A plan-based causal decision theory

Gerard J. Rothfus

1. *Introduction*

Spencer and Wells (2019) pose a serious challenge for causal decision theory (CDT) in the form of their *Frustrator* problem, in which you must decide between taking one of three items: Box A, Box B or an envelope. The envelope contains $40, while the boxes contain a collective total of $100. However, the distribution of the $100 across the boxes is uncertain. In fact, it was determined by a Newcombian predictor who placed the $100 in Box A if she predicted you would take Box B and placed it in Box B if she predicted you would take Box A. If the predictor predicted that you would opt for the envelope, she split the money evenly between the boxes, placing $50 in each. Assume that you are aware of this, that your basic values align with your potential monetary gains and that you have no causal power over the past. In this case, CDT recommends taking either Box A or Box B, depending upon how likely you judge the predictor to have predicted each of these choices.

This is a remarkably counterintuitive verdict and, as such, casts some doubt on the normative adequacy of CDT. Still, the case is not decisive. *Frustrator* is kin to a number of cases in which CDT offers similarly counterintutive verdicts whose propriety has nonetheless been skilfully defended by CDT's proponents.[1] Given a fixed credal distribution over the causal hypotheses at play, opting for one of the boxes is evidently not without a causalist rationale, in spite of its seeming folly.

To turn the screw on the causalist a bit further then, Spencer (2021) has proposed *Two Rooms*, a sequentialized variant of *Frustrator* that appears significantly more worrisome for CDT. You are offered the choice to go into one of two rooms. In the first, you will receive $35 for sure. In the second, you will face the *Frustrator* problem. Which room should you enter? According to Spencer, CDT recommends entering the first room to take the $35.[2] If so, this is an unfortunate result for the causalist. *Frustrator* includes an option that, if chosen, guarantees you $40. How could an instrumentally rational agent disprefer facing such a problem to a lump sum strictly less than $40?

Spencer argues that they could not. In particular, he suggests that rationality requires satisfaction of his *Guaranteed Principle* (GP), according to which a

1 See, for example, Egan 2007 and Ahmed 2014a. For causalist replies, see Joyce 2012, 2018, Armendt 2019, and Williamson 2021.

2 Assuming that a causalist facing the problem expects to remain one throughout.

Analysis Vol. 00 | Number 0 | 2022 | pp. 1–9 https://doi.org/10.1093/analys/anab064

rational agent will always prefer a decision with an option that guarantees a minimum utility of x to a sure outcome valued at a utility less than x. (Or, at least, such will be the case for a rational agent who expects to remain rational.) Whether we embrace the GP or not, however, it is difficult to deny that there is something highly awkward about the conjunction of recommendations Spencer ascribes to CDT. In *Frustrator*, a box is chosen over \$40, while in *Two Rooms* the opportunity to take an identical box is not even valued at \$35!

Some causalists will be tempted to defend this infelicitous conjunction (see e.g. Joyce MS), but the point of the present note is to probe the prospects for resisting it. To this end, §2 rehearses Spencer's argument that CDT recommends entering the first room, at least on one popular approach to dynamic choice. §3 then sketches an alternative plan-based approach to dynamic choice that permits CDT to evade any violation of GP in *Two Rooms*. An epistemological objection to this approach naturally suggests itself but can be answered via deliberational dynamics (§4). Finally, the question is raised whether the proposed plan-based version of CDT, which eschews the asymmetric appraisal of present and future acts that caused the violation of the GP in *Two Rooms*, will respect the GP in general, resulting in a negative verdict of independent interest for the CDT debate (§5). These remarks collectively invite further consideration of the application of CDT to dynamic choice problems (§6).

2. *Spencer's argument*

According to CDT, rational agents maximize a utility function U with the following property:

> $U(X) = \sum_i P_X(O_i)U(O_i)$, for any proposition X and partition of outcome-propositions $\{O_i\}_i$.[3]

Here, $P_X(\cdot)$ is the agent's probabilistic credence function on the *causal* supposition that X. There are various ways one might understand such probabilities, for example as probabilities of appropriate non-backtracking counterfactuals, as conditional expectations of objective chance etc. Such details, however, need not detain us, as the causal probabilities at play in Spencer's argument will be sufficiently clear without settling on any particular explication. Replacing the causal probabilities with standard conditional ones results, of course, in CDT's chief rival, evidential decision theory (EDT).[4]

Why does Spencer think that CDT recommends heading to the first room in *Two Rooms*, thus violating the GP? Well, contrast the utilities you assign to the two propositions you can immediately make true at the first stage of the

3 An *outcome-proposition* is a proposition strong enough to fix everything an agent cares about.

4 On EDT, see Jeffrey 1983 [1965], Ahmed 2014b. On CDT, see Gibbard and Harper 1978, Skyrms 1980, Lewis 1981, Joyce 1999.

problem: namely, that you enter the first room (R_1) and that you enter the second (R_2). Sticking with our assumption that utilities coincide with monetary payoffs on propositions strong enough to fix them, clearly $U(R_1) = 35$. What about $U(R_2)$? Spencer sensibly assumes that you, as a sophisticated causalist agent, will employ some backward induction and realize that, were you to enter the second room, you would opt to take either Box A or Box B. Given this prediction, we may easily compute $U(R_2)$ relative to the partition of outcome-propositions given by the possible combinations of your choices and the predictor's predictions in the *Frustrator* problem. Letting A denote the proposition that you take Box A, A^* that the predictor predicted thus etc., and dropping the possibilities that yield zero utility in the sum, we have:

$$\begin{aligned} U(R_2) &= P_{R_2}(BA^*)100 + P_{R_2}(EA^*)40 + P_{R_2}(AB^*)100 + P_{R_2}(EB^*)40 \\ &\quad + P_{R_2}(AE^*)50 + P_{R_2}(BE^*)50 + P_{R_2}(EE^*)40 \\ &= P_{R_2}(BA^*)100 + P_{R_2}(AB^*)100 + P_{R_2}(AE^*)50 + P_{R_2}(BE^*)50 \\ &\approx 0 \end{aligned}$$

This last line is justified by the fact that, given the assumed reliability of the predictor, you take it to be very unlikely that the predictor would mispredict you, and a causal supposition that you will enter the second room does nothing to alter your confidence in this regard. Since the utility of R_2 is then judged to be near 0, $U(R_1) > U(R_2)$, and thus a disciple of CDT will head to the first room.

3. *Plan-CDT*

Crucial to the argument that CDT recommends entering the first room is the assumption that the set of items relative to which CDT instructs an agent to maximize her utility includes all and only her immediately available acts or *options*, that is, the strongest propositions she can now make true by will. At the start of *Two Rooms*, this set includes only R_1 and R_2, whence the inference that CDT recommends R_1. However, in sequential choice problems like *Two Rooms*, an agent's set of currently available options is not the only set that she might reasonably be expected to maximize utility with respect to. An alternative approach would have her initiate *plans* that maximize utility (subject to feasibility constraints), where a plan can be thought of as specifying a complete sequence of options that an agent facing a given sequential choice problem might take in turn.[5] *Two Rooms*, for example, involves four plans: $\{R_1, R_2A, R_2B, R_2E\}$.

Call decision theories that view the proper objects of practical deliberation in the first way, that is, as immediately available acts or options, *Act decision*

5 Generally speaking, plans must specify acts conditional upon contemplated contingencies, and so are better thought of as conjunctions of appropriately nested conditionals; see Rothfus 2020 and Huttegger and Rothfus forthcoming. However, the structure of *Two Rooms* is sufficiently simple that a conditional treatment of plans is unnecessary here.

theories, and those that opt for the latter characterization in terms of (possibly) temporally extended plans, *Plan decision theories*. This distinction allows us to contrast *Act-CDT* with *Plan-CDT*:

(Act-CDT) Given a set of immediately available acts, $\mathcal{A}$, select an option $X \in \mathcal{A}$ such that $U(X) \geq U(X')$, for all $X' \in \mathcal{A}$.

(Plan-CDT) Given a set of immediately available acts, $\mathcal{A}$, which form the initial segments of a set of available plans Π, select an option $X \in \mathcal{A}$ that maximizes $\sum_{\pi \in \Pi} P_X(\pi)U(\pi)$.

While Act-CDT agents want to select acts that maximize utility in themselves, Plan-CDT agents want to select acts that initiate plans of greatest possible utility. With this distinction in hand, we can verify (as I do below) that it is only Act-CDT, and not Plan-CDT, that is at odds with the GP in *Two Rooms*.[6]

Spencer considers a response to his argument that may appear to involve something like Plan-CDT. In particular, he claims that if the four plans available in *Two Rooms* were all available to the deliberating agent as *diachronic options*, then CDT would indeed recommend entering the second room and taking one of the boxes, failing to contravene the GP. However, Spencer denies the possibility (or at least availability) of diachronic options, and hence dismisses the objection. The proponent of Plan-CDT concurs. There are only synchronic options and diachronic plans, that is, temporally ordered sequences of synchronic options. All that Plan-CDT claims is that the choice-worthiness of a (synchronic) option is fixed not by its own utility taken in isolation, but by (the expectation of) the utility of the (diachronic) plan it initiates.

Why prefer CDT's Plan version to its Act version? Well, to the extent that we wish to avoid violating the GP in *Two Rooms*, we have some grounds for such a preference. Speaking more generally, however, Plan-CDT is attractive because it eliminates the awkward double-think that Act-CDT's approach to sequential choice requires a causalist to adopt. Considered as an option that you can carry out now, choosing Box A is judged by its utility and is worth more to you than \$40, but considered as a future possibility, choosing Box A is treated as a state of the world and judged by its desirability and hence worth less than \$35 to you. Causalists can escape this embarrassment by embracing the more unified approach to dynamic choice encoded in Plan-CDT.

One aspect of Plan-CDT that merits comment concerns the plan-probabilities it employs. How should a rational agent assess the probability

6 In trivial dynamic choice problems involving only one decision, these decision rules agree. Given the prevalence of such *static choice* problems in the early CDT literature, it is unsurprising that these alternative perspectives were never distinguished. However, with the recent introduction of dynamic choice examples into the CDT debate (e.g. in addition to Spencer, see Ahmed 2014b, Oestherheld and Conitzer forthcoming), the distinction becomes pertinent. It is worth noting as well that some dynamic choice arguments against CDT (e.g. those in Ahmed 2014b) can be directed against both Plan- and Act-CDT.

that a given act will initiate a particular plan? The principal assumption made here is that such judgements are constrained by backward induction. Restricting attention to finite sequential choice problems, a rational agent faced with a series of decisions can first apply her choice rule (e.g. Plan-CDT) to figure out what she might do at the problem's possible final stages. In light of such determinations, she can then continue to work her way backward through the problem at hand to figure out what plans her initial choices might actually initiate. A rational agent (who expects to remain rational) will thus only assign positive value to $P_X(\pi)$ if π is a plan that survives this pruning procedure.[7]

Consider a naive application of this assumption in *Two Rooms*. What would you choose if you entered the second room? Well, we know that, depending on how you spread your credence between A^* and B^*, you will opt for either A or B. Let us suppose for concreteness that you are more confident in B^* than A^*, and so you would opt for A. This tells you that opting for R_2 initiates the plan R_2A, that is, $P_{R_2}(R_2A) = 1$. So, according to Plan-CDT, R_2 should be chosen over R_1 just in case $U(R_2A) > U(R_1)$, which we know it to be. Hence, a Plan-CDT agent with the posited credences faced with *Two Rooms* will enter the second room and then opt for Box A. CDT's alleged violation of the GP has disappeared.

4. *Sophisticated learning*

This resolution may seem a bit too quick. The sort of reasoning involved in applying the method of backward induction suggested above plausibly involves an unmentioned epistemic component: when you realize that an option appears suboptimal at a future stage, you learn that you will not take it, inducing an appropriate shift in your credences. Such shifts may cause plans that initially appeared favourable to seem considerably less so, threatening the inference that an initially optimal plan that survives the pruning process of a sophisticated deliberator will continue to appear optimal at the end of the process. Suppose you are again the sophisticated Plan-CDT agent faced with *Two Rooms*. You may indeed conclude that you will find Box A most attractive if you enter the second room, rendering R_2A the plan initiated by selecting R_2 at the problem's outset. But realizing this substantially detracts from the estimated utility of plan R_2A. You only initially judged R_2A as optimal because of the relatively high credence you placed in B^* against A^*, but if you come to believe that you would choose Box A in the second room, the basis for this valuation is lost. You will rather come to see A^* as highly probable, given your confidence in the predictor's reliability. Contrasting the utility

7 This method is a version of Hammond's (1976) *sophisticated choice*. Note that Plan decision theories are *not* committed to McClennen's (1990) alternative doctrine of *resolute choice* or its cousins, for example Meacham's (2010) *cohesive expected utility theory*.

of R_2A and R_1 from this altered epistemic perspective results in the verdict to enter the first room. CDT's commitment to entering the first room seems to re-emerge once we account for the sort of learning potentially wrought by an application of sophisticated choice.

But this story is still incomplete. If backward induction reasoning has the epistemic significance suggested here, then not only will you come to see the plan R_2A as worse than R_1, you will also come to see choosing Box A in the second room as a bad choice compared with your other available options, namely, choosing either Box B or the envelope. So you have lost your basis for thinking it would be optimal to choose Box A if you were to reach the second room and therefore also the basis of your prediction that you would so choose. This suggests a need to repeat the application of backward induction with your newly revised probabilities.

There is a reasonable worry here that this reasoning is doomed to generate an endless cycle of inferences that never terminates in clear guidance for action. We can lessen this worry and achieve greater prospects for convergence by recognizing that rational agents will carry out a somewhat similar process even in contexts of static choice, like the *Frustrator* problem faced in the second room. In choice problems where your attitudes and/or decisions may provide evidence regarding the causal structure of the problem you face, it makes sense from a causalist perspective to avoid jumping to conclusions about what you should do. In such cases, your initial credences may not adequately factor in all the information about the world capable of being gleaned from features of your own decision-making process, and so it may be overhasty to simply maximize utility relative to your initial credences. Rather, you might engage in a process of *deliberational dynamics* in which you only modestly adjust your credences in your option set in light of your calculations of expected utility, recalculate those expectations in light of your altered probabilities, and repeat. (See Skyrms 1990, Arntzenius 2008, Joyce 2012.) In relatively clean cases like *Frustrator*, the assumptions needed to ensure that such a dynamical process terminates in an equilibrium are fairly weak. As has been argued elsewhere (e.g. in Joyce 2018), the most plausible end point for the causalist facing *Frustrator* involves assigning probability 0.5 to each of A and B. From this epistemic perspective, both boxes appear equally desirable, while the envelope appears less so.

If a rational agent knows that she will employ such a deliberational process before making decisions, then she will account for this in her backward induction. In *Two Rooms*, this means that you can foresee from the start of the problem that, if you enter the second room, you will wind up in a state of indifference between choosing Box A and Box B. So the information you learn as a result of this process is neither $R_2 \supset A$ nor $R_2 \supset B$ but simply $R_2 \supset (A \vee B)$. And this information is insufficient to make the utility of R_1 greater than the utility of either R_2A or R_2B. So the epistemic standpoint you arrive at does not render R_1 the optimal plan. That said, it is also no longer clear

from this perspective which plan R_2 will initiate. It might initiate R_2A or it might initiate R_2B. Assuming that your initial credences defer (conditionally) to the epistemic state you realize your deliberation would terminate in if you were to enter the second room, you will judge that $P(A^*) = P(B^*)$ and hence that $U(R_2A) = U(R_2B)$. So it does not matter which of these two plans is initiated by R_2; they each have the same utility from your standpoint, and you know it to be greater than $U(R_1\overline{E})$. And that suffices for Plan-CDT to recommend R_2.

5. *Two Drawers*

One might hope that in Plan-CDT we have found an irenic resolution of the apparent conflict between CDT and the GP. Alas, not so. Even though Plan-CDT eschews the asymmetric appraisal of present and future acts that caused the violation of the GP in *Two Rooms*, it still leaves the door open to possible violations of the GP in other problems that merit consideration in their own right.

Having lost the lease on her spacious *Two Rooms* setup, our Newcombian predictor now offers you a choice between opening one of two drawers. The first drawer holds a sure \$25, while the second contains a closed envelope lying beside \$50. If you opt to open the second drawer, you will be allowed to take either the envelope or the \$50, but not both. You are informed that the envelope contains \$100 if and only if the predictor predicted that you would choose to open the second drawer. Otherwise, it is empty. Call this the *Two Drawers* problem.

Clearly, any decision theory that recommends opening the first drawer is in tension with the GP: \$50 is surely better than \$25. However, both Act- and Plan-CDT may recommend opening the first drawer, depending upon your initial credences regarding the predictor's choices. Your initial options are open the first drawer (D_1) and open the second drawer (D_2), while your potential subsequent options are take the envelope (E) or take the \$50 ($\overline{E}$). Clearly $U(D_1) = 25$. Writing D_1^* for the proposition that the predictor predicted D_1 and D_2^* for its negation, we can also compute:

$$\begin{aligned} U(D_2) &= P_{D_2}(ED_1^*)0 + P_{D_2}(\overline{E}D_1^*)50 + P_{D_2}(ED_2^*)100 + P_{D_2}(\overline{E}D_2^*)50 \\ &= P_{D_2}(\overline{E})50 + P_{D_2}(ED_2^*)100 \end{aligned}$$

We have no general guarantee that this quantity will be greater than 25, and hence no guarantee that Act-CDT will not recommend opening the first drawer, violating the GP yet again. Whether it does so or not depends on the values of $P_{D_2}(\overline{E})$ and $P_{D_2}(ED_2^*)$, neither of which is required to reach any threshold in order, for example, to cohere with your recognition of the predictor's reliability. In fact, sophistication demands a null value for $P_{D_2}(\overline{E})$, since selecting the envelope maximizes expected utility, conditional upon having opened the second drawer. Given this, $P_{D_2}(ED_2^*) = P(D_2^*)$.[8] So, in this case, Act-CDT will recommend D_1 as long as $P(D_2^*) < .25$, that is, as long as you start out relatively confident that the predictor thought you would choose D_1.

In *Two Drawers*, unlike *Two Rooms*, Plan-CDT offers no refuge for causalists attracted to the GP to flee to. With the same credences suggested above, a sophisticated Plan-CDT agent will realize that opening the second drawer and selecting the \$50 is an infeasible plan, and hence her concern is only between the plans D_1 and D_2E, the latter of which will appear bleak if, as above, $P(D_2^*) < 0.25$. Faced with such a choice, the Plan-CDT agent will opt for the first drawer as well. Whether this case presents a serious obstacle in its own right to the normative adequacy of CDT, whether act- or plan-based, merits further reflection.

6. *Conclusion*

Spencer's *Two Rooms* underwrites a powerful objection to Act-CDT. But there is no reason that a causalist, qua causalist, should be committed to Act-CDT. The guiding idea of CDT is that an item of practical deliberation should be judged according to its utility or efficacy value. But this leaves open whether the objects that ought to be so assessed include extended plans or only isolated acts. Opting for the former view and embracing Plan-CDT enables one to avoid embarrassing discrepancies between one's evaluations of present vs future options. As *Two Drawers* shows, this move is no general panacea when it comes to fixing CDT's dynamic difficulties, but even modest progress is worth weighing as causalists assess how to apply their theory in sequential contexts.[9]

University of Konstanz, Germany
gerard.rothfus@uni-konstanz.de

References

Ahmed, A. 2014a. Dicing with death. *Analysis* 74: 587–92.

Ahmed, A. 2014b. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.

Armendt, B. 2019. Causal decision theory and decision instability. *Journal of Philosophy* 116: 263–77.

Arntzenius, F. 2008. No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis* 68: 277–97.

8 This follows from sophistication, that is, $P_{D_2}(E) = 1$, together with the causal independence of D_2 and D_2^*, that is, $P_{D_2}(D_2^*) = P(D_2^*)$.

9 Many thanks to Daniel Herrmann, Calum McNamara, Wolfgang Spohn, two anonymous referees and participants in the U.C. Irvine formal epistemology reading group for exceptionally helpful feedback and discussion on the substance and style of this paper. The grant is from the German Science Foundation (within the Project SP 279/21-1 (Project No. 420094936)).

Egan, A. 2007. Some counterexamples to causal decision theory. *Philosophical Review* 116: 93–114.

Gibbard, A. and W. Harper. 1978. Counterfactuals and two kinds of expected utility. In *Foundations and Applications of Decision Theory*, eds. J. L. C. Hooker and E. McClennen, 125–62. Dordrecht: Reidel.

Hammond, P. 1976. Changing tastes and coherent dynamic choice. *Review of Economic Studies* 43: 159–73.

Huttegger, S. and G. Rothfus. forthcoming. Bradley conditionals and dynamic choice. *Synthese*

Jeffrey, R. 1983 [1965]. *The Logic of Decision*. Chicago: University of Chicago Press.

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

Joyce, J. 2012. Regret and instability in causal decision theory. *Synthese* 187: 123–45.

Joyce, J. 2018. Deliberation and stability in Newcomb problems and pseudo-Newcomb problems. In *Newcomb's Problem*, ed. A. Ahmed, 138–59. Cambridge: Cambridge University Press.

Joyce, J. MS. Yet another refutation of causal decision theory? Unpublished manuscript.

Lewis, D. 1981. Causal decision theory. *Australasian Journal of Philosophy* 59: 5–30.

McClennen, E. 1990. *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.

Meacham, C. 2010. Binding and its consequences. *Philosophical Studies* 149: 49–71.

Oestherheld, C. and V. Conitzer. forthcoming. Extracting money from causal decision theorists. *Philosophical Quarterly* 71.

Rothfus, G. 2020. Dynamic consistency in the logic of decision. *Philosophical Studies* 117: 3923–34.

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

Skyrms, B. 1990. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Spencer, J. 2021. An argument against causal decision theory. *Analysis* 81: 52–61.

Spencer, J. and I. Wells. 2019. Why take both boxes? *Philosophy and Phenomenological Research* 99: 27–48.

Williamson, T.L. 2021. Causal decision theory is safe from psychopaths. *Erkenntnis* 86: 665–85.

Abstract

In 'An argument against causal decision theory', Jack Spencer shows that standard formulations of causal decision theory run afoul of his Guaranteed Principle. In the sequential choice problem he employs to make this case, the transgression stems from an awkward discrepancy between how causalists typically value present vs future acts. This note suggests a version of causal decision theory that avoids this incongruity and so respects the Guaranteed Principle in Spencer's problem. However, this formulation, and hence symmetric appraisal of present and future acts, is also shown to be insufficient to secure causalist satisfaction of the Guaranteed Principle in general.

Keywords: Causal decision theory, Sequential choice, Rational planning