Newcombian Tragedy

Gerard J. Rothfus

October 2021

Abstract

Philosophers' two favorite accounts of rational choice, Evidential Decision Theory (EDT) and Causal Decision Theory (CDT), each face a number of serious objections. Especially troubling is the recent charge that each of these theories is dynamically inconsistent. I argue here that, under the epistemic assumptions that validate these charges, every plausible decision theory is doomed to a similar fate. I suggest we take this as a reason to challenge these epistemic assumptions and instead embrace some fairly radical restrictions on the set of credence functions a deliberating agent may rationally adopt. I consider two such restrictions, *Full Autonomy* and its weaker cousin *Backward Autonomy*, proving that the former suffices to render CDT dynamically consistent, while the latter does the same for EDT.

1 Introduction

The choices we make often provide us with information about the world. My decision to turn the spigot valve is evidence of imminent water flow. Opting to press my neighbor's doorbell suggests that she may come and open the door. A flip of the light switch on my wall portends a change in ambient lighting conditions. Etc. In cases like these, the evidential bearing of my decision upon the relevant state of the world is underwritten by a causal connection. Turning the spigot valve is evidence that water will flow because turning the valve tends to cause water to flow. And so on. This is the typical case and generates no paradox.

More puzzling are those cases in which our choices seem to provide evidence for states of affairs that they have no tendency to cause. Game theory supplies a familiar example. My friend and I are to play a Prisoner's Dilemma. I take my practical reasoning to be similar to her's and hence my behavior to be indicative of her's. My choice to cooperate (defect) is then evidence of my friend's choice to cooperate (defect), even though our respective choices are causally isolated from one another. Here, analogical, rather than causal, reasoning seems to motivate an evidential connection between my choice and

my friend's.[1]

Natural as such reasoning may seem, the difficulties it causes for the theory of rational choice are profound. When the evidential and causal import of an agent's choices come apart, is an option's choiceworthiness to be fixed by its evidential 'news value' or by its causal efficacy in promoting good outcomes?[2] Or perhaps by some further quantity?[3] No answer so far ventured is without its drawbacks and, a vigorous 50 year debate notwithstanding, consensus has yet to emerge on which is the least unpalatable.

One particularly disturbing feature of both the evidentialist and causalist proposals is that they leave agents liable to *dynamic inconsistency* in certain well-crafted sequential choice problems. The plans that such theories recommend as optimal *ex ante* in these problems can diverge from those they license an agent to actually carry out, step by step. While some of the extant alternative proposals evade this problem, they do so only at enormous cost (e.g. by denying that rational choice under certainty always goes by utility maximization). This is no accident, as I argue below. Decision theory faces a deep problem regarding how to go about coherently prescribing dynamic plans to agents that treat those plans (or their constituent acts) as signs of causally remote states.

My suggestion for dealing with this problem is a radical one: we must reject the seemingly natural view that the evidential and causal import of an agent's acts can rationally come apart from her internal perspective. Or, slightly less radically (and more precisely): we must restrict in some way the set of permissible ways an agent may take her present or future acts to evidentially bear upon certain causally irrelevant states. The exact form this restriction must take will depend upon the shape of the decision theory we ultimately wish to endorse (e.g. evidentialist, causalist, etc.). But there is no escaping some such restriction. Only by restricting rational credence beyond the standard canons of Bayesian coherence can we hope to secure a plausible, dynamically consistent decision theory.

After introducing some necessary background (§2-§4), my case for this conclusion will come in three acts. In the first, building on previous results establishing the dynamic inconsistency of evidential decision theory, I argue that, in the absence of nonstandard epistemic constraints, any plausible decision theory that violates a causal dominance principle is doomed to be dynamically inconsistent (§5). In the second, I generalize Arif Ahmed's argument that causal decision theory is dynamically inconsistent to conclude that any

[1] This classic example is famously discussed by Lewis 1979.

[2] For sophisticated defenses of the evidentialist and causalist answers, respectively, see Ahmed 2014b and Joyce 1999.

[3] For a variety of disparate alternatives to the strictly evidentialist and causalist answers, see, e.g., Yudkowsky 2010, Wedgewood 2011, Dohrn 2015, Levinstein and Soares 2020, Gallow 2020, and Rothfus forthcoming.

plausible decision theory that satisfies the aforementioned causal dominance principle shares a similar fate (§6), again without positing additional epistemic constraints, leaving us to conclude that no plausible decision theory can satisfy dynamic consistency in the absence of nonstandard epistemic constraints. Finally, in the third, I demonstrate that a pair of epistemic constraints on priors (*Backward Autonomy* and *Full Autonomy*) suffice to render, respectively, evidential and causal decision theory dynamically consistent (§7-§8). I conclude by briefly summarizing and discussing the import of these results for rational choice theory (§9).

2 Formal Background

Speaking informally, a decision problem is a story featuring a protagonist, *the agent*, tasked with making a (possibly trivial) sequence of choices, perhaps strewn amidst various shifts in her decision relevant attitudes (e.g. her beliefs and values). If the decision problem is well posed, its telling ought to be detailed enough for its intended audience, *the decision theorist*, to identify at each point in the story which of the agent's remaining potential courses of action are maximally instrumentally valuable for her, i.e. best advance her aims. The required details must specify both the (perceived) structure of the situation the agent faces (e.g. the number of possible decisions she could face, the possible information she might gain along the way, etc.) and the agent's own decision relevant attitudes. Each of these features ought to be captured in any formal model of decision problems.

Advancing such a model, we will think of a *decision problem* as a 5-tuple, $\langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, where:

- $\mathcal{A}$ is an algebra of *propositions* or *events*, closed under the standard Boolean connectives. These propositions serve as agents' ultimate objects of belief and desire.[4]

- T is a finite *Bayesian decision tree*, that is, a tree-structured graph whose nodes, N_T, are ordered by an immediate successor function, N_+, and are partitioned into *choice nodes*, *natural nodes*, and *terminal nodes*, with one node uniquely serving as T's *initial node*, labeled 'n_0'. Every node $n \in N_T$ is associated with a proposition, $S(n) \in \mathcal{A}$, characterizing the agent's state of information at n. We assume that the event associated with a non-terminal node is always partitioned by the events associated with the node's immediate successors. I will write '$T(n)$', where $n \in N_T$, to denote the tree that is the truncation of T at n, and $T(X)$, where $X \in \mathcal{A}$, to denote the truncation of T by X, i.e. the tree formed by intersecting $S(n)$ with

[4]The formal framework assumed in this paper is thus broadly in line with that of Jeffrey 1965/1983.

X for all $n \in N_T$ and then discarding all nodes with empty information states.[5]

- $n' \in N_T$ represents a concrete position occupied by an agent within the broader decision scenario represented by T.

- $\{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}$ is a family of node-indexed *causal credence measures* defined on $\mathcal{A}$. $P_n^X(Y)$ expresses an agent's assessment of the causal probability of Y given X from her perspective at node n, i.e. the extent to which an agent takes X to causally promote Y from this vantage point.[6] We identify an agent's standard credences, P_n, capturing her unconditional degrees of belief in the members of $\mathcal{A}$ at n, with her causal probabilities at n given a tautology, i.e. P_n^T.

- $\{V_n\}_{n \in N_T}$ is a family of node-indexed *desirability functions* defined on $\mathcal{A}$. V_n measures an agent's degrees of desire at n that the various members of $\mathcal{A}$ be found to be true. I assume that for every V_n there exists a finite partition O of $S(n)$, such that for each $o \in O$ and $o' \in \mathcal{A}$ such that $o' \subseteq o$, $V_n(o) = V_n(o')$. The members of such a partition are *outcome propositions* with respect to V_n.

The limitations of this framework should be recognized at the outset. Not all decision problems in the informal sense admit suitable models of the suggested form. For example, the restriction to finite Bayesian decision trees precludes consideration of infinite decision puzzles, Moreover, the assumption that credences and desirabilities are real-valued functions precludes consideration of agents with various nonstandard attitudes that don't admit of a real-valued representation. Further, the evolution of information in Bayesian decision trees is always assumed to be propositional in form and may preclude e.g. problems involving either non-propositional evidence or *de se* uncertainty. The assumed existence of finite outcome partitions in every decision problem also rules out decision problems in which agents have infinitely fine-grained values. Finally, if attitudes beyond (causal) credences and desirabilities (e.g. risk-sensitivity a la Buchak 2013) can count as decision relevant, then the information encoded in a formal decision problem will, in some cases, be insufficient to separate rational from irrational courses of action. All such problems fall then outside the proposed model's domain of applicability.[7] Nonetheless, such limitations noted, a vast array of typical decision scenarios of interest to philosophers fall

[5] I largely follow authors like Hammond 1988, McClennen 1990, and Cubitt 1996 in my presentation of decision trees, save that I allow learning pursuant to choice nodes as well as natural nodes.

[6] On one approach, $P^X(Y)$ might be identified with the standard probability of a non-backtracking subjunctive conditional featuring X as antecedent and Y as consequent, i.e. $P(X \mathrel{\square\!\!\rightarrow} Y)$, though this would require appropriately closing $\mathcal{A}$ under this conditional operator. We might also compute $P^X(Y)$ relative to a given background partition of causal factors or *dependency hypotheses* that are causally independent of X and individually suffice to fix the causal bearing of X upon Y and its negation. Writing such a (finite) partition as '$\{K_i\}_i \subseteq \mathcal{A}$', we might define $P^X(Y) = \sum_i P(Y|XK_i)P(K_i)$.

[7] I lift this phrase from Titelbaum (2013)'s illuminating discussion of modelling frameworks.

squarely within this domain.

In fact, from the standpoint of ideal decision theory, which aims to give practical recommendations to rationally unified agents, the domain of decision problems so construed is in some sense too broad. The above characterization allows us to consider, for example, decision problems in which agents exhibit non-probabilistic credences or arbitrary shifts in their desirabilities over time. Such decision problems seem to be *incoherent* in the sense that the agents they feature fail to be rationally unified over the course of the story they encode. What exactly is required for a decision problem to be coherent is a central question of this essay, but we may begin characterizing this notion by noting several assumptions that are standardly made in this regard:

- **Probabilism**: All coherent decision problems are probabilistic, where a decision problem, $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, is *probabilistic* just in case for every node $n \in N_T$ and for every proposition $X \in \mathcal{A} \cap S(n)$, P_n^X is a probability measure on $\mathcal{A}$ satisfying $P_n^X(S(n)) = 1$.

- **Desirabilism**: All coherent decision problems are desirabilistic, where a decision problem, $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, is *desirabilistic* just in case, for all $n \in N_T$, V_n is defined on the non-null members of $\mathcal{A}$ and satisfies Jeffrey's desirability axiom, both with respect to P_n, i.e. for any pair of non-null, mutually exclusive propositions $X, Y \in \mathcal{A} \cap S(n)$, $V_n(X \vee Y) = P_n(X|X \vee Y)V_n(X) + P_n(Y|X \vee Y)V_n(Y)$.

- **Probabilistic Conditionalization**: All coherent decision problems satisfy probabilistic conditionalization, where a decision problem, $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, satisfies *probabilistic conditionalization* just in case if n_a precedes n_b along some branch of the given tree T, then, for all $X \in \mathcal{A}$: $P_{n_b}(X) = P_{n_a}(X|S(n_b)) = P_{n_a}(XS(n_b))/P_{n_a}(S(n_b))$.

- **Desirabilistic Conditionalization**: All coherent decision problems satisfy desirabilistic conditionalization, where a decision problem, $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, satisfies *desirabilistic conditionalization* just in case if n_a precedes n_b along some branch of the given tree T, then, for all $X \in \mathcal{A}$: $V_{n_b}(x) = V_{n_a}(X|S(n_b)) = V_{n_a}(XS(n_b)) - V_{n_a}(S(n_b))$.[8].

Going forward, I will take all this on board. That is, I will assume that all coherent decision problems are probabilistic, desirabilistic, and satisfy conditionalization in both its probabilistic and desirabilistic variants. I shall call all problems satisfying these properties, as well as an additional requirement introduced below excluding backward causation, *standard decision problems*. It is an open question whether the informal notion of a coherent decision problem as one whose protagonist is rationally unified is adequately explicated by the formal definition of standard decision problems. As I shall suggest, one possible lesson of this essay is that the class of standard decision problems is, in the

[8]See Bradley 2017, p. 97, for more on Conditional Desirability.

presented framework, strictly broader than the class of truly coherent decision problems.

I take the primary objects of deliberation for a rational agent facing a coherent decision problem construed as above to be *plans*. Given a tree T, a plan p specifies a unique move for every choice node in T that an agent facing T could reach, given implementation of earlier portions of p. It thus traces a unique path through the tree, given any combination of moves by nature at its nodes. To render plans suitable objects of desirability in our framework, we need to identify them formally with propositions. Since plans correspond to appropriate sets of terminal nodes (i.e. those they might terminate in, if executed perfectly), we can identify a plan with the disjunction of propositions associated with its corresponding terminal nodes (i.e. the disjunction of all events its flawless execution might terminate in). This proposal can be worked out recursively.[9]

Definition 1. Let $n \in N_T$. The set of **plans** available at n in T, denoted '$\Omega(T, n)$', is defined recursively as follows:

1. If n is a terminal node, then $\Omega(T, n) = \{S(n)\}$.
2. If n is a choice node, then
$$\Omega(T, n) = \{S(n') \wedge p(n') : n' \in N_+(n), p(n') \in \Omega(T, n')\}.$$
3. If n is a natural node, then
$$\Omega(T, n) = \left\{ \bigwedge_i [S(n_i) \supset p(n_i)] : n_i \in N_+(n), p(n_i) \in \Omega(T, n_i) \right\}.$$

We will also need the notion of a *plan continuation*. Fix a tree T and let n_a and n_b be nodes of T such that n_a precedes n_b. If p is a plan at n_a that makes arrival at n_b possible (i.e. a plan consistent with $S(n_b)$), then the continuation of p at n_b is just the conjunction of p and $S(n_b)$. If p is incompatible with arrival at n_b, then we can be content to leave $p(n_b)$ undefined. Where defined, $p(n_b)$ will itself, of course, be a plan at n_b. We let '$\Omega(T, n_a)(n_b)$' designate the set of plan continuations at n_b of plans at n_a.

With the notion of a plan defined, we can introduce the final restriction on the class of standard decision problems promised above:

[9] This definition of plans is employed also in Rothfus 2020, which raises the worry that plans so construed will often fail to form a partition, i.e. multiple distinct plans may be compatible with one another such that it may be possible for each to be ultimately implemented. I am hesitant to claim this is a problem myself, since such compatibility does not seem obviously problematic, but, if it is, the solution would involve replacing the material conditional employed in definition 1 with a non-truth-functional counterpart. This suggestion is explored a bit in Huttegger and Rothfus 2021, but merits further reflection.

- **No Backward Causation**: In all coherent decision problems, future plans are judged causally independent of epistemologically prior states, that is, where $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ is a decision problem, if n_a is any natural node in T and $p \in \Omega(T, n_a), n_b \in N_+(n_a)$, then $P_{n_a}^p(S(n_b)) = P_{n_a}(S(n_b))$.

This principle recognizes that although the ordering of nodes in a Bayesian decision tree directly represents neither the causal nor temporal ordering of events in the world,[10] the *epistemological* ordering it does represent justifies, at least in all but the most bizarre cases, certain causal assumptions. In particular, anticipated future decisions cannot be properly understood as having a causal bearing upon events whose truth value will be ascertained prior to their being taken. What I do on Tuesday cannot causally influence what I learn on Monday. To abandon this assumption is, by my lights, to pass through the looking glass into a world of decisions problems which are either logically problematic or at least normatively intractable.

Finally, a *decision theory*, D, is a (possibly partial) function that maps a decision problem $\langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ to a non-empty subset of $\Omega(T, n')$, returning those plans in $\Omega(T, n')$ that the theory judges to be of maximal instrumental utility for an agent whose epistemic and axiological attitudes are as given by $\{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}$ and $\{V_n\}_{n \in N_T}$. If two decision problems $d_1 = \langle \mathcal{A}, T, n_a, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ and $d_2 = \langle \mathcal{A}, T, n_b, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ differ only in that n_a precedes n_b along some branch of T, I will say d_2 is a *continuant* of d_1. If $n_b \in N_+(n_a)$, I will say that d_2 is an *immediate continuant* of d_1. With the necessary formal preliminaries in place, we can now turn to considering what norms a plausible decision theory ought to satisfy.

3 Normative Principles

Normative decision theorists are interested in defining a *rational* decision theory, that is, one adequate to the task of effectively guiding an ideally rational agent's practical deliberation across various potential decision problems so as to best further her ends. There are various principles that such a decision theory should plausibly satisfy, or at least satisfy across a suitable range of problems. My central normative postulate here is that decision theories ought to be *dynamically consistent* across the range of coherent decision problems:

Definition 2. A decision theory, D, is **dynamically consistent** across a domain of decision problems Z just in case D is defined on all members of Z and, for all decision problems $d_1 = \langle \mathcal{A}, T, n_a, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ and $d_2 = \langle \mathcal{A}, T, n_b, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ such that $d_1, d_2 \in Z$ and d_2 is a continuant of d_1, if $p \in D(d_1)$ and $p(n_b)$ is defined, then $p(n_b) \in D(d_2)$.

[10] For an illuminating critique of this feature of decision trees, see Spohn unpublished.

If a decision theory recommends a particular policy as optimal in a particular decision scenario, then in any continuation of that scenario compatible with the recommended policy, the theory should not undercut itself by reneging on its initial recommendation and suggesting something new. Or, rather, it shouldn't do so as long as the decision problem is coherent. Of course, if an agent is turned against herself, then the fact that a decision theory offers her contradictory recommendations (before and after the Sirens' song, so to speak) need be no strike against the decision theory. The blame for its discrepancy in advice lies squarely with the incoherence (synchronic or diachronic) of the agent's decision relevant attitudes. But if a decision problem and its continuants feature a rational agent holding to a single set of unified aims, then we are within our rights to insist that a reasonable decision theory not introduce any dynamic inconsistency.[11]

As a plausible supplemental norm, I will also assume that decision theories ought to be *independent*:

Definition 3. A decision theory, D, is **independent** just in case there exists a function f such that, for all decision problems $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ where D is defined, $D(d) = f(\Omega(T, n'), \{P_{n'}^X\}_{X \in \mathcal{A} \cap S(n')}, V_{n'})$.

An independent decision theory is one whose recommendations in a decision problem are solely a function of the plans available in the problem together with the relevant agent's epistemic and axiological attitudes at the time the decision must be made. Other features of the decision problem, e.g. the structure of the decision tree and the agent's possible preceding and succeeding attitudes, should play no role. Of course, from the perspective of a planning agent, such factors surely can matter to what currently available act the agent ought to select. But a decision theory, as I am construing matters, is not primarily in the business of giving such advice but rather purports to tell an agent which logically available course of action is most instrumentally valuable in terms of furthering her ends, whether or not such a course is feasible given, e.g. shifts in her future desirabilities, anticipated irrationality, etc. On this view of what a decision theory amounts to, independence is an eminently plausible principle. Lastly, we shall have occasion to introduce a further supplemental norm of *conservatism* in the next section, which, alongside dynamic consistency and independence, will complete our list of criteria for assessing a decision theory's normative adequacy.

Before we can state this norm, however, we must introduce the two dominant decision theories among philosophers, in terms of which conservatism

[11] Principles akin to a dynamic consistency requirement have been endorsed by Hammond 1988, Cubitt 1996, McClennen 1990, Buchak 2013, Ahmed 2014b, Huttegger and Rothfus 2021, and Weatherson MS. The dynamic consistency norm endorsed here differs from that found in some of these authors in that it applies to decision theories rather than to agents, which may enable it to sidestep some of the objections brought against diachronic norms of rationality by partisans of *time-slice rationality*, e.g. Hedden 2015. For an overview of some of the relevant historical debates regarding dynamic consistency, see Steele 2010.

will be defined.

4 EDT vs CDT

The simplest and most elegant decision theory is *Evidential Decision Theory (EDT)*.

> **Evidential Decision Theory**: If $d_1 = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, then $EDT(d_1) \coloneqq \{p \in \Omega(T, n') | V_{n'}(p) \geq V_{n'}(p'), \forall p' \in \Omega(T, n')\}$.

EDT judges the practically best plans to be the ones that are V-maximal. According to evidentialists then, rational choice simply consists in maximization of desirability. Causal considerations need never enter the story.

Alas, simplicity does not guarantee rationality. Perhaps the most famous objection to EDT is that it violates a plausible *Causal Dominance* principle:

Definition 4. Let $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ be a decision problem and $a, b \in \Omega(T, n')$. Say that a **causally dominates** b in d, written '$a \rhd_d b$', just in case there exists a finite partition $\{S_1, S_2, ..., S_k\}$ of $\mathcal{A} \cap S(n')$ such that (i) $P_{n'}(S_i) = P_{n'}^p(S_i), \forall i \in \{1, 2, ..., k\}, p \in \Omega(T, n')$, (ii) all members of $\{pS_i\}_{p \in \{a,b\}, i \in \{1,..,k\}}$ are outcome propositions relative to $V_{n'}$, and (iii) $V_{n'}(aS_i) > V_{n'}(bS_i), \forall i \in \{1, 2, ..., k\}$. Say further that a decision theory, D, respects **Causal Dominance** just in case, for all standard decision problems $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ and plans $b \in \Omega(T, n)$, if there exists $a \in \Omega(T, n)$ such that $a \rhd_d b$, then $b \notin D(d)$.[12]

What Causal Dominance requires is that if an agent judges that a particular plan a will certainly lead to a better outcome than another plan b, given any cell of a partition each of whose members is judged causally independent of a and b, then she cannot rationally implement b. To see how EDT can contradict this principle, suppose again that I am to play a Prisoner's Dilemma with my similarly inclined friend. Once I am in my causally isolated room, my decision to cooperate (C) or defect (D) has no causal bearing on whether my friend will cooperate (c) or defect (d), i.e. $P^C(c) = P^D(c)$. So, given that this is a Prisoner's Dilemma (i.e. $V(Dc) > V(Cc)$ and $V(Dd) > V(Cd)$), Causal Dominance forbids cooperation. It is easy enough to see, however, that if my choice to cooperate is strong enough evidence of my friend's choice to cooperate (i.e. if $P(c|C) - P(c|D)$ is high enough) then EDT will recommend cooperating.[13]

[12] Our previous restriction to finite decision problems becomes key here. In infinite decision problems, it is possible that every available plan might be causally dominated by another and hence it could be impossible to have a decision theory that both always returns non-empty recommendations and satisfies causal dominance in the infinite context.

[13] Just how strong the evidence needs to be of course depends upon the precise desirabilities encoded by V.

In response to the challenge posed by such *Newcomb problems*,[14] many philosophers have sought refuge in *Causal Decision Theory (CDT)*.[15] To introduce this alternative, we first need to define the *utility* or *efficacy value* of a proposition.

Definition 5. Let a family of causal probabilities, $\{P^X\}_{X\in\mathcal{A}}$, and a desirability function, V, defined, respectively, on $\mathcal{A}$ and its non-null members be given. Let $\{O_1, O_2, ..., O_k\} \subseteq \mathcal{A}$ be a partition of outcome propositions with respect to V. The **utility** or **efficacy value** of $Y \in \mathcal{A}$, $U(Y)$, is defined as: $U(Y) \coloneqq \sum_i P^Y(O_i)V(O_i)$.

> **Causal Decision Theory**: If $d_1 = \langle \mathcal{A}, T, n', \{P^X_n\}_{n\in N_T, X\in\mathcal{A}\cap S(n)}, \{V_n\}_{n\in N_T}\rangle$, then $CDT(d_1) \coloneqq \{p \in \Omega(T, n') | U_{n'}(p) \geq U_{n'}(p'), \forall p' \in \Omega(T, n')\}$.

CDT judges the practically best plans to be the ones that are U-maximal. According to causalists then, rational choice consists in maximization of utility, understood as efficacy value.

Applied to the Prisoner's Dilemma, CDT delivers an unambiguous verdict in favor of defecting, provided that I really do take my choice to be causally irrelevant to my compatriot's decision, i.e. if $P^C(c) = P(c)$. Indeed, CDT will always respect Causal Dominance. If $a \rhd_d b$, then there exists a causally independent partition $\{S_1, ..., S_k\} \in \mathcal{A}_d$ such that $V_{n_d}(aS_i) > V_{n_d}(bS_i), \forall i \in \{1, 2, ..., k\}$. (Note: with d a fixed decision problem, I employ $\mathcal{A}_d, T_d$, etc. to denote the algebra, tree, etc. associated with d.) But this implies that $P_{n_d}(S_i)V_{n_d}(aS_i) \geq P_{n_d}(S_i)V_{n_d}(bS_i), \forall i \in \{1, 2, ..., k\}$, with the inequality being strict for some i. Hence, $\sum_i P_{n_d}(S_i)V_{n_d}(aS_i) > \sum_i P_{n_d}(S_i)V_{n_d}(bS_i)$. But given the causal independence assumption, this means $U_{n_d}(a) > U_{n_d}(b)$, so $b \notin CDT(d)$.

CDT's satisfaction of Causal Dominance appears to be a significant relative advantage over EDT. Still, not all are convinced of the superiority of CDT's recommendation in Newcomb problems. Moreover, in a range of other decision problems, EDT seems to yield more intuitive verdicts than CDT,[16] resulting in an apparent dialectical stalemate. Be that as it may, proponents of both CDT and EDT (as well as some third parties) should at least agree to the following principle that I want to suggest as a final normative constraint that any plausible decision theory ought to satisfy. Specifically, just as a decision theory ought to be independent and dynamically consistent across coherent decision problems, so too it ought to be *conservative*:

Definition 6. A decision theory, D, is **conservative** just in case, for all decision problems $d = \langle \mathcal{A}, T, n', \{P^X_n\}_{n\in N_T, X\in\mathcal{A}\cap S(n)}, \{V_n\}_{n\in N_T}\rangle$, if $EDT(d) = CDT(d)$, then $D(d) = EDT(d) = CDT(d)$.

[14] Newcomb problems were first employed as an objection to EDT by Nozick 1969.

[15] CDT has taken many forms in the decision theoretic literature. Some of the most prominent are those in Gibbard and Harper 1978, Skyrms 1980, Lewis 1981, Joyce 1999, and Pearl 2009. A central difference amongst these variants concerns how they construe the nature of causal probabilities. For our purposes, however, such differences are inconsequential.

[16] See, for example, Egan 2007, Ahmed 2014a, and Spencer and Wells 2017. For causalist replies, see Joyce 2012, Joyce 2018, Armendt 2019, Williamson 2021.

Conservative decision theories are thus those that agree with EDT and CDT wherever they coincide. Prima facie, this is a plausible requirement on a normative decision theory. The cases in which EDT and CDT come apart are tricky ones in which intuitions famously diverge. But if a particular plan is judged maximally choiceworthy, on *both* evidentialist and causalist standards alike, then it seems safe to regard the plan as in fact maximally choiceworthy. Note that both EDT and CDT are trivially conservative, as well as independent. Unfortunately, neither is dynamically consistent across standard decision problems.

5 Non-Causal Decision Theories Are Dynamically Inconsistent

One reason to find violations of Causal Dominance troubling is that they seem to lead to dynamic inconsistency. This has already been argued in the specific case of EDT, but it is true in general of theories that conflict with Causal Dominance.[17] To prove this, I will assume that independence and conservatism are requirements on any normatively adequate decision theory.

Proposition 1. *Every independent and conservative decision theory that fails to respect Causal Dominance is dynamically inconsistent across the domain of standard decision problems.*

Proof. Suppose that D is an independent, conservative decision theory that fails to respect Causal Dominance and is defined (at least) on the domain of standard decision problems. Since D fails to respect Causal Dominance, there must be a standard decision problem, $d = \langle \mathcal{A}, T, n', \{P^X_n\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, such that $b \in D(d)$ but $a \rhd_d b$ for some $a \in \Omega(T, n')$. By the definition of causal dominance, there must be some partition $\{S_1, S_2, ..., S_k\} \subseteq \mathcal{A}$ such that (i) $P_{n'}(S_i) = P^p_{n'}(S_i), \forall i \in \{1, 2, ..., k\}, p \in \Omega(T, n')$, (ii) all members of $\{pS_i\}_{p \in \{a,b\}, i \in \{1,..,k\}}$ are outcome propositions, and (iii) $V_{n'}(aS_i) > V_{n'}(bS_i), \forall i \in \{1, 2, ..., k\}$. Let $\{S_1, ..., S_k\}$ be such a partition. Let T' be the decision tree that consists of an initial natural node n_0 whose k succeeding nodes, $n_1, ..., n_k$, are the initial nodes of $T(S_1), T(S_2), ..., T(S_k)$, respectively. Now consider a decision problem $d' = \langle \mathcal{A}, T', n_0, \{P^X_n\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_{T'}} \rangle$, such that $P^X_{n_0} = P^X_{n'}$, for all $X \in \mathcal{A}$, and $V_{n_0} = V_{n'}$. (Such a decision problem is standard given the assumed causal independence of $\{S_1, ..., S_k\}$ from $\Omega(T, n_0)$ according to $\{P^X_{n_0}\}_{X \in \mathcal{A}}$.) We know that $\Omega(T, n') = \Omega(T', n_0)$, hence since D is independent, it must be that $b \in D(d')$. But $b(n_i) = bS_i \notin D(d'_i)$, where d'_i is the continuant of d_i at n_i, since D is conservative and $b(n_i) \notin EDT(d'_i) = CDT(d'_i)$. But d' and d'_i are both standard decision problems. Hence, D is dynamically inconsistent across the domain of standard decision problems. □

[17] See, for example, the arguments in Arntzenius 2008, Meacham 2010, and Rothfus 2020 all of which build on the examples of Gibbard and Harper 1978, as well as the example in Wells 2018, which can be used to similar effect.

Any decision theory then that sometimes recommends a strictly causally dominated plan is liable to violate dynamic consistency in a pair of easily crafted standard decision problems. Intuitively, to achieve this result just move the unveiling of the relevant causally independent partition to the start of the problem and watch the sequential incoherence unfold. Note that such a move is licit within the domain of standard decision problems only given the assumed causal independence (lest we transgress No Backward Causation), and so cannot, for example, be used to generate fallacious dynamic consistency arguments on behalf of unrestricted dominance principles.

6 Causal Decision Theories Are Dynamically Inconsistent

In light of the difficulties arising from violations of Causal Dominance, many have turned to CDT as a safe haven amidst Newcombian troubles. Unfortunately (and perhaps more surprisingly), CDT is also dynamically consistent. The case for this has been convincingly laid out by Arif Ahmed via his *Psycho-Insurance problem*, a sequential variant of Andy Egan's famous *Psycho-Button* case.[18]

The problem comes in two stages. In the first, a Newcombian predictor offers you a choice between pressing a button and refraining from pressing it. If she predicted that you would press the button, then she has set things up so that pressing the button will cause \$1 to be debited from your bank account. If she predicted that you would not press the button, then she has set things up so that pressing the button will cause \$1 to be credited to your account. (This is Ahmed's sanitized version of Egan's case.) Let P be the proposition that you press the button. The options you face here are then P and $\overline{P}$. Let K be the proposition that the predictor had predicted you would press while $\overline{K}$ is the proposition that she predicted the opposite. Assume that desirability and dollar payoffs coincide on propositions strong enough to fix the latter, i.e. all you care about is money and you value it linearly.

You take the predictor to be fairly reliable so let's fix that your credence in K given P and in $\neg K$ given $\neg P$ are each greater than .75 (just say .9 for concreteness). After you have made your decision with regard to pressing the button, you will be offered an opportunity to bet on whether the predictor correctly predicted your choice. The bet will pay \$0.50 if the predictor was correct and cost you \$1.50 otherwise. Let B be the proposition that you bet. We assume that you judge neither P nor B to have any causal influence upon K. In our framework,

[18] Egan 2007. Ahmed notes that cases of similar structure have appeared in the literature before, for example, in Gibbard 1992. For other sequential choice problems that could be used to make a similar point, see Oestherheld and Conitzer 2021 and Spencer 2021. For one critique of some of the conclusions drawn from these examples together with dynamic consistency, see Joyce 2016.

this decision problem is captured by $PI = \langle \mathcal{A}, T, n_0, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$, where:

- $\mathcal{A} = \mathcal{P}(W)$, where W is the set of all state-descriptions constructed from P, K, B.
- T is the decision tree depicted in Figure 1.
- n_0 is T's initial node.
- $\{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}$ is a family of probability measures satisfying: (i) $P_{n_0}(K|P) = P_{n_0}(\neg K|\neg P) = .9$ and (ii) $P_n^{PB}(K) = P_n^{P\overline{B}}(K) = P_n^{\overline{P}B}(K) = P_n^{\overline{P}\overline{B}}(K) = P_n(K), \forall n \in N_T$ at which the relevant probabilities are defined.
- $\{V_n\}_{n \in N_T}$ is a family of desirability functions each of whose payoffs on atomic elements of $\mathcal{A}$, where defined, are as specified in Table 1.

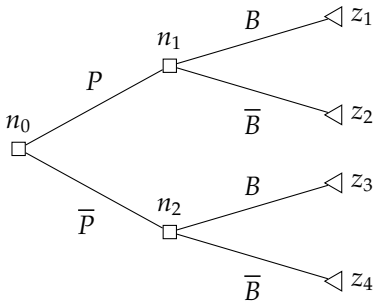

Figure 1: $T1$, the Psycho-Insurance Problem. $S(n_0) = T$, $S(n_1) = P$, $S(n_2) = \overline{P}$, $S(z_1) = PB$, $S(z_2) = P\overline{B}$, $S(z_3) = \overline{P}B$, $S(z_4) = \overline{P}.\overline{B}$.

	K	$\overline{K}$
PB	-0.5	-0.5
$P\overline{B}$	-1	1
$\overline{P}B$	-1.5	.5
$\overline{P}.\overline{B}$	0	0

Table 1: The Psycho-Insurance Problem in Normal Form

Since $\overline{P}.\overline{B} \rhd_{PI} PB$ and $P\overline{B} \rhd_{PI} \overline{P}B$, CDT's satisfaction of Causal Dominance implies that $PB, \overline{P}B \notin CDT(PI)$. So, either $P\overline{B}$ or $\overline{P}.\overline{B}$ must be in $CDT(PI)$. Suppose $P\overline{B} \in CDT(PI)$. (A similar argument will apply if $\overline{P}.\overline{B} \in CDT(PI)$.) Now consider the continuant of PI at node n_1, PI_1. At this point, $P\overline{B}(n_1) = \overline{B} \notin CDT(PI_1)$. Hence, we have a violation of dynamic consistency on the part of CDT.

It should be clear from the preceding argument that we can use Ahmed's example to craft a general argument that any conservative decision theory that respects Causal Dominance will in fact be dynamically inconsistent.

Proposition 2. *Every conservative decision theory that respects Causal Dominance is dynamically inconsistent across the domain of standard decision problems.*

Proof. Let D be a conservative decision theory defined (at least) on all standard decision problems that respects Causal Dominance. Consider the standard decision problem PI. D must be defined on PI in order to qualify as dynamically consistent across the domain of standard decision problems, so suppose that it is. By Causal Dominance and the fact that D must return a non-empty set on every input on which it is defined, we know that (i) $P\overline{B} \in D(PI)$ or (ii) $\overline{P}.\overline{B} \in D(PI)$. Now consider the standard decision problems PI_1 and PI_2 that are the continuants of PI at nodes n_1 and n_2, respectively. We know that $EDT(PI_1) = CDT(PI_1)$ and $EDT(PI_2) = CDT(PI_2)$, so conservatism requires that $D(PI_1) = CDT(PI_1)$ and $D(PI_2) = CDT(PI_2)$. But, since $P\overline{B}(n_1) = \overline{B} \notin CDT(PI_1)$ and $\overline{P}.\overline{B}(n_2) = \overline{B} \notin CDT(PI_2)$, it must be that $P\overline{B}(n_1) = \overline{B} \notin D(PI_1)$ and $\overline{P}.\overline{B}(n_2) = \overline{B} \notin D(PI_2)$. So, whichever of (i) and (ii) is true, D is dynamically inconsistent on the domain of standard decision problems. □

7 Backward Autonomy

Stitching Propositions 1 and 2 together yields the negative result:

Proposition 3. *No independent, conservative decision theory is dynamically consistent across the domain of standard decision problems.*

If dynamic consistency is a genuine constraint upon the adequacy of any normative decision theory, then we are left with few options. We could reject the rational necessity of either independence or conservatism, but each of these principles is eminently plausible. Assuming that we refuse to take this route, we are left with one avenue for preserving the possibility of a normatively adequate decision theory: restricting in some way the domain of admissible decision problems. This is the route I will explore here.

If we wish to charge the decision problems considered above with incoherence, where precisely is it to be located? The decision problems considered thus far have all been standard and so there is clearly no standard probabilistic or desirabilistic incoherence at which to lay the blame. There are, it seems, very few plausible ways to go about further restricting the domain of standard decision problems in the name of coherence. One aspect of the problems highlighted above that might be questioned concerns their unrestricted adherence to conditionalization as an update rule for revising credences and desirabilities in response to new information. Conditionalization is unobjectionable when applied to cases in which an agent passively learns the truth of a proposition

identified in the algebra of events she has beliefs over. That is, in the context of dynamic choice, pursuant to natural nodes a rational agent ought to update her credences and desirabilities by conditionalization. However, conditionalization might be a more questionable assumption when applied to cases in which the new information an agent learns is brought about by her own willful decision to bring it about. That is, it might be questionable whether conditionalization properly characterizes the sort of learning that takes place pursuant to choice nodes in dynamic choice problems.

A number of philosophers have independently challenged this assumption on grounds different than those that concern us in this essay. John Cantwell, for example, has suggested that the problem of decision instability exemplified in one-shot decision problems like Gibbard and Harper's Death in Damascus or Egan's Psycho-Button provides pragmatic motivation to adopt a bifurcated update rule requiring one to update by conditionalization when learning states and by imaging (i.e. employing causal probabilities) when learning acts.[19] Melissa Fusco arrives at similar conclusions from considerations of *epistemic time bias*.[20] This idea also seems to share something in common with the way some causal decision theorists working in the tradition of graphical causal models often talk. On one way of reading such authors, a rational decision qualifies as an *intervention* that severs any grounds for correlation between the act decided upon and temporally antecedent states of affairs. Hence, according to this approach, while learning a state goes by conditionalization, learning an act goes by an alternative mode of belief revision appropriate to causal intervention.[21]

In my view, this perspective merits deeper consideration than it has thus far received. For our purposes, the obvious question to ask is whether there are any plausible decision theories that are dynamically consistent across the domain of *almost* standard decision problems that replace Probabilisitic Conditionalization with a suitably bifurcated updating rule. However, I raise this proposal and its attendant questions only to set them aside. I wish instead to explore two different explications of the class of coherent decision problems that, while leaving probabilistic conditionalization intact as sacrosanct, each allow space for the existence of plausible, dynamically consistent decision theories. Both of these proposals suggests a modification of the admissible range of causal credence functions a rational agent may coherently adopt at a particular point in time, rather than a modification of the admissible range of functional relationships that may obtain between an agent's credence functions at disparate times. The first will be seen to render EDT dynamically consistent, while the

[19]Cantwell 2010.

[20]Fusco 2018.

[21]Some causal decision theorists view treating decisions as interventions as merely a useful fiction in the context of static decision making and hence do not view this as providing any rival updating rule to conditionalization. Others, however, do seem to intend to be writing more than fiction. See Hitchcock 2016 and the discussion in Stern 2018.

second secures the same verdict for CDT.

Beginning with the first proposal, it is natural to describe the predicament of plausible non-causal decision theories, like EDT, vis-a-vis dynamic consistency in something like the following way. In some decision problems where I anticipate learning which cell of a partition is true and then making a choice in light of that information, I might view a particular disjunctive plan as *ex ante* optimal but not view its various continuations as optimal after my learning experience. In the case of EDT, this stems from the fact that the disjunctive plan might provide me maximally good news in the form of indicating that one or more of its disjuncts are likely to be false even while I disprefer each of its continuations (i.e. indivudual disjuncts) to an another possible plan's continuations. What seems to be happening here then is that by taking information about my future acts as evidence for prior learning events, I open the door to dynamic inconsistency.

My proposal on behalf of the evidentialist committed to dynamic consistency then is to rule out such evidential judgments as incoherent. In particular, I suggest that the decision theorist so inclined restrict the range of coherent decision problems down from the standard to what I shall call the *backward autonomous*.

Definition 7. *A decision problem* $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ *is* **backward autonomous** *just in case (i) it is standard and (ii) for any natural node* $n \in N_T$*, any* $n_i \in N_+(n)$*, and any plan* $p \in \Omega(T, n)$*,* $P_n(S(n_i)|p) = P_n(S(n_i))$*.*

The corresponding partial explication of coherence now strengthens from standard Bayesian coherence to:

> **Backward Autonomy**: All coherent decision problems are backward autonomous.

The definition of backward autonomous decision problems allows us to establish not only the existence of plausible (i.e. independent and conservative) decision theories that satisfy dynamic consistency across the range of backward autonomous decision problems, but in fact the dynamic consistency of EDT itself within this scope.

Proposition 4. *EDT is dynamically consistent across the domain of backward autonomous decision problems.*

Proof. Let $d_a = \langle \mathcal{A}, T, n_a, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ be a backward autonomous decision problem and let $d_b = \langle \mathcal{A}, T, n_b, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ be a continuant of d_a such that $n_b \in N_+(n_a)$. Suppose $p \in EDT(d_a)$ and $p(n_b)$ is defined. To show that *EDT* is dynamically consistent on the domain of backward autonomous decision problems, it suffices to verify that $p(n_b) \in EDT(d_b)$, since the continuant of any backward autonomous decision problem is also backward

autonomous.

Case 1: Suppose n_a is a choice node.

This case is handled by the general proof in Rothfus 2020 that EDT's recommendations are always invariant pursuant to choice nodes.

Case 2: Suppose n_a is a natural node. Then p has the form $\wedge_i[S(n_i) \rightarrow p(n_i)]$, where the n_i's are the possible successors to n_a. Note that this is equivalent to $\vee_i S(n_i)p(n_i)$. By definition of EDT:

$V_{n_a}(\pi) \geq V_{n_a}(\pi'), \forall p' \in \Omega(T, n_a)$.

So:

$V_{n_a}(\vee_i S(n_i)p(n_i)) \geq V_{n_a}(\vee_i S(n_i)p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Applying the desirability axiom across the partition $\{S(n_i)\}_{n_i \in N_+(n_a)}$:

$\sum_j P_{n_a}(S_j|\vee_i S(n_i)p(n_i))V_{n_a}(S(n_j)\vee_i S(n_i)p(n_i)) \geq \sum_j P_{n_a}(S_j|\vee_i S(n_i)p'(n_i))V_{n_a}(S(n_j)\vee_i S(n_i)p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Since d_a is backwards autonomous:

$\sum_j P_{n_a}(S_j)V_{n_a}(S(n_j)\vee_i S(n_i)p(n_i)) \geq \sum_j P_{n_a}(S_j)V_{n_a}(S(n_j)\vee_i S(n_i)p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Which is equivalent to:

$\sum_j P_{n_a}(S_j)V_{n_a}(S(n_j)p(n_j)) \geq \sum_j P_{n_a}(S_j)V_{n_a}(S(n_j)p'(n_j)), \forall p' \in \Omega(T, n_a)$.

By Desirabilistic Conditionalization, this is equivalent to:

$\sum_j P_{n_a}(S_j)V_{n_j}(p(n_j)) \geq \sum_j P_{n_a}(S_j)V_{n_j}(p'(n_j)), \forall p' \in \Omega(T, n_a)$.

But then it must be that, for all $n_j \in N_+(n_a)$:

$V_{n_j}(p(n_j)) \geq V_{n_j}(p'(n_j)), \forall p' \in \Omega(T, n_a)$.

Since, if this failed to hold, i.e. if there were an n_j such that $V_{n_j}(p'(n_j)) > V_{n_j}(p(n_j))$ for some $p' \in \Omega(T, n_a)$ we could, contrary to supposition, form a new plan exactly similar to p except that it agrees with p' pursuant to n_j, which would be guaranteed to have greater ex ante desirability than p. Hence, no such plan exists, and the proof is complete. □

Embracing Backward Autonomy as an epistemic coherence condition thus allows the proponent of EDT to sidestep the charge of dynamic inconsistency. Having secured such a result for EDT is little comfort, however, for those convinced of the rational necessity of Causal Dominance. Clearly, Backward Autonomy, whatever its plausibility as an epistemic coherence condition, is no remedy for the dynamic choice woes of CDT and other decision theories that respect Causal Dominance. Both stages of Ahmed's *Psycho-Insurance*, for example, are backward autonomous decision problems (trivially so, since the relevant decision tree includes no natural nodes), yet they jointly suffice to establish the dynamic inconsistency of any conservative decision theory that respects Causal Dominance. Where then may a causalist turn?

8 Full Autonomy

In order to secure dynamic consistency for causalists, restricting beliefs in prior learning events conditional upon information about posterior acts is clearly insufficient. Stronger medicine is needed. In particular, cases like *Psycho-Insurance* show we must restrict at least some beliefs in causally prior states conditional upon information about posterior acts, regardless of whether the causally prior states are also epsitemically prior. The simplest and most natural suggestion for how to accomplish this is simply to insist that coherence requires *full autonomy*, where full autonomy amounts to collapsing conditional and causal credence in cases where the conditioning events are logically available plans. That is, we could insist that the domain of coherent decision problems is a subset of the domain of *fully autonomous decision problems*.

Definition 8. *A decision problem* $d = \langle \mathcal{A}, T, n', \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ *is* **fully autonomous** *just in case (i) it is standard and (ii) for any proposition* $X \in \mathcal{A}$*, any node* $n \in N_T$*, and any plan* $p \in \Omega(T, n)$*,* $P_n(X|p) = P_n^p(X)$*.*

The corresponding explication of coherence now becomes:

Full Autonomy: All coherent decision problems are fully autonomous.

Note that Full Autonomy is indeed a stronger principle than Backward Autonomy in the sense that every fully autonomous decision problem is also backward autonomous, though not vice-versa. Given No Backward Causation as a constraint on standard decision problems, if n is a natural node, $n_i \in N_+(n)$, and $p \in \Omega(T, n)$ then $P_n^p(S(n_i)) = P_n(S(n_i))$. But full autonomy requires, $P_n^p(S(n_i)) = P_n(S(n_i)|p)$. Thus, $P_n(S(n_i)|p) = P_n(S(n_i))$, in line with backward autonomy. On the other hand, *Psycho-Insurance* shows that not every backward autonomous problem is fully autonomous.

It is possible to show that Full Autonomy suffices to prove the dynamic consistency of CDT across the domain of coherent decision problems, though the same result would hold for EDT and any other conservative decision theory, since Full Autonomy is strong enough to render all such theories equivalent in the context of coherent decision problems.

Proposition 5. *CDT is dynamically consistent across the domain of fully autonomous decision problems.*

Proof. Let $d_a = \langle \mathcal{A}, T, n_a, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ be a fully autonomous decision problem and let $d_b = \langle \mathcal{A}, T, n_b, \{P_n^X\}_{n \in N_T, X \in \mathcal{A} \cap S(n)}, \{V_n\}_{n \in N_T} \rangle$ be an immediate continuant of d_a. Suppose $p \in CDT(d_a)$ and $p(n_b)$ is defined. To show that *CDT* is dynamically consistent on the domain of fully autonomous decision problems, it suffices to verify that $p(n_b) \in CDT(d_b)$, since the continuant of any fully autonomous decision problem is also fully autonomous. (For readability, I suppress subscripts referencing n_0.)

Case 1: Suppose n_a is a choice node.

We know that:

$U_{n_a}(p) \geq U_{n_a}(p^{'}), \forall p^{'} \in \Omega(T, n_a)$.

Computing utility relative to a finite set of outcome propositions O:

$\sum_{o \in O} P^{p}_{n_a}(o) V_{n_a}(o) \geq \sum_{o \in O} P^{p'}_{n_a}(o) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$.

By the full autonomy of d_a:

$\sum_{o \in O} P_{n_a}(o|p) V_{n_a}(o) \geq \sum_{o \in O} P_{n_a}(o|p') V_{n_a}(o), \forall p' \in \Omega(T, n_a)$.

But our definition of plans/continuations guarantees that plans are identical with their continuations pursuant to choice nodes:

$\sum_{o \in O} P_{n_a}(o|p(n_b)) V_{n_a}(o) \geq \sum_{o \in O} P_{n_a}(o|p'(n_b)) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$, where continuation at n_b is defined.

Which is equivalent to:

$\sum_{o \in O} P_{n_a}(o|p(n_b)S(n_b)) V_{n_a}(o) \geq \sum_{o \in O} P_{n_a}(o|p'(n_b)S(n_b)) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$, where continuation at n_b is defined.

Invoking Conditionalization:

$\sum_{o \in O: P_{n_b}(o) > 0} P_{n_b}(o|p(n_b)) V_{n_b}(o) \geq \sum_{o \in O: P_{n_b}(o) > 0} P_{n_b}(o|p'(n_b)) V_{n_b}(o), \forall p'(n_b) \in \Omega(T, n_b)$.

A second application of full autonomy guarantees:

$\sum_{o \in O: P_{n_b}(o) > 0} P^{p(n_b)}_{n_b}(o) V_{n_b}(o) \geq \sum_{o \in O: P_{n_b}(o) > 0} P^{p'(n_b)}_{n_b}(o|) V_{n_b}(o), \forall p'(n_b) \in \Omega(T, n_b)$.

By definition of U:

$U_{n_b}(p(n_b)) \geq U_{n_b}(p'(n_b)), \forall p'(n_b) \in \Omega(T, n_b)$.

Thus, $p(n_b) \in CDT(d_b)$.

Case 2: Suppose n_a is a natural node. Then p has the form $\wedge_i [S(n_i) \rightarrow p(n_i)]$, where the n_i's are the possible successors to n_a. Note that this is equivalent to $\vee_i S(n_i) p(n_i)$. So, computing utility relative to a partition of outcome propositions O, we have:

$\sum_{o \in\ O} P^{(\vee_i S(n_i) p(n_i))}_{n_a}(o) V_{n_a}(o) \geq \sum_{o \in\ O} P^{(\vee_i S(n_i) p(n_i))}_{n_a}(o) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$.

By the full autonomy of d_a:

$\sum_{o \in O} P_{n_a}(o|[\vee_i S(n_i) p(n_i)]) V_{n_a}(o) \geq \sum_{o \in O} P_{n_a}(o|[\vee_i S(n_i) p(n_i)]) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$.

Which is equivalent, by the Law of Total Probability, to:

$\sum_{o \in O} \sum_i [P_{n_a}(S(n_i) p(n_i)|\vee_i [S(n_i) p(n_i)])] P_{n_a}(o|S(n_i) p(n_i)) V_{n_a}(o) \geq \sum_{o \in O} \sum_i [P_{n_a}(S(n_i) p'(n_i)|\vee_i [S(n_i) p'(n_i)])] P_{n_a}(o|S(n_i) p'(n_i)) V_{n_a}(o), \forall p' \in \Omega(T, n_a)$.

Switching the order of the finite sums and abbreviating the plan:

$\sum_i \sum_{o \in O}[P_{n_a}(S(n_i)p(n_i)|p)]P_{n_a}(o|S(n_i)p(n_i))V_{n_a}(o) \geq \sum_i \sum_{o \in O}[P_{n_a}(S(n_i)p'(n_i)|p')]P_{n_a}(o|S(n_i)p'(n_i))V_{n_a}(o)$, $\forall p' \in \Omega(T, n_a)$.

Invoking Conditionalization:

$\sum_i \sum_{o \in O:P_{n_b}(o)>0}[P_{n_a}(S(n_i)p(n_i)|p)]P_{n_i}(o|p(n_i))V_{n_i}(o) \geq \sum_i \sum_{o \in O:P_{n_b}(o)>0}[P_{n_a}(S(n_i)p'(n_i)|p')]P_{n_i}(o|p'(n_i))V_{n_i}(o)$, $\forall p' \in \Omega(T, n_a)$.

Which is equivalent, by the full autonomy of d_a to:

$\sum_i \sum_{o \in O:P_{n_b}(o)>0}[P_{n_a}(S(n_i)p(n_i)|p)]P_{n_i}^{p(n_i)}(o)V_{n_i}(o) \geq \sum_i \sum_{o \in O:P_{n_b}(o)>0}[P_{n_a}(S(n_i)p'(n_i)|p')]P_{n_i}^{p'(n_i)}(o|)V_{n_i}(o)$, $\forall p' \in \Omega(T, n_a)$.

Pulling the first term out of the inner sum:

$\sum_i P_{n_a}(S(n_i)p(n_i)|p) \sum_{o \in O:P_{n_b}(o)>0} P_{n_i}^{p(n_i)}(o)V_{n_i}(o) \geq \sum_i P_{n_a}(S(n_i)p'(n_i)|p') \sum_{o \in O:P_{n_b}(o)>0} P_{n_i}^{p'(n_i)}(o|)V_{n_i}(o)$, $\forall p' \in \Omega(T, n_a)$.

Which is equivalent, by definition of U, to:

$\sum_i P_{n_a}(S(n_i)p(n_i)|p)U_{n_i}(p(n_i)) \geq \sum_i P_{n_a}(S(n_i)p'(n_i)|p')U_{n_i}(p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Which is equivalent to:

$\sum_i P_{n_a}(S(n_i)|p)U_{n_i}(p(n_i)) \geq \sum_i P_{n_a}(S(n_i)|p')U_{n_i}(p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Which is equivalent, since d_a is fully autonomous, to:

$\sum_i P_{n_a}^{p}(S(n_i))U_{n_i}(p(n_i)) \geq \sum_i P_{n_a}^{p'}(S(n_i))U_{n_i}(p'(n_i)), \forall p' \in \Omega(T, n_a)$.

Which is equivalent, given No Backward Causation, to:

$\sum_i P_{n_a}(S(n_i))U_{n_i}(p(n_i)) \geq \sum_i P_{n_a}(S(n_i))U_{n_i}(p'(n_i)), \forall p' \in \Omega(T, n_a)$.

But then:

$U_{n_i}(p(n_i)) \geq U_(n_i)(p'(n_i)), \forall p'(n_i) \in \Omega(T, n_i)$,

This line follows by the same reasoning employed at the end of the proof of Proposition 4. Namely, if this failed to hold, i.e. if for some i, $U_{n_i}(p'(n_i)) > U_{n_i}(p(n_i))$ for some $p' \in \Omega(T, n_a)$ we could, contrary to supposition, form a new plan exactly similar to p except that it agrees with p' pursuant to n_i, which would be guaranteed to have greater ex ante utility than p. Hence, no such plan exists. So, in particular, $p(n_b) \in CDT(d_b)$. □

Proponents of CDT thus also have an escape route by which to flee the negative verdict of Proposition 3, albeit a more demanding one than that afforded by Backward Autonomy to proponents of EDT. That is, the causalist can refute the charge of problematic dynamic inconsistencies by embracing Full Autonomy.

9 Conclusion

The happy consequences of Backward Autonomy and Full Autonomy notwithstanding, these are very strong principles that many a decision theorist will no doubt be inclined to resist as strict conditions of epistemic coherence. Full Autonomy, for example, essentially precludes ideally rational agents from ever facing Newcomb problems or from holding their own choices to be potentially correlated with causally irrelevant states of the world, e.g. the behavior of a physically remote twin. This may be a bridge too far for some, and I can't say I lack all sympathy for such skeptics. Still, the retreat from Full Autonomy is an unpleasant one. Backward Autonomy is logically weaker and does allow for the coherence of Newcomb problems, but reaping its benefits vis-a-vis dynamic consistency requires rejecting CDT in favor of EDT, which many theorists will be reluctant to do. And, in any case, it shares much of the same radical spirit as Full Autonomy anyway. Further, in light of Proposition 3, without some sort of restriction on coherence beyond the standard canons of Bayesianism, alternative decision theories also provide no safe harbor to sail to, provided we accept the eminently plausible constraints of independence and conservatism.

The rational choice theorist's dilemma then comes down to either embracing a radical epistemic principle a la Full or Backward Autonomy or else abandoning dynamic consistency as a constraint on normative decision theories.[22] Given the powerful appeal of dynamic consistency, it is worth considering whether autonomy-like principles are really so implausible as characterizations of ideal rationality. A typical human agent will no doubt violate both Full and Backward Autonomy, but then again typical human agents violate all the standard rationality postulates of Bayesian decision theories. Might there be principled epistemic reasons (beyond the brute motivation to avoid dynamic inconsistency) why an agent more collected and more reflective than we would tend to satisfy either Full or Backward Autonomy?

One route to motivating such a view might come via something like Ellery Eells' famous *tickle defense* of EDT.[23] According to Eells, a sufficient degree of self-awareness ought to screen off perceived correlations between an agent's decisions and causally irrelevant states of the world. So, for example, while my behavior might provide evidence regarding that of my causally isolated twin by way of revealing something about our shared action-guiding attitudes (e.g. beliefs and desires), once these decision inputs are fixed the evidential connection between our behaviors may be broken. Eells held that such self-awareness is partially constitutive of practical rationality since an awareness of one's own probabilities and utilities is crucial to the first-person application of decision theory. If correct, this observation would obviate the need for CDT, at least

[22] A possible third possibility, also unpleasant, would be to accept dynamic consistency, reject autonomy, and simply deny that there is any fully normatively adequate decision theory at all, a conclusion perhaps also suggested by the impossibility results of Briggs 2010.

[23] Eells 1982/2016.

in the case of ideally rational agents, since its verdicts (as well as those of any other conservative decision theory for that matter) would always agree with those of EDT. Eells seems to be getting at something close to Full Autonomy here by an entirely different route.

We might also instead view Full or Backward Autonomy as a requirement characterizing the ideal endpoint of rational deliberation. This thought seems to be in line with Eells' mature formulation of his defense of EDT in terms of *deliberational metatickles*.[24] On this view, an agent might reasonably start out deliberation with non-autonomous credences, but as she deliberates she goes through a process that includes a growing awareness of her own attitudes and dispositions that ultimately screens her choices off from all but causally downstream states of affairs. An exact dynamics that conforms to this desideratum and its implication for contexts of sequential choice are well worth investigating.[25]

While hopefully suggestive, such remarks are far from decisive. One significant concern to raise about the prospects for deliberational metatickles to guarantee convergence of an ideal agent's credences to a (fully or backward) autonomous limit, is that the Eell's story may not really justify autonomy in cases of practical indifference where a rational agent's decision is not deterministically fixed by her beliefs and desires. For example, if I am indifferent between A and B perhaps my eventual choice of A over B can tell me something about past states of the world via indicating what tie-breaking procedure I used, which may in turn be correlated with with, e.g., a friend's tie-breaking procedure, a predictor's prediction, etc. So caution is warranted in trying to use any sort of Eellsian story to justify an identification of coherent decision problems with the fully autonomous. Still, I'm not convinced that no such story can be told, though to tell it would at least require an involved and likely controversial analysis of the nature of rational choice under indifference.

Whether independent and more direct support can be found for either Full or Backward Autonomy, this much is clear: Causalist and evidentialists will be hard pressed to defend the dynamic consistency of their theories without embracing something close to these principles. To the extent that we are attracted to dynamic consistency, we are then lead to the surprising and radical consequence that, in so far as we live up to the ideals of rationality, the evidential significance of our plans is sharply constrained, in one way or another, by their causal significance.

[24] Eells 1984

[25] Note that such a dynamic reading of Full Autonomy may, unlike a static reading, leave room for practical divergence between CDT and EDT if an agent's choice of decision theory impacts which fully autonomous credal state her deliberation terminates in. For more on deliberational dynamics, albeit not versions that hope to offer convergence to Autonomy, see Skyrms 1990, Arntzenius 2008, Joyce 2012, and Lauro and Huttegger 2020.

Works Cited

Ahmed, Arif. "Dicing with Death". *Analysis* 74(4) (2014): 587–92. Print.

——.*Evidence, Decision and Causality*. Cambridge: Cambridge University Press, 2014. Print.

Armendt, Brad. "Causal Decision Theory and Decision Instability". *Journal of Philosophy* 116(5) (2019): 263–77. Print.

Arntzenius, Frank. "No Regrets, or: Edith Piaf Revamps Decision Theory". *Erkenntnis* 68 (2008): 277–97. Print.

Bradley, Richard. *Decision Theory with a Human Face*. Cambridge: Cambridge University Press, 2017. Print.

Briggs, Rachael. "Decision-Theoretic Paradoxes as Voting Paradoxes". *Philosophical Review* 119 (2010): 1–30. Print.

Buchak, Lara. *Risk and Rationality*. Oxford: Oxford University Press, 2013. Print.

Cantwell, John. "On an alleged counter-example to causal decision theory". *Synthese* 173 (2010): 127–152. Print.

Cubitt, Robin. "Rational Dynamic Choice and Expected Utility Theory". *Oxford Economic Papers* 48 (1996): 1–19. Print.

Dohrn, Daniel. "Egan and Agents: How Evidential Decision Theory Can Deal with Egan's Dilemma". *Synthese* 192(6) (2015): 1883–1908. Print.

Eells, Ellery. "Metatickles and the Dynamics of Deliberation". *Theory and Decision* 17 (1984): 71–95. Print.

——.*Rational Decision and Causality*. Cambridge: Cambridge University Press, 1982/2016. Print.

Egan, Andy. "Some Counterexamples to Causal Decision Theory". *Philosophical Review* 116 (2007): 93–114. Print.

Fusco, Melissa. "Epistemic Time Bias in Newcomb's Problem". *Newcomb's Problem*. Edited by Arif Ahmed. Cambridge: Cambridge University Press, 2018. 73–95. Print.

Gallow, Dmitri. "The Causalist's Guide to Managing the News". *Journal of Philosophy* 117(3) (2020): 117–49. Print.

Gibbard, Allan. "Weakly Self-Ratifying Strategies: Comments on McClennen". *Philosophical Studies* 34 (1992): 217–25. Print.

Gibbard, Allan and William Harper. "Counterfactuals and Two Kinds of Expected Utility". *Foundations and Applications of Decision Theory*. Edited by J. Leach C. Hooker and E. McClennen. Reidel: Dordrecht, 1978. 125–162. Print.

Hammond, Peter. "Consequentialist Foundations for Expected Utility". *Theory and Decision* 25 (1988): 25–78. Print.

Hedden, Brian. *Reasons without Persons*. Oxford: Oxford University Press, 2015. Print.

Hitchcock, Christopher. "Conditioning, Intervening, and Decision". *Synthese* 193(4) (2016): 1157–1176. Print.

Huttegger, Simon and Gerard Rothfus. "Bradley Conditionals and Dynamic Choice". *Synthese* (2021). Print.

Jeffrey, Richard. *The Logic of Decision*. Chicago: University of Chicago Press, 1965/1983. Print.
Joyce, James. "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems". *Newcomb's Problem*. Edited by Arif Ahmed. Cambridge: Cambridge University Press, 2018. 138–159. Print.
——."Regret and Instability in Causal Decision Theory". *Synthese* 187 (2012): 123–45. Print.
——."Review of Arif Ahmed: Evidence, Decision and Causality". *Journal of Philosophy* 113(4) (2016): 224–232. Print.
——.*The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press, 1999. Print.
Lauro, Greg and Simon Huttegger. "Structural Stability in Causal Decision Theory". *Erkenntnis* (2020): 1–19. Print.
Levinstein, Benjamin and Nate Soares. "Cheating Death in Damascus". *Journal of Philosophy* 117(5) (2020): 237–66. Print.
Lewis, David. "Causal Decision Theory". *Australasian Journal of Philosophy* 59 (1981): 5–30. Print.
——."Prisoners' Dilemma is a Newcomb Problem". *Philosophy and Public Affairs* 8(3) (1979): 235–40. Print.
McClennen, Edward. *Rationality and Dynamic Choice*. Cambridge: Cambridge Univeristy Press, 1990. Print.
Meacham, Christopher. "Binding and Its Consequences". *Philosophical Studies* 149 (2010): 49–71. Print.
Nozick, Robert. "Newcomb's Problem and Two Principles of Choice". *Essays in Honor of Carl G. Hempel*. Edited by Nicholas Rescher. Reidel: Dordrecht, 1969. 107–33. Print.
Oestherheld, Caspar and Vincent Conitzer. "Extracting Money from Causal Decision Theorists". *Philosophical Quarterly* 71(4) (2021). Print.
Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2009. Print.
Rothfus, Gerard. "A Plan-Based Causal Decision Theory". *Analysis* (forthcoming). Print.
——."Dynamic Consistency in the Logic of Decision". *Philosophical Studies* 117(12) (2020): 3923–34. Print.
Skyrms, Brian. *Causal Necessity*. New Haven: Yale University Press, 1980. Print.
——.*The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press, 1990. Print.
Spencer, Jack. "An Argument Against Causal Decision Theory". *Analysis* 81 (1) (2021): 52–61. Print.
Spencer, Jack and Ian Wells. "Why Take Both Boxes?" *Philosophy and Phenomenological Research* (2017). Print.
Spohn, Wolfgang. *Reflexive Rationality: Rethinking Decision and Game Theory*. unpublished. Print.
Steele, Katie. "What Are the Minimal Requirements of Rational Choice? Arguments From the Sequential-Decision Setting". *Theory and Decision* 68 (2010): 463–487. Print.

Stern, Reuben. "Diagnosing Newcomb's Problem with Causal Graphs". *Newcomb's Problem*. Edited by Arif Ahmed. Cambridge: Cambridge University Press, 2018. 201–220. Print.
Titelbaum, Michael. *Quitting Certainties*. Oxford: Oxford University Press, 2013. Print.
Weatherson, Brian. "Indecisive Decision Theory" (MS). Print.
Wedgewood, Ralph. "Gandalf's Solution to the Newcomb Problem". *Synthese* 14 (2011): 1–33. Print.
Wells, Ian. "Equal Opportunity and Newcomb's Problem". *Mind* (2018). Print.
Williamson, Timothy Luke. "Causal Decision Theory is Safe from Psychopaths". *Erkenntnis* 86 (3) (2021): 665–685. Print.
Yudkowsky, Eliezer. "Timeless Decision Theory". *Technical Report, Machine Intelligence Research Institute (MIRI)* (2010): www.intelligence.org. Print.