John Cantwell, *Synthese*, 2012, DOI : 10.1007/s11229-012-0197-5.

Conditionals in Causal Decision Theory

John Cantwell

Abstract

This paper explores the possibility that causal decision theory can be formulated in terms of probabilities of conditionals. It is argued that a generalized Stalnaker semantics in combination with an underlying branching time structure not only provides the basis for a plausible account of the semantics of indicative conditionals, but also that the resulting conditionals have properties that make them well-suited as a basis for formulating causal decision theory.

> *Decision theory (at least if we omit the frills) is not an esoteric science, however unfamiliar it may seem to an outsider. Rather it is a systematic exposition of the consequences of certain well- chosen platitudes about belief, desire, preference and choice. It is the very core of our common-sense theory of persons, dissected out and elegantly systematized.* (David Lewis (1974), p.337)[1]

> *A small distortion in the analysis of the conditional may create spurious problems with the analysis of other concepts. So if the facts about usage favor one among a number of subtly different theories, it may be important to determine which one it is.* (Robert Stalnaker (1980),p.87)

1 Introduction

David Lewis' work on causal decision theory and on conditionals was each in its own right ground-breaking, but he never succeeded in unifying the two by formulating causal decision theory in terms of conditionals (such as *If I do A (if I were to do A), the outcome will be (would be) o*). Instead Lewis developed

[1]I want to thank the participants at the 2007 Synthese conference devoted to David Lewis for valuable comments and suggestions.

his version of causal decision theory in terms of *causal dependency hypotheses*, a technical notion that, while adequate for its purpose, puts a certain distance between decision theory and the ordinary language we – as Lewis puts it – use to express the platitudes of decision theory. The aim of this paper is to investigate whether this gap can be closed.

Clearly there is no general reason to think that a systematic exposition of anything should be easily and precisely restateable in everyday language. But to the extent that decision theory can be viewed as a theory of what constitutes a coherent combination of certain propositional attitudes (degrees of belief, preference, intentions to act) and to the extent that we take agents to actually manage to reason their way to correct decisions, it is a reasonable question whether we can formulate a decision theory by invoking attitudes towards familiar entities such as conditionals (or the propositions that they express) rather than theory laden concepts such as possible worlds or causal dependence hypotheses, and without reference to 'unfamiliar' propositional attitudes such as *imaged* degrees of belief (c.f. Lewis (1981); Joyce (1999)). It will be argued that if some care is taken to the choice of interpretation of conditionals (recalling Stalnaker's injunction above that small distortions in the analysis of conditionals can give rise to spurious problems elsewhere), they can be invoked in the formulation of causal decision theory. It will argued – somewhat summarily – that the proposed interpretation of conditionals has linguistic motivation: one plausible way of interpreting our actual use of a certain kind of conditional results in a conditional that can be used in the formulation of causal decision theory, thus bringing our platitudes of decision theory a bit closer to a systematic exposition. It will also be argued that the underlying model provides a plausible refinement of standard causal decision theory and helps us to better address some of the criticisms that have been leveled at it (see in particular Section 5.2 below).

Section 2 gives some background and an informal exposition of the proposal; Section 3 contains a semi-formal sketch of the semantic structures on which the analysis of the conditional is based, Section 4 shows how the resulting semantics can be incorporated into a probabilistic representation of credal states and Section 5 contains a sketch of how a conditional with such a semantics can be integrated into causal decision theory.

2 Background

2.1 Indicative and subjunctive conditionals

Distinguish metaphysically *live* possibility—the future is open, many futures are possible—from metaphysically *dead* possibility—some futures that once were

live possibilities, are no longer live. *Epistemic* possibility—a possibility that is consistent with what I, or you, we or they know—cuts across this distinction; something may, for all you know, be a live possibility, but in actual fact be a dead possibility.

The English language, and several related languages, is well equipped to distinguish the live possibilities from the dead and both of these from the epistemic possibilities. The thesis pursued in this paper is that modal and conditional claims in the *indicative mood*, claims like "It is possible that the coin will land heads", "If Jane jumps from the cliff she will die" and "It is possible that the coin landed heads", semantically discern among live possibilities. By contrast, modal and conditional claims in the *subjunctive mood*, claims like "The coin could have landed heads" and "If Jane had jumped from the cliff she would have died" semantically discern among the possibilities that are dead or live (the subjunctive conditionals thus semantically discern among a larger space of possibilities than the indicative conditionals).

The thesis that the indicative mood is the key linguistic device used to speak of live possibilities while the subjunctive mood is a device to speak of dead possibilities, is a thesis of semantics not of metaphysics. If there are no metaphysically dead possibilities and if the semantic thesis is correct, then many claims that we hold plausible, for instance the subjunctive "Back in 1942 it was still possible that Hitler would win the war", are false (if Hitler's winning the war was not a live possibility in 1942, then it isn't the case that in 1942 it was still possible that Hitler would win the war). The semantic thesis does not, however, by itself entail any metaphysics.

Some deny that possibility claims and conditional claims in the indicative mood (henceforth 'indicatives') even have truth values (Edgington (1995); Levi (1996); Bennett (2003)), others think they have truth values, but that they speak of and discern among epistemic possibilities (Stalnaker (1975); Egan (2007a); MacFarlane (2011)), still others think that the indicative conditional is truth functional and that it has the truth conditions of the material conditional (Jackson (1979); Lewis (1986)). The thesis that indicatives speak of live possibilities is in opposition to all of these views, but is perhaps closest in spirit to the material analysis: it implies that indicatives have truth values and that their truth conditions do not depend on the epistemic state of anyone.

A claim like "If I flip the coin it will land heads" is verifiable: I just flip the coin and see what happens. True, the claim may not be verifiable at the very moment of utterance, it speaks of the future and like almost any claim about the future we have to wait and see before the matter can be settled. But once the coin is flipped all we have to do is wait and see, and we will soon discover whether the claim was true or not. So I take this to be be correct: the claim "If I flip the coin it will land heads" (proposition expressed by the utterance of the conditional) is

falsified—discovered to be *false*—when the coin is tossed and found to land tails, but is verified—discovered to be *true*—when the coin is tossed and found to land heads.

The difference between the indicative and subjunctive conditional shows up most clearly when we speak of possibilities that may or may not be dead, and typically this occurs when we are speaking of the past. Thus we have classic couples like:

(1) If Oswald didn't murder Kennedy, someone else did. (Past indicative)

(2) If Oswald hadn't murdered Kennedy, someone else would have. (Past subjunctive)

I would analyze the difference as follows. When looking towards the past there is only one live possibility – there is only one past – (although we may not know which this is), but the past is replete with now dead possibilities; as indicative conditionals (such as (1)), as opposed to subjunctive conditionals (such as (2)), are semantically sensitive only to the live possibilities, this difference can be used to explain why the past indicative and the past subjunctive say very different things. Before the time of the murder it was possible that the murder wouldn't occur, indeed that Kennedy would not be murdered at all, which is why (2) is false; but this possibility is now gone and so is inaccessible to the semantic evaluation of (1). The truth value of (1) is more problematic; no one doubts that if it has a truth value, then it is true, but there is a tradition (see e.g.Belnap (1970); Bradley (1998)) in which indicative conditionals with determinately false antecedents are taken to lack truth value. Within this tradition, semantics alone is not taken to be sufficient to account for the assertability conditions of indicative conditionals (the linguistic data is generally taken to suggest that the assertability of an indicative conditional $A \rightarrow B$ in the past tense goes with its conditional probability $Pr(B \mid A) = Pr(A \wedge B)/Pr(A)$; this is sometimes referred to as *Adam's thesis* after Adams (1975), see also Edgington (1995); Bennett (2003)). This is also the position taken in this paper (see Cantwell (2008) for a more extensive discussion). The indicative (1) thus lacks truth value (if it was Oswald who murdered Kennedy) and its high degree of assertability derives from the high conditional probability that someone other than Oswald murdered Kennedy, given that it wasn't Oswald.

The future we typically think of as open and replete with live possibilities which is why pairs like the following apparently convey the same message:

(3) If you open the window, the room will cool. (Future indicative)

(4) If you were to open the window, the room would cool. (Future subjunctive)

It is also this feature that explains the close relationship between the pair:

(5) If Oswald doesn't murder, Kennedy some else will. (Future indicative)

(6) If Oswald hadn't murdered Kennedy, someone else would have. (Past subjunctive)

The future indicative (5), as uttered before the murder of Kennedy, would be true if in every then live possibility in which Oswald doesn't murder Kennedy, someone else does. The past subjunctive (6), as uttered after the murder of Kennedy, would be true if in every possibility that was live before Oswald murdered Kennedy (but that, due to Oswald's act, is no longer live) where Oswald didn't murder Kennedy, someone else did. The future indicative and the past subjunctive speak of the same set of possibilities, the difference being that due to the passage of time the once live possibilities are now dead.

This feature of the passage of time also explains why the following pair of indicatives make so different claims even though they only involve a shift of temporal perspective:

(7) If Oswald doesn't murder, Kennedy some else will. (Future indicative)

(8) If Oswald didn't murder Kennedy, someone else did. (Past indicative)

Before the murder there were many live possibilities, but after the murder all but one (as regards the murder of Kennedy) are dead. The future (7) and past (8) indicatives express very different propositions as they are talking about very different sets of possibilities.

So the difference between indicatives and subjunctives shows up most clearly when speaking of the past, but in some cases the difference shows up when speaking of the future as well:

(9) If we are going to fire Jane at tomorrow's board meeting, we didn't raise her salary last week.

(10) If we were going to fire Jane at tomorrow's board meeting, we wouldn't have raised her salary last week.

(11) If the dam breaks within the next 24 hours, our engineers have discovered cracks in it.

(12) If the dam had been going to break within the next 24 hours, our engineers would have discovered cracks in it.

Someone who has forgotten what the board has decided about Jane's salary, or who wasn't present when that decision was made, might assert (9), while someone who knows that Jane's salary was raised could make the prediction (10). The

conditional (12) might be asserted by the proud chief-engineer at the dam site: the engineers have discovered no cracks and so there is no way that the dam will break within the next 24 hours; it may weeks ago have been a live possibility that the dam would break tomorrow, but it no longer is as it isn't cracked now. However, the indicative counterpart (11) would in this situation not be assertable, rather it would probably be replaced by something like "If the dam breaks within the next 24 hours, our engineers are less skilled than we thought".

These brief comments are only meant to provide some basic evidence in favour of the thesis that the indicative conditionals depend for their truth and falsity on the *live* possibilities (the issues involved are many and complex and a thorough treatment is beyond the scope of this paper). The connection to decision theory comes when we combine this thesis with the thesis that what is important in making a decision are the live possible outcomes, i.e. that what is important when deliberating about how to act is what *can* happen in the future, not what *could have* happened in the future.

2.2 Conditionals in decision theory

The *evidential decision theory* of Richard Jeffrey (1983) (somewhat simplified) says that one should decide on the action A that maximizes the sum:

$$\text{EEU}(A) = \sum_{o \in \mathcal{O}} \mu(o \wedge A) Pr(o|A).$$

(Here $\mathcal{O}$ is a partition of the possible outcomes, and $\mu(o \wedge A)$ is the value of the outcome o together with the action A, and $Pr(o|A)$ is the subjective probability that o given A). Combine this with the influential view commonly associated with Earnest Adams (1975) according to which the probability of an indicative conditional is its conditional probability:

Adams' Thesis $Pr(A \to B) = Pr(B|A)$.

Together we get:

$$\text{EEU}(A) = \sum_{o \in \mathcal{O}} \mu(o \wedge A) Pr(A \to B).$$

Causal decision theory grew forth as a response to the problem that the conditional probability $Pr(o|A)$ is sometimes the *wrong* probability on which to base a decision to act, for the conditional probability is not always the subjective probability that one will get the outcome o if one does A, but rather the subjective

probability that o is the outcome if one finds out that one A:ed. So what is the right probability on which to base causal decision theory? Gibbard and Harper (1978) formulated their version of causal decision theory by explicitly invoking the probability of a conditional in their definition of the *causal expected utility* of the action A:

$$\text{CEU}(A) = \sum_{o \in \mathcal{O}} \mu(o \wedge A) Pr(A \Rightarrow B).$$

The key here is that Gibbard and Harper do not assume the identity $Pr(A \Rightarrow B) = Pr(B|A)$ to hold, indeed they explicitly invoke a subjunctive reading of their conditional with a semantics that guarantees that Adams' Thesis will not in general be satisfied.

I think the choice of a subjunctive reading of the conditional was unfortunate. For instance in Newcomb's problem—one of the problems that first triggered the need for a causal as opposed to evidential decision theory—the fact that subjunctive conditionals can invoke dead possibilities becomes a major source of confusion. We *are* entitled to say things like "If the future had been such-and-such the past would or might have been different from the actual past" (compare sentences (10) and (12) above), thus we cannot out of hand dismiss conditionals like "If I were to choose both boxes the predictor might have predicted this and not placed money in the second box" (suggesting that if the future would have been different, the past would have been different too) . Indeed, this backtracking interpretation has been a response from those who have resisted 'two-boxing' in Newcomb's problem Horgan (1981). There is no corresponding problem for indicative conditionals, we cannot say "If the future is such-and-such, the past is or might be different from the actual past".

A fundamental problem with Gibbard and Harper's account is that their conditional is based on Robert Stalnaker's (1968) first account of conditionals (I will refer to this as the 'simple Stalnaker semantics') that involves the implausible assumption that for any world w and proposition A there is precisely one A-world that is most similar to w. Without this assumption the account breaks down. The alternative of using a Lewis-style interpretation, allowing that there may be multiple A-worlds that are maximally similar to w, does not allow us to formulate a coherent decision theory. On Lewis' (1973) account it can happen that $A \Rightarrow o_1$ and $A \Rightarrow o_2$ are both false while $A \Rightarrow (o_1 \vee o_2)$ is true. This means that (even for logically exclusive o_1 and o_2) we will not in general have the identity

$$Pr(A \Rightarrow (o_1 \vee o_2)) = Pr(A \Rightarrow o_1) + Pr(A \Rightarrow o_2).$$

As a result any decision theory based on Lewis' interpretation of the conditional would be intolerably sensitive to how the outcomes are partitioned.

Lewis' own formulation of causal decision theory is not made in terms of conditionals, but he keeps the door ajar for a conditional based formulation of decision theory. Supposing that there are such things as objective chances for particular events and not only long term frequencies of types of events, Lewis suggests that one can build causal decision theory on judgements such as "If A were the case, the chance that o would be r". This gives us a notion of 'Chance-based Causal Expected Utility' along the lines:

$$\text{CCEU}(A) = \sum_{o \in \mathcal{O}} \mu(o) \sum_{0 \leq r \leq 1} r \times Pr(A \Rightarrow (Ch(o) = r)).$$

A problem for Lewis, however, is that there is nothing in his account of the semantics of the subjunctive conditional that suggests that in all the closest A-worlds the chance of a particular outcome is the same. So if $Ch(o) = .4$ is true in one of the closest A-worlds, while $Ch(o) = .5$ is true in another, then $A \Rightarrow (Ch(o) = .4)$ and $A \Rightarrow (Ch(o) = .5)$ are both false. Lewis is forced to assume, "not without misgivings" (Lewis (1981),p.27), that this cannot happen. In the absence of a plausible account of the semantics of conditionals that backs Lewis' assumption, one may very well suspect that it amounts to no more than the assumption that there is always precisely one most similar A-world.

Having come thus far we may conclude that the ambition of finding a conditional based formulation of decision theory is in tatters. However, I will argue that there is still room for a conditional based formulation of decision theory if one invokes what might be called a *generalized Stalnaker semantics* following Stalnaker's (1980) refined semantic analysis of the conditional. A generalized Stalnaker semantics invokes a *set* of simple Stalnaker-style selection functions. On the revised semantics a conditional is *determinately* true (false) if it is true (false) with respect to *every* simple Stalnaker-style selection function in the set (in line with van Fraassen's super-valuational account to deal with vagueness). The generalized Stalnaker semantics avoids stipulating that there is a determinate unique most similar A-world to any world; in fact, it provides more structure than a standard Lewis-style semantics (every Lewis-style selection function can be represented as a set of Stalnaker-style selection function, but the converse relation does not hold).

David Lewis (1981) argues that a conditional based on a generalized Stalnaker semantics will be unsuitable as a basis for a decision theory. In particular, he argues that in a chancy world, the account will break down. Say that it is (objectively) indeterminate whether the coin, if tossed, will land heads or tails. Then, according to Lewis, the conditional "If the coin is flipped it will land heads" is, as he puts it, *flatly and determinately false* (p.26). If Lewis is right then, of course,

Stalnaker's generalized semantics is incorrect as it leaves the conditional with an indeterminate truth value (it is neither determinately true nor determinately false).

I do not think we need – or should – follow Lewis on this semantic issue. Clearly, in an indeterministic world it would be inappropriate to *assert* the conditional "If the coin is flipped it will land heads", but this can be explained by the fact that in an indeterministic world it would be impossible to know that the conditional is true before the coin is flipped (thus, to assert the conditional would be to violate the knowledge norm of assertion). However, if the coin is heavily biased towards heads one might venture the *prediction* or *educated guess* "If the coin is flipped it will land heads" and be fairly confident that the prediction will end up true, even though it is not determinately true prior to the flip of the coin. One needs, of course, to make sense of the idea that an educated guess can turn out to be correct, even though the sentence was not determinately true when it was uttered. But this does not appear to be an impassable obstacle (see Belnap, Perloff, and Xu (2001) for an extensive discussion on the posterior evaluation of claims about the future). An earlier utterance/guess/prediction of the conditional "If A will be the case, then B will be the case" turns out to be correct now if it *was* true in every *now* live (as opposed to every *then* live) possible world or history. In particular, assuming that the coin was tossed and turned up heads, the previously live but now dead possibility that the coin would be flipped and land tails is not relevant when assessing the correctness of the earlier guess/prediction "If the coin is flipped it will land heads".

Again, the issues are many and complex and in need of a more systematic treatment, but I think a case can be made that Stalnaker's generalized semantics comes with the additional semantic structure needed to incorporate it in a formulation of causal decision theory and that it also yields a plausible semantic account of conditionals.

3 The framework semi-formalized

My treatment of branching time and on how best to build a semantics on such a structure, is heavily indebted to the extended treatment given in Belnap, Perloff and Xu's *Facing the future* (2001). They provide extensive justifications for a number of the delicate methodological choices that need to be made in such a treatment, justifications that are omitted here as is most of the technical detail.

3.1 Basic concepts of branching time

A *moment* is a 'snap shot' of the world (the state of the world at a moment) and a *history* is a linearly ordered set of moments. Let $\mathcal{M}$ be the set of moments and

$\mathcal{H}$ be the set of histories. Histories can share moments and then branch off into the future, but it is assumed that if two histories share the same moment, they also share the same past as of that moment (so the past is determinate). Figure 1 illustrates the basic concepts .

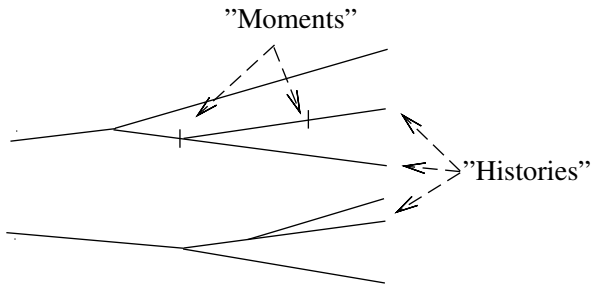

Figure 1: Basic concepts of the branching time framework. The histories evolve from left to right.

Let Alt(m) denote the set of histories that share the moment m, these are the alternative ways in which the world can evolve at the moment m – the *live* histories at m.

A selection function *Sel* takes a history h and a moment m and a set H of histories and, if defined on H, returns an element of H. Let $\mathcal{S}$ denote a set of selection functions. It is assumed that selection functions satisfy (for any set H of histories):

(Selection) If $\mathit{Sel}_{h,m}(H)$ is defined, then $\mathit{Sel}_{h,m}(H) \in H$.

(Centering) If $h \in H$, then $\mathit{Sel}_{h,m}(H) = h$.

(Non Vacuity) If $H \cap \text{Alt}(m)$ is non-empty, then $\mathit{Sel}_{h,m}(H)$ is defined.

(Live Histories Only) If $\mathit{Sel}_{h,m}(H)$ is defined, then $\mathit{Sel}_{h,m}(H) \in \text{Alt}(m)$.[2]

3.2 Points of evaluation

Sentences are evaluated as true or false relative to a *point of evaluation*. The set of points of evaluation will be denoted by $\mathcal{E}$. For our present needs the point at which a sentence is evaluated consists of a triple (h, m, Sel) containing a history h, a moment m, and a selection function *Sel* (selection functions will be discussed

[2]The constraint Live Histories Only is one of the main differences between the present framework and that of Thomason and Gupta (1980), where indicative conditionals are also analyzed within a branching time framework.

below) chosen from a set of selection functions $\mathcal{S}$, with the constraint that $h \in \text{Alt}(m)$.

In an indeterministic framework there may, at any given moment, be several histories passing through that moment and there is no fact of the matter that determines which history will eventually turn out to be the actual one (at the moment, there is no one 'actual' history, only several *possible* histories). Thus at any given moment it will not be possible to uniquely determine the point of evaluation relative to which an utterance or a claim is to be evaluated: claims may lack *settled* truth values.

A is *settled true* at m if and only if for every $h' \in \text{Alt}(m)$ and $\mathit{Sel}' \in \mathcal{S}$: A is true at (h, m, Sel').

A is *settled false* at m if and only if for every $h' \in \text{Alt}(m)$ and $\mathit{Sel}' \in \mathcal{S}$: A is false at (h, m, Sel').

3.3 Truth a point of evaluation

The main language that we will be dealing with consists of propositional atoms $p, q, r, \ldots$, closed under the standard sentential connectives $\neg$, $\wedge$, and $\vee$. In addition I will assume that the language is closed under the operators $\Diamond A$ ('it is (historically) possible that A') with its complement, $\Box A$, defined $\neg\Diamond\neg A$, as well as the conditional $A \rightarrow B$.

Let V be a valuation function that assigns a set of history-moment pairs to each propositional atom p (propositional atoms are taken to be true or false independently of the space of alternate histories and of the selection function). For any point of evaluation $e = (h, m, \mathit{Sel})$:

p is *true at* e if and only if $(h, m) \in V(p)$.

p is *false at* e if and only if $(h, m) \notin V(p)$.

$\neg A$ is *true at* e if and only if A is false at e.

$\neg A$ is *false at* e if and only if A is true at e.

$A \wedge B$ is *true at* e if and only if both A and B are true at e.

$A \wedge B$ is *false at* e if and only if either A or B is false at e.

$\Diamond A$ is *true at* e if and only if there is a $h' \in \text{Alt}(m)$ such that A is true at (h', m, Sel).

$\Diamond A$ is *false at* e if and only if for every $h' \in \text{Alt}(m)$: A is false at (h', m, Sel).

Let $H(A, m, \textit{Sel})$ denote the set of histories at which A is true at the moment m (relative to the selection function *Sel*), i.e.:

$$H(A, m, \textit{Sel}) = \{h \mid A \text{ is true at } (h, m, \textit{Sel})\}.$$

I will use the notation $\textit{Sel}_{h,m}(A)$ to denote $\textit{Sel}_{h,m}(H(A, m, \textit{Sel}))$.

For any point of evaluation $e = (h, m, \textit{Sel})$, let the A-image of e be:

$$e(A) = (\textit{Sel}_{h,m}(A), m, \textit{Sel}),$$

when $\textit{Sel}_{h,m}(A)$ is defined, otherwise $e(A)$ is also undefined.

Note that if $e(A)$ is defined, then A is true at $e(A)$; for A is true at $(h', m, \textit{Sel})$ for every $h' \in H(A, m, \textit{Sel})$, and $\textit{Sel}_{h,m}(A)$, if defined, selects an element of $H(A, m, \textit{Sel})$.

We can now give the truth conditions for the indicative conditional:

$A \rightarrow B$ is *true at* e if and only if $e(A)$ is defined and B is true at $e(A)$.

$A \rightarrow B$ is *false at* e if and only if $e(A)$ is defined and B is false at $e(A)$.

Note that conditionals can on this account lack truth value. If the antecedent of a conditional is *settled* false so that it is false at every live alternative history, then the conditional lacks truth value: not only does it not have a *settled* truth value at the given moment, it has no truth value even relative to the given point of evaluation. The thesis that conditionals with a settled false antecedent lacks truth value is a semantic thesis and has nothing to do with the indeterminacy of the future. Indicative conditionals allow us to discriminate among live histories and explore what is or will be the case in those live histories where the antecedent is true; if there are no such live histories, the conditional lacks truth value.

Note also that when A and B have settled truth values, $A \rightarrow B$ is a truth functional connective.

THEOREM 1
If A and B have settled truth values at moment m and $e = (h, m, \textit{Sel})$ is a point of evaluation, then:

1. $A \rightarrow B$ is true at e if and only if $A \wedge B$ is true at e,
2. $A \rightarrow B$ is false at e if and only if $A \wedge \neg B$ is true at e,
3. $A \rightarrow B$ lacks truth value if and only if A is false at e.

Proof: Assume that A and B have settled truth values at m. (1) Assume that $A \wedge B$ is true at $e = (h, m, \textit{Sel})$, thus both A and B are true at e. As A has a settled truth value at m it is settled true at m; so $H(A, h, m) = \text{Alt}(m)$. By Centering:

$\mathit{Sel}_{h,m}(A) = h$. So as B is true at (h, m, Sel), B is true at $(\mathit{Sel}_{h,m}(A), m, \mathit{Sel})$. So $A \rightarrow B$ is true at (h, m, Sel).

Assume that $A \rightarrow B$ is true at $e = (h, m, \mathit{Sel})$. Thus $e(A)$ is defined and B is true at $(\mathit{Sel}_{h,m}(A), m, \mathit{Sel})$. But then $\mathit{Sel}_{h,m}(A)$ is defined, so there is an $h' \in \text{Alt}(m)$ such that A is true at (h', m, Sel). As A' and B's truth are settled at m, both A and B are true at (h, m, Sel), thus $A \wedge B$ is true at (h, m, Sel).

The proofs of (2) and (3) are analogous.

□

So, in particular, when A and B are claims about the past and thus have a settled truth value, the conditional $A \rightarrow B$ has the same truth value as B when A is true, and lacks truth value if A is false.

4 Representing credal states

A standard way of representing the credal state of an agent is by means of a probability measure on the language, that is, a function Pr from sentences of the language to the real numbers, satisfying, for any truth determinate sentences A and B (sentences that are either true or false at every point of evaluation):

1. $0 \leq Pr(A) \leq 1$.
2. If A and B are true (false) at the same points of evaluation then $Pr(A) = Pr(B)$.
3. $Pr(\neg A) = 1 - Pr(A)$.
4. $Pr(A \vee B) = Pr(A) + Pr(B)$, if there is no point of evaluation where A and B are both true.

The present language, however, has conditional sentences that may be neither true nor false at a point of evaluation. Thus the above list of axioms need to be supplemented by the following law of *non-bivalent* probability, the law that the probability of A is the probability that A is true *given that it has a truth value*:[3]

5. $Pr(A) = Pr(Tr(A))/Pr(Tv(A))$, provided $Pr(Tv(A)) > 0$.

This axiom requires that the language contains the truth operator $Tr(A)$ ('It is true that A') and the truth-value operator ('It is either true or false that A').[4]

[3] See Cantwell (2006) for a Dutch book argument supporting the claim that (1-5) are laws of non-bivalent probability.

[4] I.e. $Tr(A)$ is true at e iff A is true at e; $Tr(A)$ is false at e iff A is not true at e; $Tv(A)$ is true at e iff A is true or false at e; $Tv(A)$ is false at e iff A is neither true nor false at e.

However, my main interest here is a slightly less general way of representing credal states: by assignments of probabilities to points of evaluation (probability masses). I assume that one can represent the credal state of an agent by means of a real-valued function d taking the set of points of evaluations as its domain, with the restriction that $\sum_{e \in \mathcal{E}} d(e) = 1$. Intuitively, $d(e)$ represents the agent's degree of belief that e is the correct point of evaluation. The *measure of subjective probability* based on d is then defined:

$$Pr_d(A) = \frac{\sum\{d(e)|A \text{ is true at } e\}}{\sum\{d(e)|A \text{ has a truth value at } e\}}.$$

Whenever $\sum\{d(e)|A \text{ has a truth value at } e\} = 0$, $Pr_d(A)$ is undefined.

It is trivial to show that $Pr_d(A)$ satisfies the above axioms (1-5).

4.1 Generalized Imaging

The *image of* d *under* A, in symbols $d * A$, defined when $Pr_d(\Diamond A) > 0$, is a function that takes a probability mass d and a sentence A and yields a new probability mass:[5]

$$(d * A)(e) =_{df} \frac{\sum\{d(e') \mid e'(A) = e\}}{\sum\{d(e') \mid e'(A) \text{ is defined}\}}$$

That is, we get $d * A$ by 'moving' the probability of each point of evaluation e' to its image $e'(A)$ and normalizing (to make sure that $d * A$ sums to 1).

Taking the image of d under A induces a distinct form of imaging relative to Pr_d:

$$Pr_d(B \,||\, A) =_{df} Pr_{d*A}(B).$$

Imaging was first introduced by Lewis (1976) who applied it to a simple Stalnaker semantics for conditionals. Gärdenfors (1986) later generalized this to a Lewis-style semantics by adding extra structure (see also Joyce (1999)). In the present case the added structure is obtained by emplyoing a generalized Stalnaker semantics (with sets of singleton selection functions) to generate the image. The following result, which we do not in general obtain for a Lewis-style semantics, illustrates that this added structure is not in vain:

THEOREM 2

A. The probability $Pr(A \rightarrow B)$ is defined if and only if the probability $Pr(B \,||\, A)$ is defined.

[5]Note that when $Pr_d(\Diamond A) > 0$ there is some $e = (h, m, \mathit{Sel})$ such that $\Diamond A$ is true at e and $d(e) > 0$. As $\Diamond A$ is true at e, $e(A)$ is defined (non-vacuity), hence $\sum\{d(e') \mid e'(A) \text{ is defined}\} > 0$.

B. If the probabilities $Pr(A \rightarrow B)$ and $Pr(B \,||\, A)$ are both defined, then $Pr(A \rightarrow B) = Pr(B \,||\, A)$.

Proof:

(A) $Pr(A \rightarrow B)$ is defined if and only if $\sum\{d(e) \mid A \rightarrow B$ has a truth value at $e\} > 0$ if and only if $\sum\{d(e) \mid B$ has a truth value at $e(A)\} > 0$ if and only if $\sum\{(d * A)(e) \mid B$ has a truth value at $e\} > 0$ and $\sum\{(d * A)(e) \mid e(A)$ is defined$\} > 0$ if and only if $Pr(B \,||\, A)$ is defined.

(B) Note:

$$Pr(B \,||\, A) = \frac{\sum\{(d * A)(e) \mid B \text{ is true at } e\}}{\sum\{(d * A)(e) \mid B \text{ has a truth value at } e\}} \tag{1}$$

Note:

$$\sum\{(d * A)(e) \mid B \text{ is true at } e\} = \frac{\sum\{d(e) \mid B \text{ is true at } e(A)\}}{\sum\{d(e) \mid e(A) \text{ is defined}\}} \tag{2}$$

Note:

(3)

$$\sum\{(d*A)(e) \mid B \text{ has a truth value at } e\} = \frac{\sum\{d(e) \mid B \text{ has a truth value at } e(A)\}}{\sum\{d(e) \mid e(A) \text{ is defined}\}}$$

Thus (as (2) and (3) have a common denominator):

$$Pr(B \,||\, A) = \frac{\sum\{d(e) \mid B \text{ is true at } e(A)\}}{\sum\{d(e) \mid B \text{ has a truth value at } e(A)\}} \tag{4}$$

Note:

$$Pr(A \rightarrow B) = \frac{\sum\{d(e) \mid A \rightarrow B \text{ is true at } e\}}{\sum\{d(e) \mid A \rightarrow B \text{ has a truth value at } e\}} \tag{5}$$

That is:

$$Pr(A \rightarrow B) = \frac{\sum\{d(e) \mid B \text{ is true at } e(A)\}}{\sum\{d(e) \mid B \text{ has a truth value at } e(A)\}} \tag{6}$$

Thus (combining (4) and (6)):

$$Pr(B \,||\, A) = Pr(A \rightarrow B) \tag{7}$$

□

This result establishes a connection between 'traditional' formulations of causal decision theory (e.g. Joyce's (1999) formulation of causal decision theory – arguably the most developed to date – makes explicit use of imaging probabilities) and the one that will be presented below.

4.2 Adams' Thesis

As noted in Section 2.1, the linguistic data is generally taken to suggest that the assertability of an indicative conditional $A \rightarrow B$ in the past tense goes with its conditional probability $Pr(B \mid A)$ (*Adam's thesis*). The linguistic data accords with the present account, for we have:

THEOREM 3
If A and B contain no conditionals and $Pr(A) = Pr(\Box A) > 0$, then $Pr(A \rightarrow B) = Pr(B \mid A)$.

Proof: Assume that $Pr(A) = Pr(\Box A) > 0$. Note that as A is non-conditional it has a truth value at every point of evaluation, as does $\Box A$. Note also that if $\Box A$ is true at e, then A is true at e. It follows that at every point of evaluation such that $d(e) > 0$, if A is true at e, then $\Box A$ is true at e. Thus, $X = \{e \mid A \rightarrow B \text{ is true at } e\}$ is the set of points of evaluation where A and B are both true, i.e. $X = \{e \mid A \wedge B \text{ is true at } e\}$. Furthermore, as B is non-modal (and has a truth value at every point of evaluation) $Y = \{e \mid A \rightarrow B \text{ has a truth value at } e\}$ is the set of points of evaluation where A is true (for if A is false at e, then so is $\Box A$, and so $e(A)$ is undefined), i.e. $Y = \{e \mid A \text{ is true at } e\}$. So (as $Pr(A) > 0$ and hence $Pr(\Diamond A) > 0$):

$$Pr(A \rightarrow B) = \frac{\sum_{e \in X} d(e)}{\sum_{e \in Y} d(e)} = \frac{Pr(A \wedge B)}{Pr(A)} = Pr(B \mid A).$$

□

COROLLARY 1
If A and B contain no conditionals and $Pr(A) = Pr(\Box A) > 0$, then $Pr(B \mid\mid A) = Pr(B \mid A)$.

By corroborating Adams' thesis, this result partially vindicates the claim that the proposed semantics for the indicative conditional has linguistic plausibility; for claims about the past have settled truth values, and so the probability for any conditional with a past tensed antecedent should satisfy Adams' Thesis. This result, which seemingly is at odds with Lewis' (1976) famous 'triviality' or 'impossibility' result, is made possible because the underlying probability measure is a measure on propositions that need not be bivalent.

The result has the further implication that if the world is deterministic and so both past and future tensed sentences have settled truth values, Adams's thesis applies to future tensed sentences as well (this will have repercussions in Section 5.2).

5 Outlines of a Causal Decision Theory

5.1 Causal Expected Utility

One should first note that our probability measure has a property that makes it suitable to serve as the basis for a causal decision theory.

A finite set of *categorical* (non-conditional) sentences $\mathcal{O}$ is a *logical partition* if (i) for any $o, o' \in O$, if $o \neq o'$, then there is no point of evaluation in which they are both true, and (ii) for any point of evaluation some element of $\mathcal{O}$ is true at that point.

THEOREM 4
If $\mathcal{O} = \{o_1, \ldots, o_n\}$ is a logical partition and $Pr(A) > 0$:

1. $Pr(A \rightarrow (o_i \vee o_j)) = Pr(A \rightarrow o_i) + Pr(A \rightarrow o_j)$, if $i \neq j$.
2. $Pr(A \rightarrow o_1) + \cdots + Pr(A \rightarrow o_n) = 1$.
3. $Pr(A \rightarrow \neg o_i) = 1 - Pr(A \rightarrow o_i)$.

Proof: This follows directly from the fact that $Pr_d(. \,||\, A)$ is a non-bivalent probability measure and that $Pr_d(A \rightarrow B) = Pr_d(B \,||\, A))$.

□

Let $\mathcal{O}$ be a logical partition and assume that μ is a real valued utility function with categorical sentences as its domain. Intuitively, the elements of $\mathcal{O}$ are the *possible outcomes* and $\mu(o)$ assigns the numerical utility of the outcome o. Let $\mathcal{A}$—the *available actions*—be a logical partition. We can define:

For any $A \in \mathcal{A}$ such that $Pr(A) > 0$:

$$\text{CEU}(A) = \sum_{o \in \mathcal{O}} Pr(A \rightarrow o)\mu(o \wedge A).$$

Or, equivalently:

For any $A \in \mathcal{A}$ such that $Pr(A) > 0$:

$$\text{CEU}(A) = \sum_{o \in \mathcal{O}} Pr(o \,||\, A)\mu(o \wedge A).$$

This should be contrasted with Evidential Expected Utility defined:

For any $A \in \mathcal{A}$ such that $Pr(A) > 0$:

$$\text{EEU}(A) = \sum_{o \in \mathcal{O}} Pr(o \,|\, A)\mu(o \wedge A).$$

5.2 Applying the theory

Through Theorem 2 (that relates the probability of a conditional to the corresponding imaged probability) we have realized the project of formulating causal decision theory in terms of probabilities of conditionals. However, one of the main themes of this study has been the added semantic structure introduced by the distinction between 'live' and 'dead' possibilities where the truth and falsity of the indicative conditionals (the conditionals used to formulate the decision theory) are taken only to depend on the former. This added structure makes the resulting causal decision theory more fine grained allowing for non-equivalent representations of familiar decision problems. In particular, it becomes sensitive to the distinction whether acts are taken to be *metaphysically open* (i.e. indeterministic) or whether the 'openness' of how one will act is due to mere ignorance of what one will do (i.e. deterministic).

Consider two different representations of Newcomb's problem (Nozick (1969)) as seen in Figure 2 and Figure 3. In the first representation (Figure 2)

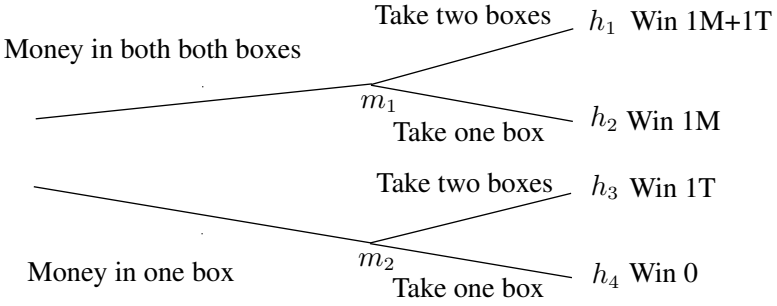

Figure 2: The indeterministic version of Newcomb's problem

the choice of taking one box ($T1$) or two boxes ($T2$) is represented as an indeterministic event; for at both m_1 and m_2 both choices are live possibilities and neither $T1$ nor $T2$ have determinate truth values. By contrast, at the time of choice the Predictor has either already placed money in both boxes or already placed money in only one box (thus "There is money in both boxes" has a determinate truth value at both m_1 and m_2). When represented in this way the current formulation of casual decision theory recommends taking two boxes. For here are the four relevant points of evaluation and the conditionals that hold true at them (reference to the selection function is omitted):

(m_1, h_1) $T1 \rightarrow \$1M$ (true); $T2 \rightarrow (\$1M + \$1T)$ (true).

(m_1, h_2) $T1 \rightarrow \$1M$ (true); $T2 \rightarrow (\$1M + \$1T)$ (true).

(m_2, h_3) $T1 \rightarrow \$0$ (true); $T2 \rightarrow \$1T$ (true).

(m_2, h_4) $T1 \rightarrow \$0$ (true); $T2 \rightarrow \$1T$ (true).

Given these it follows from simple dominance reasoning that taking two boxes ($T2$) is the better option; for "If I take one box, it will contain one million" has the same probability as "If I take two boxes, it will contain one million", just as the conditionals "If I take one box, it will be empty" has the same probability as "If I take two boxes, one will be empty".

The second representation (Figure 3) is quite different however. Here the

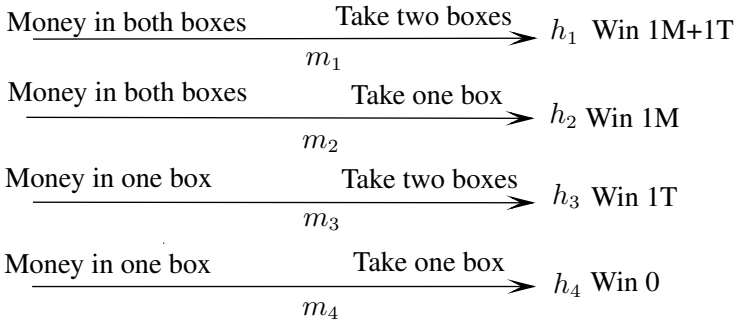

Figure 3: The deterministic version of Newcomb's problem

choice of taking one or two boxes is represented as a deterministic event. The second representation also has four relevant points of evaluation, but they differ in the truth values assigned to the conditionals:

(m_1, h_1) $T1 \rightarrow \$1M$ (lacks truth value); $T2 \rightarrow (\$1M + \$1T)$ (true).

(m_2, h_2) $T1 \rightarrow \$1M$ (true); $T2 \rightarrow (\$1M + \$1T)$ (lacks truth value).

(m_3, h_3) $T1 \rightarrow \$0$ (true); $T2 \rightarrow \$1T$ (lacks truth value).

(m_4, h_4) $T1 \rightarrow \$0$ (lacks truth value); $T2 \rightarrow \$1T$ (true).

Here there is no dominance argument, but given standard Newcomb probabilities for these four points of evaluation (the first and fourth points of evaluation have very low probabilities, the second and third have very high probabilities) we find that the current version of causal decision theory recommends (just as evidential decision theory recommends) taking only one box. For then "If I take one box, it will contain one million" will have a high probability, while "If I take two boxes, the oblique box will contain one million" will have a low probability.

We find the same duality in the examples that Andy Egan (2007b) has put forward as counterexamples to causal decision theory. Here is the 'Psychopath Button' example:

> *Paul is debating whether to press the 'kill all psychopaths' button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world with psychopaths to dying. Should Paul press the button?* (p.97)

We can represent Paul's predicament either as in Figure 4 (the indeterministic version) or as in Figure 5 (the deterministic version). Again, in both cases, there are four points of evaluation corresponding to each of the different histories. We are to assume that Paul is quite confident that he is not a psychopath and that he is quite confident that only a psychopath would press the button. Together this implies that the point of evaluation corresponding to h_4 has a very high probability (as compared to the other points of evaluation) while the point of evaluation corresponding to h_3 has a very low probability (as compared to the other points of evaluation) (the probability mass might be $d((m_1, h_1)) = .09$, $d((m_1, h_2)) = .09$, $d((m_2, h_3)) = .002$, $d((m_2, h_4)) = .8$).

For the indeterministic case (Figure 4) these probabilities imply that the con-

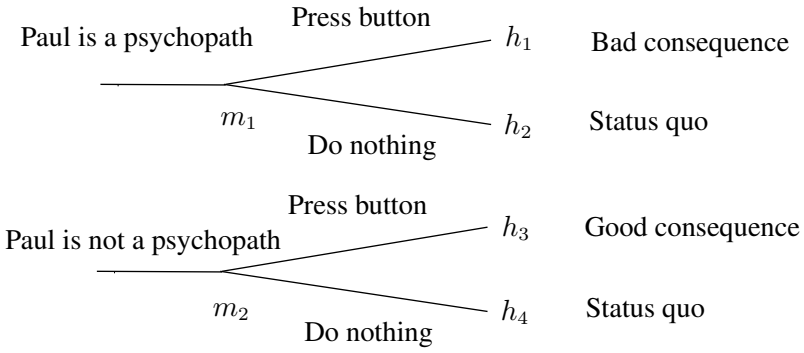

Figure 4: Indeterministic version of psychopath button.

ditional "If Paul presses the button he will get a good outcome" has a very high probability (as it is true at both (m_2, h_3) and (m_2, h_4), the latter having a very high probability) while, correspondingly, "If Paul presses the button he will get a bad outcome" has a low probability (as it is true only at (m_1, h_1) and (m_1, h_2), having low probabilities) and so, if the utility numbers are properly calibrated, the present causal decision theory will recommend pressing the button.

For the deterministic case (Figure 5) the probabilities imply that the conditional "If Paul presses the button he will get a good outcome" has a very low probability as it has a truth value only at (m_1, h_1) and (m_3, h_3) and it is false at (m_1, h_1) which has a far greater probability than (m_3, h_3). Correspondingly, the

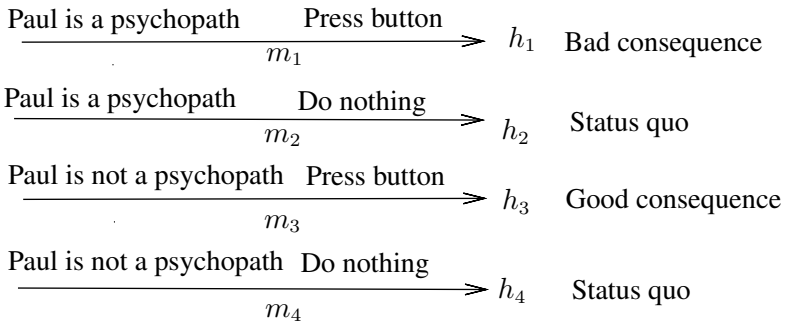

Figure 5: The deterministic version of the psychopath button.

conditional "If Paul presses the button he will get a bad outcome" has a very high probability, and so under this representation the present causal decision theory will recommend (just as evidential decision theory recommends) not pressing the button.

So, as in Newcomb's problem, the present version of causal decision theory does not uniquely recommend a course of action on the basis of a standard brief description of the decision situation – the situation is on this account underdescribed. However, if we add to the psychopath button scenario the information that Paul takes the conditional "If Paul presses the button he will get a good outcome" to have a *high* probability, this will select for the indeterministic representation and in this case the unambiguous recommendation will be to press the button.

Egan presents the psychopath button scenario as a counterexample to causal decision theory as standard causal decision theories (as opposed to the one presented here) unambiguously recommend pressing the button. It is a counterexample, he contends, as it flies in the face of our intuitions and it shows that

> causal decision theory endorses...an irrational policy of performing *the action which one confidently expects will cause the worse outcome* (p.97).
>
> [causal decision theory is] blind to features of the agents beliefs to which it should be sensitive namely, *the agents confidence that a particular course of action, if undertaken, is doomed to fail*, and bring about a worse outcome than the alternative. (p.100) [Emphasis added.]

In Cantwell (2010) I raised the issue whether these diagnoses are consistent with Egan's description of the psychopath button scenario. My worry was that if it is true that Paul is not a psychopath, then it is true that pressing the button will

cause the *best* outcome and it is true that Paul will get a *good* outcome if he presses the button. So, as Paul considers it highly likely that he is not a psychopath, he should consider it highly *un*likely that pressing the button will cause the worse outcome and highly *un*likely that he will get a bad outcome if he presses the button. But this is the exact opposite of how Egan describes the situation. My contention there, as it is here, was that if one (contrary to Egan) takes it to be a part of the description of the psychopath button scenario that Paul considers it highly likely that pressing the button will cause the best outcome and highly likely that he will get the best outcome if he presses the button, the description looses much of its intuitive force as a counterexample to causal decision theory. On the other hand, if these are *not* taken to be part of the description of the decision situation (as Egan's comments suggest) then the present causal decision theory recommends *not* pressing the button and so respects Egan's intuitions for the case.

6 Concluding remarks

By adding the structure of branching time one gets sufficient structure for a more rigorous implementation of the constraint that the conditionals used when formulating causal decision theory should not be backward tracking. It also allows us to capture the important constraint that what is important when deliberating about how to act is what *can* happen in the future, not what *could have* happened in the future. By taking the indicative conditional to be semantically linked only to the still open possibilities, one gets a linguistic carrier of the information that we wish to employ in a causal decision theory. Stalnaker's generalized semantics, in turn, guarantees that the conditional has the logical properties required for a direct implementation in causal decision theory without placing unreasonably strict constraints on the underlying similarity relation (c.f. Gibbard and Harper (1978)).

The main technical result (Theorem 2) provides a link to more developed accounts of causal decision theory (i.e. Joyce (1999)) and also provides an interpretation of generalized imaging (imaging that cannot be modeled by a single Stalnaker-style selection function) in terms of probabilities of conditionals rather than in terms of the underlying possible-worlds semantics. This is important as part of the point of formulating causal decision theory in terms of probabilities of conditionals allows for a conceptual separation of causal decision theory from an underlying metaphysics of possible worlds.

As remarked in Section 5.2 the added structure of branching time makes the decision theoretical framework more fine-grained, allowing for non-equivalent representations of familiar decision problems. The metaphysical issues here are deep and complex and well beyond the scope of a proper treatment in this paper. However, I think the results warrant the conclusion that re-introducing condition-

als into causal decision theory is not only feasible, but can also help illuminate aspects of the metaphysical commitments of causal decision theory.

References

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: Reidel.

Belnap, N. (1970). Conditional Assertion and Restricted Quantification. *Nous IV*, 1–12.

Belnap, N., M. Perloff, and M. Xu (2001). *Facing the Future*. Oxford University Press.

Bennett, J. F. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University press.

Bradley, R. (1998). A Representation Theorem for a Decision Theory with Conditionals. *Synthese 116*, 187–229.

Cantwell, J. (2006). The laws of non-bivalent probability. *Logic and Logical Philosophy 15*, 163–171.

Cantwell, J. (2008). Indicative conditionals: Factual or Epistemic? *Studia Logica 88*, 157–194.

Cantwell, J. (2010). On an Alleged Counterexample to Causal Decision Theory. *Synthese 173*(2), 127–152.

Edgington, D. (1995). On Conditionals. *Mind 104*, 235–329.

Egan, A. (2007a). Epistemic Modals, Relativism, and Assertion. *Philosophical Studies 133*(1), 1–22.

Egan, A. (2007b). Some Counterexamples to Causal Decision Theory. *The Philosophical Review 116*, 93–114.

Gärdenfors, P. (1986). Belief Revision and the Ramsey Test for Conditionals. *Philosophical Review 95*, 81–93.

Gibbard, A. and W. Harper (1978). Counterfactuals and two kinds of expected utility. In Hooker, Leach, and McClennen (Eds.), *Foundations and application of Decision theory*, pp. 125–162. Dordrecht, Holland: D. Reidel.

Horgan, T. (1981). Counterfactuals and Newcomb's problem. *Journal of Philosophy 78*, 331–356.

Jackson, F. (1979). On Assertion and Indicative conditionals. *Philosophical Review 88*, 565–589.

Jeffrey, R. (1983). *The Logic of Decision* (2nd ed.). Chicago: University of Chicago Press.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Levi, I. (1996). *For the Sake of the Argument*. Cambridge: Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge: Harvard University Press.

Lewis, D. (1974). Radical Interpretation. *Synthese 23*, 331–344.

Lewis, D. (1976). Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review 85*, 297–315.

Lewis, D. (1981). Causal Decision Theory. *Australasian Journal of Philosophy 59*, 5–30.

Lewis, D. (1986). Probabilities of Conditionals and Conditional Probabilities II. *Philosophical Review 95*, 581–589.

MacFarlane, J. (2011). Epistemic Modals Are Assessment-Sensitive. In B. Weatherson and A. Egan (Eds.), *Epistemic Modality*, pp. 144–178. Oxford University Press.

Nozick, R. (1969). Newcomb's Problem and two Principles of Choice. In e. a. N. Rescher (Ed.), *Essays in honor of Carl G. Hempel*, pp. 114–146. Dordrecht: Reidel.

Stalnaker, R. (1968). A Theory of Conditionals. In *Studies in logical theory*. Oxford: Blackwell. No. 2 in American philosophical quarterly monograph series.

Stalnaker, R. (1975). Indicative conditionals. *Philosophia 5*, 269–286. Reprinted in *Ifs*, ed. W. L. Harper, R. Stalnaker and G. Pearce 1976 by Reidel.

Stalnaker, R. (1980). A Defense of Conditional Excluded Middle. In W. L. Harper, G. Pearce, and R. Stalnaker (Eds.), *IFs*, pp. 87–104. Dordrecht and Boston: Reidel.

Thomason, R. and A. Gupta (1980). A Theory of Conditionals in the Context of Branching Time. *Philosophical Review 89*(1), 65–90.