Impartiality and Causal Decision Theory[1]

Brad Armendt

Ohio State University

Causal decision theory (CDT) is the best theory of rational choice now available.[2] I intend to provide some support for that claim in part I of this paper by responding to two criticisms of CDT. The first criticism says that CDT is superfluous, since it does no better in *the problems that matter* than does evidential decision theory (EDT) at recommending correct choices. A second criticism says that CDT by itself is flawed: according to this view, there are problems in which CDT makes bad recommendations, unless it is supplemented with an additional deliberation mechanism, either involving metatickles or screening for ratifiable choices.[3] I will argue in response to the first criticism that CDT is genuinely better than any EDT of the most sophisticated sort: there are problems where CDT gives better recommendations, and they are problems that *do* matter. In response to the second line of criticism, I will argue that CDT is not flawed: CDT does not make bad recommendations in the examples that have been put forward. I would be among the first to grant that ratifiability and metatickles are important and interesting subjects in the theory of rational choice. But CDT needs no extra principle or extra screening procedure to avoid counterexamples.

CDT's immediate ancestor is EDT. It is important to understand the exchanges between advocates of the two kinds of theory in order to fully understand the theories. But it is also worth studying and comparing the foundations of the theories. Part II of this paper is devoted to a particular question concerning the foundation of EDT: How are its assumptions about preference violated in examples like the Newcomb problem, the Twin Prisoners' Dilemma, and others? I will briefly discuss the Jeffrey/Bolker foundation for EDT, and I will show where violations of the Jeffrey/Bolker axioms for preference occur in familiar counterexamples to naive EDT.

I. A. CDT is Not Superfluous

The view that CDT is superfluous, that EDT always gives correct answers in problems where a theory of rational choice is relevant, has been defended most thoroughly by Eells (1982), (1985), and Eells & Sober (1986). The details of the general defense are intricate; they cannot all be provided in the space available. Key ideas in the defense are:

1. A rational agent's choices are functions of his beliefs and desires and his rational deliberation alone, and the rational agent knows this;

PSA 1988, Volume 1, pp. 326-336

2. A rational agent is fully aware (or fully-enough aware) of his belief and desires and deliberation; (it may be that he becomes aware of them through self-observation of his deliberation dynamics);

3. A rational agent's choices need not be perfectly efficacious in determining his actions, and he need not believe that they are. But if he takes his decision problem to be one in which use of rational decision theory has a point, he must take that problem to be one in which any tendencies of his choices to fail to produce corresponding actions are *symmetric*.

In the references mentioned, it is argued that the rational agent's awareness of (1) and (2) above enable him to screen off the desired/undesired states that are not caused by the actions from the actions themselves. (3) is required for the argument, since the screening-off guaranteed by (1) and (2) holds between the desired/undesired states and rational *choices*, but to achieve screening-off of the states from the *actions*, there must also be a suitable correlation between choices and acts. I shall not dwell on (1) or (2), though I believe that neither is an obvious principle of practical rationality.[4] I want instead to address (3).

Consider van Fraassen's version of the Twin Prisoners' Dilemma (see Jeffrey 1983, p. 20), an example where fallibility in executing choices becomes relevant to the EDT vs. CDT issue: The causal story is a standard Prisoner's Dilemma with a standard payoff matrix. It is a Twin PD in that each prisoner believes the other's choices are highly correlated with, but not causally influenced by his own. If we grant the basic version of Eells' application of EDT to the problem so far (ignoring the possibility of slips in executing choices), the agent's awareness of his beliefs and desires relevant to the problem (let R be the summary of them), and of his deliberation, enables him to screen off the other prisoner's choice (and action) from his own choice (and action). He can use EDT and arrive at the correct choice (ratting). But in van Fraassen's version of the example it is further stipulated that each prisoner believes it possible that he or his opponent might fail to execute a choice (i.e. choose ratting but end up cooperating, or vice versa), *and* that his tendency to make such a slip is highly correlated with, but not causally influential over, his opponent's tendency to do the same. Since the agent does not believe that his slipping is a function of his beliefs and desires and deliberation alone, no amount of self-knowledge about them will enable him to screen off the reintroduced correlation between his action and the other prisoner's action that leads EDT to recommend the wrong decision.[5]

The first part of (3) is clearly correct. A rational human agent will not execute all his choices perfectly, and should recognize that fact. (In ordinary circumstances this has no particular effect on the values of his choices, and it can be ignored.) Does rationality require that his beliefs about his fallibility in executing his choices satisfy some interesting condition like the one labeled "symmetric fallibility" in Armendt (1985)? I think not, but let's look for a constraint that will help the ED theorist.[6]

The need for a constraint on (the agent's beliefs about his) fallible action arises from the ED theorist's desire to achieve screening-off of states from actions, from the screening-off of the states from choices that (1) and (2) are supposed to insure. Following the notation of Eells (1982), let **CC** and **~CC** abbreviate the states, **A** and **~A** be the possible acts, and **R** be the proposition describing the agent's beliefs and desires. The ED theorist wants

(SCR-A) **Pr(CC / A & R) = Pr(CC / ~A & R), and similarly for ~CC.**

If we grant the screening-off by choices, i.e.,

(SCR-Ch) **Pr(CC / chooseA & R) = Pr(CC / choose~A & R),**

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

then the natural condition is

$$(SF) \quad \frac{Pr(A \,/\, choose{\sim}A \,\&\, R \,\&\, CC)}{Pr({\sim}A \,/\, chooseA \,\&\, R \,\&\, CC)} = \frac{Pr(chooseA \,/\, R \,\&\, CC)}{Pr(choose{\sim}A \,/\, R \,\&\, CC)};$$

or equivalently,

$$(SF') \quad Pr(A \,\&\, choose{\sim}A \,\&\, R \,\&\, CC) = Pr({\sim}A \,\&\, chooseA \,\&\, R \,\&\, CC).^{7}$$

(Similarly for ~CC.) This condition says that, loosely, the agent believes that under the conditions that the state CC holds and that he has beliefs and desires R, his tendencies to choose A and ~A are exactly mirrored by his tendencies (under worldly influences) to arrive at A and ~A by slips after his choice. I call this a condition of *symmetric fallibility*. Notice that it is violated in the van Fraassen example: Letting CC be the other prisoner's ratting and A be the agent's ratting, if the agent's slips are highly correlated with the other prisoner's, there is every reason to expect the left hand side of (SF') to be greater than the right hand side:

> **Pr(I rat & he rats & R & I choose not ratting) >**
> **Pr(I don't rat & he rats & R & I choose ratting)**

Why should we accept (SF) or some similar condition as a principle of practical rationality, the sort of principle on which it is safe to build a theory of rational choice? Eells & Sober (1986) present a different version of symmetric fallibility (see note 6), but it is worth examining some of their remarks in defense of such a principle:

> "We suggest only that the agent believes that the *direction* of a slip, in the event of a slip, is random with respect to the outcome O, given RA or given ~RA. The motivation for this is two-fold. First, it is entirely appropriate for us to assume that the agent believes that the problem confronting him is one that is *appropriate* for the application of standards of rational decision. And second, he should *not* believe this if he believes that, *regardless of the outcome of his deliberation*, his act will be caused to be in accordance with the correlation between A and O, that is, to agree with the one of O and ~O that actually obtains..."
>
> "Although CDT may give correct answers even if the decision maker does believe that the correlation is, in part, enforced by a factor that sometimes causes the irrational act, this poses little threat to EDT. In this kind of case, the causal theory fares better than the EDT to the extent that the decision situation is not one in which the agent should find it appropriate to apply standards of rational decision in the first place." [Eells & Sober 1986, p. 240, 241]

The two points listed as motivation for the assumption that slips are random are quite plausible. If the agent believes applying standards of rational choice to a given problem is inappropriate, then his problem is not one on which a subjective theory of rational choice need founder, should it give him bad recommendations. And if he believes that no matter how his deliberation goes, there's only one action he can end up performing, then his problem is not one in which applying subjective standards of rational choice has much point. But the important point here is that only in *extreme* violations of symmetric fallibility will he be in such a situation. Asymmetric fallibility can fall far short of yielding: inevitable action no matter what the course of deliberation.

Eells & Sober are aware of this: the real defense of symmetric fallibility lies in the second half of the quoted passage. The idea of the defense is that the propriety of applying principles of rational choice comes in degree, that it is proportional to the degree of the

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

violation of symmetric fallibility, and that since the degree of failure of EDT is also proportional to the degree of violation of symmetric fallibility, such failure is no flaw of EDT. I doubt that the propriety of applying rational choice theory comes in the sort of degrees imagined; if violation of symmetric fallibility falls at all short of producing inevitable action, I think the use of rational choice theory is appropriate. But even if the propriety is a matter of degree, the fact that EDT's failure is proportional to it does not remove the fact that EDT fails in problems for which it is appropriate to some (perhaps considerable) degree to apply principles of rational choice.

I. B. CDT is Not Broken

I shall now briefly discuss the second of the criticisms of CDT mentioned at the beginning of the paper. My remarks here will be brief because of space limitations, but also because my disagreement with the criticism seems to boil down to a clash of intuitions about correct answers to certain decision problems. The issue surrounding the problems is interesting, however, and I want to register my disagreement.

Eells (1985) presents a problem in which CDT recommends a choice that is not ratifiable, over one that is. He takes the ratifiable choice to be the correct answer; after considering and rejecting the idea of combining CDT with a screening for ratifiable choices, he says that CDT must be supplemented with attention to metatickles to avoid counter-example. In his comments, Harper (1985) agrees with Eells about the correct solution to the problem, but argues that adding to CDT a screening for ratifiable choices is workable and preferable to a commitment to deliberation with metatickles. I disagree with both; I say that CDT does not recommend an incorrect choice in the first place. The example I shall discuss is similar to the problem in Eells (1985), but it is a simpler version presented by Eells and Harper (1987): There are three available options, A, B, and C. As in standard Newcomb problems, there is a predictor who has made a forecast about what the agent will do, the agent believes that the predictor is very accurate, and he believes his choice in no way causally influences the forecast. In this case, the predictor simply makes a prediction about whether or not the decision maker will choose act C. So there are two relevant possible states: FC (predictor forecast C) and F~C (predictor forecast ~C). The utility matrix for the problem is this:

	FC	F~C
A	5	1
B	2	3
C	4	2

If we take the agent's degree of belief that the predictor forecasts correctly to be nearly 1 no matter which choice is made, a straightforward calculation shows that CDT's recommendation depends on the agent's initial degree of belief **Pr(FC)** as follows: if **Pr(FC) < 1/3** , CDT recommends B; if **1/3 < Pr(FC) < 1/2**, CDT recommends C; if **Pr(FC) > 1/2**, CDT recommends A. But the only ratifiable choice is B: under the hypothesis that A is chosen, B looks better; under the hypothesis C is chosen, A looks better. Eells & Harper take the true solution of the problem to be B, and fault CDT for recommending A or C when it does.

But now I disagree. It strikes me as perfectly plain that if the agent enters the problem with a degree of belief in FC that yields a recommendation of A (or of C) then he should choose A (or C). I can think of three sorts of reasons for questioning this: the first are reasons for not U-maximizing like those appealed to by one-boxers in a standard Newcomb problem. But just as those are bad reasons in the Newcomb problem, so are they bad in this example.

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

The second sort of reason is more interesting: Notice that an agent who observed the course of his deliberation and learned from the observations might become dissatisfied with the choice initially recommended by CDT. His subsequent reevaluations of the problem might shift and stabilize on a different choice. How this would go depends upon details about the values of many of his degrees of belief, and about the method by which he learns from his deliberations. I see no reason to think that *every* plausible way of filling in those details yields a recommendation of, in this case, B. More to the point, there is no incompatibility between CDT and learning from deliberation: an agent who prior to his choice acquires new information should use it. He can and should employ CDT to do so. Whether or not he gets this information and uses it is a different issue from whether or not he should use CDT to evaluate his options. It is no flaw of CDT that agents who do not learn from their deliberation choose differently from those that do.

The third sort of reason is the one I believe Eells and Harper have in mind. It is simply that correct choices are ratifiable choices, when there are ratifiable choices available. Since CDT may not recommend the only available ratifiable choice B, CDT is flawed. I simply do not think this is so: a correct choice maximizes causal expected utility [Armendt (1986, 1988)]; a ratifiable choice is one I would remain happy with were I to make it. It is a pleasant state of affairs when these coincide, as they usually do. But when they differ, ratifiability is seen to be a flawed criterion of rational choice.[8]

II. Violations of Impartiality

I now turn to my remarks on the preference axioms of EDT. I shall quickly review the axioms for the Jeffrey/Bolker foundation for EDT, and then focus on one of them, the Impartiality axiom. Finally, I shall illustrate how Impartiality is violated in the standard Newcomb problem, and indicate how that illustration can be imitated and generalized for other similar decision problems.

The Jeffrey/Bolker axioms.[9] In Jeffrey (1983), the agent's preferences are assumed to satisfy the following axioms:

1. The elements of the agent's preference ordering form a complete, atom-free Boolean algebra (of propositions);

2. The preference relation $>$ is continuous: when the supremum or infimum of an implication chain lies between **A** and **B**, so does the tail of the chain;

3. The preference relation $>$ is transitive and trichotomous;

4. Averaging: When **A** and **B** are incompatible,

$$\mathbf{A} > \mathbf{B} \text{ implies } \mathbf{A} > (\mathbf{A} \vee \mathbf{B}) > \mathbf{B}, \text{ and}$$
$$\mathbf{A} \sim \mathbf{B} \text{ implies } \mathbf{A} \sim (\mathbf{A} \vee \mathbf{B}) \sim \mathbf{B}.$$

5. Impartiality: When **A** and **B** are incompatible and $\mathbf{A} \sim \mathbf{B}$, if there is a **C** that is

   a. incompatible with both **A** and with **B**, and
   b. either $\mathbf{C} > \mathbf{A} \sim \mathbf{B}$, or $\mathbf{A} \sim \mathbf{B} > \mathbf{C}$, and
   c. $(\mathbf{A} \vee \mathbf{C}) \sim (\mathbf{B} \vee \mathbf{C})$,

then for all **D** incompatible with **A** and with **B**,

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

(A v D) ~ (B v D).

Preference orderings satisfying these axioms are representable (and representable uniquely, up to the conditions given in Jeffrey (1983)) by pairs of functions *prob* and *des* that obey the EDT expected utility rule. Axioms 1-4 seem quite innocuous, except perhaps to those who have general reservations about the richness assumptions required by any foundation for a theory of rational choice under risk. But as I shall explain in a moment, I am not here concerned with reservations of that sort. I want instead to focus on the Impartiality axiom.

I shall not attempt to give here a complete account of the Impartiality assumption and its role in the Jeffrey/Bolker theory. I simply note Jeffrey's remark:

> In chapter 7, we tested equiprobability of incompatible propositions A, B that were ranked together by using a test proposition C, incompatible with A and with B and not ranked with them: the test showed equiprobability in case the disjunctions A v C and B v C were ranked together. [Impartiality] stipulates that the choice of different test propositions C cannot make the test yield different results. [Jeffrey 1983, p. 147]

While working on a foundation for CDT that is quite different from the Jeffrey/Bolker theory, I became interested in applying the Jeffrey/Bolker theory to CDT.[10] This is worth doing, I believe, because the Jeffrey/Bolker theory has *virtues* not present in other theories (acts, states, and consequences appear in a unified set of propositions; extraneous lotteries are not required). It appears possible because, for decision-making under a fixed dependency hypothesis K, the Jeffrey/Bolker theory is *correct*.

In thinking about this, I used the following line of reasoning: (1) The Jeffrey/Bolker theorem implies that for any preference ordering satisfying their axioms there exists pairs *prob*, *des* that represent the ordering; and (2) The *des* functions are order-preserving; but (3) In the examples where naive EDT goes astray *prob* and *des* misrepresent the rational agent's preference ordering; e.g. in the Newcomb problem, A2 (taking both boxes) is rationally preferred to A1 (taking only the opaque box), while *des*(A2) < *des*(A1); so (4) In such examples the agent's preferences must violate the Jeffrey/Bolker axioms. Where does the violation occur? In Armendt (1988) I speculated that it is Impartiality that is violated, but could not say how. I hoped then, and still do, that understanding what the violation is might contribute to developing a Jeffrey/Bolker-style foundation for CDT. Well, I cannot yet produce the new foundation, but I can now identify the violation (or possibly, one of the violations) of the axioms. It is indeed a violation of Impartiality, and it is illustrated below.[11]

Consider a standard Newcomb problem with a reliable but fallible predictor. Adopt the following abbreviations:

A1:	I take only the opaque box.
A2:	I take both boxes.
F1:	The predictor forecasts that I take only opaque box.
F2:	The predictor forecasts that I take both boxes.
MT:	I collect $1,001,000.
M:	I collect $1,000,000.
T:	I collect $1000.
0:	I collect $0.

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

For simplicity suppose the agent is certain that the conditions of the problem as standardly described obtain; i.e. he is sure of the causal structure of his problem. For example, **prob(opaque box has $M / F1) = 1**, and **prob(MT / A2 & F1) = 1.** In the agent's preference ranking, **MT > M > A2 > A1 > T > 0.**

Now it is difficult to illustrate the violation of Impartiality in the Newcomb problem if we confine our attention to only the acts, states, and consequences typically mentioned in the statement of the problem (that is, if it can be done I don't know how). The trouble is that the available pairs of *incompatible* propositions *between which the agent is indifferent* are few. But an agent confronted with a Newcomb problem is an agent with preferences for other propositions as well. I shall introduce one further proposition P and consider the agent's preferences for it and for some logical combinations of P and the propositions listed above. In doing this, I intend *no alteration* of the Newcomb problem described above: as will be clear in a moment, no changes in the possible choices or causally relevant states or outcomes of the Newcomb problem are introduced. Instead, the problem is considered in slightly less isolation from the agent's other preferences. Also, note that unless I say otherwise, the preferences and utilities I describe below are the agent's preferences-for-action, not his preferences-for-news (hence A2 is preferred to A1): the agent's preferences of the latter sort will *not* violate the Impartiality axiom or any other Jeffrey/Bolker axiom if the Jeffrey/Bolker theory succeeds, as I think it does, in capturing preference-for-news. It is when "news value", measured by V, does not correspond to "act value", measured by U, that EDT goes astray. (Of course, the two sorts of preference typically coincide to a great extent.)

As mentioned above, the rational (U-maximizing) agent prefers A2 to A1. He also prefers A2 & F1 to A1 & F1. The former is equivalent, under the assumption of certainty about the workings of the game made above, to MT, and the latter is equivalent to M. Let P be the agent's proposition, "I receive $x tomorrow from out of the blue, in no way correlated (causally or statistically) with my choice, or with the predictor's forecast, or with the outcomes of the Newcomb problem." The agent might imagine finding the money on the street and having it go unclaimed or whatever. Let the number x be such that the agent is indifferent between (A2 & F1 & ~P) and (A1 & F1 & P). Presumably x = 1000, since that's the difference between U(A2 & F1) and U(A1 & F1), but it doesn't matter whether it is or not. Let P also be such that the agent's degree of belief Pr(P) is quite small. In particular, whatever Pr(A2) is, suppose Pr(P) is considerably less.

Notice that both (A2 & F1 & ~P) and (A1 & F1 & P) are preferred to F2, since the latter leads to an empty opaque box, and also to (A1 & F1 & ~P), since this misses out on both the $1000 in the transparent box and the extra windfall. The Impartiality axiom is violated if (1) is satisfied and (2) is not:

(1) **(A2 & F1 & ~P) v F2 ~ (A1 & F1 & P) v F2**, and

(2) **(A2 & F1 & ~P) v (A1 & F1 & ~P) ~**

(A1 & F1 & P) v (A1 & F1 & ~P) .

I see no reason to doubt that (1) holds. What about (2)? When the disjunctions are simplified, we see that it is equivalent to

(2') **(F1 & ~P) ~ (A1 & F1)**, in other words

(2") **(M & ~P) ~ (M & ~T).**

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

But (2") does *not* hold: the million dollar prize in conjunction with not getting the windfall (but with perhaps getting the thousand through choice A2) is *preferred* to the million dollar prize in conjunction with not getting the extra thousand dollar prize (and with perhaps getting the very unlikely windfall). Recall that Pr(P) is considerably less than Pr(A2).

Analogous illustrations of the violation of Impartiality can be given for other standard examples in which EDT and CDT diverge.[12] The method for constructing them in the general case when simplifying assumptions (e.g. certainty in the causal structure of the decision problem) are relaxed may be complicated. (The examples themselves tend to get complicated.) I take it that illustrations like this show where conflicts between the Jeffrey/Bolker assumptions about rational preference (as implemented in naive EDT) and true rational preference can arise.[13] I leave the story incomplete here; I have not yet exploited this to provide a new foundation for CDT.

Notes

[1]I am grateful to the Ohio State University College of Humanities for financial support. In thinking about this paper I have benefited from conversations with Ellery Eells, William Harper, Don Hubin, Paul Humphreys, Richard Jeffrey, George Schumm, and Brian Skyrms.

[2]For an account of different versions of CDT, see Lewis (1981); another good source is Eells (1982); for evidential decision theory, see Jeffrey (1983) or Eells (1982). EDT is sometimes known as V-maximization decision theory, CDTs as U-maximization theories.

[3]This second criticism is tied to the first: the principles that are allegedly needed to make CDT satisfactory are like those that defenders of EDT incorporate into their sophisticated versions of EDT, in order to avoid the familiar counterexamples to naive EDT.

[4]The self-knowledge requirement (2) is quite strong; but in expressing reservations about it I do not object that an agent cannot have or acquire a great degree of self-knowledge. And I believe that the study of deliberation dynamics is important; an agent who acquires information through self-observation should use it. See Skyrms (1986). But I do object to the assumption that a rational agent must have all this self-knowledge. Requirement (1) is more plausible, but it seems to me violated in cases like the following: An agent has all the other decision-making virtues we might expect, but he believes (perhaps correctly) that in a random, unpredictable 1% of the problems he encounters, his choice (i.e. his intention to act formed at the end of deliberation) is influenced by space aliens in ways that do not wholly depend on his beliefs and desires. Assuming that the influence is not detectable prior to the choice, (1) is violated, even in the 99% of the problems where he is on his own and is as rational as you please: he does not know that his choice is a function of only his beliefs and desires. Can the 99% of the problems where he is on his own be excluded from the domain where it is appropriate to use rational decision theory? (See discussion below.) Surely not. Can any particular problem, before a choice is made and the question of outside influence is determined, be excluded from that domain? Again, no--especially when the frequency of outside influence is believed as low as I have described it.

[5]Humphreys (1988) presents another example, involving Klinefelter's syndrome, that also illustrates the effects of slips between choice and act on EDT's recommendations. I am here agreeing with most of his criticisms of EDT in that paper; Humphreys does not pursue a precise characterization of the symmetric fallibility constraint required by EDT.

My remarks on this subject are derived from Armendt (1985), comments on an early version of Humphreys' paper.

[6]The condition I give is phrased in the terms used in Eells (1982, 1985). Eells and Sober (1986) give a fuller treatment for cases in which the agent believes his decision problem to be one involving interactive causal forks. This treatment is illuminating; however, its presentation here would require more space than I have, and the bottom line remains the same: Eells & Sober explicitly appeal to requirement (3) in defense of EDT. (Moreover, they require this of an agent who is imagined to be learning from observing his deliberation; such learning must continually preserve disbelief in asymmetric fallibility.) Their reasons for endorsing (3) are interesting; more about this below. The Eells & Sober symmetry condition (p. 234, 239) is adapted to the interactive fork discussion and it differs from the one I present here in that the beliefs are taken to be conditional on the choice, beliefs and desires, and outcomes, rather than the choice, beliefs and desires, and state CC, as in (SF). My criticisms of symmetric fallibility in the text apply to both versions. In discussing the justification of symmetric fallibility I quote passages from Eells & Sober since the argument is more fully set out in that article than it is elsewhere--Eells (1985, p. 184) for example.

[7]This is the most natural *sufficient* condition, and it captures the defense given in the references mentioned above (see note 6):

Given (SCR-Ch), $Pr(CC/chooseA \ \& \ R) = Pr(CC/choose{\sim}A \ \& \ R)$,

$$\frac{Pr(CC \ \& \ chooseA \ \& \ R \ \& \ A) + Pr(CC \ \& \ chooseA \ \& \ R \ \& \ {\sim}A)}{Pr(chooseA \ \& \ R \ \& \ A) + Pr(chooseA \ \& \ R \ \& \ {\sim}A)} =$$

$$\frac{Pr(CC \ \& \ choose{\sim}A \ \& \ R \ \& \ A) + Pr(CC \ \& \ choose{\sim}A \ \& \ R \ \& \ {\sim}A)}{Pr(choose{\sim}A \ \& \ R \ \& \ A) + Pr(choose{\sim}A \ \& \ R \ \& \ {\sim}A)}$$

So

$$\frac{Pr(CC \ \& \ chooseA \ \& \ R \ \& \ A) + Pr(CC \ \& \ choose{\sim}A \ \& \ R \ \& \ A)}{Pr(chooseA \ \& \ R \ \& \ A) + Pr(choose{\sim}A \ \& \ R \ \& \ A)} =$$

$$\frac{Pr(CC \ \& \ chooseA \ \& \ R \ \& \ {\sim}A) + Pr(CC \ \& \ choose{\sim}A \ \& \ R \ \& \ {\sim}A)}{Pr(chooseA \ \& \ R \ \& \ {\sim}A) + Pr(choose{\sim}A \ \& \ R \ \& \ {\sim}A)}$$

by substitution in accordance with (SF'). Then by the assumption $Pr(choose{\sim}A) = Pr({\sim}chooseA)$, which is introduced in Eells (1982),

$$\frac{Pr(CC \ \& \ R \ \& \ A)}{Pr(R \ \& \ A)} = \frac{Pr(CC \ \& \ R \ \& \ {\sim}A)}{Pr(R \ \& \ {\sim}A)}.$$

So $Pr(CC/R \ \& \ A) = Pr(CC/R \ \& \ {\sim}A)$, i.e. (SCR-A).

Under the assumption (SCR-Ch), the most interesting *necessary* condition for (SCR-A) I have found is:

(N) $Pr(A \ \& \ {\sim}chooseA \ \& \ CC \ \& \ R) - Pr({\sim}A \ \& \ chooseA \ \& \ CC \ \& \ R) =$
$Pr(A / R)*Pr(CC \ \& \ R) - Pr(chooseA / R)*Pr(CC \ \& \ R);$

I.e., when $Pr(CC \ \& \ R) > 0$,

(N') $Pr(A \ \& \ {\sim}chooseA / CC \ \& \ R) - Pr({\sim}A \ \& \ chooseA / CC \ \& \ R) =$
$Pr(A / R) - Pr(chooseA / R);$

https://doi.org/10.1086/psaprocbienmeetp.1988.1.193000 Published online by Cambridge University Press

and similarly for ~CC. I am aware of no arguments for thinking that (N) generally holds that are any better than the arguments for (SF').

[8]It is well known that in certain examples that disallow mixed strategies, the correct choices are not ratifiable, and the ratifiable choices are incorrect. See Skyrms (1984, p.84).

[9]Foundations for EDT are given by Bolker in (1965) and (1967), and by Jeffrey in (1983) and (1978). The versions in Bolker (1965) and Jeffrey (1978) are more mathematically sophisticated, but they include axioms that are close to those given in Jeffrey (1983). The axioms I describe are those in Jeffrey (1983), but it is straightforward to recast my remarks that follow to address the other versions.

[10]This has already been done in similar ways by Jeffrey (1981) and Skyrms (1982), but in both cases the application involves an assumed prior specification of appropriate dependency hypotheses. It would be better to provide an application in which appropriate dependency hypotheses are detected by their behavior in the agent's preference ordering. This is accomplished in Armendt (1986, 1988), but that foundation for CDT is quite different from, and lacks some of the virtues of, the Jeffrey/Bolker foundation for EDT.

[11]I am indebted to George Schumm and Don Hubin for help with this illustration.

[12]In the smoking gene example, let P be such that the agent is indifferent between (smoking & ~P & not having the gene) and (refraining & P & not having the gene); let having the gene, G , and not having the gene, ~G , play the role that F1 and F2 play in the illustration for the Newcomb game. In the Twin Prisoners' Dilemma, let P be such that the agent is indifferent between (I rat & ~P & he cooperates) and (I cooperate & P & he cooperates); let his ratting and his cooperating play the role that F1 and F2 play. Of course, in the different examples P can always be taken to be the same kind of good (pleasure, prison sentence) as the goods that are possible outcomes of the available choices.

[13]This example addresses only the Jeffrey/Bolker assumptions as implemented in naive EDT; if sophisticated EDT agrees with CDT in a given problem, the agent's preference ordering is not represented by *prob*, *des* pairs that make, e.g. *des*(A2) < *des*(A1). The arguments for sophisticated EDT are arguments that, contrary to first appearance, such preference orderings do not occur for rational agents. I would like to avoid building into a foundation for CDT all the assumptions that are required to make sophisticated EDT mimic CDT. The alternative will be the use in the foundation for CDT of nontrivial *conditional preferences* along the lines discussed in Armendt (1986, 1988).

References

Armendt, B. (1985). "Comments on 'Non-Nietzschean decision making,'" presented at 1985 Central Division meeting of the American Philosophical Association.

_____. (1986). "A foundation for causal decision theory." *Topoi* 5: 3-19.

_____. (1988). "Conditional preference and causal expected utility." In *Causation in Decision, Belief Change, and Statistics*, Volume II. Edited by Harper and Skyrms. Dordrecht: Reidel.

Bolker, E. (1965). *Functions Resembling Quotients of Measures*. Dissertation. Harvard University.

_____. (1967). "A simultaneous axiomatization of utility and subjective probability." *Philosophy of Science* 34: 333-340.

Eells, E. (1982). *Rational Decision and Causality.* Cambridge: Cambridge University Press.

_____. (1985). "Causal decision theory." In *PSA 1984* volume two. Edited by Asquith and Kitcher. Philosophy of Science Association.

Eells, E. and Harper, W. (1987). "Ratifiability and the principle of independence of irrelevant alternatives." Presented at the Ohio State University Conference on Social Choice and Rational Bargaining, October, 1987.

Eells, E. and Sober, E. (1986). "Common causes and decision theory." *Philosophy of Science* 53: 223-245.

Harper, W. (1985). "Ratifiability and causal decision theory: comments on Eells and Seidenfeld." In *PSA 1984* volume two. Edited by Asquith and Kitcher. Philosophy of Science Association.

Humphreys, P. (1988). "Non-Nietzschean decision making." In *Probability and Causality.* Edited by Fetzer. Dordrecht: Reidel.

Jeffrey, R. (1978). "Axiomatizing the logic of decision." In *Foundations and Applications of Decision Theory,* Volume I. Edited by Hooker, Leach, and McClennen. Dordrecht: Reidel.

_____. (1981). "The logic of decision defended." *Synthese* 48, 473-492.

_____. (1983). *The Logic of Decision,* 2nd edition. University of Chicago Press. First edition published in 1965.

Lewis, D. (1981). "Causal decision theory." *Australasian Journal of Philosophy* 59: 5-30.

Skyrms, B. (1979). *Causal Necessity.* Yale University Press.

_____. (1982). "Causal decision theory." *Journal of Philosophy* 79, 695-711.

_____. (1984). *Pragmatics and Empiricism.* Yale University Press.

_____. (1986). "Deliberational equilibria." *Topoi* 5: 59-67.