More Problematic than the Newcomb Problems:

Extraordinary Cases in Causal Decision Theory and Belief Revision

Daniel Listwa
4/01/15

John Collins
Adviser

Senior Thesis
Submitted to the Department of Philosophy at Columbia University

TABLE OF CONTENTS

"…it seems that we are only setting aside some very special cases—cases about which I, at least, have no firm views. (I think them much more problematic for decision theory than the Newcomb problems.)"

-David Lewis (1981)[1]

I. INTRODUCTION

A decision theory is a formal tool for determining what action, given your desires and beliefs, is most preferable. Evidential Decision Theory (EDT) recommends doing the action you would most like to hear that you have done. As such, it appears a good candidate for a decision theory. On further reflection, it proves faulty, as it endorses "an irrational policy of managing the news," since it may not be that the action you would prefer to have heard that you have done is the action that would bring about, given your causal beliefs, the best result. The reason the best-news-producing action may differ from the best-result-producing action is that an action may provide evidence about the agent's causally relevant circumstances. This distinction can lead to bad recommendations, as famously displayed in the Newcomb Problem. To address such issues Causal Decision Theory (CDT) has been developed, a decision theory that is intended to recommend only those actions which bring about the best results.

Different versions of CDT have been proposed, but the form often taken for granted, at least partially for its intuitiveness and simplicity, is the version proposed by David Lewis. For reasons about to be made clear, I will refer to Lewis' proposal as K-CDT.[2] Lewis' version addresses EDT's apparent fault by screening off whatever recommendation-distorting evidence an action may provide about the agent's causal circumstances. It does this first by defining a dependency hypothesis, K, as "a maximally specific proposition about how the things he cares

[1] Lewis, "Causal Decision Theory," 18.
[2] Ibid.

about do and do not depend causally on his present actions." These Ks can also be thought of as conjunction of counterfactuals of the form A□→B, i.e., 'if I were to A, then B'.

How K-CDT differs from EDT can be easily seen in reference to a standard decision matrix. In the matrix below, a basic Newcomb Problem is represented. Options are listed along the rows and possible circumstances are listed along the columns. K-CDT differs from EDT in that it restricts the types of circumstances that may be listed at the top of the columns. If each circumstance/option combination represents a possible world, K-CDT restricts relevant circumstances to those such that in every world in which that circumstance holds the set of counterfactuals is the same. Further, while both K-CDT and EDT calculate an action's expected utility through a weighted average of the utilities of each of the possible worlds in which you do the action under consideration, the weights used are different. EDT uses the probability of each circumstance given that the action is chosen, while K-CDT uses the unconditional probability of the circumstance, reflecting the idea that the action does not alter the circumstance in which it is performed.

Agent\Predictor	Predicts One-Box (H_1)	Predicts Two-Box (H_1)
One-Box	$1,000,000	$0
Two-Box	$1,001,000	$1,000

This form of CDT, however, is not the only one. Another form is one in which imaging, rather than dependency hypotheses, is taken to be primitive. Such a theory was proposed by Sobel [1986], and I will refer to it as I-CDT.[3] In the general case, a function is an imaging function iff it assigns to every argument pair (*w*, *X*), where *w* is a world and *X* is a proposition, a probability distribution over the set of possible worlds, such that positive probability is only

[3] Sobel, "Notes on Decision Theory."

assigned to *X*-worlds, i.e., worlds were the proposition X is true. Sobel utilizes a specific imaging function, which I will call a 'genuine imaging function', that is intended to reflect the similarity relations between worlds directly. In particular, it shifts the probability assigned to each world *w* to the most similar *X*-world. Following Lewis's analysis of counterfactuals, the genuine imaging function, by reflecting the similarity relations that hold at each possible world, also reflects the counterfactuals that hold at each world, thus indicating an underlying relationship with K-CDT.

In fact, Lewis argues that these two theories are, for all intents and purposes, equivalent. The equivalence argument requires the introduction of two constraints on Sobel's theory, which I introduce in some detail in section VII. The constraints are plausible and, for the most part, have gone unchallenged. The exception is Rabinowicz [1982], who points to violations of both these constraints. [4] The scenarios Rabinowicz considers are challenging in that they are only schematic, resisting the application of intuition or deeper thought about the implications of each theory's recommendation. As a result, he is able to note that the two theories come apart, but is unable to suggest which theory provides the most plausible approach. While his observations are able to provide reason to prefer I-CDT to K-CDT when one makes a number of highly contested assumptions regarding chancy universes, no resolution is offered in the boarder case.[5] Similar remarks can be found from Lewis, who, as will be discussed in section VII, was aware of the types of counterexamples to his claim regarding the equivalence of K-CDT and I-CDT, but was unable to find a compelling reason to argue, in these cases, for one over the other.

II. OVERVIEW OF THE ARGUMENT

[4] Rabinowicz, "Two Causal Decision Theories: Lewis vs Sobel."
[5] Here I refer to Rabinowicz's comments regarding the compatibility of K-CDT and centered, chancy worlds. What it means for a world to be centered is discussed in section V; the content of Rabinowicz's argument is discussed in section XI.

In this paper, I turn toward the types of cases regarding which I-CDT and K-CDT come apart. In particular, I present two novel examples, which, I argue, display I-CDT to be superior to K-CDT. The first, the "Chancy Dog Problem," follows the form of a scenario that is explored in Rabinowitz [2009] and shows I-CDT to give what I argue to be the correct recommendation in a chancy universe, where K-CDT gives the incorrect one.[6] I explain this difference as stemming from the fact that conditionalization shifts credences too much, adjusting beliefs as though the oracle provides more information that it actually does. Specifically, conditionalization leads one to revise one's beliefs as though the oracle's information tells you not only what will happen, but also what will happen were you to do something other than what you will actually do. Imaging by shifting credence directly to the nearest compatible world avoids this fault by making the minimal adjustment to one's beliefs to maintain consistency. The other example, the "Faulty Signal Problem," is structurally similar to a problematic scenario discovered by Collins [forthcoming], which reveals K-CDT to fail to provide a recommendation even in certain non-chancy universes, particularly those that exhibit what I refer to as "counterfactual asymmetries."[7] I explain why cases of counterfactual asymmetry present a particular difficulty for K-CDT, as a theory which utilizes conditionalization as a method of updating on the information provided by taking a certain action-proposition to be true.

In both evidential and causal decision theory, one considers an action by revising one's belief to suppose the action-proposition is true. The evidential decision theorist does this revision by hypothesizing that she receives, in an indicative fashion, news that the proposition is true. One who uses K-CDT does the same, but holds fixed certain causal beliefs. The Newcomb Problem and other such cases have shown, I believe, EDT to be wrong, but the problems I will

[6] Rabinowicz, "Letters from Long Ago: On Causal Decision Theory and Centered Chances."
[7] Collins, "Decision Theory After Lewis."

present will show this second suggestion to be wrong as well.[8] Instead, the correct form of decision theory requires an agent to subjectively place herself in different circumstances, even if such circumstances do not agree with facts she believes to be true. Deliberating in this subjunctive sense, as I discuss, requires imaging.

III. LEWIS' PROPOSAL: K-CDT, A PRIMER

While in the first section I provided a brief overview of the nature of Lewis' proposal, in this section and in the two to follow, I will give a more detailed presentation of what I have referred to as K-CDT and I-CDT, and the relationship between the two. To begin, take the agent's desires to be represented by a value function $v(\cdot)$ on the set of possible world *W*. For convenience and simplicity, we assume the set of worlds to be finite. Let $CR(\cdot)$ be a credence function representing the agent's beliefs as a probability distribution over *W*. The agent possesses a set of options expressed as propositions, $O_1, O_2, O_3 \ldots O_N$, which together form a partition of *W*.

EDT recommends choosing the option with the highest expected value, which is defined as the sum of the values of each of the possible worlds weighted by the conditional credence in the respective world given the option under consideration. We can represent this definition of expected value as:

$$V(O) = \sum_{w \in W} CR(w|O)v(w)$$

In Lewis' theory, a set of dependency hypotheses, $\mathcal{K} = \{K_1, K_2, K_3 \ldots K_N\}$, form another partition of *W*. Only one dependency hypothesis holds in each world, specifying the "relevant relations of causal dependence that prevail there".[9] The agent's credence is spread over the subset of dependency hypotheses that the agent considers candidates for accurate descriptions of

[8] In this discussion, I take it for granted that EDT is misguided. I recognize that this is not a settled debate and unlikely to become settled soon. Still, I believe my arguments can be informative regarding the nature of deliberation, even if one does not endorse the two-boxer position.
[9] Lewis, "Causal Decision Theory," 11.

the world's causal structure. K-CDT privileges $\boldsymbol{\mathcal{K}}$ as the partition of worlds, calculating the expected utility by using the unconditional assignment of credence to dependency hypotheses, and thus holding fixed the agent's beliefs about the causally relevant circumstances. K-CDT recommends choosing the option with the highest expected utility, here defined by the function U_K:

$$U_K(O) = \sum_{w \in W} \sum_{\mathrm{K} \in \mathcal{K}} CR(K) CR(w|OK) v(w)$$

It is at this point that we can clearly draw the distinction between EDT and K-CDT. Since the set of dependency hypotheses $\boldsymbol{\mathcal{K}}$ form a partition of *W*, we can represent the expected value function $V(\cdot)$ in the following way, which is equivalent to the mathematical representation of EDT above:

$$V(O) = \sum_{w \in W} \sum_{\mathrm{K} \in \mathcal{K}} CR(K|O) CR(w|OK) v(w)$$

The double sum, though unwieldy and resulting in a number of terms whose values trivially turn out to be zero, illustrates the underlying difference between EDT and K-CDT. In both EDT and K-CDT, $CR(w|OK)v(w)$ represents the expected value of a world assuming some causal structure—given by dependency hypothesis *K*—to be true. In K-CDT, this term is weighted by $CR(K)$, the credence that the assumed causal structure is in fact true. EDT weighs the term by $CR(K|O)$, which is the credence that the assumed causal structure is true in the actual world @ conditionalized on the supposition that action *O* is made true. In other words, EDT involves considering the action you choose to provide evidence regarding the causal state of the world. Since it is unclear that the action you suppose taking should be informative regarding the causal structure of the world, particularly because your action, at least in most

cases, does not effect this structure, this shifting of credence gives clear reason to be suspicious of EDT.[10]

IV. K-CDT AND THE GENERAL IMAGING FUNCTION

Lewis [1981], in preparing for a comparison with Sobel's theory, presents a reformulation of K-CDT in terms of an imaging function.[11] An imaging function assigns to each world-proposition pair (w, X) a credence function $w_X^{\#}(\cdot)$, such that $w_X^{\#}(X) = 1$. A general imaging function works by directing the probability assigned to world *w* by the agent's original credence function to *X*-worlds, with $w_X^{\#}$ being the image of *w* on *X*. The imaging function can be extended to credence functions $CR(\cdot)$ via the definition:

$$CR_X^{\#}(w) = \sum_{u \in W} CR(u)\, u_X^{\#}(w)$$

This credence function can be understood as the result of imaging $CR(\cdot)$ to accept *X*.[12] Given this formulation we can define a general schema for expected utility, in which the expected utility of an option *O* is the weighted value of the possible worlds, with the weights being the agent's credence function after imaging:

$$U(O) = \sum_{w \in W} CR_O^{\#}(w) v(w)$$

Note here that the imaging function $v_X^{\#}(\cdot)$ has so far been underspecified, i.e., I have not stated how the credence originally assigned to *v* be shifted to other worlds, other than that those

[10] I will not be considering here the ongoing debate between decision theorists regarding the correct response to these types of concerns. Instead, I will take for granted that we are interested in developing a causal decision theory. EDT will only reappear in this paper for what I take to be suggestive comparisons. I do not intend to express an opinion regarding how other decision theories may address the questions posed. Still, I conclude with remarks on how these examples can inform our conception of deliberation more generally. These comments I hope can be of interest to those who do not accept the move away from EDT.

[11] Lewis, "Causal Decision Theory.", 15. Here I draw extensively from the more lucid explications in Collins, "Decision Theory After Lewis" and Rabinowicz, "Letters from Long Ago: On Causal Decision Theory and Centered Chances."

[12] Collins, "Decision Theory After Lewis," 4.

worlds be X-worlds. One possible way of specifying the imaging function is to define $u_X^{\#}(\cdot) =_{def} CR(\cdot\,|XK_u)$, where K_u is the dependency hypothesis that obtains in world u. By noting that $CR(K) = \sum_{u\in K} CR(u)$, it becomes clear that K-CDT is a special case of the general schema for expected utility in which the imaging function is defined in terms of conditionalizing on the dependency hypotheses:

$$U_K(O) = \sum_{w\in W}\sum_{u\in W} CR(u)\,u_X^{\#}(w)v(w)$$

$$= \sum_{w\in W}\sum_{u\in W} CR(u)\,CR(w|OK_u)v(w) = \sum_{w\in W}\sum_{K\in\mathcal{K}} CR(K)CR(w|OK)v(w)$$

V. SOBEL'S PROPOSAL: I-CDT

While defining the imaging function in terms of dependency hypotheses results in one definition of expected utility, we can maintain the general expected utility schema but modify the specifics of the imaging function to result in an alternative. Sobel [1978] offers such an alternative by defining an imaging function that is primitive in the sense that it does not rely on the notion of dependency hypotheses. Up until this point, I have been loose as to my use of the label 'imaging'. The concept of an imaging function was first introduced by Lewis [1976] to address the assignment of probabilities to subjunctive conditionals.[13] The imaging function Lewis defines, and Sobel adopts, works in the following way. As previously introduced, $CR_A^{\#}$, is the image of CR on A, i.e., the modifications of the set of credences to accept proposition A, i.e., such that the agent's credence is only spread over A-worlds. The image on A is formed by shifting the credence assigned to each possible world v to the "closest" or "most similar" world (or worlds) to v where A is true. Here 'closest' and 'most similar' are understood to refer to Lewis-style similarity semantics for subjunctive conditionals. The question of what this

[13] Lewis, "Probabilities of Conditionals and Conditional Probabilities."

similarity relation implies will be returned to later in the paper, but for now it is enough to understand that similarity here refers to characteristics of worlds such as past history and causal structure.

It may not be the case that for every world there is a single most similar *A*-world. In the case where there is such a closest world, the entirety of the credence assigned to *v* will be assigned to that world, say, *x*, such that $v_A^{\#}(x) = 1$. These are cases in which imaging is said to be *sharp.* If *X* is some proposition true in *x*, then the sharp imaging would imply the truth of the following subjunctive conditional in world *v*: 'if it were that *A*, then it would be that *X'.* In other cases, more than one world will be most similar, in which case $v_A^{\#}(x) < 1$. Whereas the sharp imaging corresponded to a 'would-conditional', this example of *blurry* imaging implies a 'might-conditional'. Again taking *X* to be some proposition true in *x,* we can say that the blurry imaging implies that the following is true in *v*: 'if it were that *A,* then it might be that *X'.* A natural thing to say in the case of a deterministic, i.e., non-chancy, universe is that such 'would-conditionals', and thus sharp imaging, would hold. This is because in such a universe exact knowledge of the laws of nature and the facts of the world are sufficient to inform you of exactly what will occur in the future. Thus, if, in the context of decision making, you were to consider some counterfactual in which you act differently than you do in @, you can still determine precisely how the world would be different by simply considering the appropriately altered world and evolving it forward in time according to the deterministic laws. This would suggest that 'might-conditionals' be reserved for non-deterministic worlds, i.e., those in which there is genuine chance. In such a world, one cannot know exactly what would happen were some contingent matter of fact be different, since the chancy nature of the laws leaves open multiple possible futures given a specified past.

An objection to the reservation of 'might-conditionals' for cases of genuine chance is that you could believe that even in a deterministic world it may be impossible to determine a single most similar world. For example, consider the counterfactual 'If Bizet and Verdi were compatriots, Bizet would be Italian'.[14] Such a counterfactual appears false, since it seems equally plausible that the closest possible world in which Bizet and Verdi were compatriots is one in which Verdi were French. This would suggest that correct counterfactual is 'If Bizet and Verdi were compatriots, Bizet might be Italian', and that the imaging is blurred between the world in which he is French and the one in which he is Italian. A response to this example is that the indeterminacy regarding which of the worlds is actually the most similar is not a reflection of the fact both world are equally close to the actual world, but rather that we simply do not know enough to determine the true similarity ordering in this case. A sufficiently precise set of similarity criteria along with all the relevant information about the particular worlds under consideration would, in principle, be able to determine—perhaps on the basis of facts about the travel histories of Bizet and Verdi's ancestors—which of the worlds is closest. While this resolution leaves open many issues for philosophy of language, I will consider it sufficient justification for bracketing such concerns in the present case and reserving the use of 'might-conditionals' and blurry imaging for worlds involving genuine chance.[15]

When we can quantify the 'might' with a specific chance value, with $'CH(X) = k'$ meaning 'the chance of X being true is k', then we can say 'if it were that A, then $CH(X) = k$' is true in v iff $v_A^{\#}(x) = k$. A concern here may arise from the time indexing of claims about chance. For example, one may intuitively take the proposition 'the coin lands heads' as having a 50%

[14] Lewis, "Counterfactuals and Comparative Possibility."
[15] For more regarding the hotly debated issue of 'might-conditionals', including a further discussion of the type of resolution I have suggested above, see Harper, "A Sketch of Some Recent Developments in the Theory of Conditionals."

chance of being true prior to flipping the coin but a 100% chance after the coin has been flipped and in fact landed heads. I will take propositions about the likelihoods of outcomes of a chancy mechanism to be evaluated from the perspective of being prior to the activation of the mechanism. As such, '$CH(X) = .5$' is true, when X is the proposition 'the coin lands heads', if the proposition 'the coin has a 50% chance of landing heads' is true prior to the flipping of the coin. I consider the chance of some outcome prior to activation of the chancy mechanism in world *w* a fact of that world throughout its history, including after the mechanism's activation. This implies that propositions like 'the coin has a 50% chance of landing heads' have some notion of a particular time embedded in it. I intend for this account to be understood in such a way that is neutral regarding the truth of such counterfactuals as, 'if you had bet heads, you would have won', in the case of a declined invitation to bet on a coin toss that in fact comes up heads.[16] I do this in order to remain neutral on the issue of centering, which is discussed at the end of this section.

The truth conditions specified for the 'would' and 'might' conditionals above provide the specification of a particular imaging function. This is the imaging function I have been referring to as a genuine imaging function, since it corresponds to the precise movement of credence across worlds in accordance with a similarity function. For the remainder of this paper I will use there term 'imaging function' to refer to this genuine imaging function and reserve the # symbol for only such a function. I distinguish this from the non-genuine imaging function Lewis defines in terms of the dependency hypotheses.

Before moving on, I will add two more definitions that will later be relevant. An imaging function is said to be 'weakly centered' iff for all *v* and *A*, such that *v* is an *A*-world, $v_A^{\#}(v) > 0$,

[16] For a discussion of this type of example, attributed to Sydney Morgenbesser, see Slote, "Time in Counterfactuals."

which is to say that every world is at least as similar to itself as any other world is. An imaging function is 'centered' iff for all *v* and *A*, such that *v* is an *A*-world, $v_A^{\#}(v) = 1$, which is to say that every world is more similar to itself than any other world is. A satisfactory imaging function should meet at least the weaker of these criteria, but it is unclear whether the stronger holds without having a more robust response to certain questions about chance (including those relating to the coin toss referenced above). It should be noted that an imaging function for a deterministic universe, i.e., one in which all imaging is sharp, as justified above, would satisfy centering, but that centering is also compatible with a chancy universe.

VI. Are K-CDT and I-CDT Equivalent?

As is suggested by Lewis' attempt to reformulate K-CDT in terms of an imaging function, Lewis thought his and Sobel's accounts of expected utility to be essentially equivalent.[17] To demonstrate this proposed equivalence, let us say that two worlds *v* and *w image alike* iff, for all the agent's options, on each option *O* the following condition holds: $v_O^{\#} = w_O^{\#}$. Noting that imaging alike creates an equivalence class of worlds, if, for every world *v*, we define a set of worlds which image alike to *v*, then we will have a set of equivalence classes that partition *W*. Lewis then notes that two worlds will image alike iff the same dependency hypothesis obtains in those worlds; thus, the equivalence classes defined by imaging alike are identical to dependency hypotheses. Having noted this identity, Lewis feels justified in asserting the following thesis: $v_A^{\#}(w) = CR(w|AK_v)$. The acceptance of this thesis, which requires the reduction of conditional chances to conditional credences, would make Sobel and Lewis's theories technically equivalent, but as Lewis notes, Sobel is unwilling to accept this extra constraint, i.e., the reduction of chance to credence, on the imaging function. While the

[17] Lewis, "Causal Decision Theory," 17.

constraint seems unproblematic in what Lewis refers to as "ordinary cases," he acknowledges that some "extraordinary cases" would lead to a potential divide between his and Sobel's theory.[18] Before moving on to consider what these extraordinary cases are, we can now note the two constraints underlying Lewis's claim that I-CDT and K-CDT are equivalent. These two constraints are:

(C1) The imaging behavior of all the worlds under consideration can be correctly described by equivalence classes defined by imaging alike.

(C2) Chance can be reduced to credence.[19]

As I will argue, beginning with C2, both of these constraints are violated by certain cases and in such cases only I-CDT gives the right answer.

VII. Extraordinary Cases

The type of extraordinary case that Lewis and Sobel have in mind concerns "an agent who thinks he may somehow have foreknowledge of the outcomes of chance processes."[20] Such scenarios are difficult for K-CDT, because they challenge Lewis's reduction of chance to credence, by leading the agent to form beliefs about the probability of specific outcomes of some chancy process that differ from the chances she assigns to those outcomes' realizations. These cases, Lewis says, are "much more problematic for decision theory than the Newcomb problems."[21] Stating that he has no opinion on them, Lewis is prepared to set them aside and to

[18] Ibid.
[19] Note that this constraint is justified by what Lewis calls "The Principal Principle," which states that a rational agent's credences are to conform to the chances. Specifically, it states that if *E* is some proposition about times up to and including time *t* and *E* entails that the chance of *A* occurring at *t* is *x,* then the agent's credence in *A* at *t* given *E* must be *x.* Lewis refers to any evidence about the world beyond time *t* as 'inadmissible', and restricts the validity of his principle to only those agents that do not accept inadmissible evidence into their belief set. Such inadmissible evidence includes oracle predictions and time-traveling information. As we will see, it is precisely in the case of an agent who takes such evidence seriously that we will have a violation of this constraint. For more on the Principal Principle, see Lewis, "A Subjectivist's Guide to Objective Chance."
[20] Lewis, "Causal Decision Theory.", 18.
[21] Ibid., 18.

assert the equivalence of his and Sobel's theories. Rabinowicz (1982) argues against Lewis' dismissal of such abnormal cases of premonitions, saying that Lewis himself had stated that a decision theory should be applicable to an agent regardless of whether her beliefs are irrational or unfounded.[22] In a letter to Rabinowicz, Lewis agrees, saying he wants his theory to be applicable to imperfectly rational agents and further he believes "in the logical possibility of time travel, precognition, etc."[23] Lewis continues by giving an example of such a problematic decision problem that would divide his and Sobel's theories. The scenario involves a man deciding whether to invest in armor when he knows, because of an oracle, that he is not going to die in battle. Lewis suggests that there are arguments for both buying and not buying the armor, but emphasizes that he does not think "the appeal of not buying the armour is just a misguided revival" of evidential decision theory-type intuitions.[24] Having indicated slight support for the recommendation his own theory would make, Lewis ultimately throws up his hands, suggesting that "maybe the very distinction between rational and irrational conduct presupposes something that fails" in such a case.[25]

While Lewis was cautious in stating his intuitions, Egan [2007] is bolder, offering an oracle scenario structurally similar to the armor scenario referenced above.[26] Egan suggests a case in which an oracle tells you that you will be bitten by a rabid dog. Faced with this fact, he says (1) it would be irrational to try to run away from the dog, since you know you will be bitten anyway. He goes on to argue that (2) such an irrational response, of running, is what K-CDT would recommend. Referring to the oracle and a similar time travel example, he claims "they provide stark examples of cases where CDT endorses performing an action that one confidently

[22] Rabinowicz, "Two Causal Decision Theories: Lewis vs Sobel."
[23] Rabinowicz, "Letters from Long Ago: On Causal Decision Theory and Centered Chances."
[24] Ibid.
[25] Ibid.
[26] Egan, "Some Counterexamples to Causal Decision Theory," 101.

expects will bring about a worse outcome than some alternative."[27] What Egan attempts to do is provide a counterexample to K-CDT by establishing an intuition and then showing that K-CDT recommends something other than what is intuitively rational. In the following sections, I will attempt to do something similar, but by disagreeing with both of Egan's claims. Instead, I will show, by specifying the example in a particular way such that I-CDT and K-CDT come apart, that K-CDT suggests not running, while I-CDT suggests running. I will then argue, in contrast to Egan's intuition, that running is precisely the rational act. In demonstrating these points, I address precisely the type of case of premonition that Lewis sets aside. Further, I will suggest that the point at which Lewis threw up his hands is not, as he suggests, the limit of the boundary between rationality and irrationality, but rather that of his own theory, a limitation that Sobel's theory does not possess.

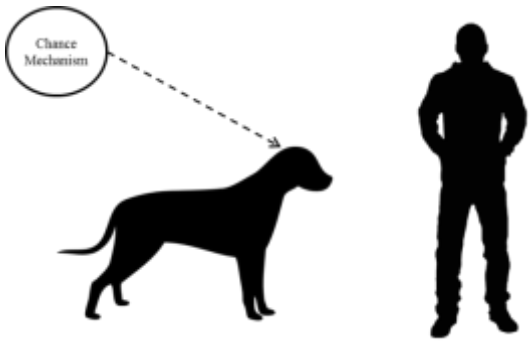
VIII. The Chancy Dog Problem

Imagine there is a ferocious dog standing in front of you that is poised to attack. Looking around, you have no available options but to run or stand still. You are sure that if you do not move, the dog will bite you, but the prospect of trying to outrun it is not particularly appealing either. You assign a utility of -10 to (B) being bitten and 0 to ($\neg B$) not being bitten. You are also not in the best shape and will be exhausted if you run, so you assign a utility of -1 to (R) running and 0 to ($\neg R$) standing still.

Action\Outcomes	B	$\neg B$
R	-11	-1
$\neg R$	-10	0

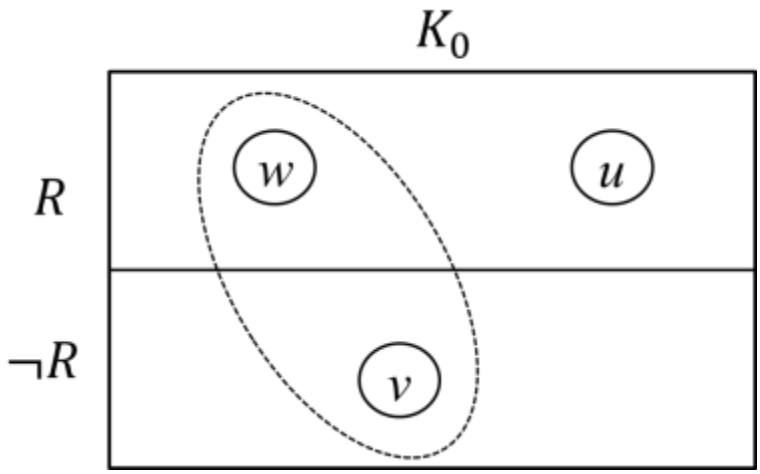
[27] Ibid.

You also believe that this dog's mind is genuinely chancy. If you run, there is a chance of x that he will decide it is not worth chasing after you, but there is a chance $1 - x$ that he will run after you. He is guaranteed to succeed in biting you if he chases after you. He will also definitely bite you if you don't run away.



Given your certainty in this scenario, you assign all you credence to the dependency hypothesis K_0, which is the conjunction of the two conditionals (using Lewis' notation): $R\Box \rightarrow [CH(\neg B) = x]$ and $\neg R\Box \rightarrow B$. Since $CR(K_1) = 1$, CDT will recommend running if $x > 0.1$.[28] Next, for the sake of concrete-ness, let's assume $x = \frac{1}{3}$ and $CR(R) = \frac{1}{2}$. We have the following diagram, Where *w* and *v* are both B-worlds and $CR(w) = \frac{1}{3}, CR(u) = \frac{1}{6}, CR(v) = \frac{1}{2}$:

[28] To see this, we can apply K-CDT. Take *w* to be the $R\&B$-world, *u* to be the $R\&\neg B$-world, and *v* to be the $\neg B\&R$-world.
Expected utility of R: $U_K(R) = \sum_{w \in W} CR(K_0)CR(w|RK_0)v(w) = CR(w)v(w) + CR(u)v(u) = (1 - x)(-11) + x(-1) = 10x - 11$
Expected utility of $\neg R$: $U_K(\neg R) = \sum_{w \in W} CR(K_0)CR(w|\neg RK_0)v(w) = CR(v)v(v) = -10$
K-CDT will recommend running when $U_K(R) > U_K(\neg R)$ i.e., $10x - 11 > -10$, so $x > .01$.



Now imagine an oracle tells you that you will be bitten. Because the oracle is infallible, your credence that you are bitten becomes $\mathrm{CR}'(B) = 1$.[29] Having received the oracle's evidence, you conditionalize on your updated belief by shifting all your credence to B-worlds.[30] You set your credence in u to 0 and normalize your credence over w and v, while holding fix the original ratio of credences. Let $CR'(\cdot)$ refer to your credence after updating on the oracle's evidence. Your credences become $CR'(w) = \frac{2}{5}, CR'(u) = 0, CR'(v) = \frac{3}{5}$.

IX. K-CDT'S RESPONSE TO THE CHANCY DOG PROBLEM

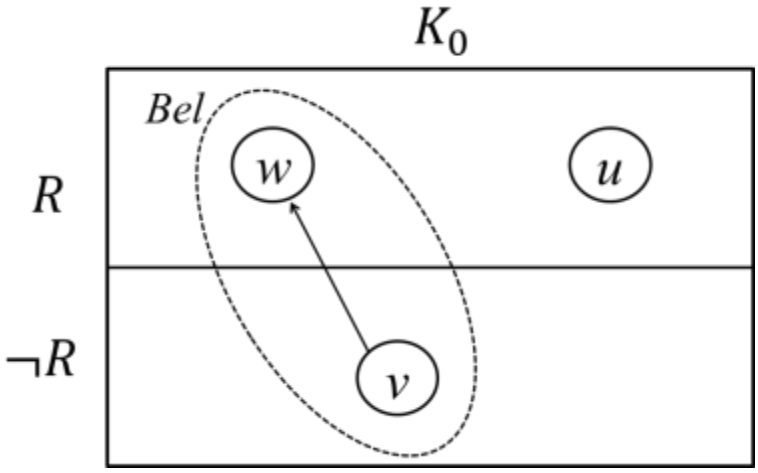
In the Chancy Dog scenario, once we have updated on the evidence from the oracle, the credences in the *R*-worlds are (as stated above) $CR'(w) = \frac{2}{5}, CR'(u) = 0$. When we calculate the credences conditionalized on *R* in the way described above, we do so by distributing our

[29] The fact that the oracle is considered infallible is not essential for the example given here. While the same result holds for the case in which the oracle is just highly reliable, I will assume infallibility for the mathematical and conceptual simplicity it provides.

[30] I make what I take to be the uncontroversial assumption that the correct way to modify one's beliefs given the oracle's prediction is via conditionalization. Conditionalizing is appropriate because the oracle is reporting on the (future) state of the actual world. This is, of course, assuming that one should consider the oracle's prediction as admissible in the first place. While clearly not obviously correct, this is what it means to take seriously the prediction, which is something Lewis suggests a decision theory should be able to do. See note 17.

credence over these two worlds in such a way that the ratios are maintained and $CR'(R|RK_0) = 1$. Therefore, we have $\mathrm{CR}'(w|RK_0) = 1$ and $\mathrm{CR}'(u|RK_0) = 0$. [31]

K-CDT only allows for credence to be shifted to worlds that are within the agent's belief state. This is due to the fact that K-CDT involves conditionalizing within each dependency hypothesis. Conditionalization holds fixed the ratio of probabilities assigned to the worlds that receive positive credence. A world outside the agent's belief state will have zero credence assigned to it initially, and thus will not receive any credence upon conditionalization. Since world *u* has zero credence assigned to it, it is not a live epistemic possibility and receives none of the credence moved over from *v*. This corresponds to the following diagram.



So, we have

$$U'_K(R) = \sum_{w \in W} CR'(K_0)CR'(w|RK_0)v(w)$$

$$= CR'(w|RK_0)v(w) + CR'(u|RK_0)v(u) + CR'(v|RK_0)v(v)$$

$$= (1)(-11) + (0)(-1) + (0)(-10) = -11$$

[31] Note that the ratio is maintained since $\frac{CR'(u)}{CR'(w)} = 0 = \frac{\mathrm{CR}'(u|RK)=0}{\mathrm{CR}'(w|RK)=1}$

$$U'_K(\neg R) = \sum_{w \in W} CR'(K_0)CR'(w|\neg RK_0)v(w)$$

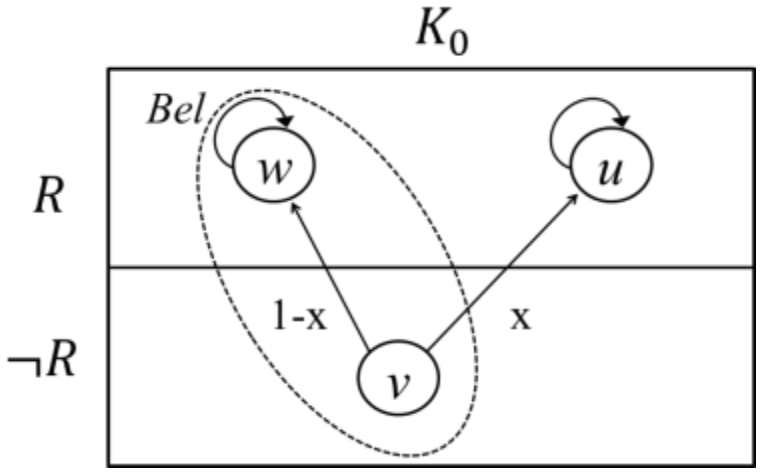
$$= CR'(w|\neg RK_0)v(w) + CR'(u|\neg RK_0)v(u) + CR'(v|\neg RK_0)v(v)$$

$$= (0)(-11) + (0)(-1) + (1)(-10) = -10$$

Therefore, $U'_K(\neg R) > U'_K(R)$ and K-CDT recommends not running.

X. A CENTERED ACCOUNT OF I-CDT'S RESPONSE TO THE CHANCY DOG PROBLEM

Imaging on the options rather than conditionalizing in way described by K-CDT differs in that that it does result in shifting some credence to worlds that are not in the agent's belief state. In particular, imaging on A maps the credence assigned to any world *w* to the closest *A*-world. In the case of a chancy situation described above in which the imaging on R is not sharp for *v* but blurry, imaging on R results in 2/3 of $CR'(v)$, i.e., $\frac{2}{3}CR'(v) = \frac{2}{3}\left(\frac{3}{5}\right) = \frac{2}{5}$, being shifted to *w* and 1/3 of $CR'(v)$, i.e., $\frac{1}{3}CR'(v) = \frac{1}{3}\left(\frac{3}{5}\right) = \frac{1}{5}$, being shifted to *u*. We will for the moment assume centering, which implies that imaging on R does not shift any of the credence already on *w* and *u*, respectively. This corresponds to the following diagram:



Therefore, where $CR'^{\#}_R(w)$ is the credence of *w* (after hearing the oracle) imaged on R,

$$CR'^{\#}_R(w) = CR'(w) + \frac{2}{3}CR'(v) = \frac{2}{5} + \frac{2}{5} = \frac{4}{5}$$

And

$$CR'^{\#}_{R}(u) = CR'(u) + \frac{1}{3}CR'(v) = 0 + \frac{1}{5} = \frac{1}{5}$$

Using these to calculate expected utility

$$U'_I(R) = \sum_{w \in W} CR'(K_1)CR'^{\#}_{R}(w)v(w) = CR'^{\#}_{R}(w)v(w) + CR'^{\#}_{R}(u)V(u) + CR'^{\#}_{R}(v)v(v)$$

$$= \left(\frac{4}{5}\right)(-11) + \left(\frac{1}{5}\right)(-1) + (0)(-10) = -9$$

$$U'_I(\neg R) = \sum_{w \in W} CR'(K_1)CR'^{\#}_{\neg R}(w)v(w) = CR'^{\#}_{\neg R}(w)v(w) + CR'^{\#}_{\neg R}(u)v(u) + CR'^{\#}_{\neg R}(v)v(v)$$

$$= (0)(-11) + (0)(-1) + (1)(-10) = -10$$

Hence $U'_I(R) > U'_I(\neg R)$, so centered I-CDT recommends running, in conflict with Lewis's account.

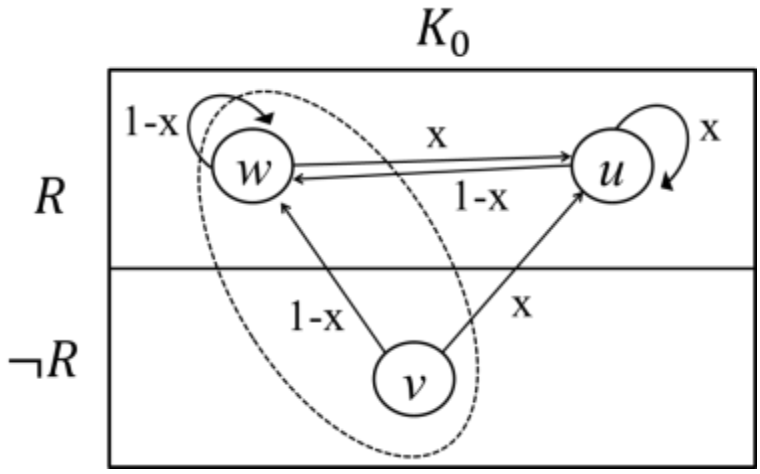
XI. WHY THE CONFLICT?: A LIMITED EXPLANATION

It should come as no surprise that a centered version of centered I-CDT does not give the same recommendation as K-CDT. Recall that Lewis's claim that his and Sobel's theories are equivalent rest on the assumption that worlds *image alike*, i.e., that for all the agent's options A, for any two worlds *v* and *w*, $v^{\#}_{A} = w^{\#}_{A}$. In the case above, however, the imaging alike constraint clearly does not hold. Imaging on option R, $v^{\#}_{R}(w) = 1 - x$ and $w^{\#}_{R}(w) = 1$. This is an example of what Rabinowicz [1982] proved for the general case: K-CDT and centered I-CDT will always be incompatible in universes that are not fully deterministic with respect to the agent's possible action. The reason for this conflict does not have to do with the reduction of chance to credence,

so we will not focus on it here.[32] Instead, let's look at the recommendation made by a weakly centered version of Sobel's theory.

XII. Weakly centered I-CDT's Response to The Chancy Dog Problem

Using a weakly centered imaging function, it is no longer the case that the $w_R^{\#}(w) = 1$. Instead, world *u* is considered just as close to *w* as *w* is to itself and vice versa. Therefore, we have $w_R^{\#}(w) = 1 - x$, $w_R^{\#}(u) = x$, $w_R^{\#}(v) = 0$. Further, it is the case that all the worlds in dependency hypothesis K_0 do in fact image alike. The situation corresponds to the diagram below:



Therefore, where $CR'^{\#}_R(w)$ is the credence of *w* (after hearing the oracle) imaged on R,

$$CR'^{\#}_R(w) = \frac{2}{3}CR'(w) + \frac{2}{3}CR'(v) + \frac{2}{3}CR'(u) = \frac{2}{3}(\frac{2}{5} + \frac{3}{5} + 0) = \frac{2}{3}$$

And

$$CR'^{\#}_R(u) = \frac{1}{3}CR'(u) + \frac{1}{3}CR'(v) + \frac{1}{3}CR'(w) = \frac{1}{3}(0 + \frac{3}{5} + \frac{2}{5}) = \frac{1}{3}$$

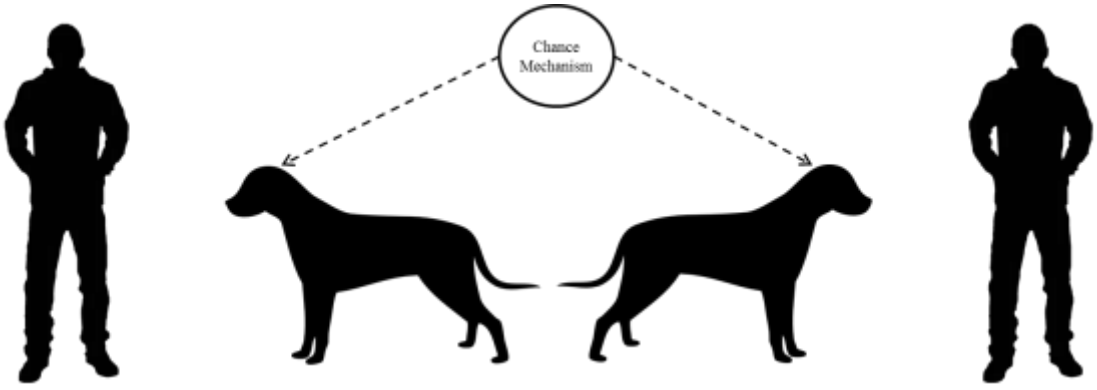
Thus, the credences are the same as they were without the evidence provided by the oracle and the weakly centered version of I-CDT recommends running.

[32] Rabinowicz, "Two Causal Decision Theories: Lewis vs Sobel." I discuss this conflict further in section XXII of this paper. Ultimately, it will become clear that centered, chancy worlds, like all violations of imaging alike, are cases of what I later call 'causal asymmetries'. What this means will be explained in section XXIII.

Returning to the original comment that motivated this discussion, Egan suggests that running after hearing the evidence from the oracle is irrational. While this intuition would support K-CDT, I believe that the true response is the reverse—the rational choice is running. Consider a variant of the problem discussed above.

XIII. TWIN CHANCY DOGS

You and your friend are confronted with a chancy dog and his twin. Their actions are both controlled by the same chance mechanism, and it links their brains together. One of these dogs has focused on you, while the other is concentrating on your friend. Like in the previous example, if you do not move you will be bitten, but if you run, the dog may or may not chase after you. In state 1 (S_1), the dog will chase after you if you run. In state 2 (S_2), the dog will not chase after you if you run. Due to the mechanism linking their brains, the dogs will either both be in S_1 or both in S_2.



You hear the oracle's prediction that you will be bitten. You friend, however, lacks the necessary sixth sense and hears no prediction at all. You cannot communicate with your friend over the barking of the ferocious dogs. Ostensibly, this scenario, from your perspective, is no different from the one chancy dog scenario. Given that your friend does not hear the oracle, you know that she (as a good CDT-user) is going to run. You on the other hand may or may not run depending on whether you use K-CDT or I-CDT. Let's say that you follow K-CDT and do not

run and, of course, the predictor is right—you are bitten. According to Egan, you have made the rational choice, but your friend disagrees. She has run away and escaped unharmed! In this case, it seems that the rational thing would have been to run.[33]

XIV. PLEASE, YOU FIRST: COLLAPSING CHANCE INTO UNCERTAINTY

Thinking more closely about the Twin Chancy Dog case, one notices that there are some counterfactuals which seem to hold in the Twin scenario that do not hold in the case of only one dog. For example, once you see that your friend has run and not been bitten, you can say to yourself "if only I had run, I would not have been bitten!" This is because, in the case that your friend is not bitten, you know that both dogs were in S_2.

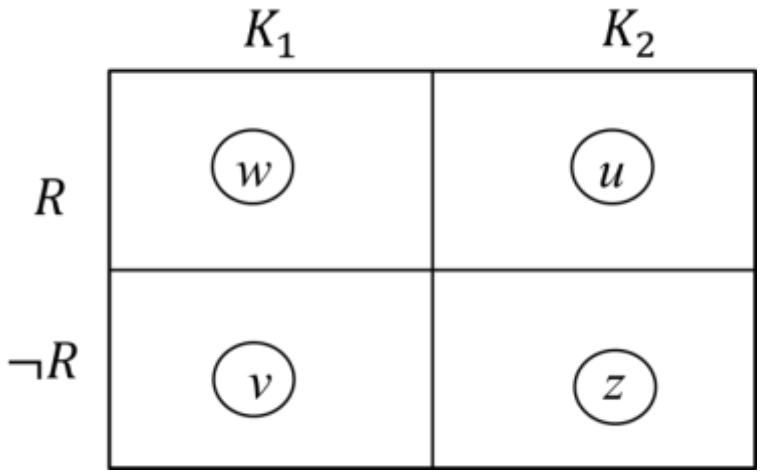
In fact, we can think of the Twin Chancy Dog scenarios as actually containing two different possible scenarios. In one, if you run, you run before your friend, activating the chancy mechanism. In the other scenario, she runs first, either getting bitten or not.[34] In this first situation, which I will call YOU-FIRST, you are faced, at the time of deliberation, with a genuinely chancy situation. Since some chancy mechanism is activated between your point of decision and the consequence (i.e., B or $\neg B$), it is possible for credences and chances to come apart. In the other situation, which I will call FRIEND-FIRST, the friend activates the chancy mechanism prior to your point of decision, so at the time of deliberation the situation is no longer genuinely chancy. Further, while your friend has activated the mechanism, you do not know whether or not the dog has chased after her (since you are too focused on the dog in front of you to look at you

[33] By saying the rational thing would have been to run, I am saying that because it is what is recommended by ignoring the predictor's evidence. If your friend would have been bitten, I *still* think running would have been the rational choice—but explaining this is a separate, yet likely related, question

[34] Your friend will always run, because she does not hear the oracle. As such, you will always find yourself in one of these two situations. In the first, you run before your friend does. In the second, you friend runs before you, either because you were slower to begin running or decided not to run. Note that even if your friend runs first, we will assume you cannot see whether the dog chases after her or not before deciding your action.

friend). As a result, your beliefs about chance will have "collapsed," in some sense, into uncertainty, allowing for Lewis's reduction of chance to credence.

In the FRIEND-FIRST scenario, we thus have an unproblematic case that can be described as a situation of pure uncertainty, rather than chance. In particular, your uncertainty can be described as spread over two dependency hypotheses: K_1, the one that holds when the chance mechanism sends out a signal that (S_1) the dog should chase after the human if s/he runs, i.e., $R\Box \rightarrow B$ and $\neg R\Box \rightarrow B$, and K_2, the one that holds when the mechanisms signals that (S_2) the dog should not give chase, i.e., $R\Box \rightarrow \neg B$ and $\neg R\Box \rightarrow B$. This corresponds to a different diagram of worlds from the one described in relation to the original chancy dog situation, where *w*, *v*, and *z* are B-worlds.



Assuming the same credences as in the original scenario (i.e., that you take either action to be equally likely and that you assign chance of 2/3 to the scenario in which you run and you are bitten (i.e., $CH(K_1) = \frac{2}{3} = CR(K_1)$), then, if the predictor were infallible such that $CR'(\neg B) = 0$, K-CDT would lead to an undefined result. The undefined result is due to the fact that zero credence is assigned to world *u*, which is the only $R\&K_2$-world. Since Lewis's theory utilizes conditionalizing on the conjunction of each option *O* and each dependency hypothesis K,

it requires that, for every *O* and *K*, $CR(OK) > 0$.[35] If, in order to apply K-CDT, we just think the predictor is very likely to be right but not infallible such that $CR'(\neg B) \approx 0$, then K-CDT would recommend running, in contrast to the recommendation in the one dog scenario. I-CDT would also recommend running (regardless of whether the predictor is infallible or not), just as it would in the one dog scenario.

If K-CDT deals with both the YOU-FIRST and FRIEND-FIRST scenarios by reducing chance into credence, why do the recommendations not agree in the two cases? The answer is that, in each case, the evidence provided by the oracle is tacitly interpreted differently. In the FRIEND-FIRST case, your beliefs regarding the outcome of the chancy mechanism and the oracle's prediction come prior to your deliberation. As such, you spread your credence over the possible worlds so that the credence assigned to the dependency hypothesis represents your credence regarding chance. On the other hand, you represent the prediction by concentrating (nearly) all the credence assigned to K_2 in the only $\neg B$-world in that dependency hypothesis, which is an $\neg R$-world. In other words, the prediction is integrated into your credences by raising $\neg R$. The oracle is implicitly interpreted as saying nothing about the outcome of the chancy process, rather it is interpreted as saying, essentially, that if your friend runs and is not bitten, you will not run.

In the YOU-FIRST scenario, the implied interpretation of the oracle's prediction is different. Upon receiving the prediction, the credence assigned to *u* is eliminated, and the credence assigned to the remaining worlds in your belief state, *w* and *v*, is renormalized. In shifting your credence in this way, your belief state only contains those worlds which are contained in K_1 in the FRIEND-FIRST case. This modification of your belief state means that, in the YOU-FIRST scenario, K-CDT implicitly leads you to interpret the oracle's prediction as saying

[35] This restriction is returned to in section XXII.

that the outcome of the chancy mechanism will be such that the dog will chase after you. This implicit interpretation suggests that by reducing chance to credence you lose the ability to be neutral regarding the meaning of the oracle's prediction. Instead, you are forced to interpret it terms of credence regarding dependency hypotheses, your actions, or both.[36] This limitation is a problem for K-CDT, but not I-CDT. As I show in the following sections, the difficulty arises because of the way conditionalizing treats new or supposed information differently from imaging. So far, I have primarily defended the use of imaging by appealing to, in the spirit of the classic Newcomb problem, examples that are challenging for K-CDT. I will now turn to an argument by analogy, relating the type of supposing involved in weighing different options to the type of belief revision that is known to call for imaging. This involves a short detour through formal models of belief revision.

XV. CONTEXT OF BELIEF CHANGE: IMAGING VS CONDITIONALIZING

Katsuno and Mendelzon [1992] convincingly argue that there are two distinct kinds of beliefs change context, each of which requires a different rule. [37] The first context is one in which the agent receives new information about an environment that is, in some sense, static. By static, I mean that the new information received at t_1 is informative about the state of the world at t_0. For example, say that at t_0 I know that my closed lunchbox contains an apple (A), a banana (B), or both (AB).[38] I thus spread my credence about the contents of the lunchbox at t_0 over those three propositions and my belief state contains $\{AB, A\neg B, \neg AB\}$. Excited to know what is in the lunchbox, I open it just a bit, look inside, and, at t_1, see no yellow. Since I know that I

[36] The FRIEND-FIRST scenario could, alternatively, be presented as one in which you interpret the oracle as saying something both about the probability of your actions and the likelihood of particularly dependency hypotheses being true. In this case, you will still choose to run.
[37] Katsuno and Mendelzon, "On the Difference Between Updating a Knowledge Base and Revising It." Note that they use the labels "revising" to refer to "conditionalizing" and "updating" to refer to "imaging.
[38] This example and the one that follows are adapted from Cozic, "Imaging and Sleeping Beauty."

would have seen yellow if a banana were in the lunch box, I can revise my beliefs such that my entire credence is on the proposition that the box contains only the apple. Since I do not believe that anyone tampered with the contents of the lunchbox between t_0 and t_1, I can also use my new information to conclude that at t_0 there was also no banana in the box. The new information allows me to narrow my credence onto a smaller set of worlds (in this case one) from among those in my original belief state. This sort of revision, Katsuno and Mendelzon demonstrate, calls for conditionalization.

Contrast this scenario to one when I have, between t_0 and t_1, given my lunchbox to a friend to hold. This friend is a banana fiend who eats any banana that comes his way and I am certain that if there were any bananas in the lunchbox at t_0, then there are none in there now at t_1. As in the previous case, I know that there are now no bananas in the box, but I cannot alter my belief state in the same way I had previously. First, I cannot draw any conclusions about the content of the lunchbox prior to my friend's intervention, so my beliefs up until the moment of the intervention should remain fixed. Second, I cannot revise my beliefs by simply removing credence from any world in my original belief state in which the lunchbox contained a banana, since doing so would lead me to incorrectly conclude that the box must now contain an apple, when in fact it may be empty. Instead, I must consider each world *w* to which I originally assigned positive credence, imagine altering the world *w* such that any banana is removed, call this altered world *w#*, and shift the credence originally assigned to the world *w* to *w#*. This method of updating corresponds to imaging.

XVI. SUBJUNCTIVE SUPPOSING IS IMAGING

At this point, I call attention to the similarity between the process by which one updates one's beliefs after the banana fiend's intervention and the subjunctive consideration of different

acts involved in deliberation. When I am choosing between different actions, say O_1 and O_2, I imagine, or suppose, that I do one of the actions, consider the expected consequences, and then compare those to what I expect to result when I suppose that I do the other action. What does it mean to suppose that I do some action? Here Lewis' [1979] account of subjunctive conditionals is helpful.[39] To suppose I do O_1, I consider each world *w* in my original belief set, alter that world *w* just before the moment I act such that I do O_1, call this altered world *w#*, and shift the credence originally assigned to the world *w* to *w#.* As should be clear, this shifting of credence is exactly the operation that corresponds to imaging as described above. The only difference is that while in the previous case it was a banana fiend responsible for altering world *w* such that it came to resemble *w#*, in the case of the subjunctive consideration of different actions, some supposed "divergence miracle" is responsible for altering the world.[40] Thus, if one takes deliberating over different option to involve subjunctive supposing of the sort Lewis describes, decision theory requires imaging.

XVII. Invalid Inferences: Moving Credence Too Much

By referencing the above distinction between contexts for conditionalizing and imaging, we can see clearly how K-CDT leads to the wrong recommendation in the Chancy Dog Problem. In the imaging context, we hold the facts of the world fixed until the point of possible intervention, whether actual (as in the case of the banana fiend) or supposed (as in the case of deliberation), at which time we alter the belief state to accommodate the new information. If we were instead to conditionalize on the new information, we would be invalidly modifying the belief state with regard to the world prior to the point of intervention. When the possible

[39] Lewis, "Counterfactual Dependence and Time's Arrow."
[40] The different between *w* to *w#* need not be due to a "divergence miracle" as Lewis describes, but I find the term suggestive. What is essential is that you imagine altering *w* in whatever is the "most appropriate" way, such that it is an O_1-world. Which way is "most appropriate" will depend on one's preferred account of counterfactual dependence.

intervention is an actual one, as in the lunchbox case, what is meant by "the point of possible intervention" is clear. It means the temporal point at which the change is made. Thus the world prior to some intervention at t_1 is simply the world at t_0. In the case of a subjunctive intervention, like those involved in deliberation, the strictly temporal interpretation is not necessary.

Consider, for the moment that the actual world is *v*, which is a $\neg R\&B$-world. Let's say that while in this world, you suppose that you counterfactually run, rather than not run. To reason counterfactually in this way, you begin with world *v* and introduce some intervention such that it becomes a R-world. In this case, the world prior to the point of intervention is *v* and the world that results from the change is the world (or, in this case, worlds) to which *v* images. Given this analogous relationship, if you were to conditionalize, rather than image, on the supposed new information that you run, you would be potentially, invalidly modifying what you take to be true in the world prior to the intervention, i.e., world *v*. Recall that the oracle tells you that in the actual world you will be bitten. More specifically, it should be interpreted as saying that, given what action you will choose, you will be bitten. It does not tell you whether, in the actual world, you would be bitten were you to do some action other than the action you choose to do. If, for the sake of simplicity, we continue to assume that the actual world is *v*, this means that the oracle does not tell you what would happen in world *v* were you to run. Thus, the counterfactuals true at *v* continue to be: $R\Box \rightarrow [CH(\neg B) = x]$ and $\neg R\Box \rightarrow B$. When you subjunctively suppose that you do run in *v*, you imagine some change in the world such that you run. Among the ways the world changes, in this example, is that in this altered world the oracle's prediction that you will be bitten may be false, i.e., you might not be bitten. Such a change would not contradict your belief that the oracle is correct in the actual world and you should not expect the oracle's

infallibility to extend beyond the actual world.[41] Other things about the world should remain fixed. Specifically, the counterfactuals which take as the antecedent the action which you are supposing to do should be the same in both the actual world and the altered world. If these sort of counterfactuals were not the same, then the altered world could not be seriously considered informative regarding what would have happened in the actual world were you to do otherwise. It is this intuition which Lewis attempts to capture through the concept of the dependency hypotheses, since dependency hypotheses ensure that credence is only shifted among worlds where the same subjunctive conditionals hold. The issue, however, is that in the case where credence and chance come apart, conditionalizing may alter the counterfactuals within the dependency hypotheses.

Recall that in the YOU-FIRST scenario, the credence assigned to *u* is eliminated and only *w* and *v* remain within the belief state. Given this epistemic state, when we conditionalize, say, on R, the credence assigned to *v* is all transferred to *w*. If this movement of credence corresponded to a genuine imaging function, it would imply that $v_R^{\#}(w) = 1$. Now recall the truth conditions for the genuine imaging function, which state, if *w* is an B-world, then $v_R^{\#}(w) = 1$ iff, at *v*, $R\Box \rightarrow B$ holds. Thus, conditionalizing leads you to shift credence as though $R\Box \rightarrow B$, rather than $R\Box \rightarrow [CH(\neg B) = x]$, where $1 > x > 0$, is true. If you take *v* to be the actual world, then you can understand conditionalizing as causing you to modify your beliefs about, what I have called, the world prior to the intervention. We can then clearly see the relationship between deliberation and Katsuno and Mendelzon's contexts for conditionalizing vs. imaging.

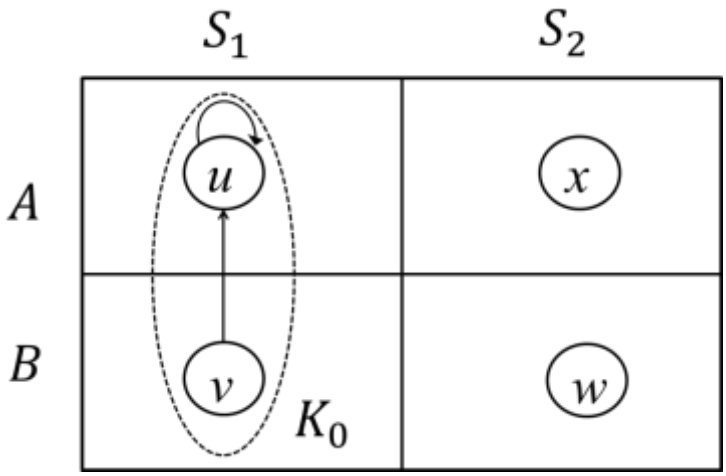
[41] The only facts that you want to maintain as true in the non-actual closest world are those which do not depend on the action you perform in the actual world. The oracle's prediction is correct because of how you act, thus it depends on your action. Note that a proper similarity relation holds fix only those propositions not dependent on the action performed in the actual world, but not necessarily all such propositions.

Conditionalization only allows new information to be inserted into one's belief state if it is compatible with one's initial beliefs. The belief that the oracle's prediction may be false is incompatible with the initial belief that the oracle is correct. Thus, if you were to conditionalize on the supposed new information provided by the subjunctive change in the world, you would have to modify your belief such that even when you run the predictor remains correct and you are bitten. This amounts to believing that, in world *v*, if you were to run, you will be bitten—a proposition whose truth value is neither part of your original beliefs nor supported by the oracle's prediction. While conditionalizing leads you to invalidly modify your beliefs, imaging does not. To put it differently, while imaging leads to "no gratuitous movement of" credence, conditionalizing leads you to shift your credence *too much.*[42]

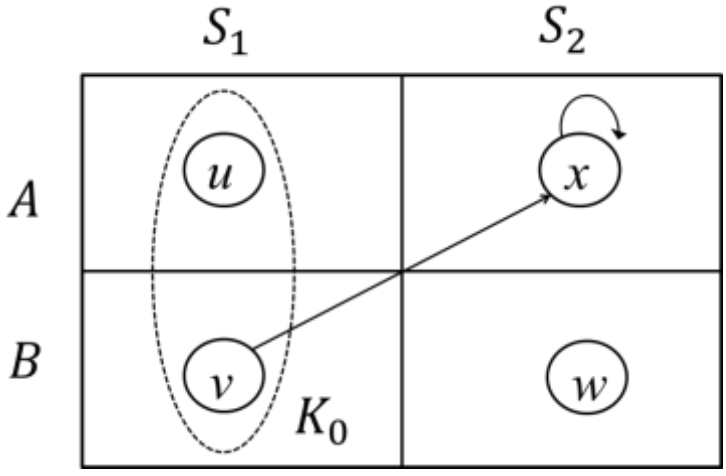
XVIII. Dependency Hypotheses as Channels for Moving Credence

The Chancy Dog scenario reveals that K-CDT fails in the case of a world in which an individual receives evidence about the outcome of a chancy mechanism. In other words, K-CDT fails when constraint C2, that chance can be reduced to credence, is violated. More deeply, we can see that K-CDT's failure is due to the fact that in such cases conditionalizing on the dependency hypotheses fails to move credence in accordance with how credence is shifted by an imaging function. A useful image is to picture dependency hypotheses as channels through which credence flows. When the "channel" created by the dependency hypothesis is correctly shaped, the credence moves just as it would were imaging used rather than conditionalizing, and K-CDT and I-CDT align. This is the scenario illustrated below, where the dotted line represents the dependency hypothesis and the arrows represents the flow of credence as given by the image of *v* and *u* on A.

[42] Lewis, "Probabilities of Conditionals and Conditional Probabilities," 141.



K-CDT and I-CDT come apart when the movement of credence as channeled by the dependency hypotheses ceases to align with how the imaging function would shift the credence between worlds. In the Chancy Dog scenario, such misalignment was seen as the oracle's prediction effectively modified the shape of the dependency hypotheses, causing it to be such that the image of *v* on *R* required shifting some credence not just along the channel of the dependency hypothesis but also "in" and "out" of that channel. A general case of such imaging "out" of the dependency hypotheses is illustrated below.



Every violation of the equivalence of K-CDT and I-CDT, whether it be due to a violation of C2, such as the Chancy Dog problem, or of C1, that the imaging behavior of all the worlds under

consideration can be described by equivalence classes defined by imaging alike, manifests itself as a case in which the dependency hypotheses does not channel the credence in accordance with the imaging function. Through the Chancy Dog problem, I attempted to show that in cases where C2 violations occur, one should favor I-CDT. Such violations can occur, however, only in chancy worlds.[43] In the next example, I will offer a violation of C1, by considering a non-chancy world in which imaging-alike fails to create dependency hypotheses that properly channel credence in accordance with the imaging function. In this case, as will be made clear, one must favor I-CDT. Before examining this example, consider the following unproblematic case.

XIX. A BACKWARDS CAUSATION "NEWCOMB" PROBLEM

Bill finds himself in a situation that is ostensibly similar to the classic Newcomb Problem. He must choose to one-box or two-box by pressing button A or B, respectively, on a remote control attached to a computer. The resemblance to an actual Newcomb Problem, however, is only superficial, because the predictor is a causal mechanism. Pressing A sends a signal back in time that causes the computer to place $1,000,000 in the opaque box (the A-state) while pressing B sends a signal back in time to not put the money in the box (the B-state). Observe that the following pay off matrix holds:

Actions/state	A-State	B-State
A	$1,000,000	$0
B	$1,001,000	$1,000

[43] Note that there are also forms of C2 violation in which the failure arises not because credence must image in or out of a dependency hypothesis, but because conditionalizing does not distribute credence within the dependency hypothesis in the same way that imaging does. This would be the case, for example, when the oracle is not thought of as infallible or simply suggests probabilities of specific outcomes different from those given by "objective" chance.

Given the implied causal structure of the problem as described and the fact that Bill takes this causal structure to be true, we can intuitively say that Bill believes himself to be in a world where the following counterfactual conditionals hold (Note, I will use A to refer to pressing the A button and A_S to refer to the A-state, with parallel notation for B): $A\Box \rightarrow A_S$ and $B\Box \rightarrow B_S$. As discussed earlier, the counterfactual is taken to reflect a semantic of similarity, i.e., $P\Box \rightarrow Q$ is true at actual world @ iff some $P\&Q$-world is more similar to @ than any $P\&\neg Q$-world. Lewis [1979] defends the intuition that the correct similarity relation in most cases is one that reflects agreement over past facts.[44] In particular, he defines a set of similarity criteria, which define similarity such that "some P-world that matches @ over all matters of particular fact *until shortly before* the time of P is closer to @ than any P-world that does not."[45] This similarity relation is tailor-made for cases in which there is no backward causation, but leads to clearly fallacious conclusions when applied to a case like the one considered here.[46]

Given that the usual criterion of holding the past fixed should not apply, we can simply take the counterfactuals stated earlier to be true. Still, for the sake of clarity, we can better justify to ourselves these counterfactuals and the similarity relations they imply by placing ourselves in the position of Bill, setting us up to explore the variant of this scenario I will later suggest. Additionally, I take this to be a worthwhile detour as it illustrates the type of subjective supposing I argue must underlie decision making. Imagine Bill finds himself in the situation described and presses A. He would be in the following world:

World *u*:

---

[44] Lewis, "Counterfactual Dependence and Time's Arrow."
[45] Ahmed, "Causal Decision Theory," 290.
[46] For greater discussion of Lewis' truth-conditions for the counterfactual and related problems, including those relating to backward causation, see Collins, Hall, and Paul, "Counterfactuals and Causation: History, Problems, and Prospects."

Times	t	$t + \Delta t$
Events	A-Signal	Chooses A

What would have been if Bill had instead pressed B at $t + \Delta t$? Given that in world *u* the A-signal is pressed, the usual account of similarity would say that the nearest possible B-world is the following:

World *v*:

Times	t	$t + \Delta t$
Events	A-Signal	Chooses B

While this world is quite favorable to Bill, it also entails a violation of the causal belief that pressing B would bring about a B-state. Since we are interested in a causal decision theory, it is essential that the underlying counterfactuals reflect these causal beliefs. As such, a satisfactory similarity function must be such that the closest B-world is:

World *w*:

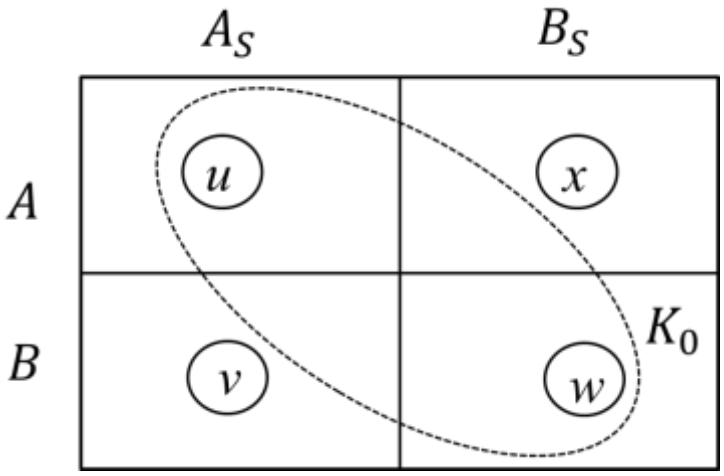
Times	t	$t + \Delta t$
Events	B-Signal	Chooses B

In a similar manner, we could consider what would be if Bill started from a world in which he presses B (i.e., world *w*) and asked what would happen if Bill had instead pressed A. Following a similar logic to the one above, we could think that the closest possible world is the following:

World *x*:

Times	t	$t + \Delta t$
Events	B-Signal	Chooses A

Such a similarity relation, as before, would defy Bill's belief that choosing A would cause an A-signal to be sent. Thus, it follows that the closest world to *w* is *u*. From this analysis, it is clear that worlds *v* and *w* represent the worlds where counterfactuals $A\Box \rightarrow A_S$ and $B\Box \rightarrow B_S$ hold. Thus, we can say that the conjunction of $A\Box \rightarrow A_S$ and $B\Box \rightarrow B_S$ form the dependency hypothesis K_0. Further, given Bill's certainty of the underlying causal model in this situation, it holds that $CR(K_0) = 1$.[47] We can illustrate this simply though the diagram below, where all the credence is concentrated within the region labeled K_0.

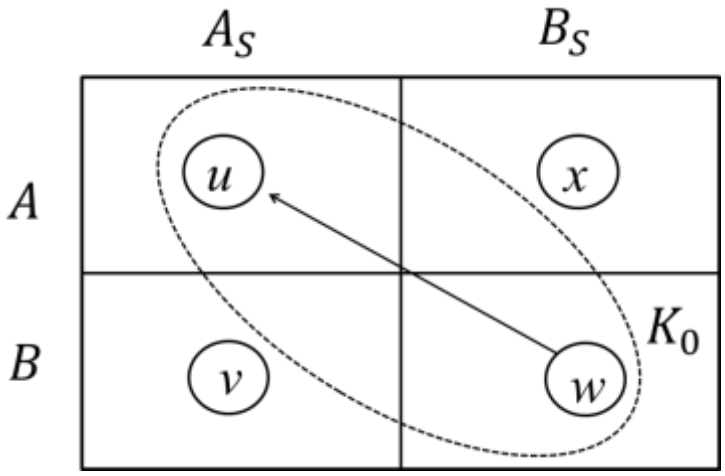


XX. Applying K-CDT to the Backward Causation "Newcomb" Problem

We now have a simple scenario in which we can apply K-CDT. Recall that when applying K-CDT, we multiply the value of each world by the probability of that world conditionalized on the conjunction of the dependency hypothesis and the action under consideration. So, when considering the utility of A, we would transfer all the probability from

[47] I will not now attempt to define the similarity criteria that yields the counterfactuals discussed here, but I will take these counterfactuals to be true and proceed by assuming the similarity relations they imply.

each non-A-world to the A-worlds within the relevant dependency hypothesis, holding fixed the original ratio of credences of those A-worlds. In this scenario, such belief revision corresponds to shifting the credences as depicted in the diagram below.



Therefore, to calculate the expected utility of A:

$$U_K(A) = \sum_{w \in W} CR'(K_0)CR'(w|AK_0)v(w) = (1)(1{,}000{,}000) + (0)(1000) = 1{,}000{,}000$$

Similarly, calculating the expected utility of B yields:

$$U_K(B) = \sum_{w \in W} CR'(K_0)CR'(w|BK_0)v(w) = (0)(1{,}000{,}000) + (1)(1000) = 1000$$

Thus, $U_K(A) > U_K(B)$ and K-CDT recommends pressing A. This recommendation is intuitively correct as pressing A causally promotes Bill receiving the million. K-CDT supplies the correct response.

XXI. The Faulty Signal Problem

Now let's consider a variant of the problem described above. Everything is the same except the computer has developed a peculiar glitch. Nothing is different if the A-signal is received. However, if a B-signal is received (initiating the B-state), the remote will short circuit causing both the A and B button to send a B-signal back in time. The situation is logically

possible. We should thus expect to be able to apply a proper similarity relation over possible worlds to determine the relevant counterfactuals. As in the previous case, the Lewisian criteria will not provide a satisfactory set of truth conditions for the counterfactuals in this case. However, we can explore the underlying causal model given by the description of the problem above to tease out what counterfactuals would be entailed by a proper semantic. Since the situation is the same when the A button is pressed, it would still be the case that, if Bill presses A, he would find himself in world *u*.

World *u*:

Times	t	$t + \Delta t$
Events	A-Signal	Chooses A

Similarly, if he were to consider what world he would find himself in if he were, instead, to have chosen to press B, he would be in world *w*. Now, however, things are a bit different, as in this *w* world, the B-signal short circuits the remote.

World *w*:

Times	t	$t + \Delta t$
Events	B-Signal (short circuiting the remote)	Chooses B

Now suppose that Bill is in world *w*. He could consider what would happen if he had pressed A at $t + \Delta t$. Given that the remote has been short-circuited, it seems that if he had pressed A, a B-signal would still be sent back in time. Therefore, he would find himself in the following world:

World *x*:

Times	t	$t + \Delta t$
Events	B-Signal (short circuiting the remote)	Chooses A

It can also be seen, without difficulty, that if Bill were to begin in world *x* and consider what world he would have found himself in had he done B, the world in question would be *w*. We can use the analysis described above to determine the counterfactuals that hold at each world:

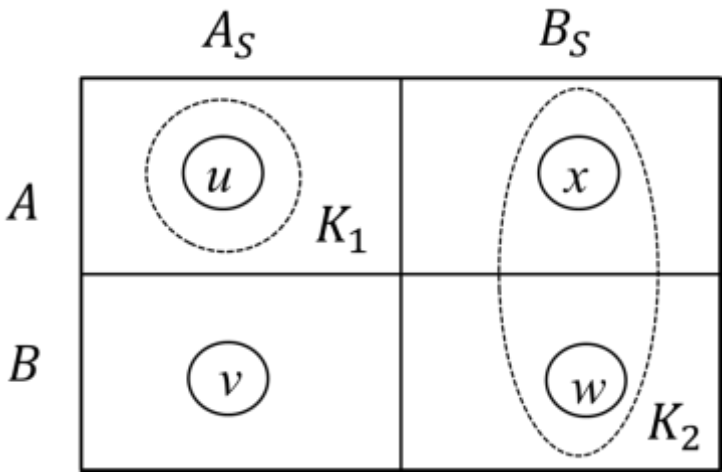
At *u*: $A\Box \rightarrow A_S$ and $B\Box \rightarrow B_S$

At *w*: $A\Box \rightarrow B_S$ and $B\Box \rightarrow B_S$

At *x*: $A\Box \rightarrow B_S$ and $B\Box \rightarrow B_S$

All of these worlds are possible given the story described and thus Bill will assign positive credence to each of them. The imaging behavior of all the worlds under consideration can be correctly described by equivalence classes defined by imaging alike. The world *v* is a world in which he presses B and an A-state resulted. In this world, $B\Box \rightarrow A_S$ holds, but this is a counterfactual to which Bill, who takes the causal story told in the description of the scenario to be true with certainty, takes to be false. Thus he assigns zero credence to world *v* and concentrates all his credence on worlds *w*, *x*, and *u*.[48] We can use the counterfactuals above to define the dependency hypotheses to which Bill assigns positive credences. One is K_1, the conjunction of $A\Box \rightarrow A_S$ and $B\Box \rightarrow B_S$. K_1 holds at *u* and only *u*. The other is K_2, the conjunction of $A\Box \rightarrow B_S$ and $B\Box \rightarrow B_S$, which holds at both *x* and *w*. This is displayed in the diagram below.

[48] He need not assign zero credence to world *v*, but, since $B\Box \rightarrow A_S$ is only true in *v*, it will lie within its own dependency hypothesis. This dependency hypothesis will be assigned zero or near zero credence.



XXII. DOES NOT COMPUTE: K-CDT FAILS

As mentioned earlier, Lewis' theory requires that, for every O and K, $CR(OK) > 0$. Lewis justifies this restriction by stating "Absolute certainty is tantamount to firm resolve never to change your mind no matter what, and that is objectionable" and goes on to argue that such resolve is irrational.[49] Rabinowicz [1982] points out the flaw in this argument:

> "I would be glad to accept this argument if not for one thing: there are cases when the assignment of zero credence to [O&K] is not dictated by rashness but by necessity. This happens whenever [O&K] is empty."[50]

In other words, Lewis's theory requires the assumption that no $O\&K = \emptyset$. Rabinowicz goes on to show that this assumption implies that any centered world is deterministic (i.e., non-chancy), thus implying that one cannot take K-CDT to be correct if one holds that worlds can be both chancy and centered.[51] The Faulty signal problem, however, shows that

[49] Lewis, "Causal Decision Theory," 14.
[50] Rabinowicz, "Two Causal Decision Theories: Lewis vs Sobel," 312.
[51] The proof from Rabinowicz [1982] is as follows: suppose that worlds are genuinely chancy, and the similarity relation over such worlds is centered. If $O\&K_w = \emptyset$, where K_w is the dependency hypothesis true in world *w*, then there must be some world *v* such that *v* is a O-world and K_w is true at *v*. Given that the similarity relation is centered, $v_O^{\#}(v) = 1$. Since *v* is within the dependency hypothesis K_w, it follows, from imaging alike, that $w_O^{\#}(v) = v_O^{\#}(v)$. So, $w_O^{\#}(v) = 1$. Thus, for any world *v* within dependency hypothesis K_w, $w_O^{\#}(v) = 1$ and for any world *u* outside of K_w, $w_O^{\#}(u) = 0$. In other words, centering implies that the worlds are deterministic, i.e., non-chancy. As will be

Rabinowicz's conclusion does not go far enough. Even if one denies that chance is centered, one cannot take K-CDT to be correct as long as one takes it to be the case that a theory of rational decision should apply to cases of backwards causation. The failure of the theory is clearly seen when one attempts to determine the expected utility of pressing B.

$$U_K(B) = \sum_{w \in W} \sum_{K \in \mathcal{K}} CR(K)CR(w|BK)v(w)$$

$$= C(K_1)CR(B\&K_1)v(B\&K_1) + C(K_2)CR(B\&K_2)v(B\&K_2)$$

Since $B\&K_1 = \emptyset$, $V(B\&K_1) = CR(B\&K_1)v(B\&K_1)$ is undefined and the expression is ill-formed. K-CDT thus fails to provide any recommendation.

XXIII. Counterfactual Asymmetry: C1 Does Not Hold

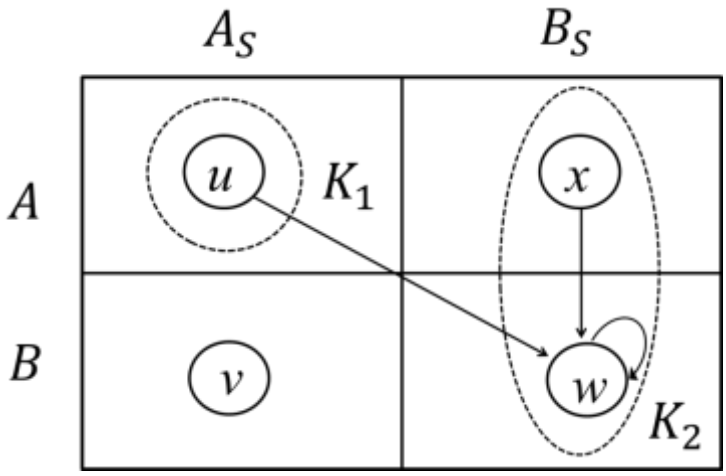
The Faulty Signal Problem is just one example of a class of decision problems to which K-CDT cannot be applied. These are scenarios that exhibit what I refer to as 'counterfactual asymmetries'. Take a decision scenario to be described by an imaging model, which is a collection of worlds and a similarity relation that describes the relative closeness of each world to the other. Take *v* to be an O-world within this imaging model and take *w* to be the closest $\neg O$-world to *v.* Such decision scenario would exhibit a counterfactual asymmetry if it were the case that the set of counterfactuals true at *w*, call this K_w, is not the same as that true at *v.* Since the set of counterfactuals true at a world *w* make up the dependency hypothesis true there, K_w, we can alternatively say that 'counterfactual asymmetry' amounts to having an imaging function such that the image of world *w* lies in dependency hypothesis other than K_w. We can interpret this condition as being applied to scenarios in which it is not the case that every option is actually a

---

made clear in the next section, this problem for K-CDT is, in fact, just a specific example of the types of cases K-CDT cannot address, i.e., those in which imaging alike fails.

live epistemic possibility in every dependency hypothesis, even though every option is truly available in each world.[52]

The problem arises from the first constraint (C1) Lewis imposes on the imaging models in order to define equivalence between K-CDT and I-CDT. Recall that C1 requires that worlds, within a dependency hypothesis, image alike, i.e., for all the agent's options A, for any two worlds *v* and *w*, $v_A^{\#} = w_A^{\#}$. In the case described here, imaging alike, yields two equivalence classes of worlds $\{u\}$ and $\{w,x\}$. Taking these two equivalence classes as our dependency hypotheses, we have the same K_1 and K_2 as before, which led to an ill-formed expression for calculating the expected utility using Lewis' method. These dependency hypotheses partition *u* and *x*, severing the flow of credence between these worlds. The dependency hypotheses thus fail to channel credence in the manner prescribed by imaging. This is illustrated in the diagram below, where the arrows represent the shifting of credence as required by imaging.

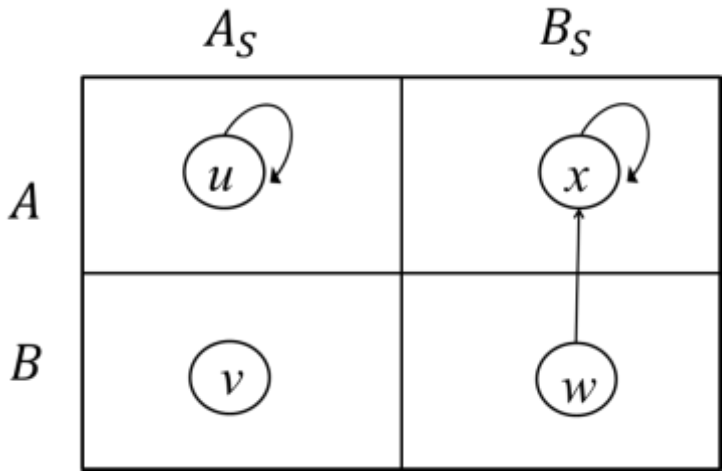


XXIV. I-CDT APPLIED: PRESS A

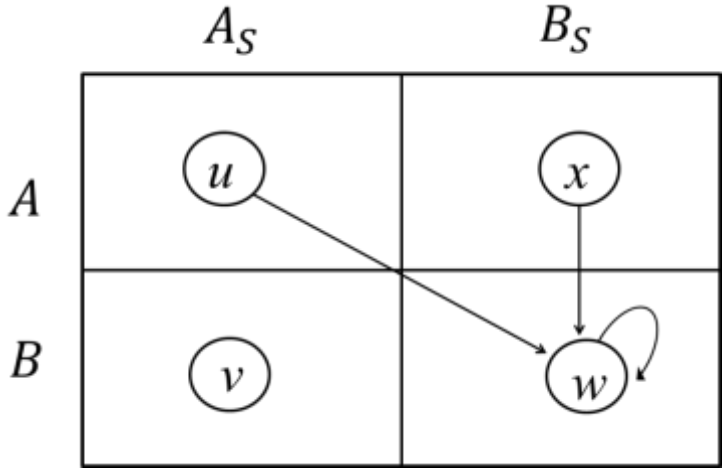
K-CDT's difficulty with cases of counterfactual asymmetry displays a serious weakness with the theory. Fortunately, it is not an insurmountable problem for CDT more generally.

[52] We can understand 'available' in this context to mean that if the agent had the desire to choose the option, she would be able to do so.

Sobel's I-CDT avoids the fault described above, since it does not require that Lewis' C1 hold. To see how it would deal with such a case, consider the diagram below, which depicts imaging on A.



Let's say that $\mathrm{CR}(u) = p, CR(x) = q, CR(w) = l$, and $CR(v) = 0$.[53] Further note, $p + q + l = 1$. Imaging on A yields ${CR'}^{\#}_{A}(u) = p$, ${CR'}^{\#}_{A}(x) = q + l$, ${CR'}^{\#}_{A}(w) = 0$. Similarly, we can image on B, which yields ${CR'}^{\#}_{B}(u) = 0$, ${CR'}^{\#}_{B}(x) = 0$, ${CR'}^{\#}_{B}(w) = 1$. These imaging relations are depicted below.



[53] Note that in the diagram above (and in the one to follow), I am unconcerned with world *v*, since it is assigned zero credence. It receives zero credences because it is, in a sense, a nomological impossibility. If we wanted to assign it some positive credence, we could alternatively say $C(v) \approx 0$ with no relevant change.

Using these to calculate expected utility

$$U_I(A) = (p)(1{,}000{,}000) + (l + q)(0) + 1{,}000(0) = p1{,}000{,}000$$

$$U_I(B) = 0 * 1{,}000{,}000 + 0 * 0 + 1{,}000 * 1 = 1{,}000$$

Thus $U_I(A) > U_I(B)$ when $p > \frac{1}{1000}$. In words, this means that Sobel's theory recommends pressing A unless the agent is nearly certain that the computer is in a B-state. This recommendation, I think, aligns with intuition and is correct. Thus, Sobel's theory provides a correct recommendation where Lewis's theory provided none at all, giving reason to consider I-CDT the superior proposal.

XXV. K-CDT AND THE INDICATIVE MODE

On a technical level, it is very easy to see why I-CDT escapes the limitation faced by K-CDT in this example. Imaging works by shifting credence from a world *w* to the nearest possible world *w#*, whatever that world may be. Conditionalizing also shifts credence to the nearest world, but only if all the counterfactuals true in *w#* are the same as those true in *w*. In the context of the Faulty Signal Problem, this restriction seems arbitrary and indefensible; but it falls out naturally from the picture of K-CDT when compared to EDT. When assessing your options from the perspective of EDT, you are primarily concerned with resolving uncertainty. You have a set of possible worlds and your credence is spread over them. In most cases of EDT, you imagine receiving news of your action, which narrows down the set of epistemically possible worlds; and then you calculate expected value. The action is taken as informative about this world, helping you determine which of the possible worlds is the actual one and thus better determine its expected value to you. By modifying your beliefs as though the action is informative about the actual world, you insert into your belief state the proposition regarding your potential action in

what Joyce [1999] refers to as the *indicative* or *matter-of-fact mode.*[54] In this mode, you accept the evidence as true in a way that maintains the truth-value of all propositions regarding which you are certain. For example, if you are certain that the set of counterfactuals K_1 is true in the actual world, then the evidence provided by inserting the proposition that you choose option O_1 will cause you to focus all your credence on $K_1 \& O_1$-worlds.

As discussed earlier (section III), EDT and K-CDT are formally very similar, differing only in the fact that K-CDT screens off any evidence provided by the action regarding which set of counterfactuals is true. Still, K-CDT involves revising your belief state in the indicative mode, beginning with a set of epistemically possible worlds and ruling out those that are not compatible with the new evidence. To screen off the evidence provided by option O_i about which set of counterfactuals, i.e., which dependency hypothesis, is true, when applying K-CDT, you essentially do the following: (1) for each dependency hypothesis K_j, you assume that K_j is true in the actual world and revise in the indicative mode by moving the credence to only $O_i \& K_j$-worlds, (2) calculate expected value assuming K_j to be true, and then (3) average these values over all the dependency hypotheses according to their prior weights. Utilizing this process, K-CDT is incompatible with the possibility of shifting credence from one world to another where different counterfactuals hold.

By utilizing the indicative mode, K-CDT assumes that you conceive of decision-making as a process of reducing uncertainty about the actual world. You know some things about the actual word are true and you deliberate by considering which action reduces the remaining uncertainty in the most favorable way. This sort of deliberation, of course, requires uncertainty. If you are absolutely certain which action you are going to take, then you cannot consider what

[54] Joyce, *The Foundations of Causal Decision Theory*, 181.

the outcome would be were you to do something different. Lewis responds by saying that you should never be certain. While we might consider this satisfactory, it will not work when assessing decisions made, for example, in the past from your present knowledge base. To this apparent problem, one could respond that it is simply nonsense to try to deliberate in the past, and I concede that that is probably true. Still, I could want to, knowing what I know now, look back at a past decision and consider how things could have been had I done otherwise. While one may simply write this off as something different from decision making and thus not within CDT's purview, it is clearly very similar to a genuine decision problem—particularly one like those considered here, i.e., oracle and time-travel cases. The indicative mode cannot deal with these cases—but the subjunctive mode can.

XXVI. Deliberation as a Subjunctive Activity

The indicative mode is concerned with how things are. For example, consider the question: 'if Shakespeare did not write *Hamlet*, was it written?'[55] If I asked this in the indicative mode, I would be asking the rather uninteresting question of if, in this world, it was discovered that not Shakespeare, but rather some other writer—perhaps Christopher Marlowe, after all—was the true author, was *Hamlet* written? The answer to this question is obviously *yes*. If, on the other hand, I asked this in the subjunctive mode—which would be more properly phrased, 'if Shakespeare had not written *Hamlet*, would it have been written?'—then I would be asking you to imagine some intervention in this world's past such that Shakespeare never writes his Danish tragedy.[56] To answer my question, you would have to consider whether *Hamlet* is ever written in this counterfactual world—so the response is probably *no*. As has been already discussed

[55] Ibid., 182. The example can be originally found in Bennet, Jonathan, 1988, "Exposition to the Phlogistron Theory of Conditionals.

[56] What this intervention should be is a point that would be clarified by a similarity relation. I suspect the correct similarity relation would be one that reorganizes the world in a way that maintains the causal structure of the actual world and is informed by structural equation models.

(section XVI), imaging corresponds to belief revision in the subjunctive mode. The examples explored here make a strong case that the correct method of belief revision for decision making is imaging and, thus, decisions are to be made in the subjunctive mode.

By recognizing that deliberation is a subjunctive activity, we can develop a much clearer picture of what it involves. The standard description of deliberation imagines the decision-maker as conceiving of a host of possible world, each associated with an available choice of action. EDT and K-CDT assumes the decision-maker is in a state of epistemic uncertainty with regard to these possibilities and uses her choice of action to zero in on a single world. With I-CDT, the decision-maker is again presented with a set of worlds. Among these worlds is the actual world, though she may or may not know which it is. For each available option, there is a world, reachable from the actual world, in which that option is realized. These worlds are reachable in the sense that they are made true by introducing some intervention into the actual world. While she does not consider herself able to cause these interventions, she retains a genuine ability to freely choose any of her available options—including those other than the one she chooses.[57] The decision-maker regards her different options as possibilities in a metaphysical, rather than epistemic sense. In fact, epistemic uncertainty plays a much more secondary role when deliberating in the subjunctive mode. Thus, taking deliberation to be subjunctive activity

[57] How this ability is retained is explained by Lewis [1981]: The intervention changes either the laws or the past history in a way that retains your causal beliefs and leads your action to be other that what is was in the actual world. Suppose that determinism is true and I have just put my hand down on my desk and have refrained from raising it. I want to say that it was a free, but predetermined act. By this I mean I could have done otherwise, by raising my hand. In other words, raising my hand was an option. The fact that I did not raise my hand is entailed jointly by the distant past and the laws of nature. Therefore, if I had raised my hand, either the past or the laws would have had to have been different. This, however, does not mean that by raising my hand I would have caused either the past or the laws to be different, since that would be an obvious impossibility. Rather, something else would have caused one of the two to be different. Therefore, if I had raised my hand, either the past or the laws would have had to been different, but I would not have caused them to be different. This illustrates that the ability to do otherwise—in this case raise my hand, when I in fact did not—does not entail that I cause either the past or the law to be different. Lewis, "Are We Free to Break the Laws?"

involves considering one's choices not simply as evidence-makers, as EDT and K-CDT do, but as genuine causes. This is what we should expect from a causal decision theory.

XXVII. Conclusion

Given your beliefs and your desires, a causal decision theory is supposed to recommend the rational choice of action. Lewis' K-CDT purports to be such a theory, but, as I have shown, it fails in comparison to a version of Sobel's I-CDT. Criticizing a decision theory on the basis of failing to recommend the most rational choice is not an uncontroversial exercise, since, to an extent, the rational choice is constituted by what would be recommended by an ideal decision theory. In the Chancy Dog Problem offered here, I have appealed to intuitions to criticize the recommendation of K-CDT, in favor of that offered by I-CDT. Specifically, I appealed to the intuition that the oracle's recommendation can only be treated as true in the actual world rather than in the counterfactual world. For this reason, you have good reason to believe that if you follow K-CDT's advice and not run, you will be giving up a genuine opportunity to escape the dog's bite. A self-fulfilled prophecy offers no assurance of inevitability. As is suggested by the continuing debate between one-boxer and two-boxers, the appeal to intuition is unlikely to provide the final word. Less controversial is the Faulty Signal Problem, which renders a clear verdict. Basing a decision theory on conditionalizing, as K-CDT does, severely limits the theory's application. While I-CDT is able to recommend an intuitively plausible action in this difficult example, K-CDT is forced to remain silent.

Both of the examples offered here are highly unusual, surely counting among what Lewis would call the "extraordinary cases." For this reason, one may be tempted, like Lewis himself, to set them aside, focusing instead on the fact that K-CDT appears to provide the

correct recommendation in the type of cases we actually encounter in our daily lives. A stopped watch may only be right twice a day, but that is enough if those are the only times at which you ever glance at your watch. Clearly the watch analogy is hyperbolic and silly, while Lewis' response is legitimate; but there is something suggestive there. A clock does not just point to a number; it represents the passage of time.

A decision theory is more than a formal tool. It is also a representation of what is involved in deliberation. As I have discussed, the Chancy Dog and Faulty Signal Problems help clarify how one ought to think about the process of weighing different options. What they show is that deliberation is a subjunctive activity, requiring one to evaluate options by moving between different, sometimes epistemologically incompatible world. By referencing the Katsuno-Mendelzon account of belief change, I suggest that imaging is the mode of belief revision that captures this subjunctive mode. These examples demonstrate the subjunctive character of decision making, because of the need to utilize imaging as the method of belief revision in order to arrive at the right recommendation. I-CDT, which utilizes imaging, correctly represents the decision-making process, making it the ideal candidate for a decision theory. In contrast, K-CDT is locked into the indicative mode, painting a misleading picture of deliberation that is bound to the actual world and inapplicable to the types of cases explored here. Perhaps the one-boxers and two-boxer will never agree, but for those interested in a genuinely causal version of decision theory, I-CDT presents the most promise.

BIBLIOGRAPHY

Ahmed, Arif. "Causal Decision Theory: A Counterexample." *Philosophical Review* 122, no. 2 (April 1, 2013): 289–306. doi:10.1215/00318108-1963725.

Collins, John. "Decision Theory After Lewis," Forthcoming.

Collins, John, Ned Hall, and L.A. Paul. "Counterfactuals and Causation: History, Problems, and Prospects." In *Counterfactuals and Causation*. MIT Press, 2004.

Cozic, Mikal. "Imaging and Sleeping Beauty: A Case for Double-Halfers." In *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge*, 112–17. TARK '07. New York, NY, USA: ACM, 2007. doi:10.1145/1324249.1324266.

Egan, Andy. "Some Counterexamples to Causal Decision Theory." *The Philosophical Review* 116, no. 1 (January 1, 2007): 93–114.

Harper, William L. "A Sketch of Some Recent Developments in the Theory of Conditionals." In *IFS*, edited by William L. Harper, Robert Stalnaker, and Glenn Pearce, 3–38. The University of Western Ontario Series in Philosophy of Science 15. Springer Netherlands, 1980. http://link.springer.com/chapter/10.1007/978-94-009-9117-0_1.

Joyce, James M. *The Foundations of Causal Decision Theory*. Cambridge University Press, 1999.

Katsuno, H., and A. O. Mendelzon. "On the Difference Between Updating a Knowledge Base and Revising It." In *Belief Revision*, 183–203. Cambridge University Press, 1992.

Lewis, David. "Are We Free to Break the Laws?" *Theoria* 47, no. 3 (1981): 113–21.

———. "A Subjectivist's Guide to Objective Chance." In *IFS*, edited by William L. Harper, Robert Stalnaker, and Glenn Pearce, 267–97. The University of Western Ontario Series in Philosophy of Science 15. Springer Netherlands, 1980. http://link.springer.com/chapter/10.1007/978-94-009-9117-0_14.

———. "Causal Decision Theory." *Australasian Journal of Philosophy* 59, no. 1 (March 1, 1981): 5–30. doi:10.1080/00048408112340011.

———. "Counterfactual Dependence and Time's Arrow." *Noûs* 13, no. 4 (November 1, 1979): 455–76. doi:10.2307/2215339.

———. "Counterfactuals and Comparative Possibility." In *IFS*, edited by William L. Harper, Robert Stalnaker, and Glenn Pearce, 57–85. The University of Western Ontario Series in Philosophy of Science 15. Springer Netherlands, 1973. http://link.springer.com/chapter/10.1007/978-94-009-9117-0_3.

———. "Probabilities of Conditionals and Conditional Probabilities." In *IFS*, edited by William L. Harper, Robert Stalnaker, and Glenn Pearce, 129–47. The University of Western

Ontario Series in Philosophy of Science 15. Springer Netherlands, 1976. http://link.springer.com/chapter/10.1007/978-94-009-9117-0_6.

Rabinowicz, Wlodek. “Letters from Long Ago: On Causal Decision Theory and Centered Chances.” In *Logic, Ethics, and All That Jazz - Essays in Honour of Jordan Howard Sobel*, edited by Lars-Göran Johansson, Jan Österberg, and Rysiek Sliwinski, 56:247–73. Uppsala Philosophical Studies, 2009. http://lup.lub.lu.se/record/1458836.

———. “Two Causal Decision Theories: Lewis vs Sobel.” edited by T. Pauli. Uppsala, 1982.

Slote, Michael A. “Time in Counterfactuals.” *The Philosophical Review* 87, no. 1 (January 1, 1978): 3–27. doi:10.2307/2184345.

Sobel, Jordan Howard. “Notes on Decision Theory: Old Wine in New Bottles.” *Australasian Journal of Philosophy* 64, no. 4 (December 1, 1986): 407–37. doi:10.1080/00048408612342621.