know if she is to be able to assign probabilities to her predictions concerning the outcome of her acts are the causal regularities operating in the decision situation she faces; i.e. what world she inhabits and what its laws are. Otherwise, this agent will have no reason to believe anything in particular about the outcome of any act open to her.

In any case, whatever we might say metaphysically about the nature of causality, the agent's knowledge of what the possible consequences of an act are and the laws governing such connections cannot merely be assumed. It would entail that in any decision situation the agent is aware of all relevant causal laws operative in that decision situation, and the agent may well be ignorant of, or wrong about, these crucial facts. This may be true in a particular case, i.e. within a decision situation d_k, and so the agent may be wrong about which act leads to which outcome in that situation, or it may be true in general of the agent and so the agent may be confused about such causal connections across many cases. Classical decision theorists simply assume that in every case the agent is cognizant of all such connections. But, the failure to represent the agent's ignorance concerning the causal laws, concerning the functions $f(s)$, operative in the possible world that the agent inhabits leads to obvious cases where a Bayesian decision-maker can satisfy A1 and A2, and can follow P1 by applying (EU), but Bayesian decision theory will end up yielding radically different results concerning what one ought to do based on the facts of the matter concerning the causal connections between acts, states, and outcomes in that situation. So, it would seem that if one *really* wants to maximize expected utility, then one must know something about these causal connections. But, as we have seen, classical decision theory says nothing about the agent's knowledge of these important features of the world that can make a radical difference in our calculations of expected utility.

Consider again Savage's original decision situation concerning the eggs noted above. In his presentation of decision situations it is simply presupposed that a

particular act and world-state combination will give rise to a particular outcome. So, classical and Bayesian decision theory simply assumes on the part of the agent that, for example, if she were to break the sixth egg into the bowl and if, in fact, the egg is rotten, then she will not get an omelet and she will have ruined five eggs. But why should we simply assume that the agent knows that this will be the outcome of her act if the egg is rotten? Is it not, at least, logically possible that she lives in a world in which admittedly odd causal laws hold such that the result of such an act when the egg is rotten will be a six-egg omelet and a saucer to wash? Is it not, at least, logically possible that the agent is just wrong and that if she were to break the sixth egg into the bowl and if, in fact, the egg is rotten, then she will get a five-egg omelet? The former alternative is not even mentioned in the decision matrix proposed by Savage, and the latter possibility is dismissed without comment. Surely this crucial knowledge concerning the connections between acts, states, and outcomes is relevant to what the agent might decide to do, but it cannot just be assumed on the part of the agent. It is, in point of fact, safe to say that no actual agent possesses anything like complete knowledge of this kind of information. But, the agent's knowledge or ignorance of the laws of nature must be accounted for in a decision situation. It would be highly unrealistic to assume that agent's are in possession of such knowledge, and one might be begin to wonder how Bayesian decision theorists can regard their prescriptions as applicable to real agents if such knowledge is just assumed on the part of all agents.

Consider the following modification of Savage's decision situation. Assume, as before, that you arrive home and find five eggs broken into a bowl and that there is a sixth egg as well which must be used or thrown away. But, in your world, w_i, the following two causal laws turned out to actually hold true:

(CL1) When a rotten egg is broken and added to a bowl of eggs with five eggs in it, you get a good six-egg omelet.
(CL2) When six good eggs are placed in a bowl the omelet is ruined and all the eggs are destroyed.

In light of these laws, the decision matrix, d_2, for this decision situation would be as follows:

Act	State	
	Good (s_1)	Bad (s_2)
Break egg into bowl (a_1)	Six eggs destroyed (o^*_{11})	Six egg omelet (o^*_{21})
Break egg into saucer (a_2)	Six eggs destroyed and a saucer to wash (o^*_{12})	Six-egg omelet and a saucer to wash (o^*_{22})
Throw egg away (a_3)	Five-egg omelet and a good egg destroyed (o^*_{13})	Five-egg omelet (o^*_{23})

In this world, the results of applying (EU) to our decision situation will, of course, yield radically different results because some of the outcomes are different. Our agent might assign the following utilities and probabilities to outcomes and

states, respectively, as follows: $u(o^*_{11}) = 2$, $u(o^*_{12}) = 1$, $u(o^*_{13}) = 3$, $u(o^*_{21}) = 6$, $u(o^*_{22}) = 4$, $u(o^*_{23}) = 5$, $P(s_1) = .5$, and $P(s_2) = .5$. So, the expected utilities are $E(a_1) = 3.5$, $E(a_2) = 2.5$, and $E(a_3) = 4$. As such, P1 tells us that if we are in d_2, then we should do a_3.

But, the reason why classical and Bayesian descriptions of decision situation are incomplete is that the decision-maker, typically, will not really know which world she inhabits, w_i or w_j, let alone the vast number of other possible worlds we might inhabit. Hence, she will not know if the causal connections in d_1 or those in d_2 obtain, and so she will not know whether her breaking the egg into the bowl when the egg is good will bring about the creation of a six-egg omelet or the ruining of six-eggs. In addition, even within the restricted context of d_1, how is it that the agent knows that breaking the egg into the bowl will have the consequence of ruining the eggs if it is, in fact rotten, rather than that doing so will cause a five-egg omelet to be brought into existence? This (possible) lack of knowledge, which clearly makes a difference in terms of the choices we should make, needs to be included in the specification of a decision situation if it is to be considered complete. So, we cannot simply assume that the decision-maker knows for certain what causal connections are truly operative in a given decision situation. Fortunately, Bayesians, following, for example Howson and Urbach (1993), might think that they have a ready answer. As Bayesians see it our knowledge of causal laws is also a matter of partial belief, and so we might be able to assign some probabilities to the relevant expressions concerning the causal connections between possible acts, states, and outcomes in a given decision situation (i.e. to the various logically possible statements of the form $f(s)$), thus making the description of a decision situation realistically complete by allowing the agent to derive probabilities for the future directed deliberative counterfactuals involved in the decision situation the agent faces.[8]

Jeffery's Digestion of Savage's Omelet

In his 1977, Richard Jeffrey proposed a solution to the problem of the completeness of decision matrices that addresses Lewis' problem, but which 'solves' it, at least apparently, by bypassing the entire issue of including causes in the description of decision situations. This is accomplished by revising what we take states to be so as vitiate the need for an agent to possess the relevant causal knowledge. As a result, it is supposed to bypass the issue of the probabilities of conditionals that arises in completing specifications of decision situations. The approach that Jeffrey advocates stems from comments made by Savage himself in his 1954/1972 and Jeffrey's point appears to be that Savage's characterizations of decision situations

[8] Shaffer (2001) provides some reasons to suspect that this is not really the case, and that Bayesian confirmation theory suffers from some serious technical problems similar to those raised here in the context of Bayesian decision theory. Other criticisms of Bayesian confirmation theory can be found in Brown (1994), Glymour (1981), Kyburg (1978), and Salmon (1990). In any case, one might be tempted to think that if one knows the relevant causal laws operating in a decision situation, then it is a trivial fact that since laws are supposed to support counterfactuals, one will be able to assign probabilities to such counterfactuals. This is by no means the case, as we shall see. It is, in point of fact, debatable whether there is any coherent sense of the probability of a counterfactual conditional.

are already complete, if only implicitly so. Things are, of course, a bit more complex, but this is the gist of Jeffrey's point. Let us examine his proposal in more detail.

Jeffery appeals to a following passage from Savage as the source for his dissolution of the problem of causal knowledge and counterfactuals in Bayesian decision theory:

> The argument has been raised that the formal description of decision that has thus been erected seems inadequate because a person may not know the consequences of the acts open to him in each state of the world. He might be so ignorant, for example, as to not be sure whether one rotten egg will spoil a six-egg omelet. But in that case nothing could be simpler than to admit that there are four states in the world corresponding to the two states of the egg and the two conceivable answers to the culinary question whether one bad egg will spoil a six-egg omelet. It seems to me obvious that this solution works in the greatest generality (Savage 1954/1972, p. 12).

Jeffrey (1977, p. 362) then notes that decision problems in Savage's framework must exhibit the following two features:

> (*Determinism*) Each act-state pair determines a unique consequence.
> (*Independence*) The probabilities of the states are the same, no matter which act is performed.

The worry Savage considers seems to be that agents may face decisions in which determinism and/or independence fail. To make things clearer let us consider the following simple decision matrix k_1 based on Jeffrey 1977:

Act	State	
	Live at least to age 65 (s_1)	Die before age 65 (s_2)
Smoke 2 packs of cigarettes a day (a_1)	Smoke and live (o_{11})	Smoke and die (o_{21})
Quit smoking (a_2)	Quit and live (o_{12})	Quit and die (o_{22})

Let us further simplify by substituting the following symbols for the expressions in in k_1: S = Smoke 2 packs of cigarettes a day; Q = Quit smoking altogether; L = Live at least to age 65; and D = Die before age 65. We then have the following less cumbersome and more generic decision matrix:

Act	State	
	L(s_1)	D(s_2)
S(a_1)	SL(o_{11})	SD(o_{21})
Q(a_2)	QL(o_{12})	QD(o_{22})

In this case we have a matrix that exhibits determinism but not independence. Given Savage's suggestion, Jeffrey than shows (1977, 365–366) that we can transform as follows to yield k'_1:

Act	State			
	[SL QL](s_1)	[SL QD](s_2)	[SD QL](s_3)	[SD QD](s_4)
S(a_1)	SL(o_{11})	SL(o_{21})	SD(o_{31})	SD(o_{41})
Q(a_2)	QL(o_{12})	QD(o_{22})	QL(o_{32})	QD(o_{42})

Call a transformation like that from k_1 to k'_1 a *Jeffrey–Savage conversion*, where '[ab cd](s_i)' is the state of the world, or set of possible worlds, i such that a, b, c, and d are true and ef(o_{ij}) is the outcome that obtains from act a_j if the world is in state s_i such that e and f are true. Such formulations exhibit *both* determinism and independence. As Jeffrey (1976) shows, this strategy is fully generalizable and that with respect to determinism and independence "… we know how to obtain those properties at will (1977, p. 367)."

What a Jeffrey–Savage conversion accomplishes is, supposedly, the obviation of the need for causal knowledge on the part of the agent by assimilating such ignorance into the specification of the states relevant to the decision situation. In such a conversion the set of states represents all and only the relevant possibilities, and so specific causal knowledge is rendered unnecessary. Jeffrey's supposes then that all that would be additionally necessary for an agent to make a rational decision in terms of Bayesian decision theory would non-causal, statistical, knowledge concerning the frequencies of the outcomes given the actions types. In the case of k_1 to k'_1 this would be the statistical probabilities $P(L|S)$ and $P(L|Q)$. He then goes on to show that the unconverted formulations of decision problems are equivalent to converted forms in terms of expected utilities and the implication is that the need for causal knowledge and the resulting problem of probabilities of deliberative counterfactuals are simply an artifact of unconverted formulations of decision problems. Given this generic equivalence Jeffrey's indicates (1977, p. 367) that there is no need, in practice, to employ Jeffrey–Savage conversion as, if the equivalence holds, then causal knowledge does not need to be incorporated into ordinary, unconverted, matricies.

This result has been examined by Gibbard and Harper, and what they show is that Jeffrey's equivalence result holds, and so Jeffrey–Savage conversion is merely a matter of selection of formalism, if and only if their Axiom 2 is true and independence holds (1981, pp. 159–163). However, as we shall see in a later section, Axiom 2 must be rejected as it presupposes the objectionable principle of conditional excluded middle. Moreover, many decision problems do not exhibit independence. As a result, as Gibbard and Harper (1981) and Lewis (1981) are acutely aware (pace Jeffrey (1977)) the choice between presentations of these formulations of decision theory is not merely a trivial issue of preferring one of the formalisms to the other. The upshot here is this. As these versions of decision theory

of not equivalent and as they do not always yield the same advice concerning what we are, rationally, to do in a situation, we must evaluate their comparative adequacy.

First, we can note that absent an acceptable proposal concerning how to assign probabilities to deliberative counterfactuals that does not assume conditional excluded middle, we are left with the conclusion that Jeffrey's proposed formalism appears to be the only sort of viable, extant, alternative open to us, even if Jeffrey himself failed to see that it really was a substantially different theory from that proposed by Gibbard and Harper (1981), et al. One attempt to resuscitate the Gibbard and Harper type of approach, the general project of causal decision theory, can be found in Lewis 1981, but it fails as it employs the device of imaging which will be criticized below. What then remains to be seen is whether Jeffrey's theory is itself acceptable. However it turns out that it is not acceptable as it leads to obviously counter-intuitive results.

We can see the problem with Jeffrey's theory by examining the simple decision situation as informally discussed by Gibbard and Harper (1981, pp. 165–166). However, first recall that if Jeffrey had been correct about the equivalence of Jeffrey–Savage converted matricies and unconverted matricies, then they should always yield the same prescription concerning which act is rational. They, in point of fact, do not do so in cases where independence fails and it is widely recognized that causal decision gets such cases correct while Jeffrey's theory yields the wrong result.[9] As a result, the obvious implication is that Jeffrey's theory is not correct.

Probabilities, Counterfactuals and Causal Decision Theory

As we have seen, what must be added to a decision situation in order to make it descriptively complete is that the agent has a well-defined probability distribution over all relevant statements of the form 'If agent β were to do act a_n and the world is in state s_p with laws L, then it would be the case that outcome o_m would obtain.' So, the decision-maker must be able to assign probabilities to a complete partition of future directed counterfactual claims concerning the connection between the acts, states, and outcomes of any decision situation if the decision matrix she is to use to make her decision is to be considered descriptively complete. In other words, the decision-maker must know, or be able to predict with some probability, what acts will cause which outcomes depending on the possible world-states relevant to the decision situation, or she must at least have some degrees of belief concerning which act will cause which outcome in a given, relevant, world-state. But, in order to be able to assign these probabilities, it would seem to be the case that the agent must know what world we inhabit and what causal laws are operative in that world.

But, consider how we would fit future directed subjunctive conditionals of the general form $A > C$, into our decision problems. Recall that a classical decision problem involves laying out a mutually exclusive and jointly exhaustive partition of

[9] See Gibbard and Harper (1981, pp. 163–166) and Lewis (1981) for details, but also see Eells (1982) for defense of Jeffrey's view, although it is ultimately unconvincing.

acts that agent may take, the outcomes of those acts, and both probability and utility assignments to those outcomes. However, as suggested here this is not a complete description of a decision problem as it leaves unanalyzed the causal connections between the selection of an act, the states of the world, and the possible outcomes. What must be added to the ontology of a classical decision problem to make it complete are just these sorts of statements about causal connections. But, what happens if we add these features into the descriptions of decision problems?

The first, and most important feature of decision theory that must be modified if we are to make decision-theoretic descriptions of decision situations complete in the manner suggested above is the formula for the calculation of expected utility, and this, of course, may have radical consequences when it come to our evaluations of what alternative is the rational one in a given decision situation as understood by a Bayesian decision theorist. For a fixed law, the most plausible way to modify the expected utility formula in light of what was discussed above seems to be something like:[10]

$$\begin{aligned}(\mathrm{EU}^*)E^*(a_\mathrm{i}) = {} & P(s_1) \times P(L) \times u(o_{\mathrm{i}1}) \times P([a_\mathrm{i} \,\&\, (s_1 \,\&\, L)] > o_{\mathrm{i}1}) \\ & + P(s_2) \times P(L) \times u(o_{\mathrm{i}2}) \times P([a_\mathrm{i} \,\&\, (s_2 \,\&\, L)] > o_{\mathrm{i}2}) + \cdots \\ & + P\big(s_\mathrm{p}\big) \times P(L) \times u(o_\mathrm{im}) \times P\big([a_\mathrm{i} \,\&\, \big(s_\mathrm{p} \,\&\, L\big)] > o_\mathrm{im}\big).\end{aligned}$$

So, the expected utility of an act is defined in terms of the probabilities of the state of the world, the probability of the law in question, the utilities of the outcomes, *and* the probabilities that if an act were to be done in a given world state governed by a law it will produce the specified outcome. But how are we to interpret probabilities like $P([a_\mathrm{i} \,\&\, (s_1 \,\&\, L)] > o_{\mathrm{i}1})$ which have the general logical form $A > C$?

The most plausible suggestion concerning how the probabilities of conditionals ought to be construed is that the probability of a conditional should be interpreted as the conditional probability of the consequent given the antecedent.[11]

$$P(A > C) = P(C \mid A) \text{ for all } A, C \text{ in the domain of } P \text{ with } P(A) > 0,$$

and,

$$P(C|A) = P(CA)/P(A) \;\; \text{provided} \;\; P(A) \neq 0.$$

Alan Hàjek has proposed the acronym 'CCCP' to refer to this account (the conditional construal of conditional probability), and this convention shall be followed throughout.

Unfortunately for the Bayesian, as David Lewis and others have demonstrated, CCCP cannot be correct on pain of triviality. Based on some rather minimal assumptions, Lewis (1976) showed that any language having a universal probability

[10] For the sake of convenience EU* is somewhat simplified as uncertainty about which laws are operating in a decision situation would require representing the various possible combination between alternative laws and alternative world-states. More importantly, there are serious problems defining not only the terms in (EU*) with the form $P(A > C)$, but also the terms $P(L)$. See Shaffer (2001) for details.

[11] A detailed and illuminating history of attempts to interpret probabilities of conditionals is presented in Milne (1997).

conditional is a trivial language, and, hence, by reductio CCCP must be rejected.[12] Furthermore, in Hájek (1989) CCCP was proved to be trivial under considerably weaker assumptions than those originally made in Lewis (1976) and so the result has proven to be resilient.

For the Bayesian decision theorists, this result becomes problematic with respect to the every decision problem. In point of fact, if one agrees with the basic point about the descriptive incompleteness of decision-theoretic decision situations and the need to specify the causal connections in those situations, then most, if not all, claims made in decision theory are suspect, especially those of normative decision theory. To drive the point home, this is because all decision situations, and prescriptions concerning decision situations, depend on facts about the causal influences of acts that have not yet been performed (and many of which will never be performed), and so decision theoretic agents must be able to specify, at least probabilistically, what acts they believe will lead to which outcomes in which states. If this is so and there is no extant suggestion for how to assign probabilities to counterfactuals, then all decision situations will be such that they cannot be made complete, and, hence, we cannot be required to be utility maximizers in virtue of the claim that epistemic "oughts" imply epistemic "cans". In attempting to make Bayesian decision matrices complete with respect to the kinds of causal connections that make d_1 and d_2 different, we have found that in doing so there will be no way to define all of the terms necessary for employing (EU*), and so we really have no grasp of what the real expected utility of any given act is.

This situation is unfortunate for the defenders of Bayesian decision theory, as there does not seem to be any extant, coherent, suggestion as to how we are to non-trivially assign probabilities to counterfactual conditionals of any sort, including future directed, deliberative, conditionals. So the problem appears to have devastating consequences for Bayesian decision theory. The defenders of such views would have to successfully argue against the claim that causal knowledge must be included in the description of decision-theoretic agents or absent such an argument they must produce an acceptable suggestion for how such probabilities are to be understood. Otherwise it would seem to be the case that Bayesian decision theory and causal decision theory must be rejected.

Prospects for a Solution

In response to Lewis' celebrated results concerning CCCP, and the extensions thereof, three major proposals have arisen concerning the nature of conditionals and the probabilities of such expressions. First, Alan Gibbard and William Harper's 1981 defense of CCCP in the context of causal decision theory will be considered. Second, Lewis (1976) has proposed a way for assigning probabilities to conditionals referred to as imaging—further been developed by Gärdenfors (1988) and Joyce (1999)—and this proposal will be considered in below. Finally, Isaac Levi, Carlos Alchourrón, Peter Gärdenfors, David Makinson, et al. have proposed various

[12] For specific details see to Lewis (1976, 1986), and McGee (1989).

accounts of conditionals which deny that conditionals are truth valued. Instead, they consider conditionals to be policies for belief revision and such policies have conditions of rational support in lieu of truth conditions. This view will be considered in below as well. What we must be immediately concerned with in these sections is, first and foremost, whether causal decision theorists can exploit one of these three suggestions in order to solve the problem of assigning probabilities to deliberative counterfactuals. After we see that they cannot do so, we will turn our attention to drawing some general conclusions about the significance of decision theory.

The Gibbard and Harper Solution

Gibbard and Harper (1981) proposed an apparently modest way to resolve the problem of deliberative counterfactuals specifically in the context of decision theory, but, as we shall see, this solution is itself, at best, controversial and it is not really acceptable. Gibbard and Harper's solution is nevertheless interesting because they are acutely aware of the general problem of assigning probabilities to conditionals, and so, in the course of formulating their own particular version of decision theory, we are presented with a formal argument in support of the claim that the CCCP *does* hold for some decision-theoretic situations. The crucial aspect of this solution involves the adoption of a particular reading of counterfactual conditionals. However, upon close examination their solution depends on the acceptance of a principle that appears to be false.

Gibbard and Harper argue as follows. First, they begin by noting that decision-making clearly involves counterfactuals of the sort indicated in earlier sections, and so Gibbard and Harper appear to agree with the point about the incompleteness of classical decision theory made in above. More specifically, they note that, "…when a person weighs a major decision, it is rational for him to ask, for each act he considers, what would happen if he performed that act (1981, p. 153)." In addition, they go on to note that, "…ordinarily, of course, a person does not know everything that would happen if he performed a given act. He must resort to probabilities: he must ascribe a probability to each pertinent counterfactual 'I do a $\Box\!\!\rightarrow$ c happens' (1981, p. 153)."[13] If decision theory is to be made complete, then we must be able to coherently assign probabilities to such conditionals. In order to accomplish this, Gibbard and Harper note that things would be rather simple if we could accept CCCP; i.e. that $P(A > C) = P(C|A)$. However, this is apparently not possible due to Lewis' triviality results. As such they acknowledge that the problem of assigning probabilities to the conditionals in decision contexts is a serious problem in decision theory. However, they propose a way to salvage CCCP, and they proceed to offer out an argument to the effect that CCCP holds in at least some decision contexts.

The Gibbard/Harper argument depends on the acceptance of two axioms about counterfactuals, and a special condition imposed on acts and outcomes. The two axioms, for all A and C, are:

[13] Rather than using the symbol '… $\Box\!\!\rightarrow$ …' for the counterfactual conditional as Gibbard and Harper do, '… $<$ …' will be employed in what follows.

AXIOM 1 (*Counterfactual Modus Ponens*) $(A \,\&\, (A < C)) \supset C$,

and,

AXIOM 2 (*The Gibbard/Harper Principle*) $(A < \neg C) \equiv \neg(A < C)$.

Axiom 1 is just the familiar principle of counterfactual modus ponens,[14] and, as Gibbard and Harper note (1981, p. 156), Axiom 2 is a principle closely related to Stalnaker's principle of conditional excluded middle. In any case, if Axioms 1 and 2 hold, then it follows that

$$A \supset [(A < C) \equiv C].$$

Gibbard and Harper refer to this derived principle as Consequence 1, and they explicitly acknowledge that neither Axiom 2 nor Consequence 1 is obviously true.

The special condition for act A and outcome O_i is:

CONDITION 1 (*Stochastic Independence*)$P(A < O_i \mid A) = P(A < O_i)$.

Condition 1 asserts the stochastic independence of $A<O_i$ and A. They then proceed to argue that if Consequence 1 is a logical truth and A and O_i satisfy Condition 1, then it follows that[15]

$$P(A > O_i) = P(O_i \mid A).$$

The details of the proof are not especially relevant here, however, as the real problem with Gibbard and Harper's solution concerns Axiom 2, thereby impugning Consequence 1 and their defense of CCCP in general.

Consider Axiom 2. This axiom asserts that there is no difference between denying a counterfactual connection between and act and an outcome and accepting that an outcome does not follow counterfactually from an act. This axiom is, of course, equivalent to the following expression:

$$[(A > \neg C) \supset \neg(A > C)] \,\&\, [\neg(A > C) \supset (A > \neg C)].$$

It is then apparent that consequence 1 is true only if the notorious principle of conditional excluded middle is generally true, and this principle is formulated as follows:

$$(\textit{Conditional Excluded Middle})(A > C) \vee (A > \neg C)$$

However, it is well known that this is a highly controversial principle that suffers from obvious counterexamples. If conditional excluded middle *were* indeed true, then the following, well known, counter-example first introduced by Quine (1950) would have to be true:

> Either Bizet and Verdi would have been Italian if they had been compatriots, or they would have been non-Italian if they had been compatriots.

However, it should be obvious that it is not true because *neither* disjunct is true. It is neither the case that if Bizet and Verdi were compatriots, then they would have

[14] In this paper Lycan's (1993) worries about the general acceptability of modus ponens will be ignored.

[15] See Gibbard and Harper (1981, pp. 155-158) for the proof.

been Italian, nor is it the case that they would have been non-Italian if they have been compatriots. Following Nute (1975), it is also neither true that if my cat were pedigree, then it would be Siamese nor is it true that if it were pedigree, then it would not be Siamese. Such counterexamples are compelling and so conditional excluded middle is very likely false, although it is true that Stalnaker (1981) has tried to defend conditional excluded middle in spite of these sorts apparently damning of counter-examples.[16] In any case, this controversy alone surely undermines the acceptability of Gibbard and Harper's Consequence 1 in undermining Axiom 2, and, thereby, their conclusion that $P(A > O_i) = P(O_i|A)$ holds in decision contexts. Gibbard and Harper's argument is then unsound or, at least, the soundness of their argument is seriously questionable. As such, we have no good reason to believe that CCCP is true in *any* decision contexts absent a compelling defense of conditional excluded middle capable of dealing with counter-examples like those examples noted above.

Lewis' Concept of Imaging

Subsequent to rejecting CCCP with respect to Stalnaker conditionals, as well as many other types of conditionals, Lewis 1976 suggested that probability conditionals should be understood as policies for *feigned* minimal belief revision, and that the probability of such a conditional should be understood to be the probability of the consequent, given the minimal revision of $P(\cdot)$ that makes the probability of the antecedent of the conditional equal to 1. More to the point, Lewis (1981) suggested that imaging could be used to solve the problem of deliberative counterfactuals. Formally, imaging tells us that,

$$P(A > C) = P'(C), \text{if } A \text{ is possible},$$

where $P'(\cdot)$ is the minimally revised probability function that makes $P(A) = 1$. Lewis tells us that we are to understand this expression along the following lines. $P(\cdot)$ is to be understood as a function defined over a finite set of possible worlds, with each world having a probability $P(w)$. Furthermore, the probabilities defined on these worlds sum to 1, and the probability of a sentence, A for example, is the sum of the probabilities of the worlds where it is true. In this context the image on A of a given probability function is obtained by 'moving' the probability of each world over to the A-world closest to w.[17] Finally, the revision in question is supposed to be the minimal revision that makes A certain. In other words, the revision is to involve only those alterations necessary for making $P(A) = 1$.

Consider how we would apply this suggestion in calculating the expected utility of an act in terms of (EU*). Substituting $P'(C)$ for each instance of the probability of a conditional in E^* we get the following formula:

[16] See Lycan (2001) and Lewis (1973) for discussion of the prevailing sentiment that conditional excluded middle is false. See Lewis (1981) for analysis of Gibbard and Harper (1981) insofar as their version of causal decision theory depends on conditional excluded middle.

[17] Lewis also assumes that there is a unique closest A-world to w.

$$(\mathrm{EU}^{*}\mathrm{sub.1})E^{*}(a_{\mathrm{i}}) = P(s_1) \times P(L) \times u(o_{\mathrm{i}1}) \times P'(\mathrm{o}_{\mathrm{i}1}) + P(s_2) \times P(L) \times u(o_{\mathrm{i}2}) \times P^{\dagger}(\mathrm{o}_{\mathrm{i}2}) + \cdots + P(s_{\mathrm{p}}) \times P(L) \times u(o_{\mathrm{im}}) \times \mathrm{P}^{\ddagger}(o_{\mathrm{im}}).$$

The superscript terms $P'(\cdot)$, $P^{\dagger}(\cdot)$, and $P^{\ddagger}(\cdot)$ are the adjusted probabilities of the conditionals in (EU*) interpreted in light of imaging, where the superscripts are intended to signify that a different probability function will arise for each application of imaging to each original expression in (EU*).

What are we to make of this expression? What is the meaning of the sum of the products of the probability of a state, the utility of an outcome, and the probability of the antecedent of our causal conditional in terms of *some other probability function*; e.g. $P(s_1) \times u(o_{\mathrm{i}1}) \times P'(o_{\mathrm{i}1})$. This seems to be especially problematic as the latter probability term is the probability one *would* assign to the consequent after making the minimal revision of one's beliefs needed to make the probability of the antecedent equal to one. The two probability terms in each component expression of (EU* sub. 1) appear, in some sense, to be incommensurate, but, more importantly, there does not really seem to be any coherent way to assign a probability to $P'(C)$. [18] The revision in terms of which $P'(C)$ is defined *does not actually occur*, as it is only a feigned revision. It only occurs counterfactually and it is not clear how in the world we are to assess what the value of $P'(C)$ should be. This is complicated by the fact that what counts as a minimal revision has not been satisfactorily fleshed out in the literature, and so, in any case, we appear to be at a loss to actual employ Lewis' solution in practice.[19]

Furthermore, Lewis' suggestion places us in a position that appears to involve a viscous regress. In order to assess the numerical value associated with the image on A of $P(\cdot)$ we must accept another counterfactual concerning what we would believe if we were certain of A. This is because the belief revision is not an actual belief revision, and in order to accept this we would need to assign a probability to the counterfactual 'If I were certain of A (if it were the case that $P(A) = 1$), then my beliefs would be $\{B\}$', where $\{B\}$ is the set of my beliefs and probability ascriptions on those beliefs. Presumably, this counterfactual must be interpreted in terms of imaging as well, and so we must accept another counterfactual about that feigned

[18] The sense in which they are incommensurate is that the revised probability function is not about what one believes at all, but about what one would believe and the probability of the relevant world-state is about what one believes. It is not all clear how the product of these probabilities is to be understood. A similar sort of problem might also be found in Jeffrey's conditionalization, a kinematical generalization of Bayesian conditionalization, where $P_{\mathrm{new}}(A) = P_{\mathrm{old}}(A|E)P_{\mathrm{new}}(E) + P_{\mathrm{old}}(A|\sim E)P_{\mathrm{new}}(\sim E)$ for the pair $\{E, \sim E\}$. As in the case of (EU sub. 2), the products of these probabilities and their sums are rather difficult to interpret. It is similar to assuming that one can make sense of the product of the probability ascriptions of two different individuals, say $P_{\mathrm{John}}(A|E)P_{\mathrm{James}}(E)$. One's partial beliefs at some time t_1 and at some subsequent time t_2 are like the beliefs of two different agents, and they are, potentially, ascribed over two different sample spaces. It does not seem to be obvious just what an expression involving products of these sorts of terms could mean.

[19] See, for example, Gärdenfors (1982, 1984, 1988), Alchourrón et al. (1985), and, especially Lindström and Rabinowicz (1989, 1990, 1992) and Joyce (1999).

revision, and so on.[20] As a result of these considerations, it does not appear as if imaging will help the Bayesian decision theorist to avoid problem in question, as imaging does not allow us to clearly specify a well defined prior probability for the kinds of future directed, deliberative, counterfactuals that a decision theoretic agents needs to know.

There have, however, been some other suggestions for how probabilities can be assigned to conditionals that do not depend on imaging. The most interesting such suggestion has been made by McGee (1994), and this sort of approach has been thoroughly discussed in Arló-Costa (2001, 2007). McGee's approach is based on appealing to a non-standard concept of probability based on Popper's (1959) controversial notion of conditional probability. They require appealing to primitive concept of finitely additive conditional probability and they do not allow us to assign probabilities to either Boolean combinations of conditionals or to nested conditionals. Issues of the acceptability of a non-standard notion of probability aside, the latter problems are surely damning for this approach when it comes to decision theory and especially to its application in artificial intelligence. The most common problems faced in A.I. planning theory involve *sequential plans* that involve both nested conditionals and Boolean combination of conditionals as is amply demonstrated in Korb and Nicholson (2004, pp. 98–103) and in Russell and Norvig (1995, pp. 337–412).

The AGM/Levi Approach to Conditionals

In the spirit of F. P. Ramsey's and Ernest Adams' accounts of the nature of conditional expressions, some philosophers and computer scientists have adopted the view that conditional expressions do not have truth values.[21] Rather, they hold that conditionals ought to be regarded as various kinds of epistemic policies for belief revision, and although their views differ with respect to various details concerning the nature of such revisions, Isaac Levi, Carlos Alchourrón, Peter Gärdenfors, David Makinson, et al. agree that conditionals of the sort we have been discussing should not be treated as assertions that have truth conditions.[22] Rather, they are to be treated as something like policies for updating or revising one's

[20] In a bit more formal presentation this problem arises as follows. If $P(A > C) = P'(C)$ by imaging, then to assess the numerical value of $P'(C)$ an agent must accept the conditional $T(A) > \{B\}$ where $T(A)$ is the belief that a particular agent is certain that A and $\{B\}$ is that agent's set of beliefs and probability distribution over those beliefs. To accept $T(A) > \{B\}$ by Lewis' own admission is to assign a (high) probability to that sentence, so the agent must be able to evaluate $P(T(A) > \{B\})$ if the agent is to be able to assess $P(A > C)$. But, by imaging, $P(T(A) > \{B\}) = P''(\{B\})$, where $P''(\{B\})$ is the agent's beliefs and probability distribution on those beliefs were the agent certain that $T(A)$. Again, according to the definition of the concept of imaging this is only a feigned revision. So, in order to assign a numerical value to $P''(\{B\})$ the agent must accept a conditional about what that agent would believe if he were certain that he were certain that A, $T(T(A)) > \{B'\}$ (where $\{B'\}$ is that agent's suitably revised beliefs and his probability distribution on those beliefs). So, the agent must assign a numerical value to $P(T(T(\mathrm{A})) > \{B'\})$ where, by imaging $P(T(T(A)) > \{B'\}) = P'''(\{B'\})$, and so on ad infinitum. Moreover, there does not seem to be any obvious, non-ad hoc, way to stem this regress that results from the nature of imaging.

[21] See Ramsey (1931), Adams (1975), Adams (1976, 1993), and Edgington (1986).

[22] See Gärdenfors (1988, 1986), Alchourrón et al. (1985), Arló Costa and Levi (1996), and Levi (1996).

beliefs relative to what one already believes; in other words they are taken to be epistemic conditionals, and in lieu of truth conditions such conditionals have conditions for rational support relative to an antecedently given belief set.[23,24]

This analysis of the sort of conditionals with which we have been concerned to this point is rather implausible and that it does not reflect our everyday deliberative reasoning very well, but we need not dwell on this alternative at length. It is enough to simply note that it is a rather implausible view when applied to conditionals like those discussed in earlier sections. As Stalnaker pointed out long ago, "many counterfactuals seem to be synthetic, and contingent, statements about unrealized possibilities" (1970, p. 42), and an agent in a decision situation needs to be able make claims about what would contingently happen were the agent to engage in each possible action and this is not merely an issue of how her beliefs might be updated. It simply does not seem that such conditionals are merely about how one's beliefs might be revised. Rather, the sorts of conditionals discussed above are synthetic and contingent claims about what outcomes would arise if a given act is done in a given world state. In any case, the AGM/Levi strategy for analyzing cannot be used to solve the problem for Bayesian decision theorists noted above as it does not allow for probabilities to be assigned to the future directed deliberative claims of the sort with which we are concerned. Certainly, if they do not have truth conditions, then they cannot have probabilities of being true. Finally, these approaches all relativize the acceptability of conditionals to belief states and so cannot issue in objective estimations of expected utility even if these other problems could be dealt with.

Responses

It would seem to be the case that absent some plausible account of how probabilities are to be coherently assigned to statements about the causal connections between acts, states and outcomes in a decision situation or some more reasonable argument to the effect that such counterfactuals do not really figure into complete decision situations, Bayesian decision theory is in real trouble. This conclusion seems warranted insofar as it is as if there is no reasonable way to simply assume that our decision-maker can just accept that, say, she is in w_i and thus faces d_1 rather than that

[23] Lindström and Rabinowicz and Hansson make a more or less sharp distinction between epistemic and ontic conditionals in Lindström and Rabinowicz (1995) and in Hansson (1995) respectively, and they hold that ontic conditionals have truth-values while epistemic conditionals have conditions of rational support. They distinguish these two types of conditionals based on the idea that the latter are accepted relative to belief systems that do not need to be complete, whereas ontic conditionals are true or false only relative to complete models. In general it is hard to see the difference between these two types of conditionals, and the distinction can be dissolved by allowing possible worlds to be partial (i.e. incomplete). However, this issue will be ignored for present purposes. For a more detailed discussion of partial worlds and conditionals see Shaffer (2000, 2001) and the essays in Doherty (1996).

[24] The epistemic implications of this view are discussed at some length in Gärdenfors (1992). In this article Gärdenfors appears to ally the belief revision tradition with coherence theories of knowledge, and this provides some explanation of the AGM theorist's views concerning conditionals, at least qua their lack of truth-values. Levi's views, which are somewhat different from those of the AGM theorists, are discussed in Arló-Costa (2007).

she is in w_j and so faces d_2 (let alone the fact that she may be in any one of a myriad of worlds) or that she even knows which act-state pairs will be matched to which outcomes within a given d_k, the issue of the incompleteness of classical, evidential or Bayesian, characterizations of decision situations is not seriously open to question. Were we to merely assume such background knowledge, then the Bayesian prescriptions concerning what a decision maker ought to do may well turn out to radically misguided, as we may simply be wrong about that information. So there seem to be two possible reactions that one might have to these results. First, one might be tempted to argue that Bayesian decision theoretical rationality is correct in some sense, even if it faces problematic anomalies. Second, one might be tempted to argue that we need to look at other accounts of decision theoretical rationality to find the sorts of prescriptions that we need to model and construct decision makers.

Idealizations and Decision Theoretical Rationality

From a more conservative perspective one might then be inclined to argue that while the technical problem of assigning probabilities to future directed, deliberative, counterfactuals does arise because decision situations—as classically described—are incomplete, we should simply regard this problem as unimportant on the basis of the claim that Bayesian decision theory is only intended to offer *idealized* representations of real decision situations. If this were the case, then we could simply stipulate that information about the causal connections between the acts and outcomes is possessed by decision theoretical agents as part of the idealizing of such situations. This assumption would then be an idealization imposed on a more realistic and correct model of decision-making in which that assumption was relaxed or eliminated and which would yield better predictions about and explanations of real decision-making. This sort of suggestion is most thoroughly worked out by Weirich (2004).

However, if this metaphysically and epistemologically naive line of argument is followed then decision theory stands in danger of becoming inapplicable to real decision-making, and, hence, rather irrelevant to substantive issues concerning real human and machine rationality. More generally, we should to be very wary of such blanket defenses of objectionable consequences of formal theories under the rubric of idealization.[25] This is obviously so because not all idealizations are acceptable in a given epistemic context with respect to a given theory and the behavior(s) of the

[25] A similar problem about appeals to idealization in the context of justifying components of formal theories arises in the context of the AGM theory of belief revision and it concerns the acceptability of the so-called recovery postulate. Hansson (2000) defends recovery as an idealization in this unacceptable manner. Such examples indicate that there may be points at which formal theories of rationality simply break down and so should be regarded with deep suspicion or simply rejected, especially when generated a priori and independent of empirical information concerning the intended domain of application of that theory (See Shaffer 2003 for further discussion). In such cases appeal to idealization without substantial justification is less than honest and amounts to the sort of ad hoc theory immunization that Popper so rightly disparaged in his 1959. More relevant to the issue at hand, Resnik (1987) précis his book on decision theory with just such a blanket apology for the philosophical and empirical challenges to that theory and Pollock (2002b) reports just this kind of reaction from decision theorists to his criticism of classical decision theory in terms of the notion of act that they accept.

systems it is intended to characterize. Appeals to idealization, by there very nature, require substantial defense in terms both of how close the supposedly idealized theory approximates the relevant real theory in terms of the (1) idealized theory's really being a simplification of the more complex, more realistic theory, and (2) how close predictions and explanations based on the idealized theory come to predictions and explanations based on the more complex and more realistic extensions of that theory.[26] (2) Is especially important here. What would need to be shown is that the relaxation or elimination of a proposed idealizing assumption yields a more accurate and less problematic characterization of the behaviors that the two theories are intended to explain. The real point at issue here is to distinguish the following sorts of cases. Surely the assumption that the masses of objects are constant in classical mechanics is a legitimate idealization. This is so because it is computationally simpler to make this assumption and because classical mechanics is ultimately derivable from the more realistic and complex equations of relativistic mechanics—where the mass of an object is dependent on its velocity. But, in Ptolemaic astronomy, the assumption that the observed orbits of planets can be accounted for by adding epicycles on to the circular motions of those bodies in terms of the motion of the fixed spheres to which they were supposed to be attached is not a legitimate idealization. This is because the assumption that there are such epicycles is totally ad hoc and is not derivable from a more realistic theory that describes the motions of the planets. This latter point is the case simply because Ptolemaic astronomy is just false.

With these warnings firmly in mind let us look more closely at Weirich's position on matters related to the discussion above. Weirich explicitly identifies his version of decision theory as a form of causal decision theory, "…as expounded by Gibbard and Harper (1978) (2004, p. 6)." This makes his claims about idealization especially relevant here for obvious reasons. What Weirich claims about the sort of descriptive incompleteness discussed above is that the assumption of full information about the outcomes that the agent's acts might produce is an idealization introduced to make decision problems tractable. He argues as follows:

> Suppose that an agent is certain of his psychological state but uncertain of the external world. Because he is not fully informed, he is uncertain of the outcome of acts he can perform. How should he pursue the goal of optimizing? Rationality still urges maximizing utility, but given uncertainty, and act's utility is an estimate of the utility of the act's outcome, the act's world … Given uncertainty, an act's utility may not be the utility of the act's world. An agent may not know the world his act produces. It might, for all he knows, produce any of many worlds. Hence, an act's utility is an estimate of the utility of the act's world. (2004, p. 25)

[26] See Shaffer (2000) for an extensive discussion of the conditions under which idealizations are acceptable in the context of physical theory and Shaffer (2002) for some discussion of idealization in the context of formal models of rationality. We should note that this particular issue is by no means the only aspect in which decision theoretic characterizations of rationality are idealizations and it seems all too typical for decision theorists to protect their theory by such appeals as Resnik (1987) and Pollock (2002b) clearly attest to.

So, the assumption of full information is an idealization of a theory of decision making in that assumption is replaced with a more realistic and more accurate account of decision-making under uncertainty.

At least two comments are in order here. First, Weirich's suggestion about the relaxing of the assumption of full information seems to be equivalent to Savage's (1954/1972) problematic equation of acts with their outcomes that was criticized above. But he is quite clear about this. He claims that, "given uncertainty, an option's utility is an estimate of the utility of the world that the option would yield if it were realized (2004, p. 28)." Expected utilities are then understood to be the estimates of the utilities of actions given the agent's assessments of the desirabilities of the outcomes of those acts on the assumption that the acts open to the agent really would produce given outcomes were they to be performed. Where there is *uncertainty* about the outcomes of a given option open to the agent we are to base decisions on our knowledge about the *possible* outcomes of each act, and an outcome's expected utility is the sum of the products of the probability and utility of each possible outcome of the act (2004, p. 28). Not only is this more realistic theory metaphysically naïve for the reasons again mentioned earlier in this paper, but it also (1) highlights the worry that utility maximization so understood is problematically subjective and (2) fails to address the issue of real agent's ignorance about what outcomes are possible relative to each option open to that agent.

First, if we allow the equation of acts and outcomes or worlds, then decision theoretical rationality becomes nothing more than a theory of wishful thinking relativized to the agent's current belief state and utility assignments, none of which may correspond to the real causal features of the world in which the agent finds herself. Moreover, what is relevant here is agent's uncertainty about statements of the form "If β were to do act a_n and the world is in state s_p, then it would result in outcome o_m", and *not* just uncertainty about outcomes o_i. What is important to decision theory is uncertainty about the *agent's ability to produce such outcomes by her actions*. Decision theory is not concerned with the probability and desirability that an outcome occurs, for it need not be the case then that such outcomes are the result of the agent's acts.

Now, of course, if this more realistic theory that relaxes the assumption of full information were correct, it would be no comfort to A.I. practitioners in particular. Who cares about rationality understood to be subjective estimation of the product of the desirability of outcomes and their probabilities independent of an agent's production of such outcomes? What is needed in A.I. is something quite different. What A.I. practitioners need is a better account of how decision making is impacted by the failure to grasp the causal laws that a world is characterized by, not an account of how agents act on the basis of their desires for outcomes and assessments of the probabilities of the occurrence of those outcomes simpliciter. Even passing familiarity with the frame problem indicates this (see Shoham 1988).

More importantly, Weirich's more realistic theory is problematic and so renders the defense of the assumption of full information by appeal to idealization suspect. Recall that he claims that where there is uncertainty about the outcomes of a given option open to the agent we are to base decisions on our knowledge about the *possible* outcomes of each act and an outcome's expected utility is the sum of the

products of the probability and utility of each possible outcome of the act (2004, p. 28). So his more realistic decision theoretical agent is assumed to know all of the possible outcomes of each option, although that agent does not necessarily know the outcome of each option. However, as the possible outcomes of any action are—from a purely logical perspective—limitless, we end up with a completely empty solution to the problem of uncertainty about the causal impact of agent's acts or we end up illicitly smuggling in the kind of causal knowledge that such agent's are not supposed to have.

If no such causal knowledge is smuggled in then every option open to an agent in every decision situation will be equivalent because they all have exactly the same set of possible outcomes, viz. every possible outcome. For each decision situation and act, state, outcome triplet of the form "If β were to do act a_n and the world is in state s_p, then it would result in outcome o_m" we will get the same set of possible o_m. So every act will be causally equivalent and will essential be equivalent to the following statement: "If β were to do act a_n and the world is in state s_p, then it would result in some outcome o_m where o_m is a member of the set of all possible outcomes". This of course only means that "If β were to do act a_n and the world is in state s_p, then it would result in some outcome," and all such claims have a probability of 1 because they are trivial. If causal knowledge is smuggled in by assuming that agent's know at least some of the causally possible outcomes of their acts in various world states then they must be able to establish probabilities for those deliberative counterfactuals. This, as we have seen, is not so easy to do in a reasonable way.

So, the behavior to be expected from Weirich's allegedly more realistic agents—so defined—still does not appear to remotely correspond to the behavior of real decision-makers. But it is the elimination of the assumption of full information—full causal information—that is of crucial importance for the understanding of real decision-making. The agent must know for each possible act and outcome pair *what the probability of the corresponding deliberative counterfactual is*. In any case, it is hard to see how Weirich's assumption of full information can count as a legitimate idealization because the more realistic theory that it is intended to be an idealization of is itself so problematic that there is a real worry that it also does not describe the behavior of actual decision-makers. First, it illegitimately equates acts with outcomes and so fails to deal with the problem of agent's causal knowledge, and even if we ignore this it illegitimately smuggles modal knowledge of possible outcomes of each act open to an agent into the model. Recall then that the acceptability of an idealized theory is judged by the closeness of its predictions/explanations to the behaviors of the systems it is intended to apply to relative to the more realistic versions of that theory. Certainly in this case, then Weirich's assumption of full information does not meet this desideratum. The behavior of agent's endowed with incomplete causal knowledge will not be remotely like that of Bayesian decision makers and so the assumption of full information cannot easily be legitimized by appealing to idealization. All of this then should start to suggest that we are really facing something like Ptolemy's epicycles.

So as far as things currently stand then there does not seem to be an extant defense of this sort with respect to Bayesian decision-theoretical rationality in the

context of the problem of deliberative counterfactuals and causal knowledge. So appealing to idealization as a panacea to this problem is a hollow tactic at best, at least as things stand. What this demonstrates is that there is some reason to suppose that the Bayesian decision theoretic account of rationality is simply false rather than an idealized case of a more complex, realistic and true theory of human and machine rationality. This is especially problematic as Bayesian decision theory was explicitly developed as a formal and mathematical theory independent of real, psychological based, accounts of decision-making. Nevertheless, Bayesian decision theorists have had some success in dealing with some problems involving rational planning in artificial intelligence and so this point needs to be explained, especially if we are inclined to advocate on behalf of non-Bayesian alternatives. But this is not hard to do. Bayesian decision theory assumes that agents are aware of the causal structure of their environments and so it is no surprise that when deployed in A.I. the theory results in success. Of course, if one programs such knowledge into an artificial agent and then sets it loose in a relatively simple, controlled and pre-structured environment characterized by the causal features pre-programmed into that agent then the artificial agent will often be successful. This is primarily why simulations of expert systems often perform well. Feed them all of the relevant causal knowledge before making decisions and of course they will make reasonably good decisions. But this is totally unlike the sorts of ignorance that real decision-makers face and upon analysis these successes are far less impressive than they initially seem. Real decision makers have to act without knowledge of the casual structure of their environments and must effectively learn by making mistakes and acting on the basis of radically incomplete knowledge about the causal laws of the world they inhabit. Passing familiarity with the failures of robotics supports this assessment quite effectively. It is easy to get reasonably good performance in computer simulations in "block-" or "toy-" worlds, but it is quite another thing to get good performance when setting an agent loose in the causal complexity of the real world.

Bounded Rationality and Decision Theory

Another, considerably less conservative, way of potentially dealing with the problem of deliberative counterfactuals and causal knowledge is suggested by work done under the rubrics of bounded rationality and ecological rationality. The main point of interest here is that in the context of bounded and ecological rationality, it is *not* necessarily assumed that the agent knows all of the details of the causal information about her environment. In fact, given this more organic understanding of decision-making decision agents can start with virtually no causal knowledge of their environments and can still—using only very simple heuristics—achieve remarkable successes in decision tasks. What is yet more interesting is that such successes also do not require much at all in way of deliberation. This is particularly intriguing because by the very nature of the approach we get the sorts of results that are not practically achievable in the context of Bayesian decision-theoretical approaches to A.I. Moreover, this approach does not require us to solve the problem of assigning probabilities to deliberative counterfactuals because such conditionals play no role in this account of rational decision-making.

To illustrate the ideas of bounded and ecological decision-making—first suggested by Herbert Simon (1956)—let us consider the most well known of the fast and frugal reasoning heuristics for problem solving and some of its close relatives. In particular the work of Gerd Gigerenzer et al. (1999, 2000) is of special importance here. The core idea behind the concept of bounded rationality is that real decision agents do not have unlimited computational capacities, time, complete information, etc. The idea then is that we need to explore the manner in which real decisions are made by actual decision makers (viz. us) in order to see how it is that such decisions are made quickly and frugally based on our actual abilities. The second core idea relevant here is the concept of ecological rationality. The idea here is that decision-making is not the result of a generic, domain-independent, capacity to deliberate and reason. As a result, the heuristics for decision-making advocated by this approach are the results of and work only in the specific environments in which they are generated. What is then important for A.I. and the modeling of real decision-making agents is finding the right decision heuristic for a given environment.

The most well known such heuristic is known as the "Take the Best" heuristic. Gigerenzer (2000) essentially explains that given the Take the Best heuristic decision-making is supposed to follow the following algorithm: for any decision problem P and agent M there will be a body of knowledge K relevant to P possessed by M. Given a set of choice targets $\{o_1, o_2,\ldots, o_n\}$, cues $\{c_1, c_2,\ldots c_n\}$, values of those cues for each choice target, and a subjective ordering of cue validities a decision-maker is supposed to apply the following procedure in the case of a binary decision:

> Step 0 (The Recognition Heuristic): If only one of two choice targets is recognized on the basis of K, then choose the recognized choice target. If neither choice target is recognized, then choose randomly between them. If both choice targets are recognized, then proceed to step 1.
> Step 1 (Take the Best Search Rule): Choose the cue with the highest validity that has not yet been tried and look up the values of the two choice targets in terms of that cue.
> Step 2 (Stopping Rule): If one choice target has a positive cue value and the other does not, then stop the search and go to step 3. Otherwise go back to step 1.
> Step 3 (Decision Rule): Opt for the choice target with the highest cue value.

In decision situations then what Take the Best tells amounts to is the use very simple, computationally tractable, means to solve decision problems. Take the Best does not require extensive deliberation—so it is a form of bounded rationality—and thus allows for decision-making to result in action much more rapidly. Of course, there is some information that decision-makers need to make such decisions, but it is just awareness of options, choice targets (i.e. outcomes) and cues. The cues in particular are important because they are understood to be probability cues each of which has a validity with respect to its statistically associating the cue value with the target factor (Gigerenzer 2000, pp. 170–171). What cues are then are factors or variables that the discriminate targets in the agent's particular environment. Gigerenzer's (2000) example is that when deciding which of two German cities is

larger, one might use the cue "possesses a team in the Bundesliga". So such knowledge is just acquired statistical knowledge about the association of the cue and the factor relevant to the choice (i.e. in this case German city size). However, this knowledge is itself knowledge acquired by the agent in her environment and so is ecological in the relevant sense. In fact, as we shall soon see, bounded and ecological rationality allows for decision-making even when such knowledge is incomplete or totally absent. Agents can still effectively act when they do not know the validities of the cues or even what the cues are. In any case, Gigerenzer (2000); Gigerenzer et al. (1999) and Bullock and Todd (1999), among others have shown this fast and frugal decision heuristic performs very well in many environments despite the obvious simplicity of such reasoning.

Even more intriguingly, there are close relatives of the Take the Best heuristic that appear to be even better approximations of real-world decision-making. For example the Minimalist heuristic replaces step 1 with the following rule:

> Step 1′ (Minimalist Search Rule): Draw a cue randomly and look up the cue values of the two decision targets.

Notice that this heuristic decision strategy no longer assumes that the agent knows the validities of the cues. By its implementation however the agent can store knowledge of past successes in K and use this in future decisions and so can come to know the cue validities for that environment. This also suggests an additional search rule known as Take the Last. Take the Last replaces step 1 with the following rule:

> Step 1″ (Take the Last Search Rule): Given a record of which cues stopped search in previous decision problems, choose the cue that stopped the most recent search that has not been tried. Check those values and if this does not solve the problem choose a random cue.

Again, Gigerenzer (2000) has shown that such approaches are remarkably successful and this is both surprising and of great interest.

Finally, work that Terry Connolly (1999) has done under the rubrics of bounded and ecological rationality is also worthy of some attention here as well. Connolly has studied what she calls "action-first" approaches to decision-theoretical rationality. On this sort of view agent's need not do much if any deliberation prior to acting. The standard Bayesian-style rule about when to stop deliberating and when to act comes from Raiffa (1968) and it simply tells us that we should act when the expected utility of one unit of further deliberation falls below the expected utility of the positive difference in utility that further deliberation would bring. What Connolly notes is that given uncertainty about what outcomes each act will lead to and uncertainty about how to evaluate such outcomes (i.e. uncertainty about cues and/or cue values) it is often best to *simply act without deliberating*—or with minimal deliberation—and then simply see what happens. What this then allows for in combination with heuristics like Take the Best, take the Last and Minimalist is (1) real learning on the part of agents about what works in their environments, (2) action without the need to consider deliberative counterfactuals and (3) action that does not require fore-knowledge of the causal laws that characterize the agent's environment. Real agents begin with no knowledge and so must experiment by

acting and seeing what happens. So given the sorts of problems we have encountered in the analysis of the Bayesian concept of deliberation it is clear that these sorts of approaches have much promise both for the modeling of real decision-making and for the construction of artificial decision agents in A.I. Of course, there is still much work to be done here to flesh out the details of this model of decision-making, but it is at least encouraging in its parsimony and tractability.

Conclusion

So absent a solution to the problem of the incompleteness of Bayesian decision situations qua causal knowledge, what defenders of that theory are faced with is an extant and substantial problem about the role of conditionals in deliberation that decision theorists, epistemologists and logicians need to address, and their doing so will require paying serious attention to deep problems with the logic of counterfactuals and the epistemological features of their rational acceptance—hopefully without having to resort to the sorts of radical revisions of probability theory and the semantics and logic of conditionals.[27] Incidentally, this problem is no less acute for defenders of causal decision theory, as our consideration of Gibbard and Harper's and Lewis' work should make apparent.[28] The serious and live possibility that the unresolved problem of causal knowledge in decision theory more broadly suggests, however, is that non-Bayesian and non-classical decision theoretic approaches to decision may well be superior with respect to the inclusion of causal knowledge in the models of deliberative agents. As a result, approaches to decision and planning like that advocated, for example by the defenders of bounded and ecological rationality may well be more worthy of pursuit in the explanation and guidance of human decision and the construction of artificial decision makers. In effect, then the difficulties raised here for Bayesian-style decision-making can be regarded as an argument—primarily philosophical in nature—in favor of the application of bounded and ecological rationality to the modeling and creation of decision theoretical agents.

References

Adams, E. (1975). *The logic of conditionals*. Dordrecht: D. Reidel Press.

Adams, E. (1976). Prior probabilities and counterfactual conditionals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 1, pp. 1–21). Dordrecht: D. Reidel.

[27] It is worth noting that Pollock's (2002a, and 2006) new account of decision theory is also based on a non-standard concept of probability and so there is another avenue there that needs to be explored.

[28] One largely unexplored—but still live—approach to the problem of resolving the issue of probabilities of counterfactuals in decision situations—whether characterized causally or classically—stems from the work on conditionals presented in Nute (1975), Fetzer and Nute (1979, 1980). The application of Fetzer and Nute's view of conditionals, however, still needs to be examined in detail concerning how it might be applied to decision theory and it offers little in the way of explicit suggestions concerning how probabilities are to be applied to conditional statements.

Adams, E. (1993). On the rightness of certain counterfactuals. *Pacific Philosophical Quarterly, 74*, 1–10.
Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic, 50*, 510–530. doi:10.2307/2274239.
Arló-Costa, H. (2001). Bayesian epistemology and epistemic conditionals: On the status of the export-import laws. *The Journal of Philosophy, 98*, 555–598. doi:10.2307/3649472.
Arló-Costa, H. (2007). The logic of conditionals. *The Stanford Encyclopedia of Philosophy*. E. N. Zalta (Ed.). http://plato.stanford.edu/entries/logic-conditionals/.
Arló-Costa, H., & Levi, I. (1996). Two notions of epistemic validity. *Synthese, 109*, 217–262. doi:10.1007/BF00413768.
Brown, H. (1994). Reason, judgment and Bayes' law. *Philosophy of Science, 61*, 351–369. doi:10.1086/289808.
Bullock, M., & Todd, P. M. (1999). Made to measure: Ecological rationality in structured environments. *Minds and Machines, 9*, 497–541. doi:10.1023/A:1008352717581.
Connolly, T. (1999). Action as a fast and frugal heuristic. *Minds and Machines, 9*, 479–496. doi:10.1023/A:1008396500743.
Doherty, P. (Ed.). (1996). *Partiality, modality and nonmonotonicity*. Stanford: CSLI Publications.
Edgington, D. (1986). Do conditionals have truth-conditions? *Critica, 18*, 3–30.
Eells, E. (1981). Causality, utility and decision. *Synthese, 48*, 295–329. doi:10.1007/BF01063891.
Eells, E. (1982). *Rational decisions and causality*. Cambridge: Cambridge University Press.
Fetzer, J., & Nute, D. (1979). Syntax, semantics and ontology: A probabilistic causal calculus. *Synthese, 40*, 453–495. doi:10.1007/BF00413415.
Fetzer, J., & Nute, D. (1980). A probabilistic causal calculus: Conflicting conceptions. *Synthese, 44*, 241–246. doi:10.1007/BF00413408.
Gärdenfors, P. (1982). Imaging and conditionalization. *The Journal of Philosophy, 79*, 747–760. doi:10.2307/2026039.
Gärdenfors, P. (1984). Epistemic importance and minimal changes in belief. *Australasian Journal of Philosophy, 62*, 136–157. doi:10.1080/00048408412341331.
Gärdenfors, P. (1988). *Knowledge in flux*. Cambridge, Mass: The MIT Press.
Gärdenfors, P. (1986). Belief revision and the Ramsey test for conditionals. *The Philosophical Review, XCV*, 81–93. doi:10.2307/2185133.
Gärdenfors, P. (1992). The dynamics of belief systems: Foundations versus coherence theories. In C. Bicchieri & M. L. Dalla Chiara (Eds.), *Knowledge, belief, and strategic interaction* (pp. 377–396). New York: Cambridge University Press.
Gärdenfors, P., & Sahlin, N. (Eds.). (1988). *Decision, probability and utility*. New York: Cambridge University Press.
Gibbard, A., & Harper, W. (1981). Two kinds of expected utility. In W. Harper, et al. (Eds.), *Ifs* (pp. 153–190). Dordrecht: D. Reidel.
Gigerenzer, G. (2000). *Adaptive thinking*. New York: Oxford University Press.
Gigerenzer, G., Todd, P. M. and the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
Glymour, C. (1981). *Theory and evidence*. Chicago: University of Chicago Press.
Good, I. J. (1983). 46656 Varieties of Bayesians. In *Good thinking* (pp. 20–21). Minneapolis: University of Minnesota Press
Hájek, A. (1989). Probabilities of conditionals revisited. *Journal of Philosophical Logic, 18*, 423–428. doi:10.1007/BF00262944.
Hansson, S. O. (1995). The Emperor's new cloths: Some recurring problems in the formal analysis of counterfactuals. In G. Crocco, L. Farinas Del Cerro & A. Herzig (Eds.), *Conditionals: From Philosophy to Computer Science* (pp. 13–31). Oxford: Clarendon Press.
Hansson, S. O. (2000). Formalization in philosophy. *The Bulletin of Symbolic Logic, 28*, 162–175. doi:10.2307/421204.
Horovitz, E., Breese, J., & Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning, 2*, 247–302. doi:10.1016/0888-613X(88)90120-X.
Howson, C. (1995). Theories of probability. *The British Journal for the Philosophy of Science, 46*, 1–32. doi:10.1093/bjps/46.1.1.
Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court.

Jeffrey, R. (1965). *The logic of decision*. New York: McGraw-Hill.
Jeffrey, R. (1976). Savage's Omelet. *Proceedings of the Philosophy of Science Association*, pp. 361–371.
Joyce, J. (1999). *The foundations of causal decision theory*. Cambridge: Cambridge University Press.
Korb, K., & Nicholson, A. (2004). *Bayesian artificial intelligence*. Boca Raton: Chapman & Hall.
Kyburg, H. (1978). Subjective probability: Criticisms, reflections, and problems. *Journal of Philosophical Logic, 7*, 157–180. doi:10.1007/BF00245926.
Langholm, T. (1996). How different is partial logic? In P. Doherty (Ed.), *Partiality, modality, and nonmonotonicity* (pp. 3–34). Stanford: CSLI Publications.
Levi, I. (1975). Newcomb's many problems. *Theory and Decision, 6*, 161–175. doi:10.1007/BF00169104.
Levi, I. (1992). Feasibility. In C. Bicchieri & M. L. Dalla Chiara (Eds.), *Knowledge, belief, and strategic interaction* (pp. 1–20). New York: Cambridge University Press.
Levi, I. (1996). *For the sake of the argument: Ramsey test conditionals, inductive inference, and nonmonotonic reasoning*. Cambridge: Cambridge University Press.
Lewis, D. (1973). *Counterfactuals*. Cambridge: Harvard University Press.
Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *The Philosophical Review, 85*, 297–315. doi:10.2307/2184045.
Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30. doi:10.1080/00048408112340011.
Lewis, D. (1986). Probabilities of conditionals and conditional probabilities II. *The Philosophical Review, 95*, 581–589. doi:10.2307/2185051.
Lindström, S., & Rabinowicz, W. (1989). On probabilistic representation of non-probabilistic belief revision. *Journal of Philosophical Logic, 18*, 69–101. doi:10.1007/BF00296175.
Lindström, S., & Rabinowicz, W. (1990). Epistemic entrenchment with incomparabilities and rational belief revision. In A. Furhmann & M. Morreau (Eds.), *The logic of theory change* (pp. 93–126). Berlin: Springer-Verlag.
Lindström, S., & Rabinowicz, W. (1992). Belief revision, epistemic conditionals, and the Ramsey test. *Synthese, 91*, 195–237. doi:10.1007/BF00413567.
Lindström, S., & Rabinowicz, W. (1995). The Ramsey test revisited. In G. Crocco, L. Farinas Del Cerro & A. Herzig (Eds.), *Conditionals: From Philosophy to Computer Science* (pp. 147–191). Oxford: Clarendon Press.
Lycan, W. (1993). MPP, RIP. In J. Tomerilin (Ed.), *Philosophical perspectives VII: Logic and language* (pp. 411–428). Atascadero: Ridgeview Publishing.
Lycan, W. (2001). *Real conditionals*. Oxford: Clarendon Press.
McGee, V. (1989). Conditional probabilities and compounds of conditionals. *The Philosophical Review, 97*, 485–541. doi:10.2307/2185116.
McGee, V. (1994). Learning the impossible. In E. Eels & B. Skyrms (Eds.), *Probabilities and conditionals*. New York: Oxford University Press.
Milne, P. (1997). Bruno de Finetti and the logic of conditional events. *The British Journal for the Philosophy of Science, 48*, 195–232. doi:10.1093/bjps/48.2.195.
Nute, D. (1975). Counterfactuals and the similarity of worlds. *The Journal of Philosophy, 72*, 773–778. doi:10.2307/2025340.
Pollock, J. (2002a). The logical foundations of means-end reasoning. In R. Elio (Ed.), *Common sense, reasoning and rationality*. Oxford: Oxford University Press.
Pollock, J. (2002b). Rational choice and action omnipotence. *The Philosophical Review, 111*, 1–23.
Pollock, J. (2006). *Thinking about acting*. Oxford: Oxford University Press.
Popper, K. (1959). *The logic of scientific discovery*. New York: Harper and Row.
Pylyshyn, Z. (Ed.). (1986). *The Robot's dilemma*. New Jersey: Norwood.
Quine, W. V. O. (1950). *Methods of logic*. New York: Holt.
Raiffa, H. (1968). *Decision analysis*. Reading: Addison-Wesley.
Ramsey, F. P. (1931). General propositions and causality. In R. Braithwaite (Ed.), *The foundations of mathematics and other logical essays*. New York: Harcourt Brace.
Resnik, M. (1987). *Choices: An introduction to decision theory*. Minneapolis: University of Minnesota Press.
Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. New York: Prentice Hall.
Salmon, W. (1990). Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. Wade Savage (Ed.), *Minnesota studies in the Philosophy of Science* (Vol. 14, pp. 175–204). Minneapolis: University of Minnesota Press.
Savage, L. (1954/1972). *The foundations of statistics* (2nd ed.). New York: Dover Publications.

Shaffer, M. (2000). *Idealization and empirical testing*. University of Miami Ph.D. dissertation, Miami.
Shaffer, M. (2001). Bayesian confirmation of theories that incorporate idealizations. *Philosophy of Science, 68*, 36–52. doi:10.1086/392865.
Shaffer, M. (2002). Coherence, justification, and the AGM theory of belief revision. In Y. Bouchard (Ed.), *Perspectives on coherentism* (pp. 139–160). Ontario, Canada: Aylmer-Éditions du Scribe.
Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review, 40*, 129–138. doi:10.1037/h0042769.
Shoham, Y. (1988). *Reasoning about change*. Cambridge: The MIT Press.
Stalnaker, R. (1970). A theory of conditionals. In W. L. Harper, et al. (Eds.), *Ifs* (pp. 41–55). London: Blackwell Publishing.
Stalnaker, R. (1981). A defense of conditional excluded middle. In W. Harper, et al. (Eds.), *Ifs* (pp. 87–104). Dordrecht: D. Reidel.
Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy, 12*, 133–163.
Todd, P. M. (1999). Simple inference heuristics versus complex decision machines. *Minds and Machines, 9*, 461–477. doi:10.1023/A:1008335515764.
Weirich, P. (2004). *Realistic decision theory*. New York: Oxford University Press.