COUNTERFACTUAL REASONING AND PRACTICAL REASONING IN CAUSAL GRAPHS

By

Reuben Stern

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
(Philosophy)

at the
UNIVERSITY OF WISCONSIN-MADISON
2016

Date of final oral examination: May 4, 2016

The dissertation is approved by the following members of the Final Oral Committee:

Daniel M. Hausman, Professor, Philosophy (Chair)
Elliott Sober, Professor, Philosophy
Malcolm Forster, Professor, Philosophy
Michael G. Titelbaum, Associate Professor, Philosophy
Peter Steiner, Assistant Professor, Educational Psychology

Acknowledgments

It would be impossible to thank everyone who has helped me to get to this point. So I won't do that. But I would like to thank just a few individuals who have been especially helpful.

First, I want to thank my advisor, Dan Hausman. Dan's philosophical acumen is matched only by his generosity. As his student, I benefitted immensely from both. He was always willing to talk about my work, and when he talked, he said smart things. I can't imagine a better advisor.

Second, I want to thank my advisor's dog, Itzhak. Though it is hard to believe, Itzhak is as good at providing companionship as his owner is at advising graduate students. When I get sad about leaving Madison, I'm most sad that I'm leaving Itzhak.

Third, I want to thank all of the excellent professors who make up Wisconsin's philosophy department—especially Elliott Sober, Malcolm Forster, Mike Titelbaum, John Mackay, and Sarah Paul. I have been very lucky to receive feedback from such a multi-talented group of philosophers.

Fourth, I want to thank my family. The amount of support I have received from them is shocking—especially since I probably have become less relatable the more I've studied philosophy. They're excellent people.

Fifth, I should thank the people of Madison who have been willing to listen to my endless pontificating over drinks. Bearing special mention in this category are Malcolm Forster, David O'Brien, Ben Schwan, Naftali Weinberger, and Olav Vassend. I hope that we have a life full of drawing DAGs over beers to look forward to.

Sixth, I should thank my friends who have been especially available for commiseration during my time as a graduate student. Some folks who fall under this banner and who have not

been mentioned yet are Hayley Clatterbuck, Ana Macedo, Brian McLoone, Clinton Packman, Rush Stewart, and Peter Steinbauer.

Seventh, I want to thank Felix Elwert, Richard Scheines, Peter Steiner, and Naftali Weinberger for introducing me to causal graphs. It was the beginning of a love affair.

Finally, I want to thank my wife, Shanna Slank, for being a constant source of support. She, too, is responsible for a love affair. Plus she is the best at everything.

Abstract

This dissertation consists of three papers in which I use the recently developed graphical approach to causal modeling in order to make headway on some traditional philosophical problems. In the first paper, I provide a general semantics for counterfactual conditionals that (unlike other semantics on offer) capably accommodates the causal modeling semantics for "non-backtracking" counterfactual conditionals. In the other two papers, I use causal graphs to develop decision theory that provides agents with advice about what to do when uncertain about the causal structure of the world.

Table of Contents

Chapter 1: Is There a General Counterfactual Semantics That Can Accommodate Interventionist Counterfactuals?

ABSTRACT

According to the standard picture of the relationship between non-backtracking counterfactuals and other counterfactuals, non-backtracking counterfactuals share their logic with other counterfactuals, but utilize a distinct similarity ordering on worlds that generates non-backtracking verdicts when plugged into the general semantics (or logic) for counterfactuals. In recent years, a number of authors have developed a promising semantics for non-backtracking counterfactuals that makes use of the hugely influential graphical approach to causal modeling found in e.g. Pearl (2009) and Spirtes, Glymour, and Scheines (2000). Though the causal modeling (CM) semantics is often thought to supersede previous accounts of non-backtracking counterfactuals, Briggs (2012) shows that the ordering on worlds supplied by CM semantics violates weak centering—i.e. the assumption that any world w is necessarily in the set of worlds closest (or most similar) to w—when plugged into the most standard general semantics for counterfactuals (i.e. Lewis/Stalnaker semantics), and that CM semantics is therefore incompatible with Lewis/Stalnaker semantics. In this paper, I explore the possibility of abandoning Lewis/Stalnaker semantics for an alternative general semantics for counterfactuals that is better suited to accommodate CM semantics. I develop my own alternative (a revision of McGee's 1985 semantics) that capably accommodates CM semantics without sacrificing weak centering, and thereby show that CM semantics can be plausibly situated within a more general semantics for counterfactuals that is based on the nearness or similarity of worlds. In the process, I argue that my revision to McGee's semantics should be regarded as a serious competitor to Lewis/Stalnaker semantics.

1. Introduction

There are many circumstances in which it is reasonable to wonder what would have happened had someone intervened or were someone to intervene (now or in the future) on the goings-on of the actual world in order to make some counterfactual state of affairs true.

Consider Coach Pete Carroll's thoughts as he watches the game tape from Super Bowl XLIX, where his Seattle Seahawks lost in a close one to the New England Patriots. As Carroll looks back at the tape, it is reasonable for him to wonder how things would have gone had he coached differently—e.g. had he called a run play instead of a pass play on the Seahawks' last offensive play.[1] After all, Carroll might find himself in similar circumstances in the future, and, if he does, knowledge of what would have happened had he called a different play will help him to make the right choice.

There are multiple modes that Carroll's counterfactual reasoning can take on. He can, for example, wonder how the game would have gone differently *prior* to the Seahawks last offensive play were he to have chosen differently—e.g. what would have had to happen on the field before the last play in order for him to have chosen differently. But these differences do not seem relevant to Carroll's future coaching choices. Rather, if Carroll's counterfactual inquiry is motivated exclusively by his desire to call better plays, then he should be concerned with the *effects* of choosing differently, where the effects of choosing to run rather than pass are limited to only those differences that are causally downstream from Carroll's choice. By narrowing his focus to the effects of his choice, Carroll manages to hold the circumstances of his choice fixed, thereby allowing him to apply his findings to any coaching decision in which the circumstances are relevantly similar to those that he confronted in Super Bowl XLIX.

---

[1] The pass play that Carroll actually did call resulted in an interception that clinched the Patriots' victory.

[2] These counterfactuals play an important role in the development of many philosophical views. For example, causal decision theorists (e.g. Gibbard and Harper 1978) argue that agents should determine

Stories like Carroll's suggest that there is a special kind of counterfactual—often called a *non-backtracking* counterfactual—that merits its own semantics. That is, since there are contexts where it is reasonable to focus specifically on counterfactual dependencies that flow in the same direction as causation (even though there are likewise counterfactual dependencies that flow in the opposite direction), it would be ideal to have a semantics specifically for these non-backtracking counterfactual dependencies.[2] In his (1979) paper, "Counterfactual Dependence and Time's Arrow," David Lewis famously provides one. His strategy is to develop a similarity ordering on worlds that can be plugged into his (1973a) general account of counterfactual semantics (as well as any other account that utilizes comparative similarity of worlds—e.g. Stalnaker 1981) in order to deliver non-backtracking verdicts.[3] So if a counterfactual, $a > b$, is true exactly when the nearest (i.e. most similar) worlds in which a is true are worlds in which b is true, then Lewis's (1979) proposal amounts to developing a similarity ordering on worlds that gets the result that, say, the most similar worlds in which Carroll calls a run play are worlds in which nearly everything prior to Carroll's choice is the same as the actual world, but that differ with respect to the *effects* of Carroll's counterfactual choice.[4]

---

[2] These counterfactuals play an important role in the development of many philosophical views. For example, causal decision theorists (e.g. Gibbard and Harper 1978) argue that agents should determine the expected utility of x-ing by calculating a weighted average of how much one values each of the possible outcomes of x-ing, where the weights correspond to one's subjective probabilities in a partition of non-backtracking counterfactuals specifying what outcomes would obtain were the agent to x.

[3] According to Lewis (1979), the similarity relation at play in non-backtracking contexts is governed by the "system of weights or priorities" quoted below.

1) It is of first importance to avoid big, widespread, diverse violations of laws.
2) It is of second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
3) It is of third importance to avoid even small, localized, simple violations of law.
4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

[4] This semantics for counterfactuals is plausible only given what Lewis (1973b) calls the "Limit Assumption"—i.e. the assumption that "as we proceed to closer and closer a-worlds we eventually hit a limit and can go no farther." Lewis himself was not willing to make this assumption and accordingly

Lewis's (1979) proposal is justifiably controversial. Though many (e.g. Woodward 2003) agree that non-backtracking counterfactuals differ from other counterfactuals insofar as their truth-conditions make reference to a specific contextually determined similarity ordering on worlds, far fewer agree with Lewis's account of the specific ordering at play in non-backtracking contexts. Woodward, for example, argues that Lewis's ability to provide a natural account of non-backtracking similarity is significantly hampered by his (1973c) commitment to the *reduction* of causal dependence to counterfactual dependence since non-backtracking similarity is most naturally understood in causal terms.[5]

In recent years, a number of authors (e.g. Briggs 2012, Galles and Pearl 1998, Halpern 2000, Hiddleston 2005, Hitchcock 2011, Pearl 2009, Woodward 2003) have abandoned Lewis's reductive ambitions and taken up the task of developing a more natural semantics for non-backtracking counterfactuals that utilizes the hugely influential graphical approach to causal modeling found in e.g. Pearl (2009) and Spirtes, Glymour, and Scheines (2000). According to this semantics, the evaluation of a non-backtracking counterfactual, $a > b$, involves determining whether b obtains in the causal model that results from *intervening* on the actual world to make a true, where an *intervention* is roughly construed as an "off-stage" manipulation of the actual world that is by definition counterfactually and causally independent of a's (non-intervention) causes.[6]

Though the authors of causal modeling (CM) semantics have traditionally thought that their semantics can be squared with more general accounts of counterfactual semantics—i.e.

---

defended a slightly different account of the truth-conditions for counterfactuals, but since the Limit Assumption plays no important role in what follows (at least as I see things), I assume the Limit Assumption in order to make exposition easier.

[5] Woodward likewise argues (i) that Lewis's account delivers the wrong non-backtracking verdicts for some counterfactuals, (ii) that Lewis's similarity criteria are not sufficiently clear to deliver definite truth-values even once the non-backtracking context has been specified, and (iii) that it is not at all clear why people would or should use Lewis's criteria to judge the similarity of worlds.

[6] This is a very rough construal of what it means to intervene on the actual world. I am more precise in Section 2.

existing accounts that are supposed to apply regardless of whether the counterfactual at hand is non-backtracking (e.g. Lewis 1973a or Stalnaker 1981)—Rachel Briggs (2012) has recently provided some reason to doubt this possibility. Briggs first provides a natural extension of CM semantics to counterfactuals whose consequents are themselves counterfactuals and then shows that *modus ponens* is invalid in the resulting semantics. This is vexing, first, because *modus ponens* is valid in the most prominent general accounts of counterfactual semantics (Lewis 1973a and Stalnaker 1981), and, second, because of its relation to the seemingly trivially true assumption known as *weak centering*—i.e. the assumption that any world *w* is necessarily in the set of worlds closest (or most similar) to *w*. (As Briggs notes, when CM semantics is paired with either Lewis's or Stalnaker's interpretation of the counterfactual, *modus ponens* is invalid because *weak centering* is violated.)[7,8] Briggs's ambition in pointing out these features is not to undercut CM semantics or its extension, but is rather to identify aspects of the semantics that merit greater philosophical attention than they have received thus far.

Here, I direct my attention towards these aspects of the extended semantics. Specifically, I focus on whether there is a general account of counterfactual semantics that can successfully accommodate CM semantics. This is an important project not only because the best general account should be compatible with our best understanding of non-backtracking counterfactuals (which is arguably supplied by CM semantics), but also because CM semantics stands on firmer ground when situated within some general account because of the unification gained by construing non-backtracking counterfactual reasoning (of the sort modeled by CM

[7] Defenders of Lewis/Stalnaker semantics sometimes define weak centering such that it (by itself) entails the validity of counterfactual *modus ponens*. But that is too strong for my taste since there are (as we will see) semantics in which the assumption that all orderings on worlds are weakly centered does *not* entail *modus ponens*. Thus I prefer the characterization of weak centering provided here because it is more general in the sense that characterizes the assumption's role in *any* semantics that utilizes similarity orderings on worlds.

[8] In section 3, I explain why the invalidity of *modus ponens* results from a violation of weak centering in in Lewis/Stalnaker semantics.

semantics) as an instance of general counterfactual reasoning. Again, according to the standard picture of the relationship between non-backtracking counterfactuals and other counterfactuals, non-backtracking counterfactuals share their logic with other counterfactuals, but utilize a distinct similarity ordering on worlds that generates non-backtracking verdicts when plugged into the general semantics (or logic) for counterfactuals. So according to the standard picture, the job of CM semantics is to deliver orderings on worlds that (when plugged into the best general semantics for counterfactuals) capture the way in which the truth of a non-backtracking counterfactual at a world depends on the causal facts at that world. But since the orderings supplied by CM semantics must violate weak centering when plugged into Lewis/Stalnaker semantics (and since weak centering is self-evidently true), it appears that CM semantics is incompatible with Lewis/Stalnaker semantics, and that we must therefore abandon either CM semantics or Lewis/Stalnaker semantics.

In this paper, I consider the possibility of abandoning Lewis/Stalnaker semantics for some alternative general account of counterfactual semantics that is better suited to accommodate CM semantics. I initially focus on McGee's (1985) modification of Lewis/Stalnaker semantics because it seems promising insofar as it capably invalidates *modus ponens* without sacrificing weak centering. But I ultimately show that McGee's modification does no better than Lewis/Stalnaker semantics at accommodating CM semantics. The reason is that McGee universally validates import-export (i.e. the principle according to which $(a \,\&\, b) > c$ and $a > (b > c)$ are logically equivalent), but (as I show) there are counterexamples to import-export in the extended CM semantics. This motivates me to develop a variant of McGee's semantics that preserves weak centering but invalidates *both* import-export and *modus ponens*, and that accordingly seems to do a satisfactory job of accommodating CM semantics. If my arguments are successful, then I will have shown that CM semantics is considerably more

revisionary than has been previously realized (because it cannot easily be accommodated by existing accounts of general counterfactual semantics—including McGee's (1985) account, which Briggs does not consider). But I likewise will have shown that CM semantics can be plausibly situated within a more general account that is based on the nearness or similarity of worlds despite what Briggs's discussion of weak centering and *modus ponens* might have led one to think.

2. Structural Equation Models and Counterfactual Semantics

In the graphical approach to causal modeling, a hypothesis about what causes what can be represented as a directed acyclic graph (DAG). DAGs graphically represent the causal relations that obtain among a set of variables **V** as a set of directed edges (or arrows), such that no directed path forms a cycle. Suppose, for example, that we are interested in modeling the following causal system:[9]

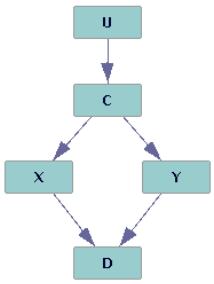
EXECUTION:

A firing squad (consisting of two executioners, X and Y) is ready to shoot a prisoner. If the court orders the execution, this will cause the captain to signal to the executioners. Each executioner will fire if, and only if, the captain signals. Either executioner's shot by itself will be enough to kill the prisoner.

If we allow U to represent whether the court orders the execution, C to represent whether the captain signals, X to represent whether Executioner X fires, Y to represent

[9] This causal system comes from an example in Judea Pearl's *Causality* and is utilized by Briggs in her paper.

whether Executioner Y fires, and *D* to represent whether the prisoner dies, then the following DAG depicts the causal relationships at play in EXECUTION.[10]



Though DAGs like this one are nothing more than pictorial representations of causal relations, interventionists make assumptions that render these graphs full of valuable information. Chief among these assumptions is the Causal Markov Condition (CMC), which entails (among other things) that *intervening* on X severs the functional dependence of X on its causes if an *intervention* on X is defined as an exogenous cause of X that is used to set the value of X to x by means that are causally independent of X's endogenous causes.[11] Though the above DAG is a good representation of what causes what in EXECUTION, it does not specify the exact way in which each variable functionally depends on its causes. For this, we need a set of structural equations.

---

[10] For dichotomous variables like these, it is customary to allow each variable to take the value of 1 when the event in question obtains, and 0 when the event in question does not obtain.

[11] It is no small task to explain how the CMC delivers this result, but there are many good places to find discussion of this. See, for example Meek & Glymour (1994), and Spirtes, Glymour, and Scheines (2000). The basic idea is that the CMC entails that any justifiably omitted cause of a variable contained within a DAG must be probabilistically independent of that variable's endogenous causes, which allows one to define an intervention on X as a justifiably omitted cause of X that (i) has a value on which one can condition to deterministically set X to x for every x in X, (ii) has a value that corresponds to *not* intervening on X (i.e. to allowing the probability distribution over X to be determined by X's causes in $\mathbf{V}$), and (iii) does not co-vary with changes in the probability distribution(s) over X's endogenous causes. Because conditioning on the intervention variable does not affect the probability distribution over X's endogenous causes but does entail that X take a specific value, the functional dependence of X on its endogenous causes is effectively severed.

$C = U$

$X = C$

$Y = C$

$D = \max\{X, Y\}$

This set of equations just expresses what we already know about EXECUTION. Since the captain shoots if and only if he is ordered to do so, the functional dependence of C on U is captured by a structural equation that says that C will take one whatever value U does. The same goes for the dependence of X and Y on C (since each executioner fires exactly when they receive instructions from the captain). And because action by either executioner is sufficient for the prisoner's death, D just takes on the maximum value that either X or Y takes.

With the basic structural equation modeling framework in place, we can now introduce the most prominent CM approach to delivering semantics for non-backtracking counterfactuals. The first step is to assign each variable its actual value, or, equivalently, the value that each variable takes at the actual world. Then, in order to evaluate whether some counterfactual, $a > b$, is true, you check whether b obtains when you *intervene* to make a true, where intervening to make a true amounts to replacing any structural equations encoding A's dependence on its causes with $A = a$ while leaving all other equations unchanged.[12]

Suppose that the court *does* order the execution, so that the captain signals, the executioners shoot, and the prisoner dies. In this case, we first assign each variable its actual value of 1 (because each of the events in question actually happened). Now suppose that we are curious whether the prisoner would have died had Executioner X not shot his weapon.

[12] It is not yet settled how CM semantics should be developed to apply to cases where the causal relationships at play are stochastic, and where the resulting submodel accordingly does not say whether b obtains, but rather says that b obtains with some probability. In this paper, I follow Briggs in sectioning off concerns that arise from stochastic causal relationships by limiting my attention to deterministic causal relationships.

According to standard CM accounts of counterfactual semantics, we should first intervene to make $X = 0$ while leaving each of the variables not causally downstream from X at its actual value. Then we should check whether this intervention results in the prisoner's death. That is, we should apply the following the three-step procedure.

1. Assign each variable in **V** its actual value of 1.
2. Obtain a submodel by intervening to make $X = 0$. (This amounts to replacing $X = C$ with $X = 1$ and leave all other structural equations unchanged.)
3. Determine whether $D = 1$ in the submodel.

In this case, when we leave U, C, and Υ at their actual values of 1, and intervene to make $X = 0$, we find that $D = 1$ (because D takes on the maximum value that either X or Υ takes on, and $\Upsilon = 1$ in the submodel). Thus the standard CM semantics delivers the intuitive non-backtracking result that the prisoner would have died even if Executioner X had not shot her weapon.

How does the CM approach to counterfactual semantics relate to standard non-CM approaches? In the parlance of causal modeling, the truth of some counterfactual, $a > b$, is a function of whether b is realized in the submodel that results from *intervening* to make a true in a model depicting the actual world. So in the parlance of Lewis/Stalnaker semantics, we may be able to regard the closest possible world(s) in which some event happens as the world(s) that result from intervening on the actual world to make that event happen.[13] Either way, by considering what follows from intervening to make some counterfactual supposition true, we identify the states of affairs that serve as truth-makers for claims about what would have happened under the counterfactual supposition. The only difference (at first pass) is that CM

[13] Strictly speaking, Stalnaker (1981) but not Lewis (1973a) would assent to this characterization because of its reliance on the Limit Assumption—i.e. the (1973b) assumption that "as we proceed to closer and closer *a*-worlds we eventually hit a limit and can go no farther." For those who find the Limit Assumption implausible, it is easy to redescribe the worlds that result from intervening to make *a* happen as closer than any worlds that would not result from intervening to make *a* happen. But as I already mentioned in footnote 4, I assume the Limit Assumption for ease of exposition.

semantics describes these states of affairs in terms of submodels, while Lewis/Stalnaker semantics describes these states of affairs in terms of the nearest (or most similar) possible worlds. And since it seems reasonable to think of the values of the variables in a submodel as designating the set of nearest worlds in which some counterfactual state of affairs is true, it may seem like Lewis/Stalnaker semantics and CM semantics make perfect companions. But alas, this is not the case. In the next section, we will discover that Lewis/Stalnaker semantics and CM semantics are incompatible when CM semantics is extended to counterfactuals whose consequents are themselves counterfactuals.

3. Counterfactual Consequents

Though the above treatment of CM semantics appears to deliver desirable results in many simple cases—e.g. when evaluating whether the prisoner would have died even if Executioner X had not shot his weapon—something more is needed in order to apply to counterfactuals whose consequents contain non-truth-functional connectives.[14] This is because it is not clear how one can check whether the consequent of a given counterfactual is true if the consequent contains a non-truth-functional connective. In the event that every connective within the consequent of a given counterfactual *is* truth-functional, we can determine whether the consequent is realized in the submodel by simply checking whether the truth-values of the atomic sentences in the submodel render the more complex truth-functional sentences true or false. (For example, if we are curious whether Executioner X and Executioner Y would shoot were the captain not to order the signal, we can simply check whether Executioner X shoots

[14] It is likewise not yet clear how one should evaluate counterfactuals with logically complex (e.g. disjunctive) antecedents since it is not clear how one can or should intervene to make logical combinations of states of affairs true. (For example, how should one intervene to make it the case that either Executioner X or Executioner Y did not shoot?) Briggs (2012) extends CM semantics to deal with such cases, but this aspect of her extension is only tangentially relevant to this paper because her treatment of disjunctive antecedents does yield any tension between general accounts of counterfactual semantics and CM semantics.

the prisoner and whether Executioner Y shoots the prisoner in the submodel that results from intervening on the actual model to make the captain not signal.) But in the event that a counterfactual's consequent is *not* truth-functional, matters aren't so clear.

In her 2012 paper, "Interventionist Counterfactuals," Briggs extends the semantics such that it delivers truth-conditions for counterfactuals with counterfactual (non-truth-functional) consequents. Her approach, as she argues, is the most natural extension of previous standard accounts of CM semantics.

Suppose that you are curious whether the prisoner would have died had Executioner X fired her gun, were the captain to have abstained from signaling—i.e. whether it is true that $\sim c > (x > d)$. According to standard CM semantics, one should, first, intervene on a model depicting the actual world in order to make the antecedent true, and, second, check the resulting submodel for the truth of the consequent. The first step can be readily applied to this counterfactual: we simply intervene to make it true that the captain does not signal. But it is not immediately clear what to do at the second step since the truth of $(x > d)$ does not follow from the truth-values of x and d in the submodel. In order to extend the semantics, Briggs suggests that we simply check the truth of $(x > d)$ in the submodel in the way suggested by the original procedure—i.e. we first intervene to make it true that x and then check to see whether d obtains in the submodel (or perhaps more appropriately the "sub-submodel"). If d does obtain, then the consequent, $x > d$, obtains in the submodel that results from intervening to make $\sim c$ true, and the target counterfactual is therefore true. And lo and behold, if we intervene to make $\sim c$ true, and then intervene to make x true in the resulting submodel, we find that d obtains in the sub-submodel, and therefore find (as expected) that it is true that $\sim c > (x > d)$.

So far, so good. But as Briggs points out, even this simple extension has vexing implications. Specifically, it entails the invalidity of *modus ponens*, and accordingly seems to entail the violation of weak centering. Consider the following inference.

1. Had Executioner X shot her gun, then the prisoner would have died had the captain not signaled—i.e. $x > (\sim c > d)$.
2. Executioner X shot her gun—i.e. x.
3. Therefore, the prisoner would have died had the captain not signaled—i.e. $(\sim c > d)$.

If we apply the extended procedure to 1, it comes out true. This is because when we intervene to make Executioner X shoot her gun, we sever the functional dependence of X on C—i.e. we replace $X = C$ with $X = 1$. Thus when we intervene to make $C = 0$ in the resulting submodel, the intervention does not exhibit any effect on X. And since the occurrence of x is sufficient for the prisoner's death, the prisoner dies in the sub-submodel—thereby securing the result that 1 is true. Finally, 2 is true by hypothesis, but 3 is false (since intervening to make $C = 0$ in the original model depicting the actual world makes both $X = 0$ and $Y = 0$). So *modus ponens* is invalid in Briggs's extension.

This may not be so bad in and of itself. As Briggs mentions, McGee (1985) makes a compelling case that there are counterexamples to counterfactual *modus ponens* in exactly those cases where *modus ponens* fails in the extension of CM semantics—i.e. cases involving counterfactuals with true antecedents and counterfactual consequents. Though our intuitions may be untutored for such counterfactuals (because, for example, most of the counterfactuals that we say have false antecedents), it is not at all clear that Briggs's extension of the CM semantics gets the *wrong* results for the above inference. Indeed, to the extent that my own intuition goes against Briggs's treatment of 1, it is because of my familiarity with

Lewis/Stalnaker semantics and my corresponding inclination to regard 1 as false.[15] But when I control for my familiarity with Lewis/Stalnaker semantics (as it seems I should in the present context), I tend to agree with Briggs's extension that 1 is true, 2 is true, and 3 is false (and therefore that *modus ponens* is invalid)—especially because 1 sounds equivalent to "Had the captain not signaled and had Executioner X shot her gun anyway, then the prisoner would have died," and it is clear that this counterfactual (whose antecedent is *not* true since the captain did not actually signal) is true. So I (like McGee) am convinced that it is not problematic to invalidate *modus ponens*.

But when the invalidity of *modus ponens* is coupled with the Lewis/Stalnaker interpretation of counterfactuals, it leads to absurd results. If a counterfactual is true only if the closest worlds in which its antecedent is satisfied are worlds in which its consequent is satisfied, then it appears that the actual world is not in the set of worlds closest to itself. At the actual world, 2 is true and 3 is false. But in order for 1 to be true, every closest possible world at which 2 is true must be a world in which 3 is true. So the actual world cannot be a member of the set of closest possible worlds in which 2 is true. But of course the actual world is in the set of worlds most similar to itself! Indeed, it is hard to see why one should go in for a semantics that utilizes the concept of similarity or closeness at all if (according to that semantics) w can fail to qualify as one of the closest/most similar worlds to w. Such a notion of closeness/similarity has no independent traction. Thus Briggs's extension appears to rest on shaky ground when coupled with the standard Lewis/Stalnaker counterfactual semantics.

4. Modus Ponens, Import-Export, and Centering

[15] According to Lewis/Stalnaker, 1 says that the closest world(s) in which Executioner X shoots are worlds in which its consequent, $\sim c > d$, is true. So by their lights, 1 is false (since the actual world is both a member of the set of the worlds closest to itself and a world in which x is true and $\sim c > d$ is false).

Though it is obvious that *weak centering* should not be violated by any semantics that utilizes closeness or similarity of worlds, McGee (1985) has argued that *modus ponens* comes with a cost—namely, any semantics that validates *modus ponens* will not validate *import-export* (i.e. treat $a > (b > c)$ and $(a \,\&\, b) > c$ as logically equivalent) and import-export *prima facie* seems at least as intuitive as *modus ponens*.

Consider McGee's (1985: p. 467) take on the matter:

> We would ordinarily say,
>
> "If Reagan hadn't won the election and a Republican had won, it would have been Anderson."
>
> Appropriately, the Stalnaker semantics, under the natural comparative similarity ordering among worlds, has this sentence come out true. As the law of exportation [import-export] predicts, we also want to say,
>
> "If Reagan hadn't won the election, then if a Republican had won, it would have been Anderson."
>
> However, the possible world most similar to the actual world in which Reagan did not win the election will be a world in which Carter finished first and Reagan second, with Anderson again a distant third, and so a world in which "If a Republican had won, it would have been Reagan" is true. Thus Stalnaker's theory wrongly predicts that in the actual world,
>
> "If Reagan hadn't won the election, then if a Republican had won, it would have been Reagan."
>
> will be true. Thus in this instance, the law of exportation [import-export] is right and Stalnaker semantics is wrong.

McGee's line of reasoning here seems spot on. For the Reagan counterfactuals, Lewis/Stalnaker semantics does not get the intuitive results because Lewis/Stalnaker semantics validates *modus ponens* rather than import-export. This suggests that there is reason to develop a general counterfactual semantics according to which import-export is valid but *modus ponens* is not. (This is the only way to preserve import-export since, as McGee proves, the only two-place connective that validates *both* import-export and *modus ponens* is the material conditional, and it is obvious that the counterfactual conditional is not the material conditional since there are many false counterfactual conditionals whose antecedents are false.)

McGee notes that it is easy enough to modify Lewis/Stalnaker semantics to deliver this result. He (1985: p. 469) writes:

> It is not hard to modify the Stalnaker semantics so that it has the right logical features. Instead of the simple notion of truth in a world, we develop a notion of truth in a world under a set of hypotheses. To be simply true in a world is to be true in that world under the empty set of hypotheses. If there is no world accessible from w in which all the members of Γ are true, then every sentence is true in w under the set of hypotheses Γ. Otherwise [i.e. when there *is* an accessible world from w in which all of the members of Γ are true] we have the following: An atomic sentence is true in w under the set of hypotheses Γ iff it is true in the possible world most similar to w in which all the members of Γ are true… Finally, $a > b$ is true in w under the set of hypotheses Γ iff b is true in w under the set of hypotheses $\Gamma \cup \{a\}$. Thus to evaluate whether $a > (b > c)$ is true under the set of hypotheses Γ, we add first a and then b to our set of hypotheses, and we see whether c is true under the augmented set of hypotheses $\Gamma \cup \{a, b\}$.

Although McGee's description of his idea is complex, the underlying thought is simple. According to Lewis/Stalnaker semantics, $a > (b > c)$ expresses the thought that $b > c$ is true of the nearest worlds in which a is true. But according to McGee's semantics, $a > (b > c)$ expresses the thought that c is true under a set of hypotheses that includes both a and b. So as McGee sees things, the function of the counterfactual conditional is to flag hypotheses under which some atomic sentence is supposed to be true (rather than to relate the antecedent and consequent by some two-place connective). So when I say that had Executioner X shot her gun, the prisoner would have died had the captain not signaled, as McGee sees things, I am saying that it is true that the prisoner dies under a set of hypotheses including both that Executioner X shoots her gun and that the captain did not signal. And since an atomic sentence is true in w under the set of hypotheses Γ if and only if it is true in the possible world(s) most similar (or closest) to w in

which all the members of Γ are true, $a > (b > c)$ is true if and only if c is true in the closest Γ worlds, where Γ must include a & b.[16]

It is easy enough to see how McGee's modification of Lewis/Stalnaker semantics can be paired with CM semantics. In the event that one is curious whether c is true under the set of hypotheses including both a and b—i.e. whether it is true that $a > (b > c)$, or equivalently that $(a \,\&\, b) > c$—we can simply intervene to make a and b true in a model depicting the actual world, and then check to see whether c obtains in the resulting submodel. At first glance, this squares with Briggs's extension. For example, the counterfactual that generates the counterexample to *modus ponens*—i.e. $x > (\sim c > d)$—appears to be equivalent to $(x \,\&\, \sim c) > d$ in her extension because the sub-submodel that results from first intervening to make x true and then intervening to make $\sim c$ true is the same as the submodel that results from intervening to make the conjunction of x and $\sim c$ true (because the order of interventions does not matter for the end result). Moreover, evaluating the counterfactual in this way does not require the semanticist to violate *weak centering* since the claim that the closest x and $\sim c$ worlds are worlds in which d obtains is consistent with the actual world's being closest to itself (even given that 2 is true and that 3 is false at the actual world) because the actual world is itself not a world at which $\sim c$ is true.

Consider the following plausible ordering of worlds where (i) the worlds consistent with the truth-values in the top row are the closest worlds to the world of evaluation, (ii) the worlds are less close to the world of evaluation the further you move down the table except for when adjacent rows are shaded the same color—denoting that those worlds are equivalently close,

[16] Because McGee limits his attention to Stalnaker (and not Lewis), he writes as though there must be one uniquely closest world in which any counterfactual supposition obtains. Since I do not make this assumption, I write as though there could be multiple nearest worlds.

and (iii) the closest worlds in which x and $\sim c$ are true are those consistent with the last row of the table—i.e. the worlds that result from intervening to make x and $\sim c$ true.

	u	c	x	y	d
1)	T	T	T	T	T
2)	T	T	T	T	F
3)	T	T	T	F	T
4)	T	T	F	T	T
5)	T	F	F	F	F
6)	F	F	F	F	F
…	…	…	…	…	…
n)	T	F	T	F	T

This ordering is weakly centered since the actual world is by hypothesis one in which each proposition is true, and each proposition is true in the set of worlds closest to the world of evaluation. Moreover, it appears to square with interventionist thought. This is because (i) the ordering is consistent with the interventionist's conviction that w_x is closer to the actual world than w_y if fewer interventions are required to bring about w_x than w_y (since row 1 depicts the actual world, rows 2 through 6 depict worlds that can be reached by a single intervention, and row n depicts a set of worlds that can be reached only via two interventions), and (ii) the ordering captures the way in which, say, the closest $\sim x$ worlds are the worlds that result from intervening to make $\sim x$ true (rather than the worlds that result from intervening to make

something upstream of $\sim x$ true that nevertheless results in the truth of $\sim x$) since rows 3 and 4 are closer than rows 5 and 6.[17]

Now, if we apply McGee's revised semantics to Briggs's counter-example to *modus ponens* given this ordering, we get the interventionist results that 1 is true (since *d* obtains in the closest x & $\sim c$ worlds), 2 is true (since x is true on row 1), and that 3 is false (since the closest $\sim c$ worlds are not worlds in which *d* obtains). So weak centering does not entail *modus ponens* in McGee's revised semantics, and it may thus seem as though the CM semantics should be paired with McGee's semantics.

But this isn't quite right. While import-export goes through for the sort of counterfactual that renders *modus ponens* invalid in Briggs's extension, it does not go through universally. Consider the following three counterfactuals.

4. Had the captain not signaled, then the prisoner would have died had the captain signaled.
5. Had the captain signaled, then the prisoner would have died had the captain not signaled.
6. Had the captain signaled and not signaled, then the prisoner would have died.

If import-export is valid in the extended CM semantics, then each of these counterfactuals should have the same truth-value. But if we use Briggs's method to evaluate these counterfactuals, it turns out that 4 and 5 do not have the same truth-values and therefore cannot both share their truth-value with 6. Moreover, even though 4 and 6 both come out true in the extended CM semantics, they appear to do so for different reasons.

Let us evaluate these conditionals in turn.

---

[17] Were the worlds represented on row 4 not closer than the worlds represented on rows 5 and 6, then the set of closest $\sim x$ worlds would contain the worlds represented in 5 and 6 which are *not* the worlds that result from intervening on the actual world to make $\sim x$ true.

4) If we *first* intervene to make *c* false, and *then* intervene to make *c* true, we effectively undo the first intervention (since we replace $C = 0$ with $C = 1$) and the counterfactual therefore reduces to its nested consequent, $c > d$, which is true.

5) If we *first* intervene to make *c* true, and *then* intervene to make *c* false, we effectively undo the first intervention (since we replace $C = 0$ with $C = 1$) and the counterfactual therefore reduces to its nested consequent, $\sim c > d$, which is false.

6) There is no possible intervention that makes *c* both true and false, so 6 appears to be vacuously true (or true for the same reasons that counterfactuals with contradictory antecedents are standardly regarded as true).

So if Briggs's extension of the CM semantics is principled (which it seems to be, apart from its relation to Lewis/Stalnaker semantics and *weak centering*), then *neither* import-export nor *modus ponens* is valid in the extended CM semantics. This may be surprising (since it may be *prima facie* reasonable to assume that the correct counterfactual semantics will validate either *modus ponens* or import-export), but, alas, it appears that neither principle of inference is valid in the extended CM semantics.

What explains the validity of import-export for counterfactuals like 1 but not 4, 5, and 6 in Briggs's extension of CM semantics? The relevant difference between these cases is that the two antecedents (i.e. the outermost antecedent and the nested antecedent) represent distinct values of the same variable in counterfactuals like 4 and 5, but not in counterfactuals like 1. This makes a difference because interventions are *modular* (insofar as they change *only* the equation for the intervened upon variable) and thus commutable so long as the relevant interventions are performed on distinct variables. But when an intervention is conducted on some variable *after* having intervened upon that very same variable, the first intervention is effectively undone. In terms of the examples, intervening first to make x true and second to

make $\sim c$ true reduces to intervening to make x & $\sim c$ true (because interventions are modular and thus commute when on different variables), but intervening first to make c true and then to make $\sim c$ true is not the same as intervening to make c & $\sim c$ true. There is no intervention that makes a contradiction obtain.

5. Revising McGee's Revision of Lewis/Stalnaker Semantics

Let us take stock. Briggs shows that *modus ponens* is invalid in her extended CM semantics. When CM semantics relies on Lewis/Stalnaker semantics for the interpretation of the counterfactual, this results from the actual world not being a member of the set of worlds closest to itself. But that's absurd, so if we must search for some other general semantics for counterfactuals in which we can embed CM semantics. At first glance, it may seem as though we should follow McGee in opting for import-export over *modus ponens* (because the resulting semantics can invalidate *modus ponens* but preserve weak centering). But as things turn out, import-export is itself invalid in the most natural extension of CM semantics. So it seems that we are left searching for some counterfactual semantics that preserves weak centering but invalidates both *modus ponens* and import-export.

Luckily, McGee's modification of Lewis/Stalnaker semantics can be easily revised so that it checks these boxes. To do this, we keep with McGee's notion of truth in a world under a set of hypotheses, but tell a different story about the way in which counterfactuals flag hypotheses that are contained in Γ. Specifically, we agree with McGee that in most circumstances $a > (b > c)$ identifies a and b as hypotheses under which c is supposed to be true, but slightly revise the view such that when we augment Γ with a hypothesis, h, that contradicts some other hypothesis (or hypotheses) already in Γ, the resulting set is not comprised of the union of Γ and h (which in this case would result in there being a contradiction in Γ), but rather

includes h but *excludes* whatever hypotheses in Γ contradict h. Thus when we evaluate 4—i.e. $\sim c > (c > d)$—if we *first* add $\sim c$ and *then* add c (as the conditional suggests we should), then the *second* addition will boot $\sim c$ from the set of hypotheses that we suppose d to be true under, and the counterfactual therefore reduces to $c > d$. It is worth noting that this wrinkle does not come up when we evaluate 6—i.e. $(\sim c \,\&\, c) > d$—since the structure of the counterfactual does not tell us to first add one of the conjuncts and then the other. Rather, if $c \,\&\, \sim c$ must be added in one fell swoop, then the counterfactual will come out as true since (according to McGee's semantics) every sentence is true in w under the set of hypotheses Γ when there is no world accessible from w in which all the members of Γ are true.

This modification appears to get things exactly as it should. Remember that even though 4 and 6 both were true in the extended CM semantics, they were true for different reasons. That is, 4 was true because intervening to make the nested antecedent, c, true undid the original intervention to make $\sim c$ true (since we replaced $C = 0$ with $C = 1$) and the counterfactual therefore seemed to reduce to its nested consequent, $c > d$ (which is true). But 6 was not true because some intervention undid another. Rather, 6 was true because there was no possible intervention to make c both true and false. This revision of McGee's semantics accounts for exactly this difference. In 4, c (but not $\sim c$) is in the set of hypotheses under which d is supposed to be true. But in 6, both $c \,\&\, \sim c$ are in the set of hypotheses under which d is supposed to be true (because they are added in one fell swoop) and 6 therefore counts as vacuously true.

The key innovation in this revision of McGee's semantics, then, is that the hypothesis flagged by some nested antecedent has priority over whatever hypotheses are flagged by the antecedent(s) of the counterfactual(s) in which it is nested. Put more simply (albeit more roughly), later antecedents take priority over earlier antecedents in the revision, but not in

McGee's semantics. Most of the time, this priority amounts to nothing (since there is no cleaning up to do when the nested antecedent is consistent with whatever is in Γ), but when a nested antecedent contradicts something already in Γ, the two semantics come apart.[18]

Given this innovation, we may wonder how we should evaluate counterfactuals when the flagged hypothesis contradicts some hypothesis that is already contradicted by some other hypothesis in Γ. That is, if we wish, for example, to evaluate $(a \,\&\, \sim a) > (a > b)$, does the latter addition of a result in $\sim a$'s exclusion from the set, and thereby render the target counterfactual not vacuously true? In order to develop the revised semantics such that it captures the verdicts of CM semantics, the latter addition of a should *not* result in $\sim a$'s exclusion from the set. This is because CM semantics asks us to intervene to make the contradiction true in order to consult the resulting submodel, but there is no intervention that makes the contradiction obtain, and therefore no way to consult any submodel (or sub-submodel). This means that we can accommodate CM semantics by requiring that $(a \,\&\, \sim a) > (a > b)$ shares its truth-value with any

[18] It may seem like the revised semantics does not deliver the desired results because import-export is supposed to fail in CM semantics whenever the nested antecedent is a different value of the same variable as the primary antecedent, and different values of the same variable need not be logically inconsistent (since variables need not be partitioned such that they represent whether a sentence is true). For example, if we partitioned *X* in EXECUTION such that it could take on three values—one for shooting left-handed, one for shooting right-handed, and one for not shooting—then intervening to make Executioner X shoot right-handed would cancel out any earlier intervention to make Executioner X shoot left-handed even though shooting left-handed and shooting right-handed are logically consistent. Thus it may seem as though import-export fails in CM semantics more often than it fails in the revised McGee semantics. Though it is true that variables need not be partitioned such that they represent whether a sentence is true—i.e. partitioned such that their only two values are a and $\sim a$—distinct values of a different variable are mutually exclusive and therefore impossibly jointly realized. This is because the set of values that a variable can take on must comprise a logical partition, where a logical partition is a set of events, or descriptions of events, that are mutually exclusive and collectively exhaustive. This means that even when we define *X* such that it is three-valued, we must regard each value of *X* as inconsistent with every other, and the revised McGee semantics therefore gets the desired results.

counterfactual whose antecedent is contradictory—i.e. by regarding $(a \,\&\, \sim a) > (a > b)$ as vacuously true.[19]

This concludes the development of the revision to McGee's semantics. The revised semantics retains McGee's ability to accommodate the verdicts of CM semantics when *modus ponens* fails (without sacrificing weak centering) but improves on McGee's semantics by likewise predicting the failure of import-export exactly when it fails in CM semantics. That is, since (i) *modus ponens* fails only when the truth-value of some counterfactual, $a > (b > c)$ differs from the truth-value of its consequent, $b > c$, and (ii) the truth-value of $a > (b > c)$ reduces to the truth-value of $b > c$ whenever import-export fails in the revised semantics, the revised semantics (like McGee's semantics) invalidates *modus ponens* while preserving weak centering by validating import-export whenever *modus ponens* fails. But since the revised semantics does *not* license import-export (unlike McGee's semantics) in the cases where import-export fails in CM semantics, the revised semantics succeeds where McGee's fails. Thus, of the options on the table, the revised semantics does the best at accommodating CM semantics.

6. Conclusion

My argument that the revised version of McGee's semantics can accommodate CM semantics is now complete. I take myself to have shown that we can accommodate the verdicts of CM semantics without sacrificing weak centering if we (i) follow McGee in interpreting the truth of a counterfactual in terms of the truth of an atomic sentence under a set of hypotheses,

[19] That said, quite apart from the revised semantics' relation to CM semantics, it is worth noting that this is a choice point in the development of the revised semantics. In the event that linguistic intuition disagrees with CM semantics, one could just as easily insist that the latter addition of a should not result in the exclusion of $\sim a$ from the set of hypotheses under which b is evaluated. Of course this involves making the revised semantics incompatible with CM semantics, but if our intuition that $(a \,\&\, \sim a) > (a > b)$ and $(a \,\&\, \sim a) > b)$ are not logically equivalent is quite strong, then it may be worth revisiting CM semantics' treatment of contradictory antecedents (since there may likewise be choice points in the development of CM semantics on this front). At any rate, I don't have strong intuitions about these counterfactuals, so I am okay leaving things as they are.

but (ii) revise McGee's conditions for set membership such that augmenting Γ with a hypothesis, *h*, that contradicts some other hypothesis (or hypotheses) already in Γ results in a set that excludes whatever hypotheses in Γ contradict *h*.

But since much of the discussion of CM semantics has been devoted to sentences that we hardly ever say, it seems prudent to reflect on why this work is worth doing. It is generally thought that the most promising general accounts of counterfactual semantics utilize similarity orderings on worlds, and that non-backtracking counterfactuals differ from other counterfactuals only with respect to the orderings on worlds at hand. But since the orderings supplied by CM semantics must violate weak centering when plugged into what is often considered the most promising general account of counterfactual semantics (i.e. Lewis/Stalnaker semantics), it may seem as though we should abandon CM semantics—perhaps especially because Lewis/Stalnaker semantics has withstood the test of time while CM semantics has not.[20] Though this line of reasoning is *prima facie* attractive, CM semantics offers what is to date the most rigorous characterization of our reasoning about what would have happened had someone intervened on the goings-on of the actual world. So it seems at least worth considering siding with CM semantics over Lewis/Stalnaker semantics—especially since the conflict between CM semantics and Lewis/Stalnaker semantics does not reveal itself until we consider sentences that we hardly ever say or think about, and it is not clear that Lewis/Stalnaker gets the right truth-values for these sentences (because it is not clear what their truth-values are).

This paper is my exploration of siding with CM semantics. Though I find siding with CM semantics attractive for the reasons stated above, it would be wrong to do so without offering some alternative general account of counterfactuals in which the orderings on worlds

[20] Lewis/Stalnaker semantics has been prominent since the 1970s, while CM semantics did not appear until Galles and Pearl (1998).

entailed by CM semantics (i.e. entailed by facts about what submodels result from intervening to make counterfactual suppositions true) can be utilized. In this paper, I have argued that my revision of McGee's semantics is up to the task. Though I have not focused much on the intuitive plausibility of my revision, it seems to inherit any intuitive confirmation conferred upon Lewis/Stalnaker semantics for its treatment of ordinary counterfactuals (because it agrees with Lewis/Stalnaker semantics about every counterfactual that does not contain an embedded counterfactual) as well as any intuitive confirmation conferred upon McGee's semantics for its treatment of import-export in ordinary cases (because it agrees with McGee's semantics that import-export is successful except in extraordinary cases).

If this is right, then CM semantics is on better footing than we might have previously thought (because there are possible general accounts of counterfactual semantics that can accommodate CM semantics without sacrificing weak centering). But it doesn't stop there. If my revision to McGee's semantics can accommodate CM semantics without thereby entailing any obviously counterintuitive results, then my revision to McGee's semantics should also be regarded as a serious competitor to Lewis/Stalnaker semantics.

Chapter 2: Interventionist Decision Theory

ABSTRACT

Jim Joyce has argued that David Lewis's formulation of causal decision theory is inadequate because it fails to apply to the "small world" decisions that people face in real life. Meanwhile, several authors have argued that causal decision theory should be developed such that it integrates the interventionist approach to causal modeling because of the expressive power afforded by the language of causal models, but, as of now, there has been little work towards this end. In this paper, I propose a variant of Lewis's causal decision theory that is intended to meet both of these demands. Specifically, I argue that Lewis's causal decision theory can be rendered applicable to small world decisions if one analyzes his dependency hypotheses as causal hypotheses that depend on the interventionist causal modeling framework for their semantics. I then argue that this interventionist variant of Lewis's causal decision theory is preferable to interventionist causal decision theories that purportedly generalize Lewis's through the use of conditional probabilities. This is because Lewisian interventionist decision theory captures the causal decision theorist's conviction that any correlation between what the agent does and cannot cause should be irrelevant to the agent's choice, while purported generalizations do not.

1. Introduction

In his 1999 book, *The Foundations of Causal Decision Theory*, Jim Joyce argues that many standard formulations of causal decision theory are flawed because they fail to apply to the "small-world" decision problems that people face in real life. Joyce's idea is, first, that causal decision theory is often formulated such that it applies only when the agent is working with incredibly rich state-descriptions, and, second, that bounded agents like us seldom (if ever) have such rich state-descriptions at our disposal. This motivates Joyce to develop a version of causal decision theory that is partition-invariant in the sense that it applies no matter how the agent carves up the world.

Meanwhile, there is pressure from elsewhere—e.g. Meek and Glymour (1994) and Hitchcock (2015)—to develop causal decision theory that utilizes the interventionist approach to causal modeling found in e.g. Pearl (2009) and Spirtes, Glymour, and Scheines (2000). The thought is that causal models provide a mathematically rigorous language in which one can represent any causal facts that are relevant to a particular decision-making context, and that it would be imprudent for causal decision theorists not to help themselves to the expressive power of this language. Though these authors have provided compelling arguments, there has been little work on developing an analysis of expected utility that makes explicit use of the causal modeling framework.

In this paper, I attempt to develop a causal-decision-theoretic analysis of expected utility that both applies to small-world decision problems and makes explicit use of the interventionist approach to causal modeling. My strategy is to tinker with David Lewis's analysis of expected utility by construing his *dependency hypotheses* as causal hypotheses that depend on the axioms of the causal modeling framework for their semantics. It may be surprising that I use Lewis's decision theory for these purposes since it is famously *not*

partition-invariant. But in what follows, I argue, first, that causal models give Lewis the tools to construe dependency hypotheses such that they are not too rich for humans to entertain, and, second, that the most obvious strategy for developing a partition-invariant version of interventionist decision theory is bound to fail.

2. Act/State Dependence

Leonard Savage (1954) famously proposed that agents should take whatever action maximizes expected utility when the expected utility of an action, A, is defined as follows:[21]

$$U(A) =^{df} \sum_S C(S)V(A,S).$$

Here, C represents an agent's subjective probability function; V represents her value function; and S ranges over states of the world. Savage's idea is that the outcome of an action at a world can be represented as what occurs when the action takes place at that world.[22] If an agent assigns probabilities to different states of the world—i.e. to different ways that the world might be—then she should determine the expected utility of her action by calculating a weighted average of the utilities she attaches to the various states, where the weights correspond to the probabilities that she assigns.

Though decision theorists agree that Savage's decision theory does well when the states at hand are independent (in some sense) of the acts, it is agreed by everyone (including Savage) that "it would be ridiculous to maximize 'expected utility' when S ranges over just any old partition [of states]."[23] A simple example illustrates this.

	Play well	**Do not play well**
Practice	Practice and play well	Practice and do not play well

[21] Savage expressed his definition of expected utility slightly differently, but the content is the same.
[22] More formally, Savage conceives of actions as functions from states to outcomes.
[23] Lewis (1981a), pp. 12-13.

Do not practice	Do not practice and play well	Do not practice and do not play well

Here, Iverson has a choice between two acts—practicing and not practicing—and is considering consequences of actions in states of the world in which he plays well and does not play well. Were Iverson to apply Savage's decision theory to these acts and states, he would quickly realize that he should abstain from practicing. This is because (given Iverson's disdain for practice) not practicing *dominates* practicing—i.e. not practicing beats practicing no matter whether the world turns out to be one in which he plays well or does not play well, thereby rendering the probabilities that Iverson assigns to the states irrelevant to his choice. Of course, it is easy to see what plagues Iverson's strand of reasoning. His deliberation goes astray because whether he plays well *depends* (to some extent) on whether he practices. It does not make sense for Iverson to deploy the unconditional probability of playing well (or not playing well) as he determines the expected utility of practicing (or not practicing) since this probability is not the same as the probability that he plays well (or does not play well) in worlds where he decides to practice (or not practice).

Because of this, the decision theorist must either (i) provide the agent with a guide for constructing partitions of states that do not yield such undesirable results (in order to render Savage's decision theory useful) or (ii) provide a conception of expected utility that is distinct from Savage's in the sense that it generalizes to cases where acts and states are dependent.[24] The decision theorist must also take a stance on what *kind* of dependence is relevant to decision theory. Newcomb problems show us that acts and states can be evidentially dependent but

[24] Brian Skyrms (1980) and David Lewis (1981a) are examples of decision theorists who take the first route. Jim Joyce (1999) and Richard Jeffrey (1983) are examples of decision theorists who take the latter route. Taking the latter route amounts to providing a *partition-invariant* analysis of expected utility and usually involves trading in Savage's unconditional probabilities of states for conditional probabilities of states given acts.

causally independent. It is likewise possible for acts and states to be evidentially independent but causally dependent.[25] The decision theorist must therefore establish whether dominance reasoning (like Iverson's) is permissible when acts and states are evidentially independent or when acts and states are causally independent—i.e. whether to be an evidential decision theorist or a causal decision theorist.

In what follows, I propose a version of causal decision theory that follows in the footsteps of Savage-style causal decision theorists—i.e. causal decision theorists who do *not* provide analyses of expected utility that purportedly generalize to cases where acts and states are causally dependent. This may seem like easy work since limiting the application of Savage's theory to all and only those partitions of states that are causally independent of the agent's options (or acts) seems to deliver causal-decision-theoretic results. But the decision theory that results from limiting the application of Savage's theory in this way is not sufficiently general because an agent can be reasonably *uncertain* about whether some state of affairs is causally independent of her options, and the resulting theory says nothing about what agents should do given this uncertainty. The decision theory I propose, on the other hand, is intended to provide

---

[25] I believe that this has gone unnoticed in the decision theory literature. In abstract, consider a case where X causes Y, and where Z is a common cause of X and Y. If $X \rightarrow Y$ offsets the probabilistic dependence between X and Y that obtains in virtue of $X \leftarrow Z \rightarrow Y$, then X and Y are causally dependent (in virtue of $X \rightarrow Y$) but evidentially independent (since $X \rightarrow Y$ offsets the evidential dependence that results from $X \leftarrow Z \rightarrow Y$). In the causal modeling literature, this sort of case is referred to as a failure of faithfulness owing to path cancellation. (See Zhang and Spirtes [2008] for more discussion of failures of faithfulness.) In concrete, consider a case in which an oracle tells you, first, that whether one gets lung cancer causally depends on whether one smokes (in the sense that smoking does something to one's body to increase the risk of lung cancer), and, second, that there exists a genetic condition that causes people to smoke and to be at reduced risk of lung cancer. Among people who possess the genetic condition, smoking increases the probability of getting lung cancer. The same goes for people who do not possess the genetic condition. But, as it happens, the unconditional probability of getting lung cancer is equivalent to the probability of getting lung cancer given that one smokes (because the prevalence of the genetic condition exactly counterbalances the probabilistic effect of smoking on the body). You do not know whether you possess the genetic condition. Should you smoke? In this case, the evidential decision theorist would seemingly think that dominance reasoning is applicable (and that you should therefore smoke), while the causal decision theorist would not think that dominance reasoning is applicable.

agents with advice even when they are uncertain about the causal structure of the world. So if the decision theory I propose is as advertised, then it not only succeeds at delivering causal-decision-theoretic verdicts, but also applies to the wide array of decision-making contexts where agents are reasonably uncertain about what their actions cause and in what particular way.

3. Causal Decision Theory

Fisher's Newcomb Problem (FNP):[26]

You must decide whether to smoke. You enjoy the pleasures of smoking, but are worried about its correlation with lung cancer. An oracle tells you that the association between smoking and lung cancer is fully explained by some genetic condition that causes people to smoke and get lung cancer. The oracle also tells you that smoking does not cause lung cancer.

According to causal decision theory, you should smoke because (i) you prefer worlds in which you smoke to worlds in which you do not smoke no matter whether the world turns out to be one in which you get lung cancer, and (ii) whether you smoke has no causal effect on whether you get lung cancer. Is causal decision theory right? This question is hard to answer in part because there are competing intuitions. But in what follows, I search for a decision theory that gets causal decision theorists what they want—i.e. a decision theory according to which dominance reasoning is applicable to FNP because there is nothing that we can do to change our genes. The hope, then, is to deliver an interventionist decision theory that is both useful to human agents and that respects the causal decision theorist's conviction that any

[26] In his (1959) *Smoking: The Cancer Controversy*, R. A. Fisher entertains (but does not espouse) the hypothesis that the correlation between whether one is a smoker and whether one suffers from lung cancer is due not to some causal influence that smoking exerts on the health of one's lungs, but rather to the causal influence that one's genetic makeup has both on whether one smokes and whether one suffers from lung cancer.

correlation between what the agent does and cannot cause is irrelevant to what the agent should do.

In his 1981 paper, "Causal Decision Theory," David Lewis proposes a version of causal decision theory according to which agents should take whatever action maximizes expected utility when the expected utility of an action, *A*, is defined as follows:

$$U(A) =^{df} \sum_K C(K)V(A,K).$$

Lewis's analysis is identical to Savage's, except for its inclusion of *K*, which ranges over *dependency hypotheses* that may or may not be true of her world. What is a dependency hypothesis? According to Lewis, it is "a maximally specific proposition about how the things ⌜the agent⌝ cares about do and do not causally depend on ⌜the agent's⌝ present actions." So dependency hypotheses not only specify what is and is not within the agent's control, but also specify, first, the objective chances of any events that lie beyond the agent's control, and, second, the effect of what the agent does on the objective chances of any events that are within the agent's control. Because these dependency hypotheses are so specific, if an agent knows that some dependency hypothesis is true, she can readily determine what world will result from her acting in a particular way. But since Lewis does not assume that the agent must be certain of any particular dependency hypothesis in order to apply his decision theory, Lewis's version of causal decision theory applies even when agents are unsure about the causal structure of the world.

Because Lewis's dependency hypotheses contain information about what is and is not within the agent's control, his decision theory capably delivers what many causal decision theorists want: the verdict that agents should smoke when confronted with FNP. So long as each of the dependency hypotheses that the agent entertains says that the chance of getting lung cancer is causally independent of whether the agent smokes, the agent has enough

information to reason her way through the dominance argument that causal decision theorists find so attractive—i.e. that no matter which dependency hypothesis is true, she is better off enjoying the pleasures of smoking.[27] But while Lewis's dependency hypotheses initially seem to give causal decision theorists what they want, they do not by themselves seem to be very useful for real-life human decision-making agents. This is because, as Joyce notes, we lack the cognitive capacity to entertain (much less form probability judgments about) propositions as rich as maximally specific dependency hypotheses.

What, then, are we humans to do? Is Lewis's decision theory useful for cognitive creatures like us? As I interpret Lewis, to the extent that he offers any advice toward this end, he suggests (following Gibbard and Harper [1978]) that we can represent less specific dependency hypotheses as non-backtracking counterfactual conditionals, where each counterfactual designates a *set* of maximally specific dependency hypotheses.[28] But Lewis's proposal is justifiably controversial.[29] In order for probabilities of counterfactuals to deliver causal-decision-theoretic verdicts, the semantics for the relevant counterfactuals must take a

---

[27] Moreover, because Lewis's dependency hypotheses are maximally specific, they are guaranteed to capture all of the agent's causal influence over the world, and are therefore guaranteed to be causally independent of what the agent does. Lewis (1981a, p. 13) provides the following proof by *reductio ad absurdum* that his dependency hypotheses are causally independent of the agent's actions: "Suppose [the dependency hypotheses are not causally independent of the agent's actions]. Consider the dependency hypothesis which we get by taking account of the ways the agent can manipulate dependency hypotheses to enhance his control over other things. This hypothesis seems to be right no matter what he does. Then he has no influence over whether this hypothesis or another is right, contrary to the supposition that the dependency hypotheses are within his influence."

[28] Lewis believes that this strategy works because, as he sees things, a maximally specific dependency hypothesis can be represented as either a complicated counterfactual whose consequent specifies everything about what would follow (and with what chance) were the agent to take any of the feasible actions, or, alternatively, a conjunction of more ordinary counterfactuals that together contain the same information. So, in the same way that we regard a conjunct as designating every conjunction with which it is consistent, we can regard a single counterfactual as designating every conjunction of counterfactuals (and therefore every maximally specific dependency hypothesis) with which it is consistent.

[29] Eells (1982) covers much of the controversy in chapter 5 of *Rational Decision and Causality*.

very specific non-backtracking form that has seemed implausible to many.[30] Moreover, as Ellery Eells has forcefully argued, even if the counterfactual approach is capable of giving causal-decision-theoretic answers, it does so for the wrong reasons. Eells (1982, p. 105) writes:

> "Giving the right answer is not the same as giving the right answer for the right reason. The crucial thing about pure, simple Newcomb situations is a certain causal feature: the absolute causal independence of the chance of the symptomatic outcome *SO* [e.g., lung cancer] (or of the chance of the common cause *CC*) from the symptomatic act *SA* [e.g., smoking]. But [the counterfactual approach] does not deal with this causal feature but rather with something, which, given a suitable theory of counterfactuals, would be a consequence of the fact that Newcomb situations have that feature: namely, the equality of $P(SA > SO)$ and $P(\sim SA > SO)$ and of $P(SA > CC)$ and $P(\sim SA >$ CC). In order to get to the causal heart of the matter . . . one must construct a suitable theory of counterfactuals which is sensitive in the right way to a state's being within or outside of the agent's influence and which has these equalities as a consequence."

When we confront FNP, we believe that the correlation between genetic makeup and whether we smoke is irrelevant to our decision because our genetic makeup is beyond our control, not because we are sure that the best semantics for the relevant counterfactuals is a particular non-backtracking semantics. Indeed, the overwhelming majority of human decision-makers have no idea whether the best semantics for the relevant counterfactuals is a non-backtracking one, but nevertheless believe that they should smoke because their actions do not cause their genetic makeup (perhaps because they believe that causes must precede their effects). It thus seems that if there were some other way of rendering Lewis's analysis of expected utility useful for human decision-makers—i.e. some way that did not rely on a particular counterfactual semantics—then it would, all other things being equal, be preferable to Lewis's own proposal.

4. Replacing Counterfactuals with Causal Hypotheses

[30] The semantics must not backtrack in the sense that counterfactual dependence must never flow in the opposite direction of causal dependence.

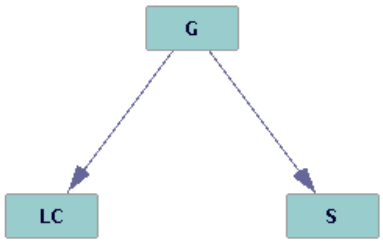
Let us take stock. Lewis's definition of expected utility in terms of maximally specific dependency hypotheses is attractive because it captures the irrelevance to deliberation of what lies beyond our control. But Joyce is right to press Lewis on whether his decision theory is useful for bounded creatures like us, since we lack the ability to entertain maximally specific dependency hypotheses. In order to get around this difficulty, Lewis suggests that we might utilize the probabilities of counterfactuals, where each counterfactual designates a set of maximally specific dependency hypotheses. It is possible that this strategy secures the desired verdicts (though I have my doubts), but I agree with Eells that even if it does, it does so for the wrong reasons. More specifically, it seems like invoking counterfactual semantics may overcomplicate things since our reasoning about Newcomb problems naturally goes through without recourse to counterfactuals.

In this section, I offer an alternative proposal for how we can render Lewis's analysis of expected utility useful. In particular, I suggest that we trade in probabilities of counterfactuals for probabilities of causal hypotheses, where causal hypotheses are ordered pairs of directed acyclic graphs and sets of objective chance distributions that obey the Causal Markov Condition (which, as Hausman and Woodward [1999] argue, is "implicit in the view that causes can be used to manipulate their effects"). But before considering how causal hypotheses may be of use to Savage-style causal decision theorists (e.g. Lewis), it is important that we first acquaint ourselves with some machinery in the interventionist's toolbox (including the Causal Markov Condition).

In the interventionist framework, a hypothesis about what causes what can be represented as a directed acyclic graph (DAG). More specifically, a DAG graphically represents the causal relations that obtain among a set of variables **V** as a set of directed edges (or arrows), such that no directed path forms a cycle. For example, the interventionist can represent the

causal relations implied by FNP in the following DAG. Allow *LC* to represent whether one gets lung cancer, *G* to represent whether one has the relevant genetic condition, and *S* to represent whether one smokes.[31]

Figure 1:



Though DAGs like this one are nothing more than pictorial representations of causal relations, interventionists make assumptions about causal relations that render these graphs full of valuable information. Chief among these is the Causal Markov Condition (CMC), which narrows down the set of probability distributions that are compatible with a given DAG.

The CMC is satisfied by a given DAG and probability distribution if and only if every variable *X* in **V** is probabilistically independent of its nondescendants conditional on its parents, where *X*'s *parents* are *X*'s most immediate causal predecessors, and *X*'s *nondescendants* are the variables in **V** that are not causally downstream from *X*.

The CMC is a bit of a mouthful, but Judea Pearl and his collaborators have neatly summarized some of its implications in graphical terms. Given the CMC, if every path between a pair of variables in **V** is *d-separated* (according to a given DAG), then the pair of variables

[31] We can regard each of these variables as dichotomous—i.e. as taking one value when the event in question obtains, and another when the event in question does not.

must be probabilistically independent of each other.[32] A path between two variables, F and G, is *d-separated* if and only if:

i. the path contains a *non-collider* that has been conditioned on, e.g., $F \rightarrow \underline{H} \rightarrow G$ or $F \leftarrow \underline{H} \rightarrow G$ (where H has been conditioned on), or

ii. the path contains a *collider* that has not been conditioned on (and whose descendants have not been conditioned on), e.g., $F \rightarrow H \leftarrow G$.[33,34]

So, from a given DAG, we can determine that certain variables must be probabilistically independent of each other. For example, given the DAG of FNP, we can determine that LC must be independent of S, conditional on any value of G.

The CMC likewise entails a version of Reichenbach's (1956) Principle of the Common Cause—i.e. that if variables F and G are correlated, then either F (directly or indirectly) caused G, G (directly or indirectly) caused F, or F and G are (direct or indirect) joint effects of a common cause. For this reason, the CMC is plausibly assumed only of variable sets that are "causally sufficient"—i.e. of variable sets for which it is the case that every common cause of any two or more variables in $\mathbf{V}$ is in $\mathbf{V}$.[35]

How does this help the causal decision theorist? With the CMC in hand, one can compute the effect of *intervening* on a causal system to set a variable to a particular value.

---

[32] Geiger and Pearl (1989) and Verma (1987) prove that the d-separation criterion characterizes all and only the conditional independence relations that follow from the CMC.

[33] A variable is a *collider* along a directed path if and only if it is the direct effect of two variables along the path. This is why H is a collider along $F \rightarrow H \leftarrow G$ but not $F \rightarrow H \rightarrow G$.

[34] The parenthetical is important because conditioning on the descendant of a collider often induces a spurious correlation between the collider's ancestors. See Elwert and Winship (2014) for discussion of this phenomenon.

[35] Why does the CMC entail Reichenbach's principle? In the event that some pair of variables are dependent and neither is a descendant of the other, there must exist some parent(s) of both variables on which one can condition to render the relevant variables independent. So it is provable of every system of variables that satisfies CMC, first, that if neither F nor G (directly or indirectly) influences one another and F and G are probabilistically dependent, then there exists a set C of variables not containing F or G but causing both, and, second, that F and G must be independent conditional on any value of C.

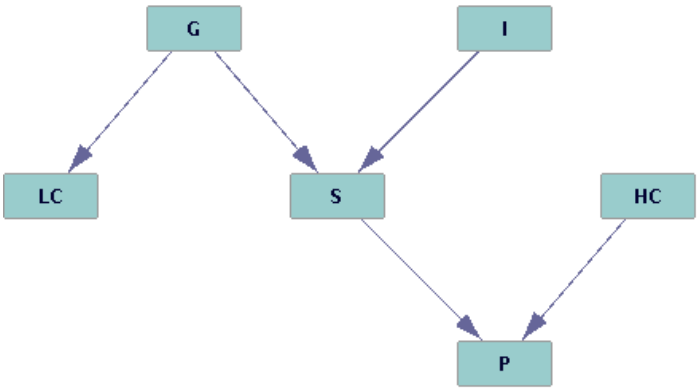
Following Spirtes, Glymour, and Scheines (2000) and Pearl (1993, 2009), allow the *intervention* on X to represent some justifiably omitted cause of X (i.e. some cause of X that is neither included in **V** nor a direct cause of any variable in **V** other than X) that can be exploited to set X to x. More specifically, allow the intervention variable, first, to represent some cause of X that is neither included in **V** nor a direct cause of any variable in **V** other than X, and, second, to be partitioned such that it contains (i) a value on which one can condition to deterministically set X to x for every x in X and (ii) a value that corresponds to *not* intervening on X (i.e. to allowing the probability distribution over X to be determined by X's causes in **V**).

Since the intervention on X must be a justifiably omitted cause of X, it must be d-separated from any of X's non-descendants in **V** (including its causal predecessors).[36] This means that we can secure the results that are sought by causal decision theorists simply by modeling decisions as interventions in Markovian models.[37] Consider the following augmented model of FNP, in which we take stock, first, of the effect that smoking has on one's popularity, *P*, second, the effect on one's popularity that is caused by one's hair color, *HC*, and, third, the effect of *intervening* to make oneself smoke, where *I* represents the intervention.

Figure 2:

[36] The intervention variable is d-separated from X's causal predecessors because X is a collider on every path from the intervention variable to X's causal predecessors. And since the intervention variable cannot be causally downstream from any variable(s) in **V** and likewise cannot be a common cause of any variables in **V**, the fact that X is d-separated from any non-descendants that are not causal predecessors of X guarantees that the intervention variable is d-separated from all of X's non-descendants.

[37] The informal discussion below should help the reader understand how assuming the CMC fuels the interventionist's ability to deliver these results, but interested readers should consult pp. 75-81 of Spirtes, Glymour, and Scheines's *Causation, Prediction, and Search*. Their Manipulation Theorem precisely characterizes the effects of conditioning on an intervention variable in Markovian models.



By the causal decision theorist's lights, *G*, *LC*, and *HC* should be irrelevant to the agent's deliberation—since the agent can do nothing to change her genetic condition, her hair color, or whether she gets lung cancer—even though *G* and *LC* are correlated with smoking. Though the CMC does not rule out any correlation between *G* or *S* and *LC*, it *does* rule out any correlation between *G* or *LC* and *I* (since *S* is a collider on every path between these variables). So even though it is consistent with the CMC that smoking raises the probability of getting lung cancer (i.e. that whether one smokes is correlated with whether one gets lung cancer), it is *inconsistent* with the CMC that *intervening* to make oneself smoke raises the probability of getting lung cancer (i.e. that whether one intervenes to make oneself smoke is correlated with whether one gets lung cancer). And since *HC* is also d-separated from *I* (and *S*, for that matter), the CMC entails that intervening to make oneself smoke is probabilistically independent from everything in **V** that is beyond the agent's control. This means that modeling the expected utility of smoking in terms of intervening to make oneself smoke—i.e., in terms of realizing a specific value of *I* (rather than *S*)—secures the results that causal decision theorists favor.[38, 39]

---

[38] I am not the first to point this out. See Hitchcock (2015), Meek and Glymour (1994), and Pearl (2009).
[39] As a referee points out, construing deliberation in terms of realizing a particular value of *I* gives the agent one more option than she would have were she to deliberate about what value of *S* to realize—namely, the option *not* to intervene (or to allow the distribution over *S* to continue to be determined by its endogenous causes). This may buck tradition, but the change seems welcome. In the context of choice, it often seems as though we not only can choose what to do (among some list of actions) but also

In order to integrate interventions into Lewis's decision theory, we allow *K* to range over causal hypotheses relative to some variable set **V** that the agent identifies as causally sufficient and apply the value function to *K* and *do*(*A*=*a*), where *do*(*A*=*a*) represents what happens when we intervene to make ourselves take the relevant action.[40,41] (The *do*-notation comes from Pearl, and is just another way of describing the effect of conditioning on some value of an intervention variable, which is standardly excluded from **V**.)[42,43] According to interventionist decision theory (IDT), then, the expected utility of acting in a particular way is defined as follows:

$$U(a) =^{df} \sum_K C(K)V(do(A = a), K).$$

Since the CMC implies constraints on a probability distribution only relative to a DAG, the *K* partition must range over ordered pairs of DAGs and probability distributions (or sets of probability distributions) that jointly satisfy the CMC. Thus when an agent entertains, say, two DAGs (call them DAG_1 and DAG_2) and four distinct probability distributions over some **V**—two of which are compatible with DAG_1 (call them P_1 and P_2) and two of which are compatible with DAG_2 (call them P_3 and P_4)—then the *K* partition consists of the following four causal hypotheses (or ordered pairs): $< DAG_1, P_1 >$, $< DAG_1, P_2 >$, $< DAG_2, P_3 >$, $< DAG_2, P_4 >$.[44] According to IDT, then, an agent should determine the expected utility of *a*-ing by calculating

---

can choose whether to do anything at all (or, equivalently, whether to interfere with the status quo). Modeling choice in terms of realizing a particular value of *I* incorporates both of these aspects of choice.

[40] If I want to use IDT to evaluate whether you made the right choice by my lights—i.e. relative to *my* subjective credence function and value function—then I should use a variable set that *I* take to be causally sufficient. But when we use IDT to assess the *rationality* of an agent—i.e. whether she does well by her own lights—we should use a variable set that the agent takes to be causally sufficient.

[41] This variable set must include a variable for the agent's action but need not include an intervention variable.

[42] Meek and Glymour (1994) describe things in terms of conditioning on some value of an intervention variable. Pearl (2009) explains how the two approaches agree.

[43] The fact that smoking raises the probability of lung cancer while intervening to make oneself smoke does not can therefore be expressed as follows: $P(LC=lc|S=s) > P(LC=lc) = P(LC=lc|do(S=s))$.

[44] The reader may wonder how to determine what DAGs an agent *should* entertain in a given context. This is not my focus here. I am instead preoccupied with what an agent should do *given* her beliefs about DAGs, no matter whether her beliefs about DAGs are reasonable.

a weighted average of the utilities that she attaches to the worlds that result from intervening to make herself *a* when each of the live causal hypotheses (or ordered pairs) is realized, where the weights correspond to the subjective probabilities that she assigns to each of the live causal hypotheses.

Whenever dominance reasoning accords with causal-decision-theoretic reasoning, IDT sanctions dominance reasoning. Consider IDT's application to FNP. Since it is stipulated that smoking does not cause lung cancer, every graph that the agent entertains should be one according to which lung cancer is *not* causally downstream from smoking.[45] And since (given the CMC) the probability of getting lung cancer given that one intervenes to make oneself smoke must be equivalent to the unconditional probability of getting lung cancer in the event that *LC* is not causally downstream from *S* (even if the probability of getting given that one smokes is greater than the unconditional probability of getting lung cancer), every probability distribution referenced in the *K* partition must be one according to which $P(LC{=}lc) = P(LC{=}lc \mid do(S{=}s))$. Thus, no matter which live causal hypothesis is realized, the chance of getting lung cancer is not affected by whether one intervenes to make oneself smoke, and the agent can thus conclude that she should smoke no matter how she spreads her confidence across the elements of the *K* partition. That is, the agent should prefer smoking to not smoking no matter which element of the *K* partition obtains (because the chance of getting lung cancer must be independent of whether one smokes no matter which of the live probability distributions obtains), and can therefore ignore her subjective probability distribution over *K* as she determines that smoking is rational.

[45] Of course it is possible that the agent *will* entertain (or assign non-zero probability to) some DAG according to which lung cancer *is* causally downstream of smoking. But doing so would seem to be irrational, given the agent's knowledge of the details of FNP. This is why I say that every graph the agent *should* entertain has this property.

5. Ironing out the Kinks

Though it is clear that IDT is helpful to Lewis insofar as it severs the backtracking dependencies between an agent's choice and the causal predecessors of the agent's act, there are still unresolved issues. For example, is it coherent to replace Lewis's maximally specific dependency hypotheses with causal hypotheses that are themselves comprised of probability distributions, given that this seems to involve probabilities of probabilities? And if it is coherent, why are causal hypotheses easier for agents to entertain than maximally specific dependency hypotheses?

IDT *prima facie* seems to require agents to adopt higher-order probability functions—i.e. probability functions that take other probability functions as arguments—and this may seem problematic because the agent who is uncertain about whether some causal hypothesis is true does not seem to be uncertain about what her subjective probability judgments are.[46] For example, it does not seem as if an agent who is 35% confident that $X \rightarrow Z \leftarrow Y$ is (at least) 35% confident that her probability function is one of the many according to which X and Y are probabilistically independent (conditional on the empty set).[47] This means that the probability functions constrained by the CMC cannot be candidate subjective probability functions, and must rather be *expert* functions of some kind—i.e. *not* functions that might (according to the agent) currently be her own, but rather functions that might (according to the agent) be rational to defer to in some circumstances.

What kind of expert function is constrained by the CMC? Here, we can take cues from Lewis. Remember that Lewis formulated his dependency hypotheses in terms of objective

---

[46] See Savage (1954) for a classic argument against such probabilities.

[47] Even given some DAG, an agent can entertain distinct causal hypotheses that are formulated in terms of distinct probability functions. The agent can be n% confident that P_1 is the probability function and 1-n% confident that P_2 is the probability function, even when she is certain that $X \rightarrow Z \leftarrow Y$ (because multiple probability functions are compatible with a given DAG according to the d-separation criterion).

chance distributions that may be true of a world, and *not* in terms of subjective probability functions. We can do the same by suggesting that DAGs impose constraints on the set of objective chance distributions that may be true of a world. This both allays the foregoing concern (since a chance function is a type of expert function), and squares well with the common belief that causal relationships are in the world, rather than between credal states in our heads. So when an agent is 35% confident that $X \rightarrow Z \leftarrow Y$, she is at least 35% confident that the true objective chance distribution is such that the chance of X is independent of the chance of Y—i.e. at least 35% confident that the world is such that she should defer to an expert function according to which X is independent of Y.

Where do these chances come from? In the causal modeling framework, the chances can be systematically recovered for a given **V** through specification of, first, the structural equations that specify the way in which each variable depends on its parents (if it has any) and an error term (which is often interpreted as representing the influence of any justifiably omitted causes), and, second, the probability (or in this case, the chance) distribution over all of the exogenous variables not including the variable that encodes whether we intervene to make ourselves act (where exogenous variables are those that do not have parents). By specifying the structural equations, we learn how chance values depend on their causes. By specifying the chance distributions over each exogenous variable other than the intervention variable, we learn the objective chances of those things that are beyond the agent's control. When you put this information together, it turns out that we learn the objective chances of what lies beyond our control, as well as the way in which the objective chances of what lies causally downstream from our action depends on our choice.

But trading in subjective probabilities for objective chances does not get us completely out of the woods. One might reasonably worry that any agent who entertains (or forms

probability judgments about) chance distributions that contain information about the chance that she will act in a particular way or the chance that she will intervene to make herself act in a particular way will know so much about what she will do that she will have no decision to make.[48] For example, if the agent who confronts FNP knows that there is a 45% objective chance that she will smoke (or a 45% chance that she will intervene to make herself smoke), it may seem as if she has no reason to deliberate, since she already knows that the world is such that there is a 45% chance that she will smoke (or will decide to smoke).

On the one hand, if **V** contains a variable for acting, but not for intervening (or deciding), then there is no trouble of this sort. This is because the chance-value of one's act that is included within a chance distribution over such a **V** corresponds to the chance of acting when one does *not* intervene—i.e. when one does not make a decision—and this does not seem to pose trouble in the context of deliberation. Consider an agent with a smoking habit. It seems entirely reasonable for her to entertain the chance that she will smoke at the bar tonight in the event that she does not make a decision—i.e. in the event that she doesn't settle for herself whether to smoke tonight. It just may not make sense for her to entertain the chance that she will smoke *as a result of deliberating* and this is not included in the chance distribution over **V**.

But if **V** *does* contain a variable representing whether the agent intervenes, then waters are a bit muddied. When an agent entertains a single chance distribution over such a **V**, the agent *does* entertain the chance that she will decide in a particular way, and this may seem problematic. There are two ways to get around this problem. The first is to bar agents from ever including the intervention variable in **V**. Though this would secure the desired results, it may seem unprincipled since it is hard to see why the agent should not regard her decision as part of the causal system that she entertains. The second is to allow for some indeterminacy by

[48] This concern is similar to Isaac Levi's famous (1989) concern that "prediction crowds out deliberation."

allowing causal hypotheses to designate *sets* of chance distributions, where each particular chance distribution (that is a member of the set constituting some causal hypothesis) must obey the independence constraints entailed by the CMC. If causal hypotheses are defined in this way, then the agent can form subjective probability judgments over disjunctions of chance distributions (instead of particular chance distributions), and thereby entertain a causal hypothesis without entertaining a particular chance-value that she will decide in a particular way.[49]

6. Improving on Lewis

Of course we still must ask how these interventionist bells and whistles help improve Lewis's decision theory.

First, the causal modeling framework allows the agent to limit her consideration to a set of variables **V** that does not include everything that is within her control and beyond her control (so long as every common cause of any two or more variables in **V** is in **V**).[50] This means that the agent can limit her consideration to those variables that she considers salient to her deliberation, which appears to get the interventionist something that Lewis hoped to get

[49] Since every member of the set—i.e. every particular distribution—must obey the constraints implied by the CMC, the set of distributions will often not be convex.

[50] This requirement may strike some as too stringent because human agents are often in no position to be sure that they attend to every common cause of any two or more variables in **V**. I am sympathetic to this objection because it is true that we often ignore common causes of variables that we attend to as we deliberate, but I nevertheless believe that the ideally rational agent would be sure to (i) include every variable in **V** that she believes might be a common cause of other variables in **V**, and (ii) spread her credences across the hypotheses according to which said variables are and are not common causes of other variables in **V**. If this proves to be too cognitively demanding, then it may be okay for the agent to sometimes ignore common causes, but this would seem to require a departure from ideal rationality. It is likewise worth noting that it seems possible to weaken the requirement that one must attend to *every* common cause of any two or more variables in **V** by requiring the agent to attend only to some important subset of the common causes any two or more variables in **V** (because, for example, the agent *can* leave out distal common causes of X and Y in the event that she has included some more proximate common cause of X and Y that screens off the distal cause). But since I am not currently sure *exactly* how causal sufficiency can be plausibly weakened, I leave this task for later.

with counterfactuals. Moreover, since the truth of a DAG over a given causally sufficient **V** is consistent with the truth of DAGs over supersets of **V** (in much the same way that Lewis's ordinary counterfactuals are consistent with more specific counterfactuals), causal hypotheses over sparse variable sets can be seen as designating disjunctions of causal hypotheses over richer variable sets.[51] The agent can thus limit her consideration to "small worlds" while making choices that are rational from the "grand world" perspective.[52]

Second, and this should be obvious, the causal modeling approach responds to Eells's and my concerns about delivering the right results for the wrong reasons. When the Lewisian causal decision theorist formulates dependency hypotheses in interventionist terms, she advocates that agents reason straight from the causal details of a given scenario to what is rational—i.e. without any recourse to some potentially obfuscating mediator like counterfactuals.[53] Moreover, because causal models provide the tools to precisely represent

---

[51] Suppose, for example, that Figure 1 is true. There is nothing in the causal modeling framework (i.e. no assumption about the DAG) that rules out the possibility that there is some intermediary cause between, say, *G* and *LC* on which one could condition in order to screen off the correlation between *G* and *LC*. (Put differently, it is consistent with Figure 1 that *G* influences *LC* only by influencing, say, one's dietary preferences, and that *G* is therefore probabilistically independent of *LC* conditional on any particular set of dietary preferences.) Nor is there anything in the causal modeling framework that entails that the variables in Figure 1 are not additionally caused by means that are causally independent of the variables contained therein (since the omission of causes that are not common causes of variables in **V** does not induce any spurious correlations, given the axioms of the causal modeling framework). It is thus plausible to regard a causal hypothesis over the variables in Figure 1 as a disjunction of more specific causal hypotheses over richer variable sets.

[52] One might worry that causal hypotheses are not guaranteed to be causally independent of the agent's options (or acts) when causal hypotheses need not be maximally specific (because non-maximally specific causal hypotheses need not include every causal relationship and therefore need not include every causal relationship between the agent's choice and any causal hypothesis), and that non-maximally specific causal hypotheses should therefore not be utilized in a Savage-style decision theory. There are two responses to this concern. First, since causal hypotheses include information about the effect of intervening to make oneself act in a particular way, and since decisions are modeled as interventions on actions, causal hypotheses *prima facie* seem to include the agent's causal influence (and therefore the agent's influence on which causal hypothesis is true, if there is one). Second, if the first response fails, one can simply require that causal hypotheses include every causal relationship between the agent's choice and any causal hypothesis without succumbing to the Charybdis of maximal specificity.

[53] Of course causal hypotheses are closely related to counterfactuals in some way (since there is a definite connection between causal relationships and counterfactuals), but they nevertheless allow us to

causal relationships, the interventionist causal decision theorist capably represents the causal features of decision-making contexts that appear to escape other frameworks. To this end, I agree with Christopher Hitchcock's (2015) defense of what he calls *causal decision metatheory*: the thesis that when engaging with a decision problem, we should use causal models to make explicit our assumptions about the causal structure of the problem, as well as the question that we are asking.[54]

7. Savage or Bust

I mentioned earlier that one can avoid difficulties that arise from act/state dependence by *either* (i) providing the agent with a guide for constructing partitions of states that do not yield undesirable results or (ii) providing a conception of expected utility that is distinct from Savage's in the sense that it generalizes to cases where acts and states are dependent. At this point, the reader may be wondering why I have chosen to follow Lewis in taking the former route rather than the latter. After all, if it is possible to build a decision theory that generalizes to any old partition of states, it may seem as if *that* decision theory would be more desirable than one that comes with instructions for building partitions of states. Might it be possible to formulate an interventionist decision theory that takes the latter form rather than the former?

When decision theorists take the second route, they usually deploy conditional probabilities in their analysis of expected utility, and not unconditional probabilities (as Savage, Lewis, and I do). By limiting one's consideration to worlds in which the agent takes the act in question, it may seem as if any trouble associated with act/state dependence dissipates. For example, were Iverson to have limited his consideration to worlds in which he practiced while

---

reason directly from causal facts—i.e. without taking up the very difficult project of specifying the exact relationship between causal facts and counterfactual facts.

[54] This idea is also found in Meek and Glymour (1994). Indeed, Hitchcock credits them with it.

forming probability judgments about the states, he seemingly would have accounted for the effect that practicing has on his performance. Newcomb problems teach us that old-fashioned conditional probabilities will not do the trick (since conditioning on smoking in FNP induces a spurious effect on the probability of lung cancer), but one might hope that we can use the interventionist framework to ground a causal notion of conditional probability that is suitable for dealing with Newcomb problems.

Judea Pearl (2009) attempts to do exactly this by recommending that we analyze the expected utility of *a*-ing by deploying $P(y|\,do(A=a))$, i.e. the probability that outcome *y* obtains given that we intervene to make ourselves *a*:[55]

$$U(a) =^{df} \sum_y C(Y = y|do(A = a))V(Y = y).$$

It is not hard to see why one would opt for such an analysis. Since intervening severs backtracking associations between acts and states, it may seem as if conditioning on $do(A=a)$ breaks exactly those act/state dependencies that have no place in causal decision theory—i.e. those that are beyond the agent's control—while leaving intact exactly the act/state dependencies that do deserve a place in causal decision theory—i.e. the dependencies that the agent can exploit to control things. So it may seem as though Pearl's suggested generalization of expected utility is a welcome addition to causal decision theory.

Though I prefer Pearl's decision theory to many other causal decision theories on the market (largely because I endorse Hitchcock's [2015] causal decision metatheory), I have two objections to his move from Savage-style interventionist decision theory to non-Savage-style interventionist decision theory.

The first is the simple methodological fact that Pearl's generalization does not actually make things easier on agents since $P(y|\,do(A=a))$ is well-defined only relative to a causal model

[55] One can think of Pearl's outcomes as the possible act/state combinations that arise from taking the act in question.

that includes A. Since agents have to construct models (that occupy the role of states in my decision theory) in order to apply Pearl's decision theory, Pearl's decision theory has no pragmatic advantage over IDT.[56]

Second, and more importantly, Pearl's decision theory does not always get the right results in cases where agents are sure about what causes what, but uncertain about what particular causal hypothesis is true. In order to see why this is true, we must reflect on Spirtes, Glymour, and Scheines's (2000) observation that mixtures of probability distributions that are compatible with a given DAG are often not themselves compatible with that DAG. The idea here is that two (or more) probability distributions can respect the independence constraints imposed by the d-separation criterion, while the probability distribution that results from their mixture often will *not* obey these constraints. Indeed, they prove that if (i) X and Y are independent conditional on Z according to two distributions and (ii) the weights in the mixture are less than 1, then X and Y are independent conditional on Z in the resulting mixture if and only if $P_2(X \mid Z)P_2(Y \mid Z) + P_1(X \mid Z)P_1(Y \mid Z) = P_1(X \mid Z)P_2(Y \mid Z) + P_2(X \mid Z)P_1(Y \mid Z)$.

This means that if an agent is certain of some DAG but entertains two or more causal hypotheses consistent with that DAG (which differ with respect to the chance distributions), and adopts the subjective probability distribution that results from mixing in correspondence with her credences in the causal hypotheses (as the Principal Principle would seem to mandate),[57] then there is no guarantee that there will not be dependencies in her subjective probability distribution that do not derive from causal dependencies. This means that when she

---

[56] For similar reasons, it is not immediately clear how to apply Pearl's decision theory when the agent is unsure about what causes what. Perhaps the agent can form her credence that y obtains given that she intervenes to x by, first, determining the chance of y given that she intervenes in every model that she entertains, and second, calculating a weighted average of these conditional objective probability estimates in correspondence with her subjective probabilities in the live causal hypotheses. But as we will see below, it is by no means obvious that this will yield the desired results.

[57] See Lewis (1980) for an explanation of how his Principal Principle imposes constraints on rational probability functions in such cases.

conditions on $do(X=x)$, there is no guarantee that she will regard only those things over which she has control as probabilistically dependent on her choice.

Of course the non-Savage style decision theorist can get around this difficulty by limiting the domain of applicability of her decision theory to cases when an agent is certain of some causal hypothesis. But this renders her decision theory considerably less useful than IDT. We are seldom, if ever, in a position to know not only what is causally relevant to what, but also the precise nature of every causal dependency at play. So it is often healthy to suspend judgment with respect to what causal hypothesis is true of some **V**. Savage-style interventionist decision theory (of the sort presented here) deals with such healthy decision-making contexts adequately; non-Savage style decision theory does not.

At this point, you may be wondering whether I have pulled the wool over your eyes. This feature of mixtures never came up in my development of IDT; perhaps it would have reared its ugly head had I given it its due there. Luckily, though, this feature of mixtures did not come up because it does not come up for the Savage-style theorist. If agents follow my advice, they should determine the expected utility of x-ing by, first, considering the value of x-ing in each of the worlds that would result were each of the live causal hypotheses to be realized, and, second, calculating a weighted average of the resulting values, where the weights accord to the subjective probabilities assigned to the individual causal hypotheses. Unlike Pearl, I never ask the agent to do anything that depends on her subjective probability that some state will occur given that she intervenes to make herself take the act in question. Because of this, it does not matter for IDT whether the agent's subjective probability function is such that she regards some events that are in fact out of her control as probabilistically dependent on whether she intervenes to take the act in question.

And it had better not matter. For there is nothing wrong with adopting a subjective

probability distribution according to which some variable that is not causally downstream from an intervention is probabilistically dependent on the intervention. Of course it is incompatible with the CMC that *some* probability distribution has this property. But since agents can be uncertain about what distribution is true among the distributions compatible with a given DAG, it seems that the CMC implies constraints on the *objective* probability distributions that enter into causal hypotheses, and *not* the *subjective* probability distributions that rational agents should adopt when entertaining multiple objective distributions. Indeed, it is often rationally impermissible for an agent to adopt a subjective probability function that treats two causally independent variables as probabilistically independent.

Suppose, for example, that you are certain of a DAG according to which X and Y are d-separated conditional on the empty set. Moreover, suppose (i) that you have eliminated all but the following two chance distributions, and (ii) that you are 50% confident in each.[58]

	x & y	**x & ~y**	**~x & y**	**~x & ~y**
Ch_1	.08	.32	.12	.48
Ch_2	.48	.12	.32	.08
Cr	.28	.22	.22	.28

By the Principal Principle, your subjective credence function should output the probability judgments listed on the third row of the table—i.e. the probability judgment that results from averaging the two chance distributions. But interestingly, these probability judgments do not satisfy the independence constraints that are satisfied by the original chance distributions. Since

[58] These numbers are taken from Seidenfeld, Schervish, and Kadane (2010). But they use this feature of mixtures to illustrate a different point—namely, that imprecise Bayesians should not impose a requirement of convexity on sets of probabilities because doing so leads to irrational choices.

the probability of x & y is equivalent to the product of the probability of x and the probability of y according to Ch_1 and Ch_2 but not according to Cr, it seems that the Principal Principle sometimes requires agents to adopt credence functions that do not satisfy the CMC.

$$\begin{array}{lll} \mathrm{Ch}_1(x) = .4 & \mathrm{Ch}_1(y) = .2 & \mathrm{Ch}_1(x \,\&\, y) = .08 \\ \mathrm{Ch}_2(x) = .6 & \mathrm{Ch}_2(y) = .8 & \mathrm{Ch}_2(x \,\&\, y) = .48 \\ \mathrm{Cr}(x) = .5 & \mathrm{Cr}(y) = .5 & \mathrm{Cr}(x \,\&\, y) = .28 \end{array}$$

It might be tempting to think that the Principal Principle leads us astray in this context, since it may seem obvious that one's credence in x should not shift as a result of learning y when x and y are causally independent. But this is not right. Though it is true that y is evidence for x according to Cr—since $.56 = \mathrm{Cr}(x|y) > \mathrm{Cr}(x) = .5$—this reflects good evidential reasoning. Why? Because the truth of y is evidence that Ch_2 is the true chance distribution (since Ch_2 outputs a higher chance-value for y than Ch_1), and this in turn is evidence that the chance of x is high (since Ch_2 likewise outputs a higher chance-value for x than Ch_1). In order to be a good evidential reasoner, then, one must sometimes adopt a subjective credence function that does not satisfy the CMC.

This appears to spell doom for Pearl's analysis of expected utility. Since Υ could very well represent the agent's choice—i.e. whether the agent intervenes to make herself act in the relevant way—conditioning on $do(x)$ does not always yield the independencies (in the agent's subjective probability distribution) that are so important for delivering the result that that any correlation between what the agent does and cannot cause should be irrelevant to the agent's

choice.[59] So, Pearl's analysis, unlike my Lewisian Savage-style analysis, does not appear to make good on causal-decision-theoretic intuitions.

Pearl could object that the agent is not a genuine decision-making context when, according to her subjective credence function, the probability of x is not independent of y.[60] But it is hard to see what licenses this view of decision. In the event that an agent is certain about what causes what (including that her decision constitutes an intervention), but uncertain about the precise ways in which the variables at play cause each other, it seems as though the agent has sufficient reason to believe that she has a decision to make. After all, she is sure that her choice is not determined by any of the factors in the model, and therefore that her choice is "up to her" in a sense. Yet according to this line of objection, such an agent has no decision to make.

In sum, when an agent is sure that her decision constitutes an intervention and that some factor is causally prior to or causally independent of her action, causal-decision-theoretic reasoning suggests, first, that the agent has a genuine decision to make, and, second, that any correlation between the relevant factor and her action should be irrelevant to her choice. Savage-style interventionist decision theory of the sort advocated here captures this conviction; purported generalizations like Pearl's do not.

8. Conclusion

At this stage, my defense of IDT is complete. I have argued that the causal decision theorist would do well to adopt a Lewisian analysis of expected utility according to which interventionist causal hypotheses play the role of Lewis's dependency hypotheses because the

---

[59] In order to talk this way, we must impose a probability distribution over the agent's choice, which, as discussed earlier, may be problematic in the context of deliberation. Even if this is so, one can simply describe the case third-personally, such that some bystander entertains the relevant chance distributions and adopts the relevant credence function. By the causal decision theorist's lights, such a bystander should judge x as irrelevant to the agent's choice, despite what Pearl's decision theory would recommend.

[60] Jim Joyce suggested this possibility in personal communication.

interventionist approach to causal modeling makes it possible, first, to represent the causal details relevant to a particular decision-making context in a rigorous mathematical language, and, second, to deliver "small world" causal decision-theoretic verdicts without overcomplicating things by invoking counterfactual semantics. I have also argued that the interventionist should prefer Lewis's analysis of expected utility (which, like Savage's, is in terms of unconditional probabilities of states) to analyses that use conditional probabilities to purportedly generalize Lewis's analysis because the move to conditional probabilities compromises the interventionist's ability to make good on the causal decision theorist's conviction that that any correlation between what the agent does and cannot cause should be irrelevant to the agent's choice.

What I have *not* argued is that the causal decision theorist is *right* that any correlation between what the agent does and cannot cause should be irrelevant to the agent's choice. Now that IDT is on the table, it is worth reflecting on exactly how IDT manages to deliver causal-decision-theoretic verdicts in order to consider what the truth of causal decision theory rides on from the interventionist perspective.

It is often thought that if the causal details of FNP are relevant to what you should do, then you should smoke. But according to interventionist reasoning (and IDT), taking stock of the causal details of FNP is *not* sufficient for delivering the verdict that agents should smoke. That is, in order for IDT to license smoking, IDT must not only take stock of the causal details by requiring the K partition to range over causal hypotheses (i.e. ordered pairs of DAGs and chance distributions), but must also model the decision to smoke as an intervention—i.e. apply the value function to the K partition and the intervention to make oneself smoke (rather than

smoking itself).[61] This underscores an important point. According to interventionist reasoning, causal-decision-theoretic verdicts are not guaranteed simply paying attention to the causal details of a given decision-making context, but rather likewise depend on a substantial thesis about the nature of choice—namely, that an agent must represent herself as intervening to make herself act whenever she makes a genuine choice.

Whether it is plausible that deliberating agents must regard themselves as intervening is outside the purview of this paper. But it nevertheless seems valuable to develop an interventionist decision theory that delivers causal decision-theoretic results for those decision-making contexts in which agents *do* represent themselves as intervening. My hope is that I have successfully developed such a decision theory in this paper.

[61] Other causal decision theories (e.g. Lewis 1980) allegedly deliver causal decision theoretic verdicts simply by partitioning the states in a particular way. As an anonymous reviewer helpfully points out, some may think that such theories are more parsimonious than IDT (since such decision theories require only one step in order to get causal-decision-theoretic verdicts, while IDT requires two). But as I see things, IDT earns its keep for reasons not pertaining to its parsimony—e.g. its integration of the independently successful interventionist approach to causal modeling and its ability to deliver "small world" causal-decision-theoretic verdicts. I also agree with Meek and Glymour (1994) that the role of interventions in IDT helps us to see that the nature of disagreement between evidential decision theorists and causal decision theorists is *not* a disagreement about the fundamental normative principles that govern rational choice, but is rather a disagreement about the nature of choice. They write (p. 1009) that we can recharacterize the dispute such that "⌜t⌝he difference between the two ⌜i.e. causal decision theory and evidential decision theory⌝ does not turn on any difference in normative principles, but on a substantive difference about the causal processes at work in the context of decision making—the causal decision theorist thinks that when someone decides… an *intervention* occurs, and the 'evidential' decision theorist thinks otherwise."

Chapter 3: Decision and Intervention

ABSTRACT

Meek and Glymour (1994) offer a new lens through which we can view the rift between causal decision theorists and evidential decision theorists. They argue (p. 1009) that we can recharacterize the dispute such that "[t]he difference between the two does not turn on any difference in normative principles, but on a substantive difference about the causal processes at work in the context of decision making—the causal decision theorist thinks that when someone decides… an *intervention* occurs, and the 'evidential' decision theorist thinks otherwise." In this paper, I adopt Meek and Glymour's framework and argue that neither causal decision theory nor evidential decision theory issues rationally defensible verdicts across all decision-making contexts. Specifically, I argue that in addition to contexts in which agents should regard themselves as intervening, there are contexts in which agents should regard themselves as not intervening, as well as contexts in which agents should be uncertain about whether they are intervening. If this is right, then a sufficiently general decision theory will issue rational verdicts not only when an agent is certain that she is or is not intervening, but also when an agent has some non-extreme degree of belief that she is intervening. My last task in this paper is to modify Stern's (2016) "interventionist decision theory" so that it delivers rational verdicts in each of these contexts.

1. Introduction

You stand before two boxes. One is transparent and contains $1,000. The other is opaque. You have a choice. You can "one-box"—i.e. take the contents of the opaque box—or "two-box"—i.e. take the contents of both boxes. The contents of the opaque box depend entirely on the earlier prediction of a remarkably successful predictor. If the predictor predicts that you will one-box, he places $1,000,000 inside the opaque box. If he predicts that you will two-box, the opaque box is empty. Should you one-box or two-box?

This is Newcomb's Problem (NP). According to the standard treatment of NP, evidential decision theory recommends that you one-box because you are expected to make more money if you one-box than if you two-box, while causal decision theory recommends that you two-box because your choice exerts no causal influence over whether the predictor puts $1,000,000 in the opaque box (since the present does not cause the past), thus ensuring that two-boxing makes you $1,000 richer than you would have been had you one-boxed.[62]

Traditionally, each camp has adopted its own version of decision theory in order to deliver its favored recommendation. Evidential decision theorists argue that agents should choose whatever action is evidence for the best outcome (i.e. whatever action maximizes conditional expected utility),[63] while causal decision theorists usually argue that agents should deploy the unconditional probabilities of non-backtracking counterfactuals as they calculate expected utility.[64] As such, the dispute between evidential decision theorists and causal decision theorists is often thought to turn on which of two irreconcilable decision norms is most intuitive. Either we should maximize the sort of expected utility that is calculated with

---

[62] There are exceptions. Spohn (2012) is a causal decision theorist who advocates one-boxing. Eells (1982) is an evidential decision theorist who recommends two-boxing.

[63] The conditional expected utility of a given action consists of a weighted average of utility across its possible outcomes, where the weights are provided by the agent's credences in each possible outcome conditional on her performing the action in question.

[64] Lewis (1981a) and Gibbard and Harper (1978) are some examples of causal decision theorists who take this route.

conditional probabilities, or we should maximize the sort of expected utility that is calculated with the probabilities of non-backtracking counterfactuals. If we go the former route, we allow backtracking probabilistic dependencies to affect expected utility, and thereby secure the evidential decision theorist's intuition that subjects should one-box. If we go the latter route, we disallow backtracking probabilistic dependencies from affecting expected utility, and thereby secure the causal decision theorist's conviction that subjects should two-box.[65]

In their 1994 paper, "Conditioning and Intervening," Christopher Meek and Clark Glymour use the interventionist approach to causal modeling to articulate a novel way of understanding the rift between causal decision theorists and evidential decision theorists. Specifically, they claim (1994: p. 1009) that we can recharacterize the dispute such that "[t]he difference between the two does not turn on any difference in normative principles, but on a substantive difference about the causal processes at work in the context of decision making—the causal decision theorist thinks that when someone decides… an *intervention* occurs, and the 'evidential' decision theorist thinks otherwise." The basic idea, as we will see later, is that if an agent models herself as intervening to make herself two-box (or one-box), then the expected utility calculation of two-boxing (or one-boxing) will agree with causal decision theory even if

---

[65] Though much ink has been spilled about which decision rule is right, there is no consensus in the literature. Causal decision theory is usually defended on the grounds that it is most intuitive to limit one's focus to factors over which one exerts causal influence, while evidential decision theory is usually defended on the grounds that it recommends whatever strategy is optimal in the long run. David Lewis, a worried causal decision theorist, describes the evidential decision theorist's line of argument as follows:

> The one-boxers sometimes taunt us: if you're so smart, why ain'cha rich? They have their millions and we have our thousands, and they think this goes to show the error of our ways. They think we are not rich because we have irrationally chosen not to have our millions. (1981b: p. 377)

The idea here is that the predictor's remarkable success means that those who two-box will lose out on a lot of money in the long run. The standard response on behalf of causal decision theory has been to grant the evidential decision theorist's point, but to claim that NP is an unfortunate scenario in which subjects are rewarded for their irrationality. But as Lewis (1981b: p. 388) notes, this has always seemed a bit unsatisfactory, for it seems to be "one more piece of two-boxist doctrine that one-boxers may consistently deny."

she uses the evidential decision theorist's rule because the causal character of an intervention entails (given the axioms of the interventionist approach to causal modeling) that intervening to make oneself one-box or two-box is not correlated with (and therefore not evidence for) the predictor's prediction. Thus, from the perspective of Meek and Glymour, the dispute between causal and evidential decision theorists turns on how agents should understand the causal processes that underlie choice rather than which of two irreconcilable norms of rational choice is most intuitive.[66]

Working within Meek and Glymour's framework, Judea Pearl (2009: pp. 108-109) argues from "common sense" that agents should represent themselves as intervening in every genuine decision-making context (where contexts are regarded as causal systems on which agents might intervene). This leads him to propose that agents should opt for whatever action x maximizes expected utility when defined as follows, where $P(y \mid do(X = x))$ corresponds to the

[66] If it is not yet clear how Meek and Glymour's position maps on to the old debate between causal decision theorists and evidential decision theorists, the following passage (1994: pp. 1014-15) might be helpful:

> "Our analysis of the dispute between causal and 'evidential' decision theory does not put us neatly on either side, exactly because we think the dispute has been misdescribed, from Nozick on. From our perspective, whether causal or 'evidential' recommendations should be followed depends only on what one believes about the causal character of a context of decision. We agree with 'evidential' decision theorists that nothing but an ordinary calculation of the maximum expected utility is required; we agree with causal decision theorists that sometimes the relevant probabilities in the calculation are not the obvious conditional probabilities… Where they recommend different decisions… is because they differ about whether an action is an intervention; whether the manipulated and unmanipulated distributions are different. If so, then a different event must be conditioned on than if not, and a different calculation results."

I hesitate to join Meek and Glymour in characterizing the dispute in terms of what event "must be conditioned upon" because I agree with Stern (2016) that the interventionist approach to causal modeling pairs better with decision theories that utilize unconditional probabilities rather than conditional probabilities in expected utility calculations. But the general point stands. That is, even when unconditional probabilities are deployed, the dispute between evidential and causal decision theorists can be recharacterized as a dispute about the nature of choice rather than about the means by which expected utility should be calculated.

probability that outcome y will obtain under the supposition that x is brought about through intervention.[67]

PEARL'S DECISION THEORY

Choose the act that maximizes expected utility when calculated as follows:

$$U(x) =^{df} \sum_{y} P(Y = y|do(X = x))V(Y = y)$$

In this paper, I argue that Pearl is wrong. More specifically, I argue that in addition to contexts in which agents should regard themselves as intervening, there are contexts in which agents should regard themselves as not intervening, as well as contexts in which agents should be uncertain about whether they are intervening. My strategy is to argue (i) that the details of NP (Newcomb's Problem) can be specified such that the subject should regard it as impossible for her to intervene, (ii) that the details of NP can be specified such that the subject should be uncertain whether she is in a position to intervene, and (iii) that each of these specifications of NP constitutes a genuine decision problem.[68]

If I am right, then it is important that our decision theory issues rational verdicts not only when an agent is certain that she is intervening, but also when an agent is certain that she is not intervening, and when an agent has some non-extreme degree of belief that she is intervening. My last task in this paper is to modify Stern's (2016) "Interventionist Decision Theory" so that it delivers rationally defensible verdicts in each of these contexts.

---

[67] Indeed, Pearl (2009: p. 108) refers to his proposal as "common-sensical decision theory" rather than causal decision theory, and writes that he "purposely avoid[s] the common title 'causal decision theory' in order to suppress even the slightest hint that any alternative, noncausal theory can be used to guide decisions."

[68] It is worth noting at the outset that my arguments apply not only to Pearl's decision theory, but also to versions of causal decision theory where it is specified that we deploy the probabilities of non-backtracking counterfactuals as we calculate expected utility. This is because I argue that there are decision-making contexts in which agents should attend to backtracking dependencies as they deliberate.

2. Modeling Newcomb's Problem

In order to understand the causal processes that underlie the NP subject's decision-making context, we should build a causal model of the sort utilized by practitioners of the interventionist approach to causation. This means searching for a directed acyclic graph (DAG) of NP that captures the features of NP that the subject should believe to be the case. The outcome in NP—i.e. the amount of money received by the subject—causally depends on both the predictor's prediction and the subject's action. These are correlated. (If they were not correlated, then the predictor could not be said to be remarkably successful.) How do these facts help us identify a plausible DAG?

The Causal Markov Condition (CMC), which is assumed in every interventionist causal model, entails something very close to the following slight variant of Reichenbach's principle of the common cause.

PCC: If variables *F* and *G* are correlated, then either *F* caused *G*, *G* caused *F*, or *F* and *G* are joint effects of a common cause.[69, 70]

It does not seem that the predictor's prediction causally depends on the subject's action because the predictor's prediction temporally precedes the subject's action. It also does not

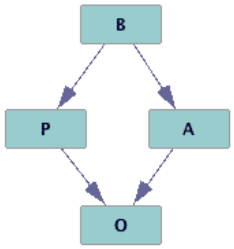
---

[69] For those interested in the nitty-gritty, Christopher Meek and Clark Glymour (1994) explain that the Causal Markov Condition is satisfied by a given causal model if and only if the following conditions obtain. Let *G* describe the causal relations among a set V of variables, and let *P* be the probability distribution over **V**. Then every variable X in **V** is probabilistically independent of its nondescendants in *G* conditional on its parents, where a variable's parents are its most immediate causal predecessors, and a variable's nondescendants are the variables in **V** that are not causally downstream from the relevant variable.
Why does this entail PCC? In the event that two variables are dependent and neither is a descendant of the other, there must exist some parent(s) of both variables on which one can condition to render the relevant variables independent. So it is provable of every system of variables that satisfies CMC, first, that if neither *F* nor *G* (directly or indirectly) influences one another and *F* and *G* are probabilistically dependent, then there exists a set *C* of variables not containing *F* or *G* but (directly or indirectly) causing both, and, second, that *F* and *G* must be independent conditional on $C = c$.

[70] One difference between Reichenbach's (1956) canonical formulation of the PCC and my own is that Reichenbach describes *F* and *G* as events. I describe *F* and *G* as variables (and not events) because the CMC is in terms of variables, not events. Of course we can represent whether some event obtains with a variable, but nothing in the CMC entails that variables *must* represent events.

seem that the subject's action causally depends on the predictor's prediction; NP depicts the predictor as a remarkably successful predictor of the subject's action, but not as a *controller* of the subject's action. So the PCC tells us we should search for some common cause. It is plausible that there is some fact that both informs the predictor's prediction and causes the subject's action. For example, the subject's behavior or mental states prior to acting may influence both the predictor's prediction and the subject's action. Allow us to represent the subject's earlier behavior with *B*, the subject's action with *A*, the predictor's prediction with *P*, and the monetary outcome with *O*.[71] The following DAG seems to give us some handle on NP.

Figure 1:



In treating Figure 1 as a DAG, I am assuming that each of its variables is a function of its parents (its most immediate causal predecessors) and its error term (if it has one).[72] Though it is often assumed that each variable has its own error term, it is acceptable to omit the representation of error terms from graphs. This is because error terms are (by stipulation) causes only of a single variable, not common causes. So their omission does not induce any spurious correlations.[73] Below is a DAG in which the presence of error terms for each variable is made explicit.

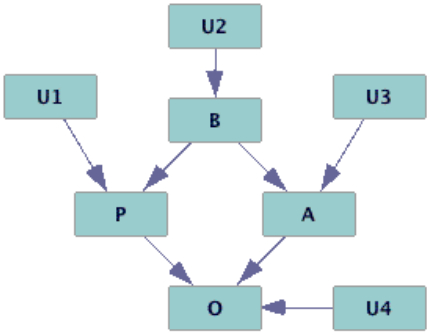
---

[71] There may be other suitable candidates for the common cause of *P* and *A*, but nothing I say in what follows depends on the particular common cause that we choose.

[72] Meek and Glymour (1994) provide a very nice exposition of the causal modeling framework, but see Pearl (2009) for a complete description of directed acyclic graphs and their use and Elwert (2013) for a good introduction.

[73] A cause *C* can be justifiably omitted from the directed acyclic graph if and only if *C* is not a common cause of any variables that appear in the directed acyclic graph.

Figure 2:



What does Figure 2 tell us about NP? The CMC not only entails the PCC, but also entails that if every path between a pair of variables in a variable set **V** is *d-separated* (according to a given DAG), then the pair of variables must be probabilistically independent of each other.[74] A path between two variables, *F* and *G*, is *d-separated* if and only if:

iii. the path contains a *non-collider* that has been conditioned on, e.g. $F \rightarrow \underline{H} \rightarrow G$ or $F \leftarrow \underline{H} \rightarrow G$ (where *H* has been conditioned on), or

iv. the path contains a *collider* that has not been conditioned on (and whose descendants have not been conditioned on), e.g. $F \rightarrow H \leftarrow G$.[75,76]

This means, for example, that every omitted cause represented by the error term, U_3 must be probabilistically independent of every endogenous variable that is neither *A* nor causally downstream of *A* because U_3 is d-separated from every such variable conditional on the empty set.

---

[74] Geiger and Pearl (1989) and Verma (1987) prove that the d-separation criterion characterizes all and only the conditional independence relations that follow from the CMC.

[75] A variable is a *collider* along a directed path if and only if it is the direct effect of two variables along the path. This is why *H* is a collider along $F \rightarrow H \leftarrow G$ but not $F \rightarrow H \rightarrow G$.

[76] The parenthetical is important because conditioning on the descendant of a collider often induces a spurious correlation between the collider's ancestors. See Elwert and Winship (2014) for discussion of this phenomenon.

With these implications of the CMC in hand, we now stand in a position to compute the effect of *intervening* on a causal system to set a variable to a particular value. Following Spirtes, Glymour, and Scheines (1993) and Pearl (1993, 2009), allow the *intervention* on A to represent some justifiably omitted cause of A (i.e. some cause of A that is covered in U_3) that can be exploited to set A to a. Since the intervention on A must be a justifiably omitted cause of A, the intervention must be d-separated from A's non-descendants in **V** (including A's causal predecessors),[77] and therefore probabilistically independent of A's non-descendants in **V**. This means that the causal character of interventions entails (given the CMC) that intervening to make oneself one-box (or two-box) is not correlated with the predictor's prediction, and that one can therefore get causal-decision-theoretic results without revising one's decision rule simply by modeling the agent as intervening on the causal system depicted at hand. That is, if the NP subject models herself as intervening, then she should obviously two-box because the predictor's remarkable success has no bearing on whether she *intervenes* to make herself one-box or two-box (even though the predictor is remarkably successful at determining whether subjects generally one-box or two-box).[78] As Pearl (2009: p. 109) puts it, interventions "*change* the probabilities that acts normally obey."

But when is it reasonable for the NP subject to represent herself as intervening? Given the above construal of what it means to intervene, the subject must believe, first, that there are justifiably omitted causes of her action (i.e. justifiably omitted causes of A) since interventions are causes themselves, and, second, that the intention to act that results from her deliberation is

---

[77] The intervention variable is d-separated from X's causal predecessors because X is a collider on every path from the intervention variable to X's causal predecessors. And since the intervention variable cannot be causally downstream from any variable(s) in **V** and likewise cannot be a common cause of any variables in **V**, the fact that X is d-separated from any non-descendants that are not causal predecessors of X guarantees that the intervention variable is d-separated from all of X's non-descendants.

[78] Sometimes people describe NP such that predictor's success is defined over the token subject confronted with the decision, and not over the type of subject who confronts NP. As I see things, when NP is described this way, the possibility of intervening on A is stipulated away. My reasons for thinking this should become clear in what follows.

among these causes. Are there decision-making contexts in which it is unreasonable to believe either of these things? This is the focus of much of what follows.

3. The Case of the Perfect Predictor

It is clear, I think, that if agents should always regard themselves as intervening when making decisions, then NP subjects should two-box. This explains why causal decision theorists who think that all genuine decision-making contexts are contexts in which one must represent oneself as intervening (e.g. Pearl) also think that you should two-box. But nothing said so far yields any reason to believe that these causal decision theorists are right that agents should always regard themselves as intervening. I believe that these causal decision theorists are wrong, and I believe that we can begin to see why if we focus on a version of NP in which it is specified that the predictor's success is perfect in a very robust sense.

Imagine that we fill in the details of NP such that the predictor is perfect across possibilia—i.e. such that $\Pr(A=\text{2-box} \mid P=\text{2-box}) = 1$ and $\Pr(A=\text{1-box} \mid P=\text{1-box}) = 1$, where the probabilities are defined over possibilia and not merely actual frequencies.[79] In this circumstance, there can be no omitted causes of P or A that can be exploited for intervention. If there were such omitted causes, then the values of P and A could not be perfectly correlated since they would come apart in circumstances where P or A is set by one of the omitted causes.[80, 81] Consider the case in which the error term for A represents (among other causes) the

[79] I do not further specify the set of possibilia (e.g. whether it consists of metaphysically possible worlds, nomologically possible worlds, etc.) because I do not want to take a stance on what sphere of modality is relevant to rational choice. The reader should feel free to insert her favorite sphere of modality, provided that her favorite sphere outstrips actuality.

[80] It should be clear that observing a perfect association between two variables in some finite sample does not imply that the two variables are perfectly correlated when the probabilities are defined over possibilia. It is possible, for example, that a predictor accurately predicts what subjects will do in each of some 100 trials, but that the predictor's prediction is not perfectly correlated with subjects' actions because there are possible scenarios in which subjects' actions come apart from whatever the predictor predicts.

subject's intention that results from making her decision. Were the subject to intervene on *A* by forming the intention to *a*, then the value of *A* would no longer be determined by *B*, and there would correspondingly be scenarios in which the values of *P* and *A* come apart. Similarly, were there some way to set *P* that exhibited no effect on the subject's prior behavior, then *P* and *A* could come apart. Thus *P* and *A* could not be perfectly correlated across possibilia were there omitted causes of *P* or omitted causes of *A*. But in the case of the robustly perfect predictor, *P* and *A are* perfectly correlated by hypothesis, so there can be no omitted causes of *P* or *A* when the predictor is perfect.[82]

When the predictor is in this sense perfect, the only possible way for the subject to influence the actual outcome is to intervene on her prior behavior, *B*. By defining the predictor as perfect, we define away omitted causes of *A*, and, in turn, define away the possibility of directly intervening on *A*. Remember that interventions are causes themselves. This means that the only way for the subject to control *A* is via the one cause of *A* that *is* included in the model: her prior behavior. Thus the subject stands to win $1,000,000 if she behaves (i) such that the predictor is *guaranteed* to predict that she will one-box and (ii) such that her later self is *guaranteed* to one-box. If she behaves otherwise—i.e. if she behaves such that the predictor is guaranteed to predict two-boxing and such that her later self is guaranteed to two-box—then she stands to win a measly (by comparison) sum of $1,000. The subject will thus be $999,000 richer if she one-boxes, but she must in effect set herself to one-box prior to the predictor's prediction.

---

[81] Some may notice, first, that the perfect correlation between *P* and *A* only entails that the probability that *P* and *A* come apart is zero, and, second, that zero probability does not entail impossibility. Though true, this does not create cause for concern in anything that follows (since each of my claims would go through with the more modest claim that the probability is zero). I speak in terms of impossibility only for ease of exposition.

[82] It may be possible that there are omitted causes of the sort that are not exploitable for intervention, so long as the probability that the omitted causes take on values which allow the predictor's prediction and the subject's action to differ is zero. Since these causes are irrelevant to anything I address in what follows, I ignore their possibility for ease of exposition.

Some decision theorists will no doubt contend that the subject does not have any decision to make when the predictor's success is specified to be perfect because it will already be determined what the subject will do at the time of her decision. According to this line of thought, the subject can decide at some early time to behave such that she is guaranteed to win $1,000,000, but the subject *cannot* decide what to do after the predictor has made the prediction, because at that point, the subject's destiny has been preordained.

Though I agree that the subject's destiny is set once the predictor makes the prediction (if not before), I believe that the subject nevertheless has a decision to make after her destiny is set. Consider the subject's epistemic standpoint. She knows that she will one-box if and only if the predictor predicts that she will one-box, but she does not know what the predictor has predicted she will do. She likewise knows that if she chooses to one-box, the predictor will have predicted that she would one-box, and that if she chooses to two-box, the predictor will have predicted that she would two-box. Though it is already metaphysically determined what she will do, there are epistemic possibilities for her to rule out. When she decides to one-box, she decides to do what she knows to be the only action that is compatible with her making $1,000,000. Put differently, she decides to take the action that epistemically necessitates (but does not metaphysically necessitate) that she behaved in the way that would make the predictor predict that she would take that action.

One might contend that the agent has a genuine choice to make only when she sees herself as capable of *metaphysically* determining her action, but as Dummett (1986) argues, ruling out merely epistemic possibilities is not unique to decision-making contexts like NP. Just as it has already been metaphysically determined what I will do in the context of the perfect predictor, there is a sense in which it has already been metaphysically determined what I will do in any decision-making context. To see this, we need not accept any controversial thesis

about whether the world is deterministic. All we must accept is that there is a truth-value right now about whether I will x in the future. Though it is already either true or false that I will x in 5 minutes, I can still deliberate about whether to x. Likewise, though it is already determined that I will one-box or two-box in the context of the perfect predictor, I can still deliberate about whether to one-box or two-box.[83]

4. Irreducibly Stochastic Success

Everything said thus far might lead one to think that the subject should one-box if the predictor is perfect, but two-box if the predictor ever gets things wrong. Likewise, everything said thus far might lead one to think that the subject should deem herself incapable of intervention only when the predictor's success is specified to be perfect. But this is not right. What matters is not whether the predictor is a perfect predictor of the subject's action *per se*, but rather whether the correlation between P and A is so robust that it rules out the possibility of intervening on A by settling on a-ing.

Imagine that the predictor is accurate 95% of the time. Imagine further that the probabilistic aspect of the predictor's success is irreducible—i.e. that the predictor's error is not explicable in terms of omitted variables. (Perhaps it helps to assimilate the interpretation the predictor's success with the ontic interpretation of the wave function.) This means that *whenever* the predictor predicts one-boxing, there is a 95% chance that the agent one-boxes. Or put differently, conditional on the value of the predictor's prediction, it is impossible to affect the probability distribution over the subject's action; *every* individual action has a 95% chance of matching the predictor's prediction.[84]

---

[83] I will return to this issue in section 6.
[84] This requires positing single-case probabilities.

This bars us from regarding the predictor's error as due to omitted causes.[85] Were there additional causes of *P* or *A*, then it would be possible to exploit these causes in order to change the probabilities that the act obeys. For example, were it possible to intervene on *A* by settling on two-boxing (so as to break the dependence between *A* and *B*), then not every individual action would have a 95% chance of matching the predictor's prediction (since the chance of the individual action resulting from intervention must be set by means that are probabilistically independent of the predictor's prediction). Thus we are left in a similar situation to that in which the predictor is perfect. The subject stands to win the most money by deciding to act in a way that entails that the predictor will have predicted one-boxing with some high objective chance.

5. Towards a Decision Theory

Contrary to the beliefs of many causal decision theorists, it appears that there are contexts in which agents should represent themselves as incapable of intervention. Moreover, in the contexts of the perfect predictor and the irreducibly stochastic predictor, it appears that old-fashioned evidential decision theory gets things right. Consider the results of applying the following analysis of expected utility, which Pearl (2009: p. 108) takes to correspond to evidential decision theory.[86]

EVIDENTIAL DECISION THEORY

Choose the act that maximizes expected utility when calculated as follows:

---

[85] An interesting question asks how we should represent this type of case in a DAG. Since we are barred from representing the error in the way that we usually represent error terms—i.e. as justifiably omitted causes—we have to find some other way to represent the stochastic element. I join Dan Hausman (1998: ch. 9) in thinking that a compelling strategy may be to incorporate objective chance-values in the variables themselves. But this is not the place to settle this matter.

[86] Pearl's statement of evidential decision theory is a somewhat crude depiction of evidential decision theory since so many sophisticated variants have emerged over the years, but I nevertheless treat Pearl's statement of evidential decision theory as canonical in what follows.

$$U(x) =^{df} \sum_{y} P(Y = y|X = x)V(Y = y)$$

In the case of the perfect predictor, were we to consider what is likely when we condition on one-boxing or two-boxing, we'd see that one-boxing guarantees winning $1,000,000 and that two-boxing guarantees winning only $1,000. Likewise, in the case of the irreducibly stochastic predictor, evidential decision theory seems to get things right. If we condition on one-boxing, our calculation yields expected winnings of $950,000. Whereas when we condition on two-boxing, our calculation yields expected winnings of $51,000.[87]

Why does evidential decision theory fare so well in these decision-making contexts? Because in the specified scenarios, the agent is certain that there is nothing she can do to make herself an exception to the evidential probabilistic relations contained in her causal model. In the case of NP, this means that there is nothing the subject can do to take herself out of the population over which the predictor makes good predictions. When NP is specified such that intervention impossible, acting in a particular way epistemically necessitates that the world was some particular way in the past (or epistemically necessitates some objective chance that the world was some particular way in the past), so we are not led astray by conditioning on our actions and inducing backtracking dependencies.

But what does this mean for our general decision theory? Should causal decision theorists take up evidential decision theory since their own rule does not appear to apply to cases where the agent is certain that she cannot intervene? Not so fast. Though the subject cannot intervene when the predictor's success is specified as perfect or irreducibly stochastic,

[87] When the irreducibly stochastic predictor is right and the subject one-boxes, the subject receives $1,000,000. When the irreducibly stochastic predictor is wrong and the subject one-boxes, the subject receives nothing. Since the irreducibly stochastic predictor is right 95% of the time, the subject's expected winnings are $950,000 if she one-boxes. Likewise, when the irreducibly stochastic predictor is right and the subject two-boxes, the subject receives $1,000. When the irreducibly stochastic predictor is wrong and the subject two-boxes, the subject receives $1,001,000. Since the irreducibly stochastic predictor is right 95% of the time, the subject's expected winnings are $51,000 if she two-boxes.

there are decision-making contexts in which the agent should be certain that she *can* intervene on the causal system at hand, and in these cases, it is causal decision theory that gets things right.

Consider Moriah, who wants to be a zombie for Halloween and believes that no zombie costume is complete without yellowish skin. Moriah believes that the probability that one suffers from fatigue given that one has yellowish skin is greater than the unconditional probability of suffering from fatigue (because iron deficiency is a common cause of both), but Moriah reasonably believes that her intention to have (or not to have) yellowish skin is causally independent of whether she suffers from an iron deficiency. If Moriah's decision-making context is specified by a causal model in which iron deficiency is a common cause of yellowish skin and fatigue, it appears irrational for Moriah to consider what is likely given that she is a person with yellowish skin since this induces spurious backtracking dependencies that are irrelevant to Moriah's decision. But if Moriah instead considers what is likely given that she *intervenes* to make her skin yellowish, she (rightly) sees that her decision to make her skin yellowish has no bearing on whether she should expect to suffer from fatigue.

So what should we do given that neither evidential decision theory nor Pearl's decision theory applies universally across all contexts? According to Meek and Glymour, we should try to establish whether we're in a situation that calls for regarding oneself as capable of intervention, and then proceed accordingly. Meek and Glymour thus seem to endorse engaging in the sort of analysis I provide of NP. By reflecting carefully on the causal system imposed by different sorts of predictive success, we are able to conclude that subjects should regard themselves as incapable of intervention when the predictor's success is perfect or irreducibly stochastic. Likewise, by considering the causal system governing the color of Moriah's skin, we

are able to conclude that Moriah's volition *is* independent of whether she suffers from an iron deficiency, and therefore plausibly *does* constitute an intervention on the system at hand.

But the fact that Meek and Glymour's advice helps Moriah and some NP subjects does not mean that it is always helpful. Consider a specification of NP in which the predictor's success is reducibly stochastic—i.e. stochastic due to the omission of causes. We have already said that this opens the door for the subject to regard herself as capable of intervention. But does it open the door all of the way? Can we infer from the mere fact that the predictor's success is reducibly stochastic that the subject should regard herself as capable of intervention? The answer is no. Even when there is error (as there almost always actually is), NP can be specified such that the agent should be certain that she is not intervening or such that she should be uncertain whether she intervening.

Suppose, for example, that the predictor is accurate 95% of the time, and that the predictor is wrong only when the air pressure fluctuates heavily. (Perhaps air pressure fluctuation is for the predictor what Kryptonite is for Superman.) Since the level of air pressure is presumably outside of the subject's control, it seems that the subject should be certain that she is not intervening, and that she should therefore follow the advice of the evidential decision theorist. Alternatively, suppose that the subject is told that the predictor is accurate 95% of the time, but is not told what causes (if any) explain the predictor's error. The subject needs to investigate the causal setup in order to determine whether she is in a position to intervene. But what if after her inquiry into the causal setup, she is still uncertain whether she can intervene? How should agents respond to uncertainty about whether their settling on *a*-ing constitutes an intervention on the causal system at hand? Meek and Glymour have not answered this important question.

At first pass, it may seem as though the agent should simply take a weighted average of the expected utility calculations that result from applying Pearl's decision theory and evidential decision theory, where the weights correspond to the agent's credences that she is intervening and not intervening. That is, it may seem as though the agent should opt for whatever action x maximizes expected utility when calculated as follows, where C represents the agent's credence that her decision constitutes an intervention on the causal system at hand.

HYBRIDIZED DECISION THEORY

Choose the act that maximizes expected utility when calculated as follows:

$$U(x) =^{df} \sum_y C(P\big(Y = y\big|do(X = x)\big)V(Y = y)) + (1 - C)(P(Y = y|X = x)V(Y = y))$$

Hybridized decision theory reduces to Pearl's when the agent is fully certain that she can intervene and to evidential decision theory when the agent is fully certain that she cannot intervene. But hybridized decision theory builds on Pearl's decision theory and evidential decision theory insofar as it applies when the agent is not sure whether she is in a position to intervene. Suppose, for example that after some fruitless inquiry into the causal setup of NP, the subject is 50% confident that she is intervening. How should she evaluate two-boxing? According to hybridized decision theory, she should calculate a weighted average between the expected utility of intervening to two-box and the expected utility of two-boxing without intervening, where the relevant weights corresponds to her credence that her decision will constitute an intervention (and her credence that her decision will not constitute an intervention).[88] Likewise, suppose that the subject knows that there are 100 subjects, ten of

[88] According to hybridized decision theory, then, the subject who is 50% confident that she is intervening should calculate the expected utility of two-boxing by averaging the expected utility of two-boxing without intervention ($1,000,000p + $1,000) and the expected utility of two-boxing by intervention ((.05)($1,000,000) + $1000), where p is the unconditional probability that the predictor puts a million dollars in the opaque box. Likewise, the subject should calculate the expected utility of one-boxing by averaging the expected utility of one-boxing without intervening (.95)($1,000,000) and

whom have been temporarily endowed with libertarian free will, and that no one who has been endowed with libertarian free will is aware of his or her gift. If the predictor is accurate 95% of the time, and if all of the error is caused by libertarian free will, how should the subject deliberate? Unlike the subject who was ignorant of the causal setup of NP, this subject would seem to have very good grounds to be 10% confident that her decision constitutes an intervention.[89] Unlike evidential decision theory and Pearl's decision theory, then, hybridized decision theory is built to provide agents with advice in contexts like these because it tracks the agent's confidence that she is intervening and purportedly identifies a rational response to uncertainty about whether one is intervening.[90] But does hybridized decision theory succeed at its job? That is, does it actually always identify the *rational* response to uncertainty about whether one is intervening?

The answer here is no. Though hybridized decision theory is on the right track, it does not stand up to scrutiny because it fails to take stock of the distinct ways that one can fail to intervene on a given causal system.[91] Suppose, for example, that we augment the causal system

---

the expected utility of one-boxing by intervention ($1,000,000p). So according to hybridized decision theory, the expected utility of two-boxing equals $500,000p + $25,000 + $1000 and the expected utility of one-boxing equals $500,000p + $475,000. Since $475,000 is greater than $26,000, hybridized decision theory recommends one-boxing over two-boxing. Indeed, hybridized decision theory recommends two-boxing over one-boxing only if the agent is roughly 99.9% confident that her intention constitutes an intervention. If hybridized decision theory recommends two-boxing over one-boxing, then $C(\$1{,}000{,}000p + \$1{,}000) + (1 - C)(\$50{,}000 + \$1000) > C(\$1{,}000{,}000p) + (1 - C)(\$950{,}000)$. By subtracting $C(\$1{,}000{,}000p)$ from both sides of the inequality, one can show that C must be greater than 899/900 in order for hybridized decision theory to recommend two-boxing over one-boxing.

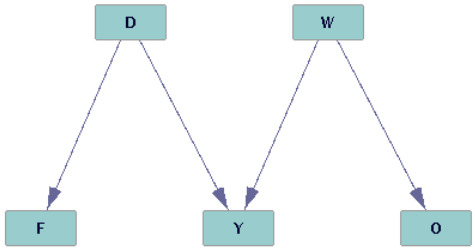
[89] According to this setup, the predictor gets lucky half of the time when predicting what the libertarians will do.

[90] The reader may have noticed that I have not given a precise recipe for building exact credences that one's intention constitutes an intervention. This is not because I have anything against building such a recipe, but is rather because it is outside of the scope of this paper to do so. Indeed, if hybridized decision theory is well motivated, then it is important that philosophers establish exactly how one's credence in one's ability to intervene should be sensitive to evidence. But even without answering this interesting epistemological question, I believe it is valuable to propose a decision rule that instructs agents what they should do *given* their credences in their ability to intervene.

[91] It may fail for other reasons, too. For example, the probability distributions cited in evidential decision theory and Pearl's decision theory are supposed to be subjective probability distributions—i.e. distributions representing agents' subjective credences—but the agent who is uncertain about whether

that Moriah considers to include the following variables: *F* for whether one suffers from fatigue, *D* for whether one suffers from an iron deficiency, *Y* for whether one has yellowish skin, *W* for whether one is weirdly obsessed with Halloween, and *O* for whether one gets ostracized at school. Now, as it happens, Moriah knows, first, that kids who are weirdly obsessed with Halloween tend to get ostracized and likewise tend to have yellowish skin, second, that having yellowish skin does not cause kids to get ostracized (even though there is a correlation between having yellowish skin and being ostracized that results from weird Halloween obsessions). So Moriah is sure of the DAG in Figure 3 (below). Moriah has no idea whether she herself is weirdly obsessed with Halloween, but does not want to be ostracized at school. Should Moriah make her skin yellowish?

Figure 3:



If Moriah is certain that she is intervening, then it is obvious, I think, that Moriah should go ahead and make her skin yellowish (since intervening to make one's skin yellowish cannot affect the probability of any variable represented in Figure 3 other than *Y*). But unlike before (when the causal system at hand excluded *W* and *O*), it is not obvious that Moriah should be certain that she intervening because it is reasonable for Moriah to entertain the possibility that her intention (or what she settles on) is partially determined by the weird

---

she is intervening does not seem uncertain about what her credences are. (She rather seems uncertain about what her credences should be.) Though this may be an objection to hybridized decision theory, we will see that it does not plague the decision theory that I ultimately defend—namely, generalized interventionist decision theory (GIDT).

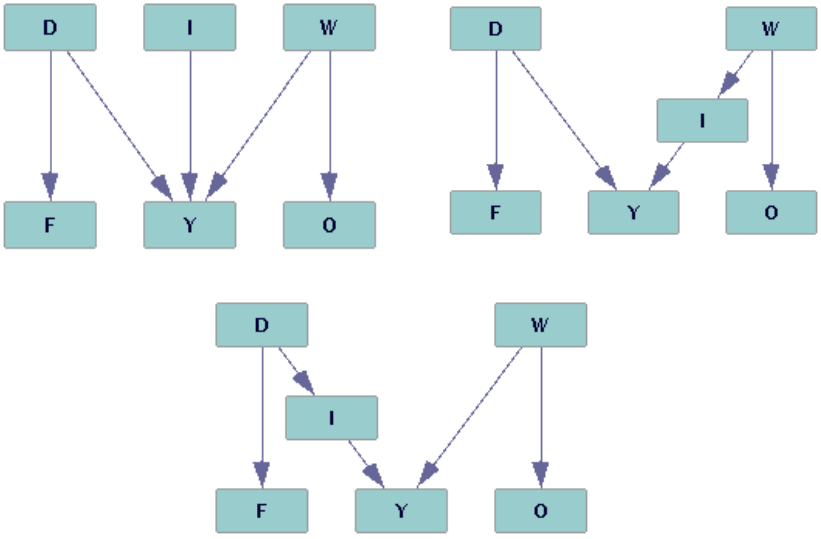
Halloween obsession that she may have. Let us say that Moriah's subjective probability that she is intervening with respect to Υ on the DAG in Figure 3 is n.

By hybridized decision theory's lights, Moriah should determine the expected utility of making her skin yellowish by, first, determining the expected utility of intervening to make her skin yellowish and multiplying that by n, second, determining the expected utility of having yellowish skin *simpliciter* and multiplying that by $(1 - n)$, and, third, adding up the results. This means that if Moriah follows the advice of hybridized decision theory, her deliberation will be sensitive to the correlation between having yellowish skin and suffering from fatigue despite the fact that Moriah knows that this correlation is irrelevant to her choice because she knows that her intention is causally independent of both whether she has an iron deficiency (D) and whether she suffers from fatigue (F). This is because the portion of the expected utility calculation that pertains to the possibility that Moriah is not intervening does not discern between the possibility that Moriah fails to intervene because her choice is a function of whether she has a weird Halloween obsession (W) and the possibility that Moriah fails to intervene because her choice is a function of whether she has an iron deficiency (D). Thus there is a backtracking dependence from Υ to D (and F) that erroneously plays a role when applying hybridized decision theory. So even though Moriah is certain that there are not fatigue-based reasons to abstain from coloring her skin, applying hybridized decision theory will get the result that there are such reasons because the probability of suffering from fatigue given that one has yellowish skin simpliciter is greater than the unconditional probability of suffering from fatigue.

What we need in place of hybridized decision theory, then, is a decision theory that is sensitive to the differences between Moriah's intention (I) being related to the causal system in each of the following three ways (in order to account for the fact that Moriah's expected utility

calculation should be sensitive to correlations between *I* and other variables that are consistent with the DAGs on the first row, but not the DAG on the second row).

Figure 4:



Luckily, Stern's (2016) "interventionist decision theory" can be easily modified to deliver the desired results.

INTERVENTIONIST DECISION THEORY

Choose the act that maximizes expected utility when calculated as follows:

$$U(x) =^{df} \sum_{K} P(K)V(do(X = x), K)$$

According to Stern, the elements of the *K* partition are causal hypotheses over the variable set at hand, where causal hypotheses are ordered pairs of DAGs and sets of objective chance distributions. Though interventionist decision theory sounds complex, the basic idea is simple. According to interventionist decision theory, the agent should determine the expected utility of *x*-ing by, first, spreading her subjective probability distribution across the possible ways that the world could causally be (i.e. across the *K* partition), and, second, using those probabilities to calculate a weighted average of the utilities she attaches to intervening to make herself *x* when

each causal hypothesis is realized. Stern's reason for construing causal hypotheses in terms of ordered pairs of DAGs and sets of objective chance distributions is simply that being *n* confident in a causal hypothesis over some variable set corresponds (from within the interventionist approach to causal modeling), first, to being *n* confident in a particular DAG over that variable set, and, second, to being *n* confident in a particular objective chance distribution (or set of objective chance distributions), where the objective chance distribution specifies the nature and strength of the dependencies represented in the DAG. Stern's idea is that if the agent spreads her confidence across every way that things could be causally related (i.e. across the possible DAGs over the variable set at hand) as well as over the exact ways in which those causal relations could obtain (i.e. across the chance distributions compatible with the live DAGs), then the agent can determine the expected utility of *x*-ing by a calculating a weighted average of the utilities that she attaches to the worlds that result from intervening to make herself *a* when each of the live causal hypotheses (or ordered pairs) is realized, where the weights correspond to the subjective probabilities that she assigns to each of the live causal hypotheses.[92]

The inclusion of Pearl's *do*-operator in interventionist decision theory means that it cannot (without revision) be used in contexts where agents are less than certain that they are intervening because it never asks the agent to consider how much she would value worlds in which she *x*'s by means other than intervening—i.e. by means other than $do(X = x)$. But this does not mean that interventionist decision theory is useless in the present context. Rather, one can revise interventionist decision theory to accommodate contexts where an agent is less than certain that she is intervening by applying the value function to the *intention* to *x* and leaving it open whether the intention is an intervention—i.e. by analyzing the expected utility of *x*-ing as

[92] In order for interventionist decision theory to be plausible, interventions (like Savage's 1954 acts) must be functions from causal hypotheses to outcomes.

follows, where i_x is the value of the intention variable that corresponds to intending to x (or settling on x-ing).[93,94]

GENERALIZED INTERVENTIONIST DECISION THEORY (GIDT)

Choose the act that maximizes expected utility when calculated as follows:

$$U(x) =^{df} \sum_K P(K)V((I = i_x), K)$$

GIDT generalizes interventionist decision theory to apply to contexts where the agent is less than certain that she is intervening by licensing x-ing when the expected utility of intending to x is greater than the expected utility of any other option on the table, where 'the intention to x' is construed broadly enough to refer both to the intervention to make oneself x and to the endogenous intention to x—e.g. such that it refers to I in *both* of the DAGs represented on the first row of Figure 4. Since the endogenous intention to x is *not* necessarily d-separated from the non-descendants of X, GIDT (like hybridized decision theory) has the benefit of incorporating backtracking dependencies when the agent is not sure that she is intervening, but (unlike hybridized decision theory) it does not incorporate too much backtracking (e.g. from the color of Moriah's skin, Y, to whether she has the iron deficiency, D)

---

[93] Some readers may worry about the fact that the agent is asked to apply the value function to worlds in which she intends to x rather than worlds in which she x's *simpliciter* (perhaps because it seems as if the agent is asked to choose what to choose, or to choose what to intend, rather than what to do). But the shift in focus to intentions is not novel to GIDT. Every decision theory where the value function is applied to *intervening* to make oneself x (rather than x-ing itself)—e.g. Pearl's decision theory and interventionist decision theory—shares this feature. This is because interventions are *not* the acts themselves, but are rather causes of the acts. And since they're the sort of causes that reliably result from deciding to x, it is plausible that they are intentions. See Paul (2009, 2012) for a defense of a view according to which (i) the intention to x very reliably results from the decision to x and (ii) the intention to x plays a causal role in the occurrence of x.

[94] Some may wonder why the left-hand side of GIDT is in terms of $U(x)$ rather than $U(i_x)$, or, equivalently, why GIDT asks agents to opt for whatever *act* maximizes expected utility rather than whatever *intention* maximizes expected utility. Just as with the last footnote, I am simply following the lead of interventionist decision theory and Pearl's decision theory since the left-hand side of both analyses is in terms of $U(x)$ rather than $U(do(x))$. But it is worth mentioning that I would have no problem with changing GIDT so that it asks agents to opt for whatever intention maximizes expected utility since it is natural to describe your decision whether to x as the decision to be such that you intend to x.

because the endogenous intention (unlike X itself) can be d-separated from some of X's ancestors. So when Moriah applies GIDT to her choice, she plausibly spreads her credences over causal hypotheses consistent with the DAGs on the first row of Figure 4, but not across the DAG on the second row of Figure 4, and therefore never entertains any causal hypotheses according to which her intention is correlated (in the candidate chance distributions) with D and F. Thus opting for GIDT over hybridized decision theory allows us to capture the sense in which the correlation between Y and D and F is irrelevant to Moriah's choice, given her causal knowledge.[95]

6. Decision without Intervention

Though GIDT generalizes interventionist decision theory such that it applies to decision-making contexts where agents are less than certain that they are intervening, some may argue that there are no *genuine* decision-making contexts in which this uncertainty matters, and that GIDT therefore does not improve upon interventionist decision theory. But recently, the relevance of this uncertainty has attracted ample philosophical attention (albeit

[95] The details of GIDT's application to the case at hand of course depend on unspecified details about the case—e.g. the properties of Moriah's subjective probability and utility functions, and the properties of the objective chance distributions that (along with the DAGs) make up the K partition. But the basic idea is simple. In order for Moriah to determine the expected utility of making her skin yellowish, she must first determine how much she values intending to make her skin yellowish in each of the live causal hypotheses. This effectively amounts to determining how much she values the worlds that result from intending to make her skin yellowish in each of the candidate chance distributions, where the effect of intending to make her skin yellowish in each live chance distribution is the effect of conditioning on intending to make her skin yellowish. In the distributions that are compatible with intervening, then, the chance of Y's non-descendants must be equivalent to what they are unconditionally (since the CMC entails that I is not correlated with Y's non-descendants). But in the live distributions that are compatible with not intervening—i.e. the distributions compatible with the second DAG in Figure 4—the chance distributions over W and O (but not D and F) can be affected by intending to make one's skin yellowish. Now, once Moriah has determined how much she values intending to make her skin yellowish in worlds where each causal hypothesis is realized (as well as how much she values intending each of her other options), she should calculate the expected utility for each of her options by calculating a weighted average over how much she values intending the options in each causal hypothesis, where the weights are provided by her subjective credences in the causal hypotheses. After making these calculations, according to GIDT, Moriah should opt for whatever option maximizes expected utility.

dressed up in different clothes) because there are some cases where even longtime causal decision theorists have intuitions at odd with the results of always representing agents as intervening.

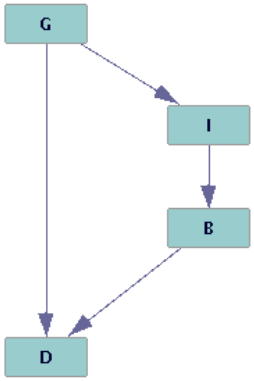
Consider, for example, the recent outpouring of literature in response to Andy Egan's (2007) purported counter-example to causal decision theory in which an agent (i) has the option of pushing a button that will kill all psychopaths, (ii) believes that only a psychopath would push such a button (perhaps because the genetic condition that causes psychopathy likewise causes button-pushing), and (iii) desires that all psychopaths be dead so long as she is not one of them.[96] Most philosophers have the intuition that the agent should refrain from pushing the button because she is aware that she will die if she does, but this is inconsistent with regarding the agent's intention as an intervention because intervening to push the button entails severing the causal and probabilistic dependence between being a psychopath and pushing the button, and thereby entails that the subject should push the button (provided that her unconditional probability that she is a psychopath is sufficiently low). The fact that many philosophers are interested in developing a decision theory that yields this verdict (even when the unconditional probability of psychopathy is very low) is therefore a sign that philosophers are interested in developing a decision theory that ranges over cases where agents are not certain that they can intervene.

If we apply GIDT to Egan's case, it delivers the results that most philosophers favor. If, as specified above, the agent believes she will push the button only if she is a psychopath, then the agent is certain that there is nothing she can do to take herself out of the population over which pushing the button has evidential bearing on her being a psychopath, and she is therefore certain that she is not intervening. Since there is only one place to put I in a plausible causal

[96] Ahmed (2012), Arntzenius (2008), and Wedgwood (2013) are examples of the literature in response to Egan.

model of the causal system at hand (Figure 5, where *G* encodes whether one has the genetic condition, *B* encodes whether one pushes the button dies, and *D* encodes who dies), GIDT straightforwardly gets evidential decision theory's result that the agent should not push the button (because she is sure that she will die if she does).[97]

Figure 5:



Likewise, if we modify the case to make the subject have some small reason to believe that she may be the exception to the rule (perhaps because 1% of the individuals who push the button simply have libertarian free will, and she does not know whether she has libertarian free

[97] Egan (2007: p. 97) presents another "counter-example to causal decision theory" about a "murder lesion." He writes:

> "Mary is debating whether to shoot her rival, Alfred. If she shoots and hits, things will be very good for her. If she shoots and misses, things will be very bad. (Alfred always finds out about unsuccessful assassination attempts, and he is sensitive about such things.) If she doesn't shoot, things will go on in the usual, okay-but-not- great kind of way. Though Mary is fairly confident that she will not actually shoot, she has, just to keep her options open, been preparing for this moment by honing her skills at the shooting range. Her rifle is accurate and well maintained. In view of this, she thinks that it is very likely that if she were to shoot, then she would hit… But Mary also knows that there is a certain sort of brain lesion that tends to cause both murder attempts and bad aim at the critical moment. If she has this lesion, all of her training will do her no good—her hand is almost certain to shake as she squeezes the trigger. Happily for most of us, but not so happily for Mary, most shooters have this lesion, and so most shooters miss. Should Mary shoot?"

Egan and others think that it is rational for Mary to abstain from shooting, but it is clear that if Mary believes that her intention constitutes an intervention, then she should shoot because she should think that her decision breaks any causal or probabilistic dependence between having the lesion and pulling the trigger. It seems as if Mary should have low credence that her intention constitutes an intervention, and, given the disutility of missing, GIDT will make the same recommendation as evidential decision theory.

will), then the subject should have some non-zero credence that her intention constitutes an intervention, and her credence in this possibility should affect her deliberation, but probably not the verdict—i.e. she should not push the button. (Although the expected utility of pushing the button should be higher when she acknowledges some possibility of intervention than when she does not, it need not be higher than the expected utility of not pushing the button.)

GIDT thus guides choice in some contexts not covered by competing decision rules. GIDT is capable of (i) issuing verdicts that take into account an agent's non-extreme credence in her ability to intervene, (ii) issuing verdicts that mirror those issued by evidential decision theory in contexts where it seems appropriate (e.g., Egan's case), and (iii) issuing verdicts that mirror those issued by causal decision theorists of Pearl's ilk (e.g., Moriah's decision whether to apply the yellow dye). But GIDT is not out of the thick yet. Though GIDT captures what seems intuitively rational in these cases, there are other contexts in which GIDT challenges the current philosophical consensus.

In his (1959) *Smoking: The Cancer Controversy*, R.A. Fisher entertains (but does not espouse) the hypothesis that the correlation between whether one is a smoker and whether one suffers from lung cancer is due *not* to some causal influence that smoking exerts on the health of one's lungs, but rather to the causal influence that one's genetic makeup has both on whether one smokes and whether one suffers from lung cancer—i.e. that the relevant correlation is accounted for by a common cause (one's genetic makeup) rather than smoking's exhibiting some causal effect on whether one gets lung cancer.

Though there is good reason to doubt the truth of Fisher's hypothesis, the philosophical consensus is that if Fisher's hypothesis were true, then it would be rational to smoke no matter how much you hate lung cancer (if you enjoy smoking). That is, while it remains controversial what subjects should do when confronted with NP (perhaps because the predictor's

preternatural success complicates things), it is almost universally agreed that if the correlation between smoking and lung cancer is due to a common cause, then the correlation should be irrelevant to what subjects decide and that it is thus rational for subjects to smoke.[98] This result is easily secured by both Pearl's decision theory and interventionist decision theory since smoking does not cause lung cancer and you enjoy smoking no matter whether you have lung cancer. But the group of philosophers who are moved to believe that smoking is obviously rational goes well beyond dyed-in-the-wool causal decision theorists. Indeed, many evidential decision theorists attempt to justify the same verdict (for example) by following Eells' (1982) lead in deploying the "tickle defense" in order to argue that the correlation between lung cancer and smoking is screened off in the context of decision.[99] GIDT, by contrast, denies that the correlation between smoking and lung cancer is irrelevant to whether it's rational to smoke. Is this a defect of GIDT?

GIDT does not get the result that so many philosophers find intuitive because it asserts that the correlation *is* relevant when the agent is not certain that her decision constitutes an intervention. That is, if the agent entertains any causal hypotheses according to which her

---

[98]Seidenfeld (1984: pp. 205-06) may dissent. He seems to object to any use of Fisher's hypothesis as motivation for causal decision theory on the grounds that Fisher's hypothesis is no less fanciful than NP. Since Seidenfeld believes that subjects should one-box when confronted with NP, he can be interpreted as thinking that Fisher's hypothesis does nothing to change his mind.

[99] The tickle defense (which its author, Ellery Eells (1984), insists should be called "the metatickle defense") consists, first, in claiming that insofar as the outcome of any agent's decision is a result of some cause other than her deliberation, it must exhibit its effect by affecting the agent's beliefs and desires, and, second, that when the agent comes to realize what her beliefs and desires are, she effectively learns something that screens off the correlation between her action and its causal predecessors. So, according to the tickle defense, once the agent feels the metatickle that smoking is or is not the action that maximizes expected utility according to her beliefs and desires, she learns something that blocks the correlation between smoking and lung cancer, and she should accordingly adjust her beliefs and desires so as to render smoking the object of choice. There is much to be said about the tickle defense and the model of deliberation that it presumes (wherein agents take the outputs of their calculations of x's expected utility as evidence as they deliberate about whether to x), but these details need not occupy us here. The important upshot in this context is that when the evidential decision theorist deploys the tickle defense, the resulting decision theory delivers the same verdicts as Pearl's decision theory since conditioning on a metatickle, like conditioning on $do(x)$, breaks any correlation between how one acts and the causes of how one acts.

intention is not an intervention on the causal system at hand (as she plausibly should, since it seems unreasonable for the agent to be certain that her intention is not partially determined by her genetic makeup), then the agent should be concerned with the evidential bearing that her intention has on the probability that she will get lung cancer in worlds where she is not intervening.

Though some philosophers may believe this implication is reason to doubt GIDT, it is hard to understand why agents should disregard the correlation between lung cancer and smoking when they are uncertain whether they can intervene (since choosing to smoke really does raise the objective probability of getting lung cancer in such worlds). But, as noted before, some philosophers may insist that the agent has a decision to make only to the extent that her decision is not determined by other factors,[100] and that the agent should correspondingly limit her consideration to what is likely under the supposition that she intervenes even if the agent believes that she may not be able to intervene. In effect, this objection amounts to the claim that agents face genuine decision problems only when they are certain that they can intervene, and that there is correspondingly no reason to move from Pearl's decision theory or interventionist decision theory to something like hybridized decision theory or GIDT. Put differently, one can argue that a necessary condition for deciding to x is full belief that your decision to x (or the intention to x that results from deciding to x) constitutes an intervention. So in the case of the perfect predictor, for example, someone might argue that the subject has no decision to make on the grounds that there is nothing she can do to change what she will

[100] Eells (1984: p. 896) may be arguing something like this when he writes: "To the extent to which you believe that your actual act is influenced by factors other than your decision (the result of rational deliberation), it would seem that, to you, rational deliberation will lose its point."

do.[101] Likewise, in Egan's case, one can argue that if the agent has a genuine decision to make, then she should push the button.[102]

I signaled a response to this objection in the discussion of the perfect predictor case, but it is worth belaboring, given the seeming prominence of the view that there is no decision to make in such contexts.[103] So long as the subject does not herself know what particular action she will take, and so long as the subject believes that she will take whatever action she decides to take, she has epistemically possible worlds to eliminate. In my view, this is sufficient for her to have a decision to make. When faced with the perfect predictor, the NP subject eliminates the epistemic (but not metaphysical) possibility that she two-boxes by deciding to one-box. Likewise, when we specify Egan's case such that the agent is certain that she cannot intervene—i.e. such that she knows she is a psychopath if she the pushes the button—the agent can *decide* not to push the button because she can rule out the epistemically live option of pushing the button and dying. Dummett (1986: pp. 164, 168) articulates this line of thought well in the following response to Mackie's (1977) treatment of the perfect predictor.

> If I think this ⌜i.e. that each person is determined by his character to adopt one strategy or the other, and that it is the psychologist's perfect assessment of this character that determines the situation, namely whether the closed box is or is not empty⌝, Mackie maintains that I must recognize that my choice of strategy

---

[101] This line of thought has been pursued in response to a version of NP that Jeffrey (1984) attributes to Bas van Fraassen. Van Fraassen's specification of NP is meant to yield a case where the tickle defense does not apply, and where evidential decision theorists like Eells and Jeffrey should correspondingly should advocate one-boxing. The basic idea is that the predictor not only predicts the subject's decision, but also whether the subject carries out her decision successfully, therefore ruling out the possibility that some tickle (or metatickle) will screen off the correlation between the predictor's prediction and the subject's action. Two-boxers of all stripes (e.g. Eells 2000 and Joyce 2007) have responded that van Fraassen's case is not a genuine decision problem because the subject's destiny has been determined (by factors other than the agent's deliberation) once the predictor makes her prediction.

[102] Ahmed (2012) argues that our intuition about Egan's case is faulty and that there is reason to push the button.

[103] This line of thought can be straightforwardly found in Mackie (1977) and Pearl (2009). I also believe that one can read Eells (2000) and Joyce (2007) as defending this thesis, but explaining this interpretation would require engaging with their dynamic model of deliberation, and this is outside the scope of this paper.

> is not free, and that I therefore do not need to make up my mind. More exactly, what he says is that the question, 'What is it reasonable for me to do?', is idle.
>
> Obviously, this is wrong: for me the question is not idle at all. Here I am, wavering between taking both boxes and taking only the closed one, and trying to decide which it is reasonable for me to do. The thought, 'My choice is not really open: the psychologist already knows which I shall do', may fill me with despair: but it will not help me to decide, and it equally will not dispense me from deciding…
>
> The only kind of freedom relevant to the paradox is that I am free to take one box or both in the sense that I can rule out the possibility that I should attempt to take only one and be unable to avoid taking both. Suppose that I believe that I am, in *that* sense, wholly constrained. I think, that is, that, if there is nothing in the closed box, I shall simply be unable, if I try, to take only it, and that, if it contains [the money], I shall similarly be unable to take both boxes. Then, indeed, the question, 'Which shall I do?', has become idle for me: it will, I believe, turn out the same whichever I decide. But, if I do *not* think that, the question is not idle, independently of any opinion I hold about determinism… My decision reasonably affects for me the probability that the closed box contains [the money] rather than nothing.

As the subject deliberates about what to do, there are two possibilities that appear open to her: either she will one-box or she will two-box. Moreover, she knows that if she decides to one-box, she will, in fact one-box, and that if she decides to two-box, she will, in fact two-box. The subject thus has a decision to make. Of course the subject would have no decision to make were she aware that she'd take some *particular* action no matter what—perhaps because she knew that the predictor predicted that she would two-box, or because she knew that she antecedently behaved in some particular way that necessitated her taking a particular action—but so long as there remain epistemic possibilities that she can capably eliminate, it seems that she has a decision to make.[104]

---

[104] Of course if causal decision theorists apply Pearl's decision theory or interventionist decision theory (or, alternatively, any rule where expected utility is calculated with non-backtracking counterfactuals) to this decision context, they will get the result that the subject that the subject should two-box in the case of the perfect predictor. They may likewise argue that the perfect predictor decision-making context is one of the unfortunate contexts in which it pays to be irrational. But this gets things backwards. Here, again, Dummett (1986: pp. 168-69) sets things straight:

> "*After* I have [one-boxed], the rules governing the assertion of counterfactuals [i.e. that they do not backtrack] may entitle me to assert, 'If I had taken both boxes, I should

Likewise, when the agent considers what is likely in the event that she cannot intervene on the causal structure imposed by Fisher's hypothesis, there are two epistemic possibilities that have not been eliminated: either she has the genetic makeup that causes lung cancer or she does not. Moreover, she knows that the probability that she has the relevant genetic makeup is higher if she decides to smoke than if she decides to abstain. So it appears that here, too, the subject has a decision to make. The agent cannot just *assume* that she is automatically in a position to intervene when she has a decision to make, since, as we have seen, this is a substantive assumption in the decision-making context.

Moreover, even if the reader is not convinced that the agent has a decision to make when she is *certain* that she is not intervening, it is hard to see how one can deny that an agent has a decision to make when an agent has *some* confidence that she is intervening. When Moriah is highly confident that she is intervening, she is highly confident that she can change the objective probabilities that her act obeys. And if Moriah is highly confident of this, it seems obvious that she should deliberate about what to do. Of course anyone who believes that we must believe we are intervening in order to decide can argue that Moriah's high confidence that she is intervening should be re-described as her confidence that she is in a decision problem, and that Moriah's confidence that she is not intervening should correspondingly not play any role in her deliberation. (The idea would be that Moriah should operate under the assumption that she is intervening for the reason that this assumption is safe in the event that she has a decision to make.) But this seems counter-intuitive. Once Moriah acknowledges the possibility that it is within her power to change the probabilities that her act obeys by intervening,

---

have got [$1,001,000]; but that is only a remark about our use of counterfactual conditionals. *Before* I make my choice, I should be a fool to disregard the high probability of the statement, 'If I take both boxes, I shall get only $1,000'. That is not merely a remark about our use of the word 'probability', nor even about our use of the word 'rational', but about what it is rational to do."

everyone should agree that Moriah should deliberate about what to do, and not merely that there exists an epistemic possibility that Moriah should deliberate about what to do. And if this much is true, it is hard to see how Moriah could justifiably ignore her uncertainty that she is intervening (since whether she is intervening affects the probabilities of the various outcomes) during deliberation. So it seems that even those who disagree with Dummett that there are genuine decision-making contexts in which agents are *certain* that they cannot intervene should agree that GIDT is well motivated since it delivers the right results when agents are uncertain whether they can intervene.

Of course it is still incumbent upon defenders of GIDT to explain why there is an asymmetry in philosophers' intuitive reactions to Egan's case and Fisher's case. One possible explanation is that humans are psychologically primed to consider decisions as interventions—i.e. to believe that they are intervening by default—and that certain cues can prompt humans to shift their focus to the possibility that they cannot intervene.[105,106]

In Egan's case, on the one hand, if the agent mistakenly assumes that it is within her power to intervene, she still must consider the probability that she is a psychopath (since the action is desirable only if she is not a psychopath even when she is intervening), and this may direct her attention to the fact that deciding in a particular way may be correlated with psychopathy. By focusing on the probability that one is a psychopath, the agent's attention may be redirected to the possible objective evidential significance of her decision—i.e. the possibility

---

[105] As we deliberate, we might be primed to think of ourselves as people who we can coerce. For example, given Fisher's hypothesis, we might ask whether we should ban ourselves from smoking, where we think of ourselves as a population that is subject to legislation. When we assume such legislative power, our decision seems to constitute an intervention, and we correspondingly find that we should *not* ban ourselves from smoking. But when we make sure to focus on what *we* should do (rather than how we should legislate some coercible population), it appears that we do not have such good reason to believe that our intention constitutes an intervention. (It may be that our genetic makeup affects our decision whether to smoke.) Perhaps we're psychologically primed to take the perspective of the legislator even when we're in a decision-making context that does not license this.

[106] I am greatly indebted to Dan Hausman for suggesting an explanation along these lines in personal communication.

that she is not intervening—and this possible evidential significance may guide her choice to abstain from pushing the button.

In Fisher's case, on the other hand, when the agent considers what she should do under the supposition that she intervenes, she recognizes that smoking beats abstaining *no matter what* her genetic makeup is, and *no matter whether* she has lung cancer. So as the agent considers what to do (under the mistaken assumption that her intention constitutes an intervention), she need not think about the probability that she has the gene, and there is accordingly nothing to redirect her attention to the possible evidential significance of her decision.

According to this line of thought, then, the asymmetry in intuition is grounded in the fact that when the agent believes she is intervening, the unconditional probability of psychopathy is relevant to the agent's choice, while the unconditional probability of lung cancer is irrelevant.[107] Or put differently, the relevant difference between the two cases is that the extent to which intervening to make oneself smoke is more desirable than intervening to make oneself not smoke does not depend on whether the agent gets lung cancer, while the extent to which intervening to make oneself push the button is (or is not) more desirable than intervening to make oneself not push the button very much depends on whether one is a psychopath. Because the unconditional probability of psychopathy matters even when the agent is (mistakenly) certain that she is intervening, the agent is primed to think about factors (including her intention) that could be evidentially relevant to whether she is a psychopath and

[107] There is another (albeit perhaps less plausible) psychological explanation of the asymmetry in intuition that is couched in terms of the relevance of the unconditional probability of psychopathy and the irrelevance of the unconditional probability of lung cancer. While I have suggested that the irrelevance of the unconditional probability of lung cancer leads people to be over-confident that they are intervening (because they never considers worlds in which they are not), one could likewise argue that people aren't mistaken about the probability that they are intervening, but are instead mistaken about what is likely in the event that they are not intervening—e.g. that endogenously settling on smoking does not affect the probability that one gets lung cancer. According to this line of explanation, we mistakenly infer that settling on smoking doesn't provide evidence about lung cancer from the fact that the unconditional probability of lung cancer is irrelevant to the extent that smoking is better than not smoking, no matter whether we're intervening.

is led to consider worlds in which she does not intervene as a result. But because the unconditional probability of lung cancer does *not* matter when the agent intervenes, the agent is never primed to think about factors that could be evidentially relevant to whether she has lung cancer and is therefore never led to consider worlds in which she fails to intervene.[108]

Though GIDT does not categorically weigh in favor of smoking in Fisher's case, it does seem to capture the way in which the unconditional probability of psychopathy plays a special role in deliberation. According to GIDT, the agent can ignore the unconditional chance (in any of the live objective chance distributions) that she will get lung cancer as she assesses what to do because the unconditional chance of lung cancer is irrelevant both when she is intervening (since intervening to smoke beats intervening to not smoke by the same amount no matter the unconditional chance of lung cancer) and when she is not intervening (since the value of the outcome of intending to smoke in worlds where she is not intervening depends on the chance that she gets lung cancer *given* that she intends to smoke but not on the unconditional chance of lung cancer).[109] In Egan's case, on the other hand, the agent can ignore the unconditional chance that she is a psychopath only if she is certain she is *not* intervening (because intervening to push the button does *not* beat intervening to not push the button no matter the unconditional chance of psychopathy).[110]

---

[108] It is important to distinguish this argument from a different one that a causal decision theorist might make—namely, that evidence about whether the agent is a psychopath is relevant to whether the causal consequences of pushing the button are better (or worse) than the causal consequences of abstaining while evidence about whether one has the gene is irrelevant to whether the causal consequences of smoking are better (or worse) than the causal consequences of abstaining. Though this might seem *prima facie* compelling, it is obviously mistaken because whether one has the gene *is* relevant to whether the causal consequences of smoking are better than the consequences of abstaining in the event that the agent is not intervening.

[109] When the agent considers how much she values intending to smoke in worlds where she is not intervening, she should consider how much she values worlds in which the objective probability of lung cancer is raised by intending to smoke. The unconditional objective probability of lung cancer plays no role in this calculation.

[110] Similarly, when the agent considers how much she values intending to push the button in worlds where she is not intervening, she should consider how much she values worlds in which the objective

So it appears that GIDT identifies an asymmetry between the two cases that may help explain philosophers' intuitive reactions. According to GIDT, Fisher's case and Egan's case *are* fundamentally different insofar as the unconditional chance of psychopathy possibly affects what is rational to do, while the unconditional chance of lung cancer necessarily does not.[111] But GIDT does not categorically weigh in favor of smoking. Instead, because GIDT asks the agent to weigh causal hypotheses according to which intending to smoke *does* raise the chance of getting lung cancer (i.e. causal hypotheses according to which the agent is not intervening), there can be lung-cancer-related reasons to abstain from smoking even given Fisher's hypothesis, just as there can be predictor-related reasons to one-box in NP.

By GIDT's lights, then, when the agent's attention is drawn to the evidential significance of her decision, the agent's attention is drawn to something on which it should have been focused all the while—namely, the backtracking dependencies that would result from her intention not constituting an intervention. This is the very point of GIDT. When an agent considers what is likely to occur given that she intends in this or that way, she should pay careful attention to the causal facts about her intention, and this entails paying attention to the possibility that her intention does not constitute an intervention. If the agent has a decision to make and acknowledges the possibility that she is not intervening, she should determine what is likely in the event that she does not intervene and should thus pay attention to the objective probabilistic dependence between smoking and lung cancer in worlds in which she is not intervening. Though philosophers' intuitions tell against GIDT as they react to Fisher's case, it

---

probability of psychopathy is raised by intending to push the button. The unconditional objective probability of psychopathy plays no role in this calculation.

[111] Of course one can construct cases where the unconditional probability of getting lung cancer does *not* exhibit the same effect on expected utility when the agent decides to smoke and decides to not smoke. All one must do is make it the case that the extent to which she prefers smoking to not smoking depend on whether she get lung cancer. But as I see things, this effectively amounts to turning Fisher's case into a case with the same structure as Egan's case. If my explanation of philosophers' intuitions is correct, then for some values of the outcomes, philosophers should think that it is irrational to smoke when Fisher's case admits to this description.

is easier to stomach GIDT's verdict once we realize, first, that philosophers' intuitions may be driven by the irrelevance of the unconditional chance of lung cancer, and, second, that GIDT provides us with the tools to take stock of this irrelevance.

7. Loose Ends

First, one may wonder whether complications are posed for GIDT by the fact that what counts as an intervention depends on the causal system at hand. Consider Moriah's decision whether to make her skin yellowish by applying the yellow dye. Though her intention to make her skin yellowish obviously constitutes an intervention on the causal system in which iron deficiency is a common cause of fatigue and yellowish skin, there are other causal systems on which her intention does not obviously constitute an intervention—e.g. the system depicted in Figure 3, or, more dramatically, a complete causal graph of the workings of the (possibly deterministic) universe. In the former case, it is plausible that Moriah should not be certain whether her intention constitutes an intervention. In the latter case, since there are no exogenous causes by hypothesis, Moriah should be certain that her decision does not constitute an intervention.

Does this somehow make trouble for GIDT? It does not. In fact, the model-relativity of 'intervention' brings out one way in which GIDT is superior to its competitors. While GIDT delivers a verdict that is plausibly rational no matter how we draw up Moriah's decision-making context—i.e. no matter which variable set Moriah considers—Pearl's decision theory and evidential decision theory deliver plausible verdicts only given certain specifications of Moriah's decision-making context. According to GIDT, when Moriah considers the causal system in which Halloween obsession is a common cause of ostracization and yellowish skin, Moriah's should compute a weighted average over the possibilities in which she intervenes and

does not intervene. When she considers the complete causal model of the universe, she should be certain that she is not intervening and her deliberation should accordingly be limited to possibilities in which she does not intervene. But according to Pearl's decision theory and Stern's interventionist decision theory, Moriah should consider what is likely upon intervening even when she considers the complete causal model of the universe. And according to the evidential decision theorist who believes that her decision theory is right across all decision-making contexts (where contexts are specified as causal systems on which the agent might intervene), Moriah should consider what is likely upon not intervening even when she considers the causal system comprised of iron deficiency, fatigue, and yellowish skin. So GIDT seems to properly account for the model-relativity of 'intervention' while Pearl's decision theory, interventionist decision theory, and evidential decision theory do not.

Second, one may ask why agents should only consider possibilities where the agent succeeds in performing the action in question, given the possibility that agents may doubt their ability to intervene to bring about the relevant action. That is, one may wonder why we should focus on possibilities in which the agent performs the act in question even when the agent fails to intervene. This underscores an important point about the nature of an agent's confidence that she is intervening. When I say that an agent is n confident that she is intervening to make herself x, I do *not* mean that the agent is n confident that her intervention to x will be successful in making her x. Rather, I take it for granted that she should be certain that she will do whatever she intends to do, and the credence therefore refers to I occupying a particular place in the causal graph at hand.

It is an interesting question whether we ever decide to do things that we do not fully believe that we will do upon deciding or intending to do them, and it is additionally interesting how we should model such decisions if there are such decisions. But it is an interesting question

that lies outside the purview of this paper. The type of uncertainty on which this paper focuses is *not* uncertainty about whether one will do the thing that one decides to do, but is rather uncertainty about whether one does what one decides to do as a result of intervention.[112]

8. Conclusion

I have argued that there are decision-making contexts in which agents should be certain that their decisions do not constitute interventions, as well as decision-making contexts in which agents should be unsure whether their decisions constitute interventions. This led me to propose a decision theory intended to improve on Pearl's decision theory by applying to each of these contexts.[113]

Though generalized interventionist decision theory (GIDT) earns its keep over Pearl's decision theory and interventionist decision theory when an agent is less than certain that her decision constitutes an intervention, it applies to contexts in which agents are certain, too. So, if after reading this paper, you remain unconvinced that there are decision-making contexts in which agents are less than certain that they are intervening, you may still apply GIDT (and secure interventionist decision theory's results). But my hope is that the discussion of Newcomb's Problem and other cases has given you sufficient reason to believe that any agent

---

[112] It is worth noting that most decision theorists do not offer recipes for what agents should do if they will not do what they decide to do, though there are exceptions (e.g. Jeffrey 1984 and Eells 1984). This has not worried many decision theorists because they have assumed that they can always redescribe an action that an agent decides to take (but is not certain that she can take) as an action that she *is* certain she can take. Consider my decision to grab a sandwich from the corner store. Since I acknowledge the possibility that I might get hit by a car en route to the corner store, I am not certain that I will grab a sandwich from the corner store even if I decide to do so. That said, I *am* certain that I will head towards the corner store in order to grab a sandwich, and the decision theorist can redescribe my decision as the decision to do *this*.

[113] I do not advertise GIDT as a competitor to IDT because Stern is careful to mention that the applicability of IDT may be limited to decision-making contexts in which rational agents should represent themselves as intervening.

who is always certain that she is intervening is less commonsensical than Pearl would have you think.

Bibliography

Ahmed, A. (2012). 'Push the Button', *Philosophy of Science*, 79: pp. 386-395.

Arntzenius, F. (2008). 'No Regrets; or, Edith Piaf Revamps Decision Theory', *Erkenntnis*, 68: pp. 277–97.

Briggs, R. (2012). 'Interventionist Counterfactuals.' *Philosophical Studies*, 160: pp. 139–166.

Dummett, M. (1986). 'Causal Loops', in R. Flood and M. Lockwood (*ed*), *The Nature of Time*. Oxford, UK: Blackwell Publishers.

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

Eells, E. (1984). 'Metatickles and the Dynamics of Deliberation', *Theory and Decision*, 17: pp. 71-95.

Eells, E. (2000). 'The Foundations of Causal Decision Theory by James Joyce', *The British Journal for the Philosophy of Science*, 51: pp. 893-900.

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

Elwert, F. (2013). 'Graphical Causal Models', in S. Morgan (*ed*), *Handbook of Causal Analysis for Social Research*. New York, NY: Springer.

Elwert, F., and Winship, C. (2014). 'Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.' *Annual Review of Sociology*, 40: pp. 31–53.

Fisher, R. A. (1959). *Smoking: The Cancer Controversy*. London, UK: Oliver and Boyd.

Galles, D., and Pearl, J. (1998). 'An Axiomatic Characterization of Causal Counterfactuals.' *Foundations of Science*, 3: pp. 151–182.

Geiger, D. and Pearl, J. (1989). 'Logical and Algorithmic Properties of Conditional Independence and Qualitative Independence.' Report CSD 870056, R-97-IIL, Cognitive

Systems Laboratory, University of California, Los Angeles.

Gibbard, A., and Harper, W. (1978). 'Counterfactuals and Two Kinds of Expected Utility.' in C. Hooker, J. Leach and E. McClennen (eds), *Foundations and Applications of Decision Theory*. Dordrecht: Riedel: pp. 125–62.

Halpern, J. (2000). 'Axiomatizing Causal Reasoning.' *Journal of Artificial Intelligence Research*, 12: pp. 317–337.

Hausman, D. (1998). *Causal Asymmetries*. Cambridge, UK: Cambridge University Press.

Hausman, D. and Woodward, J. (1999). 'Independence, Invariance, and the Causal Markov Condition.' *British Journal for the Philosophy of Science*, 50: pp. 521–83.

Hiddleston, E. (2005). 'A Causal Theory of Counterfactuals.' *Nous*, 39: pp. 232–257.

Hitchcock, C. (2001). 'The Intransitivity of Causation Revealed in Arrows and Graphs.' *The Journal of Philosophy*, 98: pp. 273–299.

Hitchcock, C. (2013). 'What is the 'Cause' in Causal Decision Theory?', *Erkenntnis*, 78: pp. 129–146.

Hitchcock, C. (2015). 'Conditioning, Intervening, and Decision.' *Synthese*, pp. 1–20.

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

Jeffrey, R. (2004). *Subjective Probability: The Real Thing*. Cambridge, UK: Cambridge University Press.

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

Joyce, J. (2007). 'Are Newcomb Problems Really Decisions?', *Synthese*, 156: pp. 537-562.

Levi, I. (1987). 'Rationality, Prediction, and Autonomous Choice.' *Canadian Journal of Philosophy*, 19: pp. 339–63.

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

Lewis, D. (1973b). 'Counterfactuals and Comparative Possibility.' *Journal of Philosophical Logic*, 2: pp. 418–446.

Lewis, D. (1973c). 'Causation.' *Journal of Philosophy*, 70: pp. 556–567.

Lewis, D. (1979). 'Counterfactual Dependence and Time's Arrow.' *Nous*, 13: pp. 455–476.

Lewis, D. (1980). 'A Subjectivist's Guide to Objective Chance.' in R. Jeffrey, ed., *Studies in Inductive Logic and Probability, Vol. II*. Berkeley: University of California Press, pp. 263–94.

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

Lewis, D. (1981b). 'Why Ain'cha Rich?', *Nous*, 15: pp. 377-380.

Mackie, J.L. (1977). 'Newcomb's Paradox and the Direction of Causation', *Canadian Journal of Philosophy*, 7: pp. 213-225.

McGee, V. (1985). 'A Counterexample to Modus Ponens.' *The Journal of Philosophy*, 82: pp. 462–471.

Meek, G., and Glymour, C. (1994). 'Conditioning and Intervening.' *The British Journal for the Philosophy of Science*, 45: pp. 1001–1021.

Paul, S. (2009). 'How We Know What We're Doing', *Philosophers' Imprint*, 9(11).

Paul, S. (2012). 'How We Know What We Intend', *Philosophical Studies*.161: pp. 327–346.

Pearl, J. (1993). 'Comment: Graphical Models, Causality, and Intervention.' *Statistical Science*, 8: 266–69.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed., Cambridge, UK: Cambridge University Press.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley, CA: University of California

Press.

Savage, L. (1954). *The Foundations of Statistics*, New York: John Wiley.

Seidenfeld, T. (1984). 'Comments on Causal Decision Theory', *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984: pp. 201-212.

Seidenfeld, T., Schervish, M., and Kadane, J. (2010). 'Coherent Choice Functions under Uncertainty.' *Synthese*, 172: pp. 157–76.

Skyrms, B. (1980). *Causal Necessity*. New Haven: Yale University Press.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd ed., New York, NY: Springer-Verlag.

Spohn, W. (2012). 'Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box', *Synthese*, 187: pp. 95-122.

Stalnaker, R. (1981). 'A Theory of Conditionals.' in W. Harper, R. Stalnaker, and G. Pearce (eds), *Ifs*. Dordrecht, NLD: Dordrecht Riedel: pp. 41–55.

Stern, R. (2016). 'Interventionist Decision Theory', unpublished manuscript.

Verma, T. (1987). 'Causal networks: semantics and expressiveness.' Technical Report R-65-I, Cognitive Systems Laboratory, University of California, Los Angeles.

Wedgwood, R. (2013). 'Gandalf's Solution to the Newcomb Problem', *Synthese*, 190: pp. 2643-2675.

Woodward, J. (2003). *Making Things Happen*. Oxford, UK: Oxford University Press.

Zhang, J., and Spirtes, P. (2008). 'Detection of Unfaithfulness and Robust Causal Inference.' *Minds and Machines*, 18: pp. 239–71.