This is an extended draft of a paper that will eventually appear as

"Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb Problems", in Arif Ahmed, ed., *Newcomb's Problem*: Oxford University Press, 2018.

Material in blue does not appear in that paper, and should be cited using this version. Material in black should be cited as coming from the Ahmed volume, once it appears.

Draft: Newcomb Problems and Pseudo-Newcomb Problems[1]

James M. Joyce
The University of Michigan, Ann Arbor

This essay defends *causal decision theory* (CDT) against some alleged counterexamples that proponents of evidential decision theory (EDT) have raised against it. I argue that sophisticated *deliberational* versions of CDT, pioneered by Skyrms (1982, 1990) and elaborated in Arntzenius (2008) and Joyce (2012), can defuse any of these counterexamples.

The paper has six sections. §1 distinguishes *Newcomb problems* from *pseudo-Newcomb problems*. §2 addresses predictability and freedom. §3 distinguishes CDT and EDT. §4 defends CDT's handling of Newcomb problems. §5 introduces the notion of deliberational equilibrium, and distinguishes picking from choosing. §6 considers "unstable" decisions and defuses counterexamples from Spencer and Wells (2018) and Ahmed (2014b).

§1 Newcomb Problems

In *Newcomb problems* choices correlate with features of the world that choosers cannot causally influence. As a result, acts that *cause* desirable/undesirable future results can also *indicate* undesirable/desirable past events, leaving agents to wonder whether to causally promote desirable outcomes or to produce news of desirable outcomes they do not control.

All Newcomb problems can be subsumed under a common rubric. Imagine an idealized agent, let's say *you*, who is now (time t_1) facing a decision, and a predictor, *Omega*, who at *past* time t_0 made a guess about how you would act on the basis of an examination of your t_0 brain-state (a common cause of both his guess and your choice). Things are arranged so that the

[1] This essay benefited greatly from discussions with Arif Ahmed, Brad Armendt, Kevin Blackwell, Simon Huttegger, Sarah Moss, Huw Price and Brian Skyrms.

outcome of your choice depends on Omega's guess, but your only evidence about this comes via your knowledge of own current beliefs and desires, which are clues to your t_0 brain-state.

For our purposes, any (ideal) Newcomb problem satisfies the following:

NP$_1$ For each possible act *A* there is an associated 'type-*A*' brain state, and you are subjectively certain that you will do *A* at t_1 iff you occupied the type-*A* state at t_0, hereafter A^τ.[2] You typically will *not* know your type until you know what you (irrevocably) choose.

NP$_2$ Omega tried to discern your type at t_0, and guessed that you would choose *A*, hereafter A^π, iff he identified you as type-*A*. Omega may or may not be a perfect identifier of types. But, for all acts *A* and *B*, you know Omega's chances of misclassifying you as type-*B* when you are really type-*A*.

NP$_3$ The past is *fixed*. You cannot now change your t_0 state, Omega's prediction, or any other past fact.

NP$_4$ Omega is *reliable* i.e., better than chance at predicting your act. Moreover, absent further evidence, your confidence in Omega's reliability is constant throughout the decision-making process.

NP$_5$ You are *free*. No obstacles prevent you from choosing whichever act you ultimately judge best. Crucially, whatever Omega predicted and however reliable he is, you have the power to falsify his prediction. You might not want to exercise this power and might be sure that you will not, but you can.

I take ***NP***$_1$-***NP***$_5$, as elaborated below, as *definitive* of Newcomb problems. People frequently mistake decisions satisfying only some of these conditions for genuine Newcomb problems, and wrongly portray solutions to such *pseudo*-Newcomb problems as answers to the real thing.

To explain ***NP***$_1$-***NP***$_5$ it helps to have a formal model of you as a decision maker.[3] You face a free choice among *acts* $\{A_1, A_2,\ldots, A_M\}$ whose *outcomes* depend on which member of a partition $\{S_1, S_2,\ldots, S_N\}$ of *states* obtains. Each act/state pair fixes an outcome $O_{m,n}$ that encompasses all relevant consequences of A_m when S_n obtains. States describe features of the world that you cannot influence, but which may affect the outcome of your act. In light of ***NP***$_3$, each S_n will

[2] We could have let you be merely highly confident about type-act correlations without altering the discussion.

[3] This is admittedly an idealization, but in the all cases we consider real agents approximate the behavior of their ideal counterparts.

specify both your type and Omega's guess, i.e., each is a conjunction A_j^{τ} & A_i^{π} &…., the ellipsis capturing any further facts that affect outcomes.

At each time *t* you are endowed with a subjective *utility* function $util_t$ that measures the desirabilities of outcomes, and a *credence function* $prob_t$ that encodes your degrees of belief in events that might influence the outcome of your choice. Since we will not consider changes in desire, $util_t$ is assumed constant. There is some dispute over the proper domain of $prob_t$, but all agree that probabilities of states *conditional on acts* are well-defined. So, you always have a definite estimate, $prob_t(S_n/A_m)$, of the probability that S_n will obtain if you do A_m.[4]

To illustrate, consider first the *Flagship* Newcomb problem: you choose between act ONE of taking an opaque box that contains $1000000 iff Omega predicted that you would take *only* that box, and act TWO of taking the opaque box plus a transparent box containing $1000. We represent your decision thus ($1000 = 1 util):

Flagship

	ONE$^{\tau}$ & ONE$^{\pi}$	ONE$^{\tau}$ & TWO$^{\pi}$	TWO$^{\tau}$ & ONE$^{\pi}$	TWO$^{\tau}$ & TWO$^{\pi}$
ONE	1000 , p	\$0 , 1 – p	1000 , 0	0 , 0
TWO	1001, 0	1 , 0	1001, 1 – q	1, q

TABLE-1

Acts are at the left, states across the top. The first entry in each cell is the utility of the outcome received there. The second is your estimate of the probability of the cell's associated state *conditional on its act*. Your estimates of the probabilities that Omega correctly guessed your choices of ONE or TWO are $p = prob(\text{ONE}^{\tau}/\text{ONE})$ and $q = prob(\text{TWO}^{\tau}/\text{TWO})$, which both exceed 0.5 (***NP***$_4$). Both are one when Omega is a *perfect predictor*.

Gibbard & Harper's (1978) *Death-in-Damascus* provides another example. You are on the highway between Damascus and Aleppo, and will suffer a fate worse than death if you fail to arrive in one of the cities by nightfall. Alas, your prospects are little better if you do arrive. Yesterday Omega (irrevocably) predicted your destination, and sent assassins to the predicted city. To go there is to die; to go the other way is to live. With 1 = life and 0 = death, your decision is:

[4] ***NP***$_1$ requires $prob_t(S_n/A_m) = 0$ whenever S_n entails A_k^{τ} for $k \neq m$. ***NP***$_4$ requires that the sum of the $prob_t(S_n/A_m)$ for which S_m entails A_m^{π} (i.e., the probability that Omega correctly guessed A_m) exceeds ½.

DD	ALEP$^{\tau}$ & ALEP$^{\pi}$	ALEP$^{\tau}$ & DAM $^{\pi}$	DAM$^{\tau}$ & ALEP$^{\pi}$	DAM$^{\tau}$ & DAM$^{\pi}$
ALEP	0 , p	1 , 1 – p	0 , 0	1, 0
DAM	1 , 0	0 , 0	1 , 1 – q	0 , q

TABLE-2

Here you have a mortality rate of p = *prob*(ALEP$^{\tau}$/ALEP) or q = *prob*(DAM$^{\tau}$/DAM) depending on whether you choose Aleppo or Damascus. Again, we can imagine Omega as perfect, in which case your mortality rate is 1 wherever you go.

We discuss these examples in detail below, but first let's better understand ***NP***$_1$- ***NP***$_5$.

§2 Reconciling ***NP***$_1$-***NP***$_5$: How You Can Be Free While Omega is Reliable?

Newcomb problems asks us to square three seemingly irreconcilable beliefs. You, the agent, must be convinced that (a) you cannot affect Omega's prediction, (b) Omega *reliably* predicts your choices, but (c) you are free to falsify his prediction. But, how can you be sure that Omega's prediction and your choice will coincide unless one affects the other?

Let's start by asking how you can know you will choose *A* at t_1 iff you occupied the type-*A* brain state at t_0, as ***NP***$_1$ requires. The answer is that, you *count* as type-*A* at t_0 when, given your *actual* circumstances at t_0 and after, you end up disposed to rank *A* among your best options at t_1 and to pick it over similarly ranked options. Dependence on *actual* circumstances is crucial. As a type-*A* you will choose *A* in the actual world, but need *not* do so in other possible worlds (though "type-*A*" denotes a different brain state there). A two-box type in Flagship will actually choose TWO (probably because it dominates ONE). But, if something had disturbed its reasoning between t_0 and t_1 it might have chosen ONE. Thus, there is only a *contingent* connection between being in the type-*A* state at t_0 and choosing *A* at t_1. Differently put, *intrinsic* features of the type-*A* state (which "being type-*A*" is *not*) could cause different choices at t_1 depending on what transpires between t_0 and t_1.

The upshot is that your act is predictable from intrinsic features of your t_0 brain-state only insofar as your actual situation is known. This helps us understand ***NP***$_2$. Omega may know enough about your situation to deduce a definite t_1 choice from each intrinsically described t_0 brain state. Unless he is perfect, however, he will not know which intrinsic t_0 state you occupy, and so might misidentify your type. You, in contrast, might know more than Omega does about

your t_0 brain-state, but you might be unable to identify your type before t_1 because you lack information about the situation between t_0 and t_1. So, you can be certain you will choose *A* at t_1 iff you are type-*A* at t_0, and yet not know your type until you (irrevocably) decide what to do.

NP$_3$ requires you to regard the world's state as *causally* independent of your act. You must be confident that, whichever state is actual and whatever act you will in fact choose, the world would still have been in that same state had you chosen otherwise. This requirement of *causal act/state independence* has been elucidated in a variety of ways. Following Stalnaker (1967), Gibbard and Harper (1978) characterize it using subjunctive conditionals. On this model, you regard *S* as causally independent of *A* just when your credences for $A \mathbin{\square\!\!\rightarrow} S$ and *S* are equal. Lewis (1981) offers a similar analysis, but uses conditionals with chance consequents. Both models interpret subjunctive conditionals to exclude "backtrackers" like, "even though Omega correctly predicted TWO, he would have predicted ONE had I chosen ONE." Pearl (2000) employs a *do-operator* that conditions on *A* while holding past facts fixed.[5] *S* is then independent of *A* when $prob_t(S|\mathrm{do}A) = prob_t(S)$. Joyce (2010) advocates *Bayesian imaging* for these purposes. All these approaches are consistent with one another (Joyce, 2010). For our purposes it does no harm to use Gibbard/Harper's approach, which is a special case of the rest. This gives us:

NP$_3$ (*Causal Independence*) For each state *S* and acts *A* and *B*, you are certain that if *A* & *S* actually holds, then *S* would still have held had you chosen *B*, so that $prob_t(B \mathbin{\square\!\!\rightarrow} S \,/\, A \,\&\, S) = 1$ at all times *t*.

This entails that states are counterfactually independent of acts: $prob_t(A \mathbin{\square\!\!\rightarrow} S) = prob_t(S)$.

NP$_3$ requires you to regard the past as beyond your present influence. Specifically, you must believe that acting differently than you in fact do would not alter either your type or Omega's prediction, so that $prob_t(B \mathbin{\square\!\!\rightarrow} C^{\pi} \,\&\, D^{\tau} \,/\, A \,\&\, C^{\pi} \,\&\, D^{\tau}) = 1$ for any acts *A*, *B*, *C*, *D*. In Flagship, you must be convinced that if you take only the opaque box and get the million, then you would still have gotten the million had you taken both. In DD you must believe that if you are in fact going to die in Aleppo, then you would have lived had you fled to Damascus. Let me stress that *you are not in a Newcomb problem if you lack these sorts of beliefs*! Newcomb choosers are not deluded about their ability to change the past!

Turning now to ***NP***$_4$, we need to know what it means to regard Omega as *reliable*. Does it mean that he is likely to be correct *given what he guessed*, or *given what you choose*? It is the second notion that matters. For Omega's reliability to remain unchallenged up through the moment of choice, reliability in the first sense would require your act-given-state credences to remain constant throughout the decision-making process. This would prevent you from seeing

[5] Glymour and Meek (1994) have roughly the same view.

yourself as having a free choice about *A* since your evidence about Omega's predictive prowess would constrain what you could believe about your own actions. In DD, if $prob(A|A^{\pi}) = 0.8$ and $prob(D|D^{\pi}) = 0.9$ are fixed, then the laws of probability dictate that $0.1 \leq prob(A) \leq 0.8$, which means you cannot be fully confident either that you will do *A* or that you won't. As Isaac Levi (1989, 2000) stresses, you cannot see yourself as controlling your fate when act probabilities are restricted this way. His response is to deny that it makes sense to assign probabilities to you own acts during deliberation.

I have argued against this view (2002, 2007), and will not recap my misgivings here. The key point is that seeing yourself as free is consistent with being confident (even certain) that Omega has correctly predicted *the act you will actually do*, but not with being confident that you will do *whatever he predicted*. Seeing yourself as free is consistent with confidence (even certainty) that Omega has correctly predicted *the act you will actually do*, but not with confidence that you will do *whatever he predicted*. In Flagship you assign high credence to both these "prediction-if-act" conditionals:

If I do choose ONE/TWO, then Omega is likely to have predicted ONE/TWO.

But, you assign low credence to at least one of these "act-given-prediction" conditionals:

If Omega predicted ONE/TWO, then I am likely to choose ONE/TWO.

To the extent that you see yourself as free, your credences for these conditionals coincide with your credences for their *consequents*, which rules out being confident in both. Suppose you come to see ONE as your best option. As a free agent you will *not* reason like this: "ONE is my best option, but I won't choose it if Omega predicted TWO, because he is so reliable." Rather, you reason: "I'll choose my best option whatever Omega predicted. So, I'll choose ONE even if he predicted TWO. But, since ONE is my best option, Omega probably predicted ONE." As a free agent you are confident that you will do what you most prefer, whatever Omega predicted, and however reliable you take him to be. But, you are also confident that Omega predicted the act you will most prefer.

This brings us to $\boldsymbol{NP}_5$, and what it means to be free. The relevant notion is familiar from compatibilist theories of free will. You see yourself as free when you are confident that (i) no constraints prevent you from choosing an act that you judge to be among your best options, and (ii) your choice is the immediate effect of your ranking the chosen act among your best options. In Newcomb problems (i) is the idea that your choice is not constrained by your past state or by Omega's prediction. Even conditional on A^{π} & B^{τ}, the *only* thing that prevents you from choosing any act on the menu is that you do not rank it among your best options. In Flagship, nothing prevents one-box types from choosing TWO, or two-box types

from choosing ONE. Agents eschew "contrary-to-type" choices not because they *cannot* make them, but because they do not see making them as being in their interests.

Clause (ii) requires your credences for actions to respond to evidence in a distinctive way. You should be confident of performing *A* only if you rank it among your best options. To a first approximation, other considerations should affect your estimate of *A*'s probability only insofar as they convey information about *A*'s merits relative to other acts. I will say more later, but the salient point now is that, when freely choosing, your beliefs about acts are driven *exclusively* by your judgments about which of them best satisfy your desires, and your belief that you will choose an option that you ultimately regard as best.

§3 Causalism and Evidentialism

In Newcomb problems acts influence the future and indicate the past. By performing *A* you bring about outcomes that *A* causes and create evidence for thinking that you occupied the "type-*A*" state at t_0. Choosing ONE in Flagship secures you whatever is in the opaque box, and creates evidence that it is not empty. Choosing ALEP in DD causes you to be in Aleppo, and creates evidence that Omega's assassins are already there. Here is the question that divides CDT and EDT: Should you choose acts exclusively on the basis of what they cause or also on the basis of what they non-causally indicate? CDT says that only causal consequences matter; EDT considers purely evidential implications as well.

This disparity emerges in calculations of expected utility. The theories agree that an act's value is a probability-weighted average of the utilities of its potential outcomes, but CDT weights each outcome by the *unconditional* probability of the state that brings it about, while EDT weights it by the probability of that state *conditional on the act*. *Under appropriate conditions* (see below), *A*'s choiceworthiness in CDT is given by its *efficacy* value $U(A_m) = \Sigma_n$ $prob(S_n)\cdot util(O_{m,n})$, while *A*'s choiceworthiness in EDT is its *news value* $V(A_m) = \Sigma_n$ $prob(S_n/A_m)\cdot util(O_{m,n})$. Differences in *U*-values reflect expected disparities in desirabilities of future outcomes that acts *cause*, while differences in *V*-values reflect expected disparities in evidence that acts provide about future *or past* facts.

Causation and indication, which usually go together, diverge in Newcomb problems. Causalists look at Flagship like this, where *x* is your credence for being a one-box type, $p = prob(\text{ONE}^{\pi}/\text{ONE})$ and $q = prob(\text{TWO}^{\pi}/\text{TWO})$:

Flagship	ONE$^{\tau}$ & ONE$^{\pi}$	ONE$^{\tau}$ & TWO$^{\pi}$	TWO$^{\tau}$ & ONE$^{\pi}$	TWO$^{\tau}$ & TWO$^{\pi}$
ONE	1000 , px	\$0 , $(1 - p)x$	1000 , 0	0 , 0
TWO	1001 , 0	1 , 0	1001 , $(1 - q)(1 - x)$	1 , $q(1 - x)$

TABLE-3

The sum of the second entries in each column is the unconditional probability of that column's state, and causal expected utilities are:

$U(\text{ONE}) = 1000 \cdot px + 0 \cdot (1 - p)x + 1000 \cdot (1 - q)(1 - x) + 0 \cdot q(1 - x) = 1000 \cdot (px + (1 - q)(1 - x))$

$U(\text{TWO}) = 1001 \cdot px + 1 \cdot (1 - p)x + 1001 \cdot (1 - q)(1 - x) + 1 \cdot q(1 - x) = U(\text{ONE}) + 1$

Thus, CDT favors TWO over ONE. EDT computes expected utilities like this:

$V(\text{ONE}) = 1000 \cdot p + 0 \cdot (1 - p) + 1000 \cdot 0 + 0 \cdot 0 = 1000 \cdot p$

$V(\text{TWO}) = 1001 \cdot 0 + 1 \cdot 0 + 1001 \cdot (1 - q) + 1 \cdot q = 1001 - 1000 \cdot q$

This favors ONE over TWO when Omega is sufficiently reliable ($p + q > {}^{1001}/_{1000}$).

Death in Damascus is more complicated. Here is CDT's picture, x being your credence in being an Aleppo type:

DD	ALEP$^{\tau}$ & ALEP$^{\pi}$	ALEP$^{\tau}$ & DAM$^{\pi}$	DAM$^{\tau}$ & ALEP$^{\pi}$	DAM$^{\tau}$ & DAM$^{\pi}$
ALEP	0 , px	1 , $(1 - p)x$	0 , 0	1 , 0
DAM	1 , 0	0 , 0	1 , $(1 - q)(1 - x)$	0 , $q(1 - x)$

TABLE-4

Then, $U(\text{ALEP}) = (1 - p)x + q(1 - x) = 1 - U(\text{DAM})$, and $U(\text{ALEP}) \geq U(\text{DAM})$ iff $x \leq (q - ½)/[(p - ½) + (q - ½)]$. The evidential utilities are $V(\text{ALEP}) = 1 - p$ and $V(\text{DAM}) = 1 - q$, and $V(\text{ALEP}) \geq V(\text{DAM})$ iff $q \geq p$. So, EDT recommends whichever city is associated with the type that Omega has the hardest time predicting *irrespective of your beliefs about your type*. CDT's recommendation, in contrast, depends on how confident you are about your type. Notice that, in addition to ranking acts using different criteria, U's values are higher than V's. This is because U, with its focus on the future, factors out the bad news of having to face DD in the first place, while V reflects this.

CDT and EDT also differ about the values of decisions as *wholes*. In any decision your time-t credences are a snapshot of your views about what you are likely to do and what then occur.

You can use these credences to attach a utility to your overall predicament, the *status quo* as Skyrms (1990) calls it, by averaging expected utilities of acts, weighting each by its probability. In CDT, $U(SQ) = \Sigma_m\, prob(A_m)\cdot U(A_m) = \Sigma_m\Sigma_n\, prob(A_m)\cdot prob(S_n)\cdot U(O_{m,n})$ is your best estimate of the improvement/decline in your fortunes that will occur as an effect of making the decision. In EDT, $V(SQ) = \Sigma_m\, prob(A_m)\cdot V(A_m) = \Sigma_m\, \Sigma_n\, prob(S_n\ \&\ A_m)\cdot U(O_{m,n})$ is the news value of making the decision. In DD, $U(SQ) = 1 - x + (2x - 1)\cdot U(\textsc{Alep})$, while $V(SQ) = 1 - (xp + (1 - x)q)$. At $x = {}^1/_3$, $p = 0.9$ and $q = 0.7$ the numbers work out so that $U(SQ) = 0.5$ and $V(SQ) = 0.233$. CDT is thus indifferent between facing DD and playing Russian roulette with three bullets in a six-shooter, while EDT recommends playing with four bullets. This difference is to be expected since CDT ignores the "bad news" of having to play Russian roulette in the first place.

Notice too that CDT assesses decisions differently when considering them *in prospect* than when they are actively being made. You consider a choice among acts $A_1,\ldots, A_M$ in prospect when you will choose among them at some *future* time, but cannot choose *now*.[6] Viewed in prospect, acts in future decisions are treated not as current *options*, but as potential *outcomes* lying causally downstream of your current choice. And, as with anything not under your *current* control, CDT assesses future acts using their current *news values*. For example, if you can decide now whether to face DD *tomorrow*, then ALEP is not a current option; it is a potential consequence whose current value is $V(\textsc{Alep})$. This value is very low: learning that you will choose ALEP is bad news, now and tomorrow. Of course, CDT says that tomorrow you should not worry about bad news that you *cannot* then control (e.g., being caught in DD), but only about how to make things that you can control. But, it also says that you should consider ALEP's news value when deciding whether to expose yourself to a future decision in which it might be chosen. For example, when deciding whether to face DD, the fact that facing it is a strong indicator of your death is highly relevant. Thus, although CDT values a decision you can now make at $U(SQ)$, it values the same decision in prospect at $V(SQ)$. As a result, CDT and EDT often agree about which future decisions to make even if not about how to make them.

§4 Pseudo-Flagship Fallacies

Opponents of CDT press two broad sorts of objections. Some, like Levi (2000), claim that the theory is ill-founded or incoherent, e.g., because it lets agents have credences for their own acts. Others, like Horgan (1981), Egan (2007), and Ahmed (2014a, b), claim that CDT gives bad advice. Since I have discussed the first worry elsewhere,[7] I will focus on the second here. First,

[6] You cannot "tie yourself to the mast" and irrevocably commit to performing these future acts. That would amount to making the future decision now.

[7] See Joyce (2002) and (2007) and Rabinowicz (2002).

I will consider arguments against CDT's two-boxing policy in Flagship. Then, after developing CDT more fully in §5, I will address more complicated cases, like DD, in §6. A unifying theme will be that many objections to CDT involve a kind of bait-and-switch in which the answer to a *pseudo*-Newcomb is presented as the solution to a genuine Newcomb problem.

For example, some proponents of one-boxing see Flagship like this, with zero-probability outcomes seen as *impossibilities*:

	ONE$^\tau$ & ONE$^\pi$	ONE$^\tau$ & TWO$^\pi$	TWO$^\tau$ & ONE$^\pi$	TWO$^\tau$ & TWO$^\pi$
ONE	1000 , p	0 , $1 - p$	∅	∅
TWO	∅	∅	1001 , $1 - q$	1 , q

Table-5

This requires you to hold act/state correlations fixed by treating ONE ≡ ONE$^\tau$ like a logical truth.[8] This is not Flagship! In Flagship one-box types have the option to two-box, and two-box types have the option to one-box. They do not exercise these options, and are certain they will not, simply because they do not see it as in their best interests to do so. In Table-5, however, contrary-to-type choices are not just undesirable; they are impossible. Since ~ONE & ONE$^\tau$ and ~TWO & TWO$^\tau$ are contradictory, you lack the power to falsity Omega's prediction — had your choice been different, his guess would have been different — a violation of ***NP***$_3$ and ***NP***$_5$. The mistake lies in treating outcomes in which you act contrary to type like impossibilities rather than events that will not occur because you will not want them to occur. You are confident of ONE ≡ ONE$^\tau$ *not* because you lack the power to falsify it, but because you see it as being in your interest to make it true. If you are a two-box type, you will eventually become convinced of ~ONE & ~ONE$^\tau$ and also of ONE $\Box\!\!\rightarrow$ ~ONE$^\tau$. So, even though you are certain of ONE ≡ ONE$^\tau$ given what you *actually* choose (TWO), you are also certain that it would be false were you to choose differently. Likewise for one-box types.

For another pseudo-Newcomb, suppose you are going to face Flagship at time t_1 with p = 0.8 and q = 0.9. But, at $t_{-1} < t_0$ (*before* Omega guesses) you can choose your *type*, and Omega will base his guess on what you choose. There is no reneging: if you choose to be a one/two-box type at t_{-1}, you will freely choose ONE/TWO at t_1 (because you will want to). Clearly, you should choose ONE$^\tau$. This will cost a utile at t_1, but that's cheap for an 80% chance at 1000 utiles. But, if you choose to be a two-box type at t_{-1}, you get the extra 1 utile, but only a 10%

[8] The dialectic here does not change appreciably if we replace logical equivalence by counterfactual equivalence ONE $\leftarrow\!\!\Box\!\!\rightarrow$ ONE$^\tau$.

chance of an added 1000. Here everyone recommends choosing to be a one-box type at t_{-1}: it maximizes *both U* and *V*.

Some think it incoherent for CDT to recommend ONE$^{\tau}$ over TWO$^{\tau}$ at t_{-1}, but TWO over ONE at t_1. This is sometimes portrayed as an unwillingness of CDT to stand behind its advice. If TWO is right at t_1, shouldn't you strive to be the type who chooses TWO then? Others see temporal inconsistency. By recommending ONE$^{\tau}$ at t_{-1} doesn't CDT implicitly endorse ONE at t_1, and then reverse itself and endorse TWO at t_1. Still others (Yudkowsky 2010) see "reflective incoherence": agents who maximize *U* at all times will wish at t_1 that they had used a decision rule that chose ONE at t_{-1}.

These worries presuppose that by choosing ONE$^{\tau}$ in ONE$^{\tau}$-vs-TWO$^{\tau}$ you somehow sanction ONE in ONE-vs-TWO. This is mistaken. You make the first choice *before* Omega's guess, when you can still influence it. You make the second choice when Omega's guess is part of the inalterable past. But, CDT's advice is consistent. It *always* says that, at any time *t*, choose *from among the options available to you at t* an act that you expect to be most efficacious at causing desirable results. CDT only seems inconsistent when we forget that ONE$^{\tau}$-vs-TWO$^{\tau}$ and ONE-vs-TWO involve different options with different causal properties exercised at different times. In ONE$^{\tau}$-vs-TWO$^{\tau}$ you can influence whether you get the million, and CDT recommends the action, from among those available at t_{-1}, that is most likely to bring this about. That's ONE$^{\tau}$. But, in ONE-vs-TWO your type is fixed, and CDT tells you to cause the best results given your t_1 options. That's TWO.

Even though TWO is what you *should* do at t_1, it is not what you *will* do if you chose *rationally* at t_{-1}. If you chose ONE$^{\tau}$ at t_{-1} then, while you are free to choose TWO at t_1, you will not, because, as a one-box type, you mistakenly favor ONE. You might favor ONE because your t_{-1} choice made you an EDTist, or misled you into thinking that you can alter the past, or maybe clouded your cerebellum with vapors of black bile. Who cares? The point is that sanctioning ONE$^{\tau}$ over TWO$^{\tau}$ is no endorsement of ONE over TWO. It is no part of your goal at t_{-1} to choose rationally at t_1. Your only goal at t_{-1} is to choose *from among the options available to you then*, the act that you expect to best promote desirable outcomes. From CDT's perspective, your t_{-1} options are "ONE$^{\tau}$-and-irrationally-choose-ONE" and "TWO$^{\tau}$-and-rationally-choose TWO". Since the former causes the best outcome, CDT recommends it even at the cost of later irrationality. The moral is that *at all times* CDT endorses the same choice in each decision: ONE$^{\tau}$ over TWO$^{\tau}$ at t_{-1}, TWO over ONE at t_1. Even though the first choice makes you botch the second, the 1 utile penalty for irrationality at t_1 is more than offset by the 70% increase in your chance at 1000 utiles you get by acting rationally at t_{-1}. Moreover, you will never wish that you had used a decision rule other than CDT. You might wish that you had different options (e.g., ONE$^{\tau}$-at-t_{-1}-and-TWO-at-t_1), but that involves no reflective inconsistency.

There are further examples of this sort one might consider, but I hope to have given readers some sense of these bait-and-switch imitations of Flagship. We turn now to more complicated cases in which CDT does not recommend a unique choice.

§5 Decision Instability

Despite its centrality in the literature, Flagship is not a typical Newcomb problem. Since two-boxing *dominates* one-boxing, CDT's recommendation does not depend on your credences. In problems like DD credences matter. In particular, information about how likely you are to perform various acts can be evidence about both what future outcomes those acts might cause and what past facts they non-causally indicate. While CDT regards the latter as irrelevant, it *requires* you to consider the first sort of data. This is why I said CDT requires *U*-maximization "under appropriate conditions." Properly understood, it has you maximize *U only after taking into account all readily available evidence about what your acts may cause*.

While I will not reargue it here, Joyce (2012) contends that you have not processed all your evidence about what your acts might cause until your credences and expected utilities reach a *deliberational equilibrium* (*prob**, *U**) in which every act of positive probability has the utility of the status quo. Here I follow a trail blazed by Skyrms (1982) and travelled by Arntzenius (2008) by modeling deliberation as an information-gathering process in which you learn the best ways to pursue desirable outcomes by comparing the efficacies of acts to that of the status quo. Acts with U_t-values higher/lower than $U_t(SQ)$ are seen as better/worse than average at promoting your aims. In general, you want acts with U_t-values exceeding $U_t(SQ)$ to have their credences increased at the expense of acts with U_t-values below $U_t(SQ)$. Yet, you cannot alter credences *ad lib*. Like any beliefs, beliefs about your acts should only change in response to evidence. But, not just any evidence. Since you see yourself as free to choose whichever act you deem best, in deliberation your credences for acts should respond only to evidence about their choiceworthiness. Other factors may affect act probabilities *only* by affecting your views about choiceworthiness. The evidential relations work this way during deliberation because (a) you see yourself as a free agent who will do what you ultimately prefer, and (b) you treat the fact that $U_t(A)$ exceeds $U_t(SQ)$ as (inconclusive) evidence that *A* will rank among your most preferred options when all evidence is in.

The details of the deliberative process are not critical, but the idea is that, at each time *t*, you acquire information about the efficacies of options by learning a conjunction $[U_t(SQ) = u \,\&_m\, U_t(A_m) = u_m]$. But, you should not *choose* on the basis of these *U*-values until you know they

incorporate all readily available information about what your acts might cause.[9] This happens when $U_t(A_m) = U_t(SQ)$ for all A_m with $prob_t(A_m) > 0$. If $prob_t$ and U_t pass this test, then you have achieved an equilibrium and all relevant evidence has been processed. If not, you must update using a belief revision rule that *seeks the good* (Skyrms) by mapping your time-t credences and utilities to time-t+1 credences and utilities in such a way that $prob_{t+1}(A) \geq prob_t(A)$ iff $U_t(A) \geq U_t(SQ)$. For this purpose I like *Bayesian dynamics*, which has $prob_{t+1}(A) = prob_t(A)\cdot[U_t(A)/U_t(SQ)]$.[10] Once act probabilities are updated, all other credences are revised via a *Jeffrey shift*: $prob_{t+1}(\bullet) = \Sigma_m\, prob_{t+1}(A_m)\cdot prob_t(\bullet/A_m)$. Since this shift satisfies $prob_{t+1}(S/A) = prob_t(S/A)$, it disturbs neither your confidence in Omega's reliability nor your news values. Note also that this process adjusts act credences *only* in response to evidence about U_t-values. This is critical: in the midst of deliberation a rational agent's beliefs about acts change only in response to evidence about the merits of those acts.

In all cases we consider a *deliberational equilibrium* ($prob^*$, U^*) will eventually be reached, and updating on U^*-values has no further effect on your credences. You are left with a set of *live* acts $\mathfrak{B} = \{B_1,\ldots, B_K\}$ such that $prob^*(B_k) > 0$, $\Sigma_k\, prob^*(B_k) = 1$, and $U^*(B_k) = U^*(SQ) \geq U^*(A_m)$ for any A_m. Any act not among the B_k is moot since you are sure you will not choose it.

When $\mathfrak{B}$ contains only one act, this is what CDT mandates. Any version of Flagship has a unique equilibrium in which you are certain you will take two boxes, confident (or certain) that Omega guessed this, and expecting to get far less than 1000 *whatever you do*.

When multiple acts survive in equilibrium CDT is indifferent among them. This happens in DD. If p and q both exceed one-half, DD has a unique equilibrium with $0 < prob^*(\text{ALEP}) = (q - ½)/[(p - ½) + (q - ½)] < 1$ and $U^*(\text{ALEP}) = U^*(\text{DAM})$. In such cases, you must *pick*. "Pick" is a term of art for a choice process which selects one from a set of equally good acts in a way that is *not* sensitive to differences in utility. (Think Buridan's ass!) Picking is inherently arational. Picking A over B does *not* imply that you have more *reason* to choose A than B.

Your picking method is a fact about your "type" that goes beyond those features of your t_0 mental state that affect which acts you deem choiceworthy at t_1. For simplicity, I assume your

[9] Even if there are modest costs to acquiring evidence my conclusions still hold. Also, if you *must* choose before you have time to process all relevant causal information, then CDT tells you to maximize efficacy value relative to your current, imperfect beliefs. CDT can say this while still insisting that you would have make a better decision if you had more time to gather informaiton. Thanks to Brad Armendt for pressing me on this.

[10] Utilities are measured on a positive scale. Utilities can always be scaled this way for decisions with only finitely many acts, or where utilities are bounded.

type *determines* your pick.[11] This means that an Aleppo/Damascus-type is someone who picks ALEP/DAM in the DD equilibrium. I want to emphasize that you will not *care* how you pick. By arriving at equilibrium you ensure that you see every live act as maximally efficacious in light of all available evidence about what your acts might cause. As far as *desires* are concerned, a pick is an irrelevant detail.

Even so, you will have *beliefs* about your picking tendencies, and these explain why Omega is better than you at predicting your choices: he has better information about how you'll pick! At t_0 he learn two sorts of facts about type: facts that help him deduce the equilibrium that you will settle at, and facts about how you will pick once there. If his evidence about these things is better than yours, he will be better at predicting your behavior. And, his evidence is better. In equilibrium, your confidence that you will pick *A* is *prob**(*A*). As stressed above, this is sensitive only to evidence about *U*-values, and in equilibrium you have taken all such data into account. *prob**(*A*) is thus your fully informed estimate of the probability that you are the "pick-*A*" type. If you think Omega has better information about your type than you do, then you will expect him to better predict your picks.

§6 Death and Damascus, and Some Variants

Many of CDT's detractors allege that it mishandles decisions in which multiple acts survive into equilibrium. These are mostly bait-and-switch arguments, albeit of a subtler variety than those already encountered.

Consider first a version of DD where $p = prob(\text{ALEP}^{\pi}/\text{ALEP})$ exceeds $q = prob(\text{DAM}^{\pi}/\text{DAM})$, say $p = 0.9$ and $q = 0.7$. Here the unique equilibrium is $prob(\text{ALEP}) = 1/3$. So, if CDT is correct, not only is it acceptable to pick Aleppo, you should have a one-in-three probability that you will. Doesn't that seem wrong, given that 10% of those who go to Aleppo survive, while 30% of those who go to Damascus survive? Shouldn't you choose a higher survival rate (as EDT says), and be certain you will?

Definitely! But, that's *not* your choice. This is the "choose your type" fallacy in new garb. Your survival rate in DD causally depends on your choice *and your type*. If you are an Aleppo-type, you choose between *Aleppo-and-dying-with-probability-0.9* and *Damascus-and-dying-with-probability-0.1* (*not 0.7*). If you are a Damascus-type, you choose between *Damascus-and-dying-with-probability-0.7* and *Aleppo-and-dying-with-probability-0.3* (*not 0.9*). Unfortunately, being unsure of your type, you cannot know which decision you face. Having arrived at the

[11] In a fuller treatment, your type would determine your *chances* of picking various ways. Your credence for *A* is then your expectation of that chance, so $prob^*(A) = \int_M prob^*(M)\, M(A)\, dM$ where *M* ranges over picking mechanisms, and *M*(*A*) is your chance of picking *A* when mechanism *M* is in effect.

*prob**(ALEP) = 1/3 equilibrium you have acquired as much data as you can, but the decision's diabolical structure — with one city offering better survival rates as the other grows more likely — makes it impossible to know which decision you face *until after you pick*. Before you pick your credence is 1/3 that you face {Aleppo & 0.9 death, Damascus & 0.1 death} and 2/3 that you face {Aleppo & 0.3 death, Damascus & 0.7 death}. Picking resolves the uncertainty. To their dismay, Aleppo-types learn that they have chosen a 90% chance of death in Aleppo, when they could have had a 90% chance of life in Damascus. Damascus-types are slightly less distressed to learn that they opted for a 70% chance of death in Damascus over a 70% chance of life in Aleppo. But nobody chooses from {Aleppo & 90% death, Damascus & 70% death}. To do that they would have to choose their type, which ***NP***$_3$ prohibits.

Going to Damascus is optimal if you can choose your type *before Omega guesses*. You will then be choosing between these options:

- ALEP$^\tau$ now and choose from {Aleppo & 90% death, Damascus & 10% death} later on when you see Aleppo as the better option.
- DAM$^\tau$ now and choose from {Aleppo & 30% death, Damascus & 70% death} later on when you see Damascus as the better option.

The second option is best. While choosing it will cause your future self to choose the worse future option, it secures you an extra 20% survival probability. This is an exact analogue of the "choosing to be a one-box type" decision. As before, it does nothing to undermine CDT.

The idea that you choose your survival rate in DD is hard to shake, as a recent paper by Jack Spencer and Ian Wells (2018) illustrates. Spencer and Wells offer a counterexample, *The Semi-Frustrater*, which allegedly undermines CDT's dominance principle. Retelling their story as a version of DD, suppose you (irrevocably) choose Aleppo or Damascus by pointing toward the chosen city with your right or left hand. The twist is that Omega is better at predicting righties than lefties, but righties get cake (0.05 utiles)! Your situation looks like this, where w, x, y and $z = 1 - (w + x + y)$ is your credence for the act in its associated row, and $p_R > p_L$ and $q_R > q_L$:

SF	ALEP$^\tau$ & ALEP$^\pi$	ALEP$^\tau$ & DAM$^\pi$	DAM$^\tau$ & ALEP$^\pi$	DAM$^\tau$ & DAM$^\pi$
ALEP$_R$	0.05 , $w{\cdot}p_R$	1.05 , $w{\cdot}(1-p_R)$	0.05 , 0	1.05 , 0
ALEP$_L$	0 , $x{\cdot}p_L$	1 , $x{\cdot}(1-p_L)$	0 , 0	1 , 0
DAM$_R$	1.05 , 0	0.05 , 0	1.05 , $y{\cdot}(1-q_R)$	0.05 , $y{\cdot}q_R$
DAM$_L$	1 , 0	0, 0	1 , $z{\cdot}(1-q_L)$	0 , $z{\cdot}q_L$

TABLE-6

For any initial beliefs that give both righty acts positive credence, SF has the same equilibrium as DD, so that $prob^*(\text{ALEP}_R) = (½ - q_R)/[(½ - p_R) + (½ - q_R)] = 1 - prob^*(\text{DAM}_R)$. Lefty acts are thus inconsequential. Their initial probabilities are moot, and it does not matter how reliably Omega predicts them. Under any conditions, their equilibrium probabilities vanish because some "cake" option always has higher equilibrium expected utility.

Spencer and Wells see this as wrong. They claim that rationality requires you to use your left hand, and permits choosing either city that way.[12] "Consistent right-handers," they write, end up poorer than consistent left-handers because they choose irrationally." (p. xx) This remark occurs as part of a discussion of the *Why Ain'cha Rich* (WAR) argument in which Spencer and Wells argue that, while WAR fails to justify one-boxing in Flagship, it does justify left-handing in SF. Echoing a well-known line, they correctly argue that the much ballyhooed fact that one-boxers end up richer than two-boxers cuts no ice against CDT because, through no merit of their own, one-box types *start out* with better options. They cannot help being rich no matter how poorly they choose, while two-box types cannot help being poor no matter how well they choose. Once we factor in this disparity in initial endowments we see that it is one-boxers who act irrationally. Endowed with terrific options ($1000000-vs-$1001000), they choose the worst, whereas two-box types respond to their paltry options ($0-vs-$1000) by choosing the best. Moral: people who choose irrationally from desirable options can end up better off than people who choose rationally from undesirable options. However, this does not apply in SF, Wells and Spencer argue, because "like consistent left-handers, consistent right-handers always make their choices [in circumstances that involve] exactly [1.05 utiles]." The point seems to be that righties have *better* options in SF: righties cannot do worse than 0.05 utiles, while lefties can end up with 0; righties can secure as much as 1.05, while lefties max out at 1. So, it *should* count against righties that they end up worse off than lefties.

This reasoning is flawed. Spencer and Wells misidentify the advantages of righty versus lefty decisions, and their claim about "consistent" right- and left-handers is only plausible for choices among *types*. For definiteness, suppose Omega correctly predicts lefty choices at a rate 0.2 lower than righty choices, and that he is correct about righties who go to Aleppo/Damascus at a rate of $p_R = 0.9/q_R = 0.7$. Making these assumptions in a real Newcomb problem means agreeing that lefty-types are 0.2 less likely to die than righty-types *whether they point with their left or their right*, a *huge* initial advantage for lefty-*types*! They enjoy a 0.2 higher survival rate even if they take cake, and righty-types face a 0.2 lower survival rate even if they refuse it. Explicitly, a lefty-type faces one of these decisions, though they know not which (being unsure of their type):

[12] Here Spencer and Wells assume $prob(\text{ALEP}^\pi/\text{ALEP}_L) = prob(\text{DAM}^\pi/\text{DAM}_L)$.

ALEP$^{\tau}$ & LEFT$^{\tau}$
ALEP$_R$ $\rightarrow$ cake, 0.7 death.
ALEP$_L$ $\rightarrow$ no cake, 0.7 death.
DAM$_R$ $\rightarrow$ cake, 0.3 death.
DAM$_L$ $\rightarrow$ no cake, 0.3 death.

DAM$^{\tau}$ & LEFT$^{\tau}$
ALEP$_R$ $\rightarrow$ cake, 0.5 death.
ALEP$_L$ $\rightarrow$ no cake, 0.5 death.
DAM$_R$ $\rightarrow$ cake, 0.5 death.
DAM$_L$ $\rightarrow$ no cake, 0.5 death.

A righty-type faces one of these decisions:

ALEP$^{\tau}$ & RIGHT$^{\tau}$
ALEP$_R$ $\rightarrow$ cake, 0.9 death.
ALEP$_L$ $\rightarrow$ no cake, 0.9 death.
DAM$_R$ $\rightarrow$ cake, 0.1 death.
DAM$_L$ $\rightarrow$ no cake, 0.1 death.

DAM$^{\tau}$ & RIGHT$^{\tau}$
ALEP$_R$ $\rightarrow$ cake, 0.3 death.
ALEP$_L$ $\rightarrow$ no cake, 0.3 death.
DAM$_R$ $\rightarrow$ cake, 0.7 death.
DAM$_L$ $\rightarrow$ no cake, 0.7 death.

Clearly, it would be better to face one of the top two choices than one of the bottom two, but that bird will have flown by the time you choose. Whatever decision you face, you should take cake since doing so has no effect on your survival probabilities. It only affects what you know about them.

We can, of course, imagine pseudo-Newcombs wherein you should refuse cake because hand-choice causally affects survival, perhaps because you choose your hand-type *before* Omega guesses. Here it is relevant (and decisive) that "consistent" right-handers (righty-types) end up poorer than "consistent" left-handers. But, in SF you choose an act, not a type, only *after* Omega guesses. Once he has guessed all advantages of being a lefty-type evaporate: pointing with your left has no differential effect except to cost you cake.

Varying this theme, it is easy to confuse the claim that you should choose a lefty *act* in DD with the claim that you should choose a lefty *decision*. Suppose you choose in stages. Initially, you (irrevocably) choose a hand to point with, and then you choose between {ALEP$_R$, DAM$_R$} or {ALEP$_L$, DAM$_L$} depending on the hand selected. At the initial stage you should clearly choose the LEFTY decision. Though this means forgoing cake, it more than compensates by offering a 0.2 better survival probability. It does not follow, however, that you should choose ALEP$_L$ or DAM$_L$ in SF, where righty-acts are options. In the context of a choice between the RIGHTY versus LEFTY decisions, SF's acts are not options, but potential *consequences*, or *acts-in-prospect* as in §3. Recall too that, like EDT, CDT assesses acts-in-prospect by their news values. When assessing

the LEFTY decision, CDT tells you to (i) figure out how likely you are to choose ALEP_L and DAM_L later if you choose LEFTY now, and (ii) use these probabilities to determine an expected news value for the decision. Similarly for RIGHTY. This yields

$$U(\text{LEFTY}) = prob(\text{ALEP}_L)\cdot V(\text{ALEP}_L) + prob(\text{DAM}_L)\cdot V(\text{DAM}_L)$$

$$U(\text{RIGHTY}) = prob(\text{ALEP}_R)\cdot V(\text{ALEP}_R) + prob(\text{DAM}_R)\cdot V(\text{DAM}_R).$$

To find the relevant probabilities you use equilibrium values for future decisions. With the reliability rates we've been using, $U(\text{LEFTY}) = 0.5 > U(\text{RIGHTY}) = 0.233$.[13] So, unless cake is better than a 0.267 increase in your mortality rate, you should choose to make the LEFTY decision. Starting with different p and q values will yield different utilities, but even for small differences in Omega's predictive abilities the cake must be terrific to make the RIGHTY decision a rational choice. But, to reemphasize the key point, this does not imply that you should point with your left in SF. In SF all four acts — ALEP_R, DAM_R, ALEP_L, DAM_L — are options and you are assessing them on the basis of their propensity to cause desirable consequences. As long as both righty acts are available, one will always win this competition. Pointing with your left arm when you can use your right is forgoing cake needlessly.

In addition to trying to justify choosing with the left Spencer and Wells argue that you cannot rationally choose on the basis of U-values unless there is some act A such that "(i) you are in a position to know of A that it maximizes [U], and (ii) conditional on A, [you] are still in a position to know of A that it maximizes [U]." (p. 18) If this is correct, then U-utilities are entirely irrelevant to decisions, like DD or SF, where being confident that you will choose any act entails being confident that some other act maximizes U. According to Spencer and Wells, CDT offers no guidance value in such cases of "decision instability."

This has affinities with the *ratificationist* idea that choiceworthy acts maximize expected utility conditional on the hypothesis that they will be chosen.[14] Spencer and Wells' account is an improvement over ratificationism — e.g., it does not imply that an act can be choiceworthy *merely* in virtue being the sole ratifiable alternative —but both views suffer from a common flaw. Both privilege assessments of acts made from the epistemic perspective that the actor will have *after* she chooses them. In DD you are meant to ask whether ALEP or DAM maximizes expected utility when expectations are computed using $prob(\bullet/\text{ALEP}^\tau)$ or $prob(\bullet/\text{DAM}^\tau)$. Since

[13] $prob(\text{ALEP}_R) = 1/3$, $prob(\text{ALEP}_L) = 0$, $V(\text{ALEP}_L) = 0.3$, $V(\text{DAM}_L) = 0.5$, $V(\text{ALEP}_R) = 0.1$ and $V(\text{ALEP}_R) = 0.3$.

[14] Ratifiability as a criterion for choiceworthiness has been endorsed by evidentialists, like Jeffrey (1983), and causalists, like Harper (1986) and Weirich (1985). Evidentialists and causalists have even joined forces to defend it, as in Eells and Harper (1991).

the answer to both questions is "no" you are meant to conclude either that both ALEP and DAM are unchoiceworthy, or that CDT provides no help in choosing between them.

This misses a core insight of Skyrms' deliberational approach. It is a mistake to assess acts like ALEP and DAM conditional their being chosen because that means assessing them from an perspective that omits relevant information about what they might cause. Reflect on the fact that one of $prob(\bullet/\text{ALEP}^\tau)$ or $prob(\bullet/\text{DAM}^\tau)$ is sure to generate a *bogus* ranking of acts. Suppose you actually are an Aleppo type. Then, conditioning on ALEP^τ will *correctly* lead you to rank DAM above ALEP, but conditioning on DAM^τ will *incorrectly* lead you to rank ALEP above DAM. So, if you are an Aleppo-type who follows the ratificationists or Spencer and Wells, then you end up relying on *false* information when you rule out DAM, or when you treat its *U*-value as irrelevant. This prevents you from making your objectively best choice.

Of course, you will not know which conditional credal state accurately ranks your options until you know how you'll choose (and thereby learn your type). What to do? Easy! This is a standard exercise in Bayesian inference. If you know you are going to learn the true member of a partition $\{E_1,\ldots, E_N\}$, and want to estimate a quantity f whose value depends on E_n, then you should: (a) gather all the free evidence you can about the E_n, (b) use this data to calculate a posterior estimate $Exp(f/E_n)$ for each n, and (c) set your estimate of f equal to your current expectation of your posterior estimate. Applying this method to DD, you should deliberate your way to an equilibrium ($prob^*$, U^*); compute $U^*(\text{ALEP}/\text{ALEP}^\tau)$ and $U^*(\text{DAM}/\text{DAM}^\tau)$; and set your estimates to

$$U^*(\text{ALEP}) = prob^*(\text{ALEP}^\tau)\cdot U^*(\text{ALEP}/\text{ALEP}^\tau) + prob^*(\text{DAM}^\tau)\cdot U^*(\text{ALEP}/\text{DAM}^\tau)$$

$$U^*(\text{DAM}) = prob^*(\text{ALEP}^\tau)\cdot U^*(\text{DAM}/\text{ALEP}^\tau) + prob^*(\text{DAM}^\tau)\cdot U^*(\text{DAM}/\text{DAM}^\tau)$$

You end up exactly where CDT recommends,[15] with your equilibrium credences at $prob^*(\text{ALEP}) = (\tfrac{1}{2} - q)/[(\tfrac{1}{2} - p) + (\tfrac{1}{2} - q)]$, and both acts have U^*-value one-half.

Contra Spencer and Wells, CDT does offer useful guidance here. It says that, in light of all the evidence you have when you choose, neither act in DD can be reasonably expected to be strictly more effective than the other at promoting your aims, so pick! It is true that you cannot rationally choose ALEP or DAM *without picking*, for that would require you to both believe that you will do the act and simultaneously see it as a *U*-maximizing option. Spencer and Wells go wrong in thinking that this makes *U* irrelevant. Having reached the equilibrium for DD (or SF), the *reason* you can pick between ALEP or DAM is that is their U^*-values are the same, and the

[15] Readers may convince themselves that, for any credence function, one has $U(\text{ALEP}/\text{ALEP}^\tau) = 1 - p$, $U(\text{ALEP}/\text{DOM}^\tau) = q$, $U(\text{DOM}/\text{ALEP}^\tau) = p$ and $U(\text{DOM}/\text{DOM}^\tau) = 1 - q$. From this it follows that both ways of computing *U*-values —directly or as an expectation of posteriors — yield $U(\text{ALEP}) = x(1 - p) + (1 - x)q$ and $U(\text{DOM}) = xp + (1 - x)(1 - q)$.

fact that you are picking means that you are *not* obliged to see your act as optimal conditional on being picked. You must see it as optimal *in light of the information you have when you pick*, but in decisions like DD and SF this is perfectly consistent with knowing that it will not seem optimal after you pick. This makes such decisions bad ones to face, but does make it a bad idea to pick acts that maximize causal expected utility when facing them.

Finally, let's consider a more menacing counterexample, due to Ahmed (2014).[16] Imagine DD with a perfect predictor and the added option, COIN, of letting your destination be settled by a fair coin (Aleppo iff heads). Omega can predict whether you toss the coin but not how it lands. If he predicted a toss, he rolled a die and sent assassins to Aleppo/Damascus if it landed even/odd. Wherever the assassins are, and whatever sent them there, your chance of avoiding death by tossing is 0.5. So, your decision looks like this:

DDC	ALEP$^{\tau}$ ALEP$^{\pi}$	*DAM*$^{\tau}$ *DAM*$^{\pi}$	*COIN*$^{\tau}$ *COIN*$^{\pi}$ *Head*	*COIN*$^{\tau}$ *COIN*$^{\pi}$ *Tail*
ALEP	0 , 1	1 , 0	0 , 0	1, 0
DAM	1, 0	0 , 1	1 , 0	0 , 0
COIN	0.5, 0	0.5 , 0	0.5 , ½	0.5 , ½

Table-7

All equilibria satisfy $0 \leq prob^*(\text{ALEP}) = prob^*(\text{DAM}) \leq 0.5$, and $U^*(\text{ALEP}) = U^*(\text{DAM}) = U^*(\text{COIN}) = 0.5$. So, CDT requires *indifference* between all acts, which means that you should refuse to pay a cent to toss the coin. In contrast, EDT has you pay any price up to 0.5 since $V(A) = V(D) = 0 < V(\text{COIN}^{-\Delta}) = 0.5 - \Delta$, where $\text{COIN}^{-\Delta}$ is the option of paying Δ utiles to toss the coin.

Ahmed calls this advice "absurd" and sees it is a reason to abandon CDT. (2014, p. 592). Many will agree. But, CDT has it right: you should not pay a cent to toss the coin because you should not see yourself as buying anything with your money. Proponents of EDT, of course, argue that you are buying an increased probability of life. Since Omega is 100% reliable, they argue, you are certain to die in Aleppo/Damascus if you choose Aleppo/Damascus. But, if you toss the coin then, wherever Omega's henchmen are, you have a 0.5 objective chance of living. It boils down to a certainty of death versus a 50% chance at life (for a pittance). Pay!

To see the flaw in this reasoning, note that in any equilibrium $prob^*(\text{ALEP}) = prob^*(\text{DAM}) = x > 0$ you do *not* believe that ALEP and DAM offer certain death. You estimate that picking ALEP gives you probability x of death (ALEP$^{\pi}$), probability x of life (DAM$^{\pi}$), and probability $1 - 2x$ of a

[16] Spencer and Wells offer an example, *The Frustrater*, which is equivalent to Ahmed's.

fifty-fifty objective chance at life (COIN$^{\pi}$). Thus, your *credences* for the causal hypotheses H_A = "Choosing ALEP will cause my death" and H_D = "Choosing DAM will cause my death" are both 0.5, not 1.0! What trips people up is that, in virtue of Omega's reliability, you *are* justifiably certain that:

H If I choose ALEP or DAM, then choosing the act I choose will cause my death.

H seems equivalent to the conjunction H_A & H_B, and seems to entail that choosing ALEP or DAM will cause sure death. This is wrong. The phrase "the act I choose" *rigidly* denotes what you *actually* choose. *H* says nothing about the act not chosen (so $H \not\equiv H_A$ & H_B). In fact, choosing the other act will cause your *survival*, which would make it a better choice than COIN$^{-\Delta}$ if you knew what it was. Unfortunately, unlike Omega, you will not know what "the act I choose" and "the other act" denote until after you pick. When you pick, you are constrained by your *current* evidence, on which you assign probabilities of x, x and $1 - 2x$ to the hypotheses that choosing ALEP/DAM will cause your objective chance of death to be 100%, 0% or 50%, respectively. So, your credence of living conditional on any of these acts is 0.5. In terms of your subjective estimates of survival probabilities, all three acts offer the same thing. So, paying to toss the coin would be paying for what you already take yourself to have.

The idea that you get something for your money has at least two possible sources. It may express an irrational form or *ambiguity aversion*, or it may be a conflation between DDC and two subtly different pseudo-Newcombs in which COIN$^{-\Delta}$ is optimal. First, consider ambiguity aversion, our well-documented preference for credences based on known objective chances. People might prefer COIN because it ensures a 0.5 *objective* chance of survival, while ALEP/DAM's 0.5 survival probability reflects uncertainty about the chances: in equilibrium, DDC is like drawing from an urn with balls marked "100% death," "0% death," "50% death" in proportions of x, x, $1 - 2x$. Choosing COIN replaces ambiguity with clarity. Though I will not argue it here, I see ambiguity aversion as irrational.[17] But, even if I am wrong, it explains COIN's appeal in a way that is consistent with CDT. The ambiguity averse agent chooses COIN not to improve expected survival probabilities, but to relieve herself of the anxiety of not knowing objective risks.

Paying to toss the coin also seems right because it is so easy to confuse DDC with pseudo-Newcombs where it is right. We will consider two examples. First, suppose you are slated to face DD with a perfect predictor later, but can now *avoid* that choice by paying Δ and going to Aleppo/Damascus iff a fair coin lands heads/tails. Omega has predicted whether you will take this deal. If he guessed that you would accept he rolled a fair die and sent assassins to Aleppo/Damascus iff even/odd. Otherwise, he executed his usual DD protocol. CDT might

[17] See Al-Najjar and Weinstein (2009), especially sections 2 and 3.

seem to advise against paying because each act in DD has an equilibrium utility of 0.5, while $U^*(\text{COIN}^{-\Delta}) = 0.5 - \Delta$. Not so – CDT has you pay up to a half utile to toss to *cause* the desirable result of avoiding the ALEP/DAM choice! Instead of choosing from { ALEP, DAM , COIN}, you get to decide between *F* = [DD later, keep Δ] and ~*F* = [avoid DD, pay Δ, 50% risk of death]. Since facing-DD-and-choosing-ALEP or facing-DD-and-choosing-DAM are causally downstream of your *current* choice, CDT treats them as acts-in-prospect to be assessed by news value. Of course, it is terrible news that you are slated to go up against a perfectly reliable Omega in DD — a sure harbinger of death — and CDT recognizes this by setting $U^*(F) = prob^*(A)\cdot V(A) + prob^*(D)\cdot V(D) = V(F) = 0$, and $U^*(\sim F) = V(\sim F) = 0.5 - \Delta$. So, like EDT, CDT says you *should* pay up to a half utile to avoid DD.

F-vs-~*F* is easily conflated with DDC. Even Ahmed seems to run them together. When he supposes that you face DD against a perfect predictor, he writes

> Everyone agrees that yours is an unfortunate situation. You are playing high-stakes hide-and-seek against someone who can predict where you will hide… There is every reason to think you will lose. (2014, p. 588)

He then imagines a "third option," $\text{COIN}^{-\Delta}$, and shows that in a choice from {ALEP, DAM, $\text{COIN}^{-\Delta}$} CDT rejects $\text{COIN}^{-\Delta}$ for any $\Delta > 0$. This is entirely correct, but Ahmed shifts focus when arguing that this advice is absurd:

> Would you rather be playing hide-and-seek against (a) an uncannily good predictor of your movements or (b) someone who can only randomly guess at them? [You are being] offering the chance to reduce [Omega] from (a) to (b). Of course you should take the offer. (2014, p. 589)

I agree! If you can choose to make a pseudo-Newcomb decision with Omega is no better than chance rather than a Newcomb decision where he is perfectly reliable, you should do it! This is why you take ~*F* over *F*. By choosing ~*F* you take ALEP and DAM off the table, thereby forcing Omega to "play on neutral turf" where his predictive powers can do you no harm. With ALEP and DAM on the table, he has a significant probability (2*x*) of guessing your choice. But, he can be no better than chance if you tie your destination to the coin toss. CDT says to take that deal! But, that deal is *not* DDC. In DDC, ALEP and DAM remain live options right up to the time you irrevocably pick, which forces you to ask how likely it is that choosing them will cause your death. The answer is 0.5, the same as your credence that choosing COIN will cause your death. So, you should pay for COIN when that takes ALEP and DAM off the table, but when both remain live options paying to toss is paying for what you already have. Thus, CDT gets both DDC and the ~*F*-vs-*F* decisions right.

The distinction between DDC and similar decisions in which CDT endorses paying turns on a subtle difference in options. DDC has three — ALEP, DAM and COIN$^{-\Delta}$ — and you can only refrain from choosing COIN$^{-\Delta}$ by choosing one of ALEP or DAM. There is no fourth alternative of deciding not to toss the coin without committing (by choice or pick) to ALEP or DAM, i.e., no disjunctive "ALEP or DAM" option. As a result, you cannot rationally choose COIN$^{-\Delta}$ unless your estimate of the survival probability caused by tossing the coin exceeds your estimates of the highest of the survival probabilities caused by Alep and Dam, which never happens. In contrast, if you have an option like ~*F* that lets you decline to toss *without* picking a city, then you assess COIN$^{-\Delta}$ relative to the menu {COIN$^{-\Delta}$, ~COIN$^{-\Delta}$}, rather than {COIN$^{-\Delta}$, ALEP, DAM}. Here, ~COIN$^{-\Delta}$ is the option of first declining COIN$^{-\Delta}$ and only later deciding between ALEP and DAM. In any decision with this bipartite structure CDT will treat ~COIN$^{-\Delta}$ & ALEP and ~COIN$^{-\Delta}$ & DAM as acts-in-prospect, and will endorse COIN$^{-\Delta}$ over ~COIN$^{-\Delta}$.

Some people may assume that ~COIN$^{-\Delta}$ is always an option. They might even think that a rational agent can always ensure its availability by making a kind of pre-decision in which the options are choosing *later* from the {COIN$^{-\Delta}$, ~COIN$^{-\Delta}$} menu or the {COIN$^{-\Delta}$, ALEP, DAM} menu. I doubt this, but even if it were true it would pose no problems for CDT. If ~COIN$^{-\Delta}$ is always an option, then Ahmed's counterexample will never arise and CDT will always rightly recommend paying to toss the coin. If agents can pre-decide between {COIN$^{-\Delta}$, ~COIN$^{-\Delta}$} and {COIN$^{-\Delta}$, ALEP, DAM}, then we have another pseudo-Newcomb. CDT will treat all entries in these menus as acts-in-prospect, and will recommend selecting the first menu, and subsequently choosing COIN$^{-\Delta}$ from it. Either way, Ahmed's example does not undermine CDT. In any version of the problem in which ~COIN$^{-\Delta}$ is an option CDT recommends paying to toss the coin. In DDC, where ALEP and DAM are options but COIN$^{-\Delta}$ is not, CDT rightly recommends not paying because paying buys nothing.

One might consider other examples, but readers should have the flavor of CDT's responses. If we focus on *real* Newcomb problems, which satisfy ***NP***$_1$-***NP***$_5$, the sophisticated version of CDT that requires choices to be made in equilibrium gets every case right. When it seems to falter, either the theory is being misapplied, options are being misidentified, or a solution to a pseudo-Newcomb problem is being passed off as a solution to a genuine Newcomb problem.

References

Arif Ahmed (2014a) *Evidence, Decision and Causality*. (Cambridge: Cambridge University Press)

_____(2014b) "Dicing with Death," *Analysis* **74**: 587-592.

Nabil I. Al-Najjar and Jonathan Weinstein (2009) "The Ambiguity Aversion Literature: a Critical Assessment," *Economics and Philosophy* **25**: 249–284.

Frank Arntzenius (2008) "No Regrets, or: Edith Piaf Revamps Decision Theory," *Erkenntnis* 68: 277–297.

Ellery Eells and William Harper (1991) "Ratifiability, Game Theory, and the Principle of Independence of Irrelevant Alternatives," *Australasian Journal of Philosophy* **69**: 1-19.

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

Allan Gibbard and William Harper (1978) "Counterfactuals and Two Kinds of Expected Utility," in *Foundations and Applications of Decision Theory*, edited by C. Hooker, J. Leach, and E. McClennen, pp. 125-62. Dordrecht: Reidel.

William Harper (1986) "Mixed Strategies and Ratifiability in Causal Decision Theory," *Erkenntnis* **24**: 25–36.

Chris Hitchcock (1996) "Causal Decision Theory and Decision-Theoretic Causation," *Noûs* **30**: 508-526.

Terry Horgan (1981) "Counterfactuals and Newcomb's Problem," *Journal of Philosophy* **78**: 331–356.

Jenann Ismael (2007) "Freedom, Compulsion, and Causation," *Psyche* **13/1** http://psyche.cs.monash.edu.au/

Richard Jeffrey (1983) *The Logic of Decision*, 2nd edition. Chicago: University of Chicago Press.

James M. Joyce (2016) "Review Essay, Arif Ahmed: *Evidence, Decision and Causality*," *Journal of Philosophy* **113**: 224-232.

_____ (2012) "Ratifiability and Stability in Causal Decision Theory," *Synthese* **187**: 123-145.

_____ (2010) "Causal Reasoning and Backtracking," *Philosophical Studies* **147**: 139-154.

_____ (2007) "Are Newcomb Problems Really Decisions?" *Synthese***, 156**: 537-562.

_____ (2002) "Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions," *Philosophical Studies* **110**: 69-102.

Daniel Kahneman & Amos Tversky (1984) "Choices, Values, and Frames," *American Psychologist* **39**: 341–350.

Greg Lauro and Simon M. Huttegger (201x), "Decision Dependence and Causal Decision Theory: A Critical Response to Hare and Heddon," forthcoming

David Lewis (1981) "Causal Decision Theory," *Australasian Journal of Philosophy* **59**: 5-30.

Isaac Levi (2000) "Review: The Foundations of Causal Decision Theory," Journal of Philosophy (97): 387-402

_____ (1989) "Rationality, Prediction and Autonomous Choice," *Canadian Journal of Philosophy* Supplemental. Vol. **19**: 339–363.

Christopher Meek and Clark Glymour (1994) "Conditioning and Intervening," *British Journal for the Philosophy of Science* **45**: 1001–1021.

Judea Pearl (2000) *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press. [Second edition, 2009]

Huw Price (1992) "The Direction of Causation: Ramsey's Ultimate Contingency," *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*: 253-267

Wlodek Rabinowicz (2002) "Does Practical Deliberation Crowd Out Self-Prediction?" *Erkenntnis* **57**: 91–122.

Brian Skyrms (1982) "Causal Decision Theory," *Journal of Philosophy*, **79**: 695–711.

_____ (1990) *The Dynamics of Rational Deliberation* (Cambridge: Harvard University Press)

Jack Spencer and Ian Wells (2018) "Why Take Both Boxes?" *Philosophy and Phenomenological Research* (forthcoming)

Robert Stalnaker (1981) "Letter to David Lewis of 21 May 1972," reprinted in *Ifs: Conditionals, Belief, Decision, Chance, and Time*, edited by W. Harper, R. Stalnaker and G. Pearce, Western Ontario Series in Philosophy of Science (Dordrecht: Reidel) **15**: 151–152.

Paul Weirich (1985) "Decision Instability," *Australasian Journal of Philosophy*, **63**: 465–472.

Eliezer Yudkowsk (2010) "Timeless Decision Theory," Technical Report, Machine Intelligence Research Institute (MIRI)