Caspar Hare Massachusetts Institute of Technology January 2015
Brian Hedden University of Sydney

Self-Reinforcing and Self-Frustrating Decisions

There is a sense of the term 'ought' according to which what a person ought to do depends not on how the world is, but on how the person believes the world to be. Philosophers typically isolate this as their intended sense of the term by talking of what people 'subjectively ought' to do. Suppose, for example, that you are offered hors d'oeuvres at a fancy party. They look delicious, you are hungry, and you wish to please your host. However, unbeknownst to you, they are riddled with a lethal strain of botulism. A philosopher may say that, in light of your beliefs, you subjectively ought to eat the hors d'oeuvres, though the consequences of your doing so will be disastrous.1

Our focus here will be on theories of the subjective *ought* that imply

> *Decision Dependence*
>
> In some cases what you subjectively ought to do at a certain time depends on what you believe you will do at that time.

We want to do three things.

The first thing we want to do is to show that, in spite of Decision Dependence being prima facie odd (consider how odd it would sound for me to say "I believe that I will do this, so I ought to do this", and consider how much yet odder it would sound for me to say "I believe that I will do this, so I ought not to do this"), the class of

[1] Philosophers typically take themselves to isolate a different sense of 'ought' by talking of what people 'objectively ought to do' – although you *subjectively* ought to eat the hors d'oeuvres, you *objectively* ought to decline them. What exactly is the relation between the subjective and objective oughts? This is a tricky question. We will not address it here.

theories that imply Decision Dependence is quite large. Among decision theorists, recent attention to Decision Dependence has been attention to Decision Dependence as a feature of *causal decision theory*. Among philosophers who work on the ethics of creation, recent attention to Decision Dependence has been attention to Decision Dependence as a feature of some *actual-person-affecting* theories. Among philosophers who think about prudence and welfare, recent attention to Decision Dependence has been attention to *actual-preference-satisfying* deontic theories. In Section 1 we will describe these three sorts of theory. In Section 2 we will formalize them and characterize them in a more general way.

The second thing we want to do is give a new, and in our view compelling, argument that Decision Dependence is false. Many philosophers have felt there to be something problematic about Decision Dependence, but the problem has proven to be exceedingly difficult to pin down. In sections 3 and 4 we will review and dismiss some unsatisfactory arguments against Decision Dependence. In Sections 5 and 6 we will give the argument that satisfies us. We will argue that a self-aware, epistemically rational agent who is guided by the theory will behave in ways that are difficult to defend, *even by the lights of an advocate of the theory*.

The third thing we want to do is to explain how our discussion of Decision Dependence bears on the classic Newcomb case, a case that has been at the center of much theorizing about practical rationality for decades. Standard causal decision theory supports two-boxing in the classic Newcomb case, and we have argued that standard causal decision theory is false – because it implies that the subjective *ought* is decision dependent in other cases. Is there a good theory that supports two boxing in the classic Newcomb case but does not imply that the subjective *ought* is decision dependent in other cases? In Section 7 we will argue that there is not. This is a happy day for one-boxers.

1. Three Theories that Imply Decision Dependence

Here is one example of a theory that says that, in some cases, what you subjectively ought to do depends on what you believe you will do.

> *Satisfy Anticipated Desires (SAD)*
> Other things being equal, you subjectively ought to strive to satisfy desires that you anticipate having. If you now believe that you will later desire that you acted a certain way now then, other things being equal, you subjectively ought to act that way now.2

If this theory is correct then the subjective *ought* will be decision-dependent in some cases in which you believe that your present decision will affect what desires you later have. For example:

> Nice Choices at the Spa
> Aromatherapy or body-wrap – which is it to be? You believe that, whichever you choose, you will be very glad you chose it. Mid-aromatherapy, the aromatherapy will seem self-evidently superior. Mid-body-wrap, the body-wrap will seem self-evidently superior.

If you believe that you will choose the aromatherapy then you believe that you will later think it most desirable that you chose the aromatherapy, so, by SAD, you

[2] This is the sort of idea that seems to underlie "I'll be glad I did it" reasoning – "I'll be glad I did it, so, other things being equal, I ought to do it." See Harman (2009).

subjectively ought to choose the aromatherapy. If you believe that you will choose the body-wrap then you believe that you will later think it most desirable that you chose the body-wrap, so, by SAD, you subjectively ought to choose the body-wrap.

Here is another example of a theory that says that, sometimes, what you ought to do depends on what you believe you will do:

> *Satisfy the Interests of Children and Kids (SICK)*
> Other things being equal, you ought to strive to do what you believe will be good for your children. If you now believe that it will turn out to have been in the interests of one of your children that you acted a certain way now, then, other things being equal, you subjectively ought to act that way.[3]

If this theory is correct then the subjective *ought* will be decision-dependent in some cases in which you believe that your present decision will affect what children you later have. For example:

> <u>Nice Choices at the Adoption Agency</u>
> Annie or Beth – who is it to be? You believe that you are a good parent. It is better for Annie that you adopt Annie, better for Beth that you adopt Beth.

If you believe that you will adopt Annie then you believe that it will turn out to have been in the interests of one of your children (Annie) that you adopt Annie, so, by SICK, you subjectively ought to adopt Annie. If you believe that you will adopt Beth then you believe that it will turn out to have been in the interests of one of your

[3] See Hare (2007) for a discussion of theories of this general kind.

children (Beth) that you adopt Beth, so, by SICK, you subjectively ought to adopt Beth.

And here is a third example of a theory that says that, sometimes, what you ought to do depends on what you believe you will do:

> *Accept What you Cannot Control, With Appropriate Regard for Dependencies (AWCCWARD)*
> If you believe that things beyond your causal influence are such that, supposing they are the way they are, it is most desirable that you act a certain way, then you subjectively ought to act that way.[4]

If this theory is correct then the subjective *ought* will be decision-dependent in some cases in which your beliefs about how things beyond your control are depend on your beliefs about what you will do. Consider:

> <u>The Nice Demon</u>
> Two opaque crates are placed before you. You get to take one and only one of them. Which should you take? You are sure that money has been placed in the boxes by a demon, on the basis of a prediction she made about which crate you would later take:

	in A	**in B**
If she predicted you would take Crate A, then she put	$1000	$0
If she predicted you would take Crate B, then she put	$0	$1000

> The demon has shown herself to be fiendishly good at making predictions. You are sure (or, at least, as close to sure as makes no difference) that it will turn out that she made the right one.

[4] This idea underlies all versions of causal decision theory. We will discuss it in detail in the next section.

If you believe that you will take Crate A then you believe that things beyond your causal influence are this way: there's $1000 in Crate A and nothing in Crate B. Supposing there's $1000 in Crate A and nothing in Crate B, it is desirable that you take Crate A. So, by AWCCWARD, you subjectively ought to take Crate A. If you believe that you will take crateB then you believe that things beyond your causal influence are a different way: there's nothing in Crate A and $1000 in Crate B. Supposing there's nothing in Crate A and $1000 in Crate B, it is desirable that you take Crate B. So, by AWCCWARD, you subjectively ought to take Crate B.[5]

2. A Formal Interlude

One achievement of twentieth century philosophy and economics was the creation of formal tools that allow us to describe theories of the subjective ought very precisely. The benefit to using these tools is accuracy. The cost is obscure technicality. If you have no patience for obscure technicality then please skip ahead to Section 3.

We will begin with AWCCWARD. Its natural formalization is famous – known as *causal decision theory*.[6]

As a background, let's suppose that your present doxastic (which is to say *belief-like*) attitudes can be represented by a credence function, C, from propositions to real numbers between 0 and 1 – the numbers representing how likely you think it

[5] There is in fact a third possibility, namely that you assign probability 0.5 to your taking Crate A and probability 0.5 to your taking Crate B. In this case, taking Crate A and taking Crate B look equally good, and so AWCCWARD permits you to take either Crate (for formal details, see the next section). Note however that the doxastic state of assigning probability 0.5 to your taking Crate A and 0.5 to taking Crate B is an unstable equilibrium, in the sense that if you become any more confident that you will take Crate A, then AWCCWARD recommends taking Crate A, and similary, *mutatis mutandis*, for Crate B. See Skyrms (1990) and Arntzenius (2008) for details.

[6] There are many different versions of causal decision theory. We will follow Lewis (1981).

that the propositions are true. And let's suppose that your present conative (which is to say *desire-like*) attitudes can be represented by a function, U, from propositions to real numbers – the numbers representing how desirable you think it that the propositions be true. And let's suppose that the ways in which you might act now can be represented by a set of propositions A. Call the propositions in A *act-propositions*.

Now let D be a set of propositions concerning how things beyond your control are. Let D be *exclusive* (no two propositions in D can both be true), *exhaustive* (all propositions about how things beyond your control are entail the disjunction of the propositions in D) and *relevant* (for all act-propositions a, all propositions d in D, and all propositions r, if aÙd is consistent with both r and Ør then *U*(aÙdÙr) = *U*(aÙdÙØr).) Call the propositions in D *dependency hypotheses*.

Where d is a variable ranging over dependency hypotheses, we define the *causal expected utility* (E_CU) of an act proposition, a, like this:

$$E_CU(a) = \sum_d(C(d).U(\text{aÙd}))$$

And we say that you subjectively ought to make true the act-proposition with highest causal expected utility.

To get a feel for how to apply causal decision theory, consider the Nice Demon case. In that case there are two act-propositions:

a_A: You take Crate A.

a_B: You take Crate B.

and two relevant dependency hypotheses[7]:

[7] This is a slight idealization. In realistic cases, the space of dependency hypotheses will need to be much more fine-grained in order to satisfy *relevance*. Here, however, the idealization is harmless.

$d_{\$1000inA}$: There's \$1000 in Crate A, nothing Crate B.

$d_{\$1000inB}$: There's nothing in Crate A, \$1000 in Crate B.

If you believe that you will take Crate A, then $C(d_{\$1000inA})=1$ and $C(d_{\$1000inB})=0$, so $E_CU(a_A)$ = U(you get \$1000) and $E_CU(a_B)$ = U(you get nothing). So, supposing that you like money, you subjectively ought to make proposition a_A true, which is to say that you subjectively ought to take Crate A. If you believe that you will take Crate B, then $C(d_{\$1000inA})=0$ and $C(d_{\$1000inB})=1$, so $E_CU(a_A)$ = U(you get nothing), and $E_CU(a_B)$ = U(you get \$1000). So, supposing that you like money, you subjectively ought to make proposition a_B true, which is to say that you subjectively ought to take Crate B.

Causal decision theory is standardly contrasted with *evidential decision theory*, which says that you subjectively ought to make true the act-proposition with highest evidential expected utility (E_EU) – defined in this way:

$$E_EU(a) = \sum_d(C(d/a).U(a\text{Ù}d))$$

Where 'C(d/a)' refers to your conditional credence in dependency hypothesis d, given that you make true act-proposition a.

Evidential theory says that in this case, irrespective of what you believe you will do, there is nothing that you subjectively ought to do. Irrespective of what you believe you will do, $C(d_{\$1000inA}/a_A) = 1$, $C(d_{\$1000inA}/a_B) = 0$, $C(d_{\$1000inB}/a_B) = 1$, $C(d_{\$1000inB}/a_A) = 0$. So, irrespective of what you believe you will do, $E_EU(a_A)$ = U(you get \$1000) and $E_EU(a_B)$ = U(you get \$1000). The two options have the same evidential expected utility.

So much for AWCCWARD. Now for the formalization of SAD. Let 'futU' refer to the proposition that your future desires will be represented by utility function U.

Formal SAD says that, other things being equal, you ought to make true the act-proposition with highest *expected-expected utility* (E^2U) – defined in this way (for those sympathetic to causalist reasoning):

$$E^2U(a) = \sum_U C(futU).E_CU(a)$$

or in this way (for those sympathetic to evidentialist reasoning):

$$E^2U(a) = \sum_U C(futU).E_EU(a)$$

where U is a variable that ranges over utility functions. In prose, the expected-expected utility of an act is the sum of the possible (causal or evidential) expected utilities of that act, given each of the different utility functions you might have in the future, weighted by your credence that you will in fact have that utility function in the future. Formal SAD says that you subjectively ought to perform the act with highest expected-expected utility.

To get a feel for how to apply formal SAD, look again at Nice Choices at the Spa. In that case there are two act propositions:

a_A: You choose the aromatherapy

a_B: You choose the body-wrap

and two utility functions that represent desires that you may later have

U_A: a function such that $U_A(a_A) > U_A(a_B)$

U_B: a function such that $U_B(a_B) > U_B(a_A)$

If you believe that you will choose the aromatherapy, then $C(futU_A) = 1$ and $C(futU_B) = 0$. It follows that $E^2U(a_A) > E^2U(a_B)$, and you ought to choose the aromatherapy. If you believe that you will choose the body-wrap then $C(propU_A) = 0$ and $C(propU_B) = 1$. It follows that $E^2U(a_B) > E^2U(a_A)$, and you ought to choose the body-wrap.

The contrast theory here, the theory that stands in the same relationship to formal SAD as evidential decision theory stands in to causal decision theory, is *future satisfactionism*. This says, roughly, that other things being equal you ought to maximize your expected future state of satisfaction. Formally, you ought to make true the act-proposition with highest expected satisfaction (ES) – defined in this way (for those sympathetic to causalist reasoning):

$$ES(a) = \sum_U C(futU/a).E_CU(a)$$

or in this way (for those sympathetic to evidentialist reasoning):

$$ES(a) = \sum_U C(futU/a).E_EU(a)$$

where U is a variable ranging over utility functions.

In this case, irrespective of what you believe you will do, $C(futU_A/a_A) = 1$, $C(futU_A/a_B) = 0$, $C(futU_B/a_B) = 1$, $C(futU_B/a_A) = 0$. So, irrespective of what you believe you will do, $ES(a_A) = U_A(a_A)$ and $ES(a_B) = U_B(a_B)$. So, irrespective of what you believe you will do, if $U_A(a_A) > U_B(a_B)$ then you ought to choose the massage, if $U_A(a_A) = U_B(a_B)$ then there's nothing that you ought to, if $U_A(a_A) < U_B(a_B)$ then you ought to choose the aromatherapy.

Finally, let's move to the formal representation of SICK. First we suppose that, just as we can represent your conative attitudes with a utility function, so we can represent the interests of the various children you might have with utility functions. Let 'futcU' refer to the proposition that your future child has interests represented by utility function U. Formal SICK says that, all other things being equal, you ought to make true the act-proposition with highest expected-expected utility for your child (E^2UK) – defined in this way (for those sympathetic to causalist reasoning):

$$E^2UK(a) = \sum_U C(futcU).E_C U(a)$$

or in this way (for those sympathetic to evidentialist reasoning):

$$E^2UK(a) = \sum_U C(futcU).E_E U(a)$$

To get a feel for how to apply formal SICK, look again at the Nice Choices at the Adoption Agency case. In that case there are two act propositions:

a_A: You adopt Annie

a_B: You adopt Beth

and two utility functions, representing the interests of Annie and Beth

U_A: a function such that $U_A(a_A) > U_A(a_B)$

U_B: a function such that $U_B(a_B) > U_B(a_A)$

If you believe that you will adopt Annie, then $C(futcU_A) = 1$ and $C(futcU_B) = 0$. So $E^2UK(a_A) > E^2UK(a_B)$, so you ought to adopt Annie. If you believe that you will adopt Beth then $C(futcU_A) = 0$ and $C(futcU_B) = 1$. So $E^2UK(a_B) > E^2UK(a_A)$, so you ought to adopt Beth.

The contrast theory for SICK may be called *Welfare Maximization*. It says roughly that you should maximize the expected well-being of your future child (where 'your future child' is read non-rigidly). Formally, you ought to make true the act proposition with highest *expected satisfaction for your child* (ESK) – defined in this way (for those sympathetic to causalist reasoning):

$$ESK(a) = \sum_U C(futcU/a).E_C U(a)$$

Or in this way (for those sympathetic to evidentialist reasoning:

$$ESK(a) = \sum_U C(futcU/a).E_E U(a)$$

As before, this theory says that what you subjectively ought to do does not depend on what you believe you will do. In this case, irrespective of what you believe you will do, if $U_A(a_A) > U_B(a_B)$ then you ought to adopt Annie, if $U_A(a_A) = U_B(a_B)$ then there's nothing that you ought to, if $U_A(a_A) < U_B(a_B)$ then you ought to adopt Beth.

3. A First Pass at Pinning Down the Worry: Decisions will be Unstable in Self-Frustrating Cases

So much for the formalization of SAD, SICK and AWKWAARD. Is it a defect in these theories that they entail that there are situations in which the subjective *ought* is decision-dependent? To get a grip on the question it will be helpful to have a way of representing and sorting the different ways in which the subjective ought might be decision-dependent in different situations. (Note that our arguments in succeeding sections of the paper will not depend on these diagrams; they are for illustrative purposes only.)

For situations in which you have two options available to you, A and B, here is a simple way to represent decision-dependence: First, let points on the unit interval represent credences that you might have concerning what you will do – with distance from the right end representing your credence that you will do A, distance from the left end representing your credence that you will do B (intuitively: the closer the point to the 'A' in the diagram, the more confident you are that you will do A, the closer the point to the 'B' in the diagram, the more confident you are that you will do B). Next, represent what a theory says about a situation by marking regions of the interval. So, for example:

Fig. 1

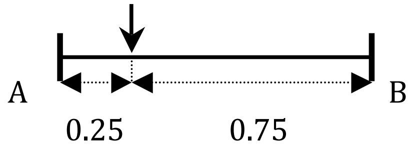
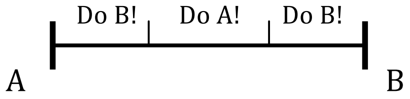


Fig. 2



The indicated point in *Fig. 1* represents the attitude of having credence 0.75 that you will do A, and credence 0.25 that you will do B. *Fig. 2* represents a theory that says of a situation, roughly, that you ought to do A unless you are confident about what you will do, in which case you ought to do B.

A similar method works for three-option cases. First, let points within an isosceles triangle, with height 1 and base-length 1, represent doxastic attitudes that you might have about what you will do – with horizontal distance from the right side representing your credence that you will do A, horizontal distance from the left side representing your credence that you will do B, vertical distance from the base representing your credence that you will do C. Next represent what a theory says about a particular case by marking regions of the triangle. So, for example:

Fig. 3

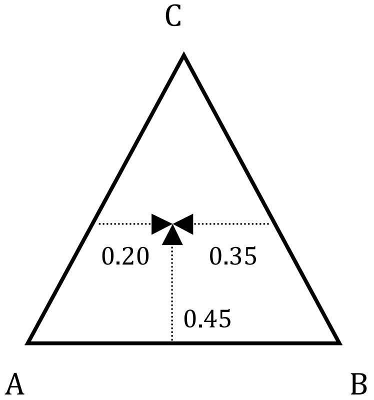
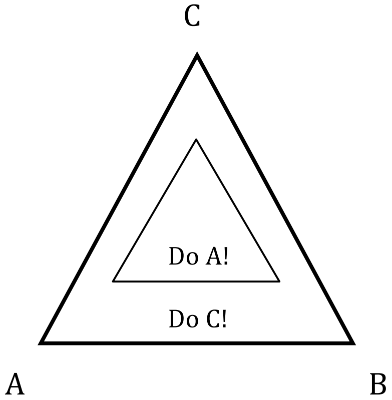


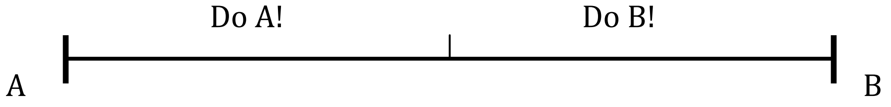
Fig. 4



The indicated point in *Fig. 3* represents the attitude of having credence 0.35 that you will do A, credence 0.2 that you will do B, and credence 0.45 that you will do C. *Fig. 4* represents a theory that says of a situation, roughly, that you ought to do A unless you are confident about what you will do, in which case you ought to do C.

Now, two forms of decision-dependence are particularly interesting. The first is *self-reinforcing* decision-dependence. This comes about when a theory says of a situation that, as your confidence that you will take any particular option increases, so it becomes the case that you ought to take that option. Whatever you believe you will do, you ought to do it. The cases we have seen so far (Nice Choices at the Spa, Nice Choices at the Adoption Agency, The Nice Demon) have all been cases of self-reinforcing decision-dependence, cases in which the recommendations of SAD, SICK and CDT look like this:

Fig. 5



The second is *self-frustrating* decision-dependence. This comes about when a theory says of a situation that, as your confidence that you will take any particular option increases, so it becomes the case that you ought *not* to take that option. Consider:

> Nasty Choices at the Spa
>
> Abdominal-acupuncture or bee-sting-therapy – which is it to be? You believe that, whichever you choose, you will wish that you had chosen the other.

Nasty Choices at the Adoption Agency

Annie or Beth – who is it to be? You believe that you are a bad parent. It is better for Annie that you adopt Beth, better for Beth that you adopt Annie.

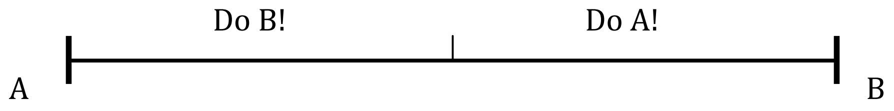
The Nasty Demon

Crate A or Crate B? This time you are sure that the demon wanted to frustrate you:

	in A	**in B**
If she predicted you would take Crate A, then she put	$0	$1000
If she predicted you would take Crate B, then she put	$1000	$0

In these cases SAD, SICK and CDT say that whatever you believe you will do, you ought not to do it. Their recommendations look like this:

Fig. 6



Now, as many philosophers (Gibbard and Harper 1978, Weirich 1985, 1988, Harper 1985, 1986, Richter 1984, Skyrms 1986, Sobel 1994) have observed, in cases like this, if you resolve to be guided by the decision-dependent theory, and you are self-aware, then any decision you make will be in a certain sense *unstable*. Whatever you decide to do, your deciding to do it will give you confidence that you will do it, and confidence that you will do it will show you that you ought to do the other thing, which (given your resolve to be guided by the decision-dependent theory) will lead you to decide to do the other thing, which will give you confidence that you will do the other thing… and so on. You will be unable to stand by your decisions.

Some philosophers[8] have taken this observation to be an objection to theories that imply decision dependence, but it is not so obvious why there is anything objectionable about it. First, it is not obvious why we should demand of a theory of the subjective ought that someone who tries to comply with its demands should always be able to commit themselves to a decision in this sense. Maybe a lack of commitment to your decisions is precisely the right attitude to have in these strange cases.[9]

Second, it is not obvious that if you try to comply with the demands of the theories in these situations, then you will be unable to commit yourself to a decision. Your decision to do A will make it the case that you subjectively ought to do B if your decision to do A gives you confidence that you will wind up doing A. But your decision to do A will give you confidence that you will wind up doing A only if you are confident that the decision is final. And in these self-frustrating cases, where no decision is stable, it is not clear that, if you are aware that you are guided by a theory that implies decision dependence, you should ever be confident that your decision is final. Granted there is something strange about deciding to do A while remaining no more confident that you will A than that you will do B. But again, this may be exactly the right attitude to have in these strange, self-frustrating cases.

[8] Richter (1984) presses this objection against CDT. Harper (1985, 1986) and Weirich (1988) propose modifications to CDT to deal with cases of decision instability, indicating that they agree with Richter that this is a problem for standard CDT.

[9] This is just to say that it is unclear why ratifiability should matter. In the terminology of Jeffrey (1983), an act is said to be ratifiable iff it looks at least as good as the alternatives even once you become certain that you will perform it (that is, iff that act looks at least as good as the alternatives according to your credences, conditional on the proposition that you perform it). We see no compelling reason to think that the true theory of rational decision-making should recommend only ratifiable acts.

Finally, while neither the decision to do A nor the decision to do B is stable, the decision to perform the so called 'mixed' act of doing A with probability 0.5 and doing B with probability 0.5 may be stable. This is because, when you are 0.5 confident that you will do A and 0.5 confident that you will do B, all of the options (A, B, and the mixed act) have the same causal expected utility; they look equally good, according to CDT. So, even if one thinks that rational people must make stable decisions, this does not straightforwardly show you cannot be rational and guided by CDT in these cases. There may be a stable decision to be made.[10]

4. Second Pass: Don't the Theories Just Say Counter-Intuitive Things About *Asymmetric* Self-Frustrating Cases?

Another worry (one that has received a good deal of attention recently11) is that causal decision theory simply says the wrong thing about what you ought to do in self-frustrating cases of a particular kind. Consider:

> The Asymmetrically Nasty Demon12
>
> Crate A or Crate B? Again you are sure that the demon wanted to frustrate you. But this time you are sure that she wanted you to be more frustrated by choosing B than by choosing A:

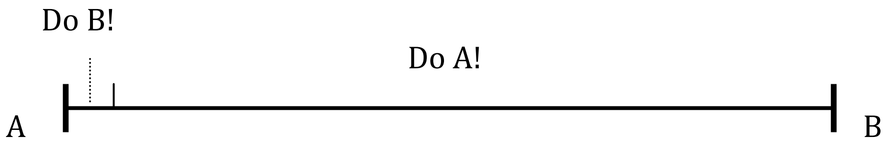
[10] Of course, invoking mixed acts requires the theorist to say something about what mixed acts are and when they are available to agents. In particular, does performing a mixed act require the agent to have a randomizing device available and to bind herself to taking the option indicated by the randomizing device. Since we are neither endorsing nor opposing the use of mix acts in decision theory, we raise this worry only to set it aside. Note also that even if mixed acts are unavailable, one might argue that there is a certain doxastic state you could be in (namely 0.5 confidence that you will take A, 0.5 that you will take B), which is a stable equilibrium. See Skyrms (1990) and Arntzenius (2008) for discussion.

[11] See especially Egan (2007).

	in A	**in B**
If she predicted you would take Crate A, then she put	$1000	$1,100
If she predicted you would take Crate B, then she put	$1000	$0

In this case we can represent the recommendations of CDT like this:

Fig. 7



CDT says that if you are certain or near-certain (to be precise: if you have credence greater than 1000/1100 = $0.\overline{90}$) that you will take Crate A, then you ought to take Crate B. But wouldn't it be crazy to take Crate B, whatever you believe? You know, coming into the situation, that whatever you do, it will turn out that you would have been better off doing the other thing. You will regret your choice, whatever you do. So why do the thing with the terrible outcome, the thing you will really, really regret? If it is irrational to take Crate B, no matter what you believe about what you will do, then CDT is wrong, since it says that you ought to take B if you are very confident that you will not do so.

Causal decision theorists have a reply to this objection.[13] If you are certain that you will take Crate A, and so you are certain that Crate A contains $1,000 and Crate B $1,100, then indeed you ought to take Crate B. Of course you ought to take Crate B – you are certain that it contains more money! Now, it is true that if you do what you ought to do on these grounds, if you take Crate B, and you know that you are a causal decision theorist, then there would appear to be something defective about

[12] We should note that Egan appeals to different cases: the 'psycho-button' case, the 'murder lesion' case, . But they share the same general form – they are asymmetric self-frustrating cases.

[13] Thanks to Bob Stalnaker for putting this reply to us in a particularly forceful way.

you.[14] We can say: "Why were you so sure of something that turned out to false – namely, that you were going to take Crate A? Didn't you know that you were a causal decision theorist? Couldn't you have anticipated that this confidence that you were going to take Crate A would lead you to take Crate B?" But, if there is a defect here, it is the defect that comes with believing something that it is not epistemically rational to believe. And it is not the job of a theory of the subjective practical *ought* to tell us what it is epistemically rational for you to believe. It is the job of a theory of the subjective, practical *ought* to tell us what, given your beliefs, you ought to do. If you believe, against all evidence, that your mother is a murdering psychopath, then you ought to leave her house immediately. If you believe, against all evidence, that you have discovered a counter-example to Fermat's Last Theorem, then you ought to alert the media. You ought to do these things no matter whether your beliefs are epistemically rational.

The general point is that it is no mark against a theory of the subjective ought that sometimes people who are self aware, *epistemically irrational*, and doing what the theory says they ought to do, behave in odd, self-destructive ways. Sometimes odd beliefs license odd behavior. That is no great surprise.

5. Our Problem

Our problem with decision-dependent decision theories is this: Followers of decision-dependent theories will in some cases behave in odd, self-destructive ways if they are also self-aware and epistemically rational. They will, by anticipating

[14] Note that one might also think that it is simply *impossible* to do one thing while being near certain that you would not do it.

features of the very decision they are in the process of making, push themselves into situations that are not desirable even by their own lights. In this way, if any of SAD, SICK, or CDT is true, then when combined with our best theories of epistemic rationality, we wind up with a very unattractive picture of how rational agents behave. So much the worse for SAD, SICK, and CDT, and for decision-dependence more generally.

We will focus here on a case in which CDT, when combined with assumptions of self-awareness and epistemic rationality, yields unattractive results. Analogous cases can be constructed for SAD and SICK. We will spare you those details. The case:

Three Crates

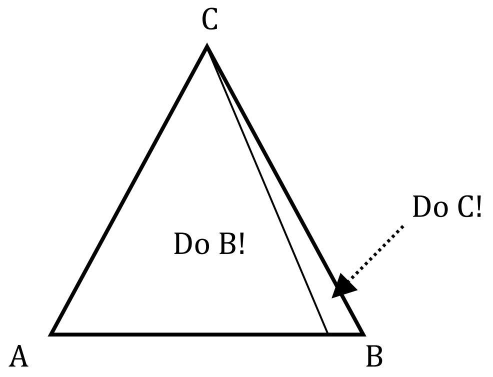
You know the demon behaved like this:

	in A	**in B**	**in C**
If she predicted A, then she put	$1,000,000	$1,001,000	$0
If she predicted B, then she put	$0	$0	$1,000
If she predicted C, then she put	$0	$0	$0

In this case the evidentialist takes Crate A, guided by her confidence that she will get $1,000,000 if she takes A, $0 if she takes B, $0 if she takes C.[15] What does the causalist do? We can represent the recommendations of CDT like this:

[15] Not all evidentialists would agree with this. Eells (1981), in arguing that evidentialism can recommend one-boxing in the Newcomb Problem (see Section 6), says that the demon's predictor and the agent's choice will have a common cause, so if the agent (by introspecting) can tell whether what that cause is. She will notice a certain 'tickle,' so to speak, which is either a common cause of an A-choice and an A-prediction, or a common cause of a B-choice and a B-prediction, etc. And then she will in effect be able to tell what the demon predicted and then take the box with the most money in it. Eells' approach, however, seems not to apply in cases where the agent cannot detect whether she has the relevant 'tickle' or in cases where it is stipulated that the agent's choice and the demon's prediction lack a common cause. In any event, we will henceforth consider only versions of evidentialism which do not appeal to Eells' so-called 'tickle defense.' Our evidentialist is of the sort who embraces one-boxing in the Newcomb Problem.

Fig. 9



What CDT recommends that you do depends on what you believe you will do. Roughly: If you are certain or near-certain that you will take Crate A or Crate B, then CDT recommends that you take Crate B. If you are certain or near-certain that you take Crate B or Crate C, and you have some confidence that you will take Crate B, then CDT recommends that you take Crate C. If you are certain that you will take Crate C then CDT says that all three options are equally desirable and permits you to take any of the three boxes. What the causalist does depends on what she believes she will do.

What does the *self-aware* causalist do? Let's make this question more precise. Suppose that you are practically rational by the standards of the causalist. In particular, suppose that

> 1. *You Respect Weak Dominance*
> If, right before you make your mind, you are sure that there is at least as much money in one crate as in another, and it is not the case that you are sure that there is at least as much in the other as in the one, then you will not take the other.

And suppose that you are self-aware. In particular, suppose that right before you make up your mind

> 2. *You are Sure that you Respect Weak Dominance*
> You are sure that 1 is true.
>
> 3. *Your Knowledge of the Contents of the Boxes is Luminous*
> If you are sure/unsure that money is distributed in the crates in a certain way, then you are sure that you are sure/unsure that money is distributed in the crates in a certain way.
>
> 4. *You are not Prone to Astounding Yourself*
> If you are sure that you will not take a particular crate then you will not take that crate.

How do you behave in this case, if all this is true of you – if you are practically rational by the standards of the causalist, and self-aware?

You take Crate C. To see why, first notice that, right before you make up your mind, you are sure that you will not take A. Argument: Suppose, for *reductio*, that you are unsure that you will not take A. So, by your confidence in the predictor, you are unsure that there is at least as much money in A as in B. So, by the description of the case, you are sure that there is at least as much money in B as in A, and unsure that there is at least as much money in A as in B. So, by 3 *Your Knowledge of the Contents of the Boxes is Luminous*, you are sure that (you are sure that there is at least as much money in B as in A, and unsure that there is at least as much money in

A as in B). So, by 2 *You are sure that you Respect Weak Dominance*, you are sure that you will not take A – but that's a contradiction.

It follows that, right before you make up your mind, you are also sure that you will not take B. Argument: You are sure that you will not take A. So, by your confidence in the predictor, you are sure that there is at least as much money in C as in B. Suppose, for reductio, that you are unsure that you will not take B. So, by your confidence in the predictor, you are unsure that there is at least as much money in B as in C. So, by 3 *Your Knowledge of the Contents of the Boxes is Luminous*, you are sure that (you are sure that there is at least as much money in C as in B, and unsure that there is at least as much money in B as in C). So, by 2, you are sure that you will not take B – but that's a contradiction.

Right before you make up your mind, you are sure that you will not take A or B. [16] So, by 4 *You are not Prone to Astounding Yourself*, you will take C. If you are rational by the standards of the causalist and self-aware then you will take C.[17]

We think that speaks very badly for rationality-by-the-standards-of-the-causalist. By taking Crate C, the self-aware causalist winds up with the princely sum

[16] We should note that Brian Skyrms and Frank Arntzenius have developed general accounts of what epistemically rational, self-aware causalists believe that they will do in situations in which the causal decision theoretic expected value of options depends on their beliefs about what they will do. Both accounts entail that you ought to end up in a *deliberational equilibrium*. Your credences about what you will do are in deliberational equilibrium iff, given those credences, all of the act-propositions to which you assign positive credence have equal causal expected utilities. Here, the only deliberational equilibrium is credence 1 that you will take Box C. Therefore, both accounts suggest that the epistemically rational, self-aware causalist will come to believe that she will take C. See Skyrms (1990) and Arntzenius (2008).

[17] Conditions 1-4 together yield what in game theory is called 'iterated elimination of (weakly) dominated strategies.' Starting with the initial 3 x 3 decision matrix, we rule out any weakly or strongly dominated acts and the proposition that the predictor predicted you would choose that dominated act. This results in a smaller 2 x 2 decision matrix. Then, we take this 2 x 2 matrix and rule out any dominated acts (along with the possibility of the predictor having predicted this action). And so on. We invoke conditions 1-4 to show *why* iterated elimination of dominated strategies is legitimate and also to highlight that it is *not* legitimate if the agent is not self-aware. (We also invoke conditions 1-4 to show that iterated elimination of weakly dominated strategies is as defensible as iterated elimination of strongly dominated strategies.)

of $0. Worse, by her own lights, taking Crate C *guarantees* her a return of $0. That is, given that she was certain she would take Crate C, she was certain that Crate C contained $0. Of course, had she thought that she might take Crate B, then she would not have been certain that Crate C was empty. But she didn't think she might take Crate B, and so she was certain that her choice of Crate C would yield $0. (Of course that is not to say that we cannot explain why she took Crate C – if she had taken any other box then she would not have been self-aware and rational. It is just to say that in explaining why she took Crate C we do not attribute to her any motivating reason to take Crate C. She did not take herself to have any reason to take Crate C, because, having convinced herself that the demon predicted she would take Crate C, she was certain that all the crates contained the same amount of money, and money, by hypothesis, is all she cared about.)

Our case against CDT (and against Decision Dependence more broadly) stops there. To be blunt: we think that the claim that a practically and epistemically rational person will take crate C in these circumstances is strongly counterintuitive and that this bears against the claim that CDT is the correct theory of practical rationality.[18]

But we can also dramatize the problem in the following way: The epistemically rational and self-aware evidentialist takes Crate A and, predictably enough, gets $1,000,000, while the epistemically rational and self-aware causalist takes Crate C

[18] To emphasize, our case against CDT (and decision-dependent theories more broadly) does not rest on considering an arbitrary case in which we have stipulated that you start out certain that you will take Box C. Rather, we have demonstrated that if you are a causalist and moreover are epistemically ideal (in the sense of being both self-aware and rational in responding to evidence), then you *must* wind up certain that you will take Box C.

and, predictably enough, gets $0. Consider how they might defend their rational honor:[19]

Evidentialist: I took Crate A and, predictably enough, got $1,000,000. You took Crate C and, predictably enough, got nothing.

Causalist: True, I am poor and you are rich. But consider what would have happened if we had behaved differently. If you had done as I did then you would have been still richer.

Evidentialist: No. I have $1,000,000. If I had taken Crate C, as you did, then I would have nothing.

Causalist: I mean that if you had *reasoned* as I did then you would be richer.

Evidentialist: You reasoned in a way that led you to the conclusion that C was the crate to take. If I had reasoned that way then I would have nothing.

Causalist: But you and I started with different beliefs about ourselves and the world. In particular, I was sure that I would respect Weak Dominance, so I was sure that I would not take Crate A, so I was sure that the demon had not predicted that I would take Crate A. You were not sure that you would respect Weak Dominance, so you were not sure that the demon had not predicted that you would take Crate A. If you had reasoned in the proper, causal decision theoretic way from there, then you would have taken Crate B, and walked away with $1,001,000.

[19] The evidentialist's charge against the causalist is, of course, the old 'Why Ain'cha Rich?' objection leveled against two-boxing in the Newcomb Problem (discussed in the next section). There, the evidentialist winds up rich and the causalist winds up poor. But in the Newcomb Problem, the causalist can respond that had she done as the evidentialist did, she would have been poorer, while if the evidentialist had done as the causalist did, he (the evidentialist) would have been richer (see Lewis 1999). As the following dialogue shows, however, in Three Boxes, neither counterfactual is true.

Evidentialist: If I had done all that then I would not have been self-aware. I would have followed causal decision theory without anticipating that I would follow causal decision theory. Self-awareness is an epistemic virtue, lack of it an epistemic defect. So, yes, if I had been epistemically sub-optimal[20], but practically optimal-by-your-standards, then I would have taken Crate B, and walked away with $1,001,000. But if I had been both epistemically and practically optimal-by-your-standards, if I had been the very model of epistemic and practical perfection, I would have taken Crate C, and walked away with nothing.

Causalist: Ok. So if you had done as I did, then you would not have been richer. But still, if I had done what you did then I would have been poorer.

Evidentialist: No, you have *nothing*. If you had done as I did then you would not have had less than nothing. You can't have less than nothing.

Causalist: Oh, right. But still, though I would not have been poorer if I had done as you did, at least I would not have been richer.

Evidentialist: Yes, you would have been no richer or poorer if you had behaved differently, but that hardly illustrates that you behaved in a uniquely rational way. Indeed, it makes your behavior puzzling. Why were you so intent on choosing box C, given that, as you chose box C, you were sure that it contained no money?

Causalist: I had to choose one of the crates. C was as good as any.

[20] Note that even if self-awareness is an epistemic virtue and lack thereof an epistemic defect, it may not be the case that one is *irrational* if one lacks self-awareness, at least so long as being epistemically sub-optimal does not entail being epistemically irrational. Note also that the type of self-awareness considered here (in 2-3 above and 2-4 below) is quite weak and does not require anything approaching complete knowledge of one's mental states and future choices.

Evidentialist: But it is not like C was a random choice. In this kind of situation you *always* choose C.

Causalist: Well, if I had not chosen C then I would not have been self-aware and rational.

Evidentialist: But you don't care about being self-aware and rational. You only care about money.

Causalist: I am self-aware and rational, so I just… do it.

Evidentialist: Curious. You consistently act in a way such that you are always sure that if you act that way then you will be pennilessness, though as you do it, you see no reason to do it. Maybe that is how lemmings feel as they dive off of sea cliffs: 'I see absolutely no reason to do this. But I am a lemming, dammit! So I just… do it.'[21]

The causalist comes off very badly in this exchange, in our view. He chose to do something that not only had no 'news value' for him (the thing that evidentialists care about and causalists do not) but also had no anticipated good consequences

[21] The causalist may remain unmoved by the 'Why Ain'cha Rich?' objection. Even if she no longer has recourse to the reply that she'd have been poorer and the evidentialist richer had each done what the other did, she might still make the reply that the rich evidentialist simply faced a good set of choices while she, the causalist, faced a bad set of choices, and she is not to be blamed for having faced a bad set of choices. But suppose that despite the causalist's 100% confidence that the demon would be accurate, the demon in fact made the wrong prediction, thinking instead that the causalist would take Crate B. In this case, the causalist takes Crate C and winds up with nothing but cannot blame her pennilessness on having faced a bad set of choices. Rather, she can only blame it on her *believing* herself to have faced a bad set of choices. But the fact that she had this belief cannot be blamed on the demon; any guilt rests entirely with herself.

(the thing that causalists care about and evidentialists do not.) This is not the behavior of a rational person.

6. Previous 'Why Aincha Rich?' Arguments

If you are familiar with the history of the debate between causalists and evidentialists, the above argument may remind you of a traditional 'Why Aincha Rich?' argument. Maybe so. But our argument is better than any previous such argument. We will explain why by talking about two of them.

The first, most famous 'Why Aincha Rich?' argument starts with the classic Newcomb case. In that case there is an opaque box and a transparent box. You have to choose between taking just the opaque box ('one-boxing') or taking both boxes ('two-boxing'). You see there to be $1,000 in the transparent box, but cannot see what is in the opaque one. You know that an unerringly accurate predictor put $1,000,000 in the opaque box if she predicted you would one-box and $0 in the opaque box if she predicted you would two-box.

The Classic Newcomb Case

One box or two boxes – which is it to be? You are sure that the unerringly accurate demon proceeded like this.

	in 1 box	**in 2 boxes**
If she predicted you would 1-box, then there is	$1,000,000	$1,001,000
If she predicted you would 2-box, then there is	$0	$1,000

Evidential Decision Theory recommends one-boxing, for your expected earnings, conditional on your one-boxing, are $1,000,000, whereas your expected earnings, conditional on your two-boxing, are $1,000. Causal Decision Theory, by contrast, recommends two-boxing, for you are certain that regardless of what the predictor predicted you would do, there is more money in both boxes combined than in the opaque box alone.[22]

When many people are placed in many Newcomb cases, those who one-box tend to wind up with $1,000,000, whereas those who two-box tend to wind up with $1,000. Moreover, this pattern is perfectly foreseeable, given the predictor's accuracy. 'So…' says the evidentialist to the causalist '…if you people are so rational, why aincha rich?'

The causalist concedes that she is poor, but blames her circumstances. 'You and I were in very different circumstances.' she tells the evidentialist 'I made the best of mine, while you made the worst of yours. The only thing we learn from your predictable riches is that it is possible for there to be mechanism that punishes people for having a disposition to behave as causalism recommends they behave. But that hardly tells against causalism. It is possible for there to be a mechanism that punishes people for having a disposition to behave as evidentialism recommends they behave – an intuitive psychopath goes around bashing the evidentialists on the head. For any decision theory it is possible for there to be a mechanism that punishes its followers.'[23]

No progress is made. And there is a good reason for this. The evidentialist and causalist may agree that a decision theory should be judged on whether it following it will, predictably, yield better results in relevantly similar cases, but they disagree on what cases count as 'relevantly similar.' For the causalist, cases are relevantly

[22] Actually, it is not quite as straightforward as this for CDT to recommend two-boxing, as we explain in the next section.

[23] Gibbard and Harper (1978), Lewis (1999).

similar when they are similar with respect to factors outside of the agent's control. So, for example, though two cases in which there is nothing in the opaque box may be relevantly similar, a case in which there is nothing in the opaque box is not relevantly similar to a case in which there $1,000,000 in the opaque box. Within classes of cases that are relevantly similar in this way causalists, predictably, tend to do better than evidentialists. For the evidentialist, cases are relevantly similar when they are similar with respect to the doxastic state of the agent. So, for example, when the evidentialist and the causalist find themselves in a Newcomb case, they are (no matter what the contents of the boxes are) in relevantly similar cases. Within classes of cases that relevantly similar in this way evidentialists, predictably, tend to do better than causalists.

A second sort of 'Why Aincha Rich' argument, this time against the evidentialist, has recently(ish) been proposed by Frank Arntzenius (2008). His goal is to imagine a case in which, in very similar betting situations, causalists come out better than evidentialists.[24]

> <u>Yankees or Red Sox?</u>
>
> The Yankees are playing the Red Sox. You know the Yankees win 90% of the time. You must bet on one team. A bet on the Yankees will win you $1 if they win, lose you $2 if they lose. A bet on the Red Sox will win you $2 if they win, lose you $1 if they lose. It would seem like betting on the Yankees is the way forward, but there's a wrinkle. Before you decide

[24] As we explain below in the next footnote, Arntzenius's case for the causalist predictably doing better than the evidentialist crucially relies on his stipulation that the causalist in his case is non-self-aware, not realizing that she is one who follows causalism. Lewis (1999) gives a proof to the effect that there cannot be a 'Why Aincha Rich?' objection leveled against evientialism, but his proof relies on the stipulation that all parties are self-aware, knowing their credences and utilities and also knowing which decision theory they follow.

> which team to bet on, a reliable predictor tells you whether you will win or lose. She says either 'You will win your next bet' or 'You will lose your next bet.'

An evidentialist in your position bets on the Red Sox, no matter what the prediction. That is because, conditional on the proposition that you next bet will win, a bet on the Red Sox wins $2, while a bet on the Yankees wins $1. Similarly, conditional on the proposition that your next bet will lose, a bet on the Red Sox loses $1, while a bet on the Yankees loses $2. So evidentialists always bet on the Red Sox, and predictably enough, do badly, since the Yankees win 90% of the time.

By contrast, according to Arntzenius, a causalist in your position bets on the Yankees, no matter the prediction. Crucially, he assumes that the caualist 'does not know he is a causal decision theorist, indeed that he has no credences about which bet he will take out when he calculates his causal utilities' (2008, fn 10). (We have serious qualms about how this stipulation affects the dialectic[25], but let us grant it for now.) Given this stipulation, the causalist effectively ignores the predictor's statement and takes the two relevant dependency hypotheses to be 'Yankees win'

[25] This stipulation is crucial, and also contentious. Arntzenius cannot get the result that the causalist will always bet on the Yankees, and hence that the causalist will predictably do better than the evidentialist in the long run, without it. For if the causalist is self-aware, then which bet he will take depends on what he starts out thinking that he will do. Suppose, for instance, that he starts off (for some reason) confident that he will bet on the Red Sox. Then, if the predictor tells him that he will win his next bet, then because he knows he is confident that he'll bet on the Red Sox, this gives him more evidence that the Red Sox will win, and also thereby gives him yet more reason to bet on the Red Sox. In effect, being told that he'll win his next bet puts the causalist in a case of self-reinforcing decision-dependence – whatever he starts off thinking he'll do, being told he'll win his next bet gives him yet more reason to do that thing. By contrast, being told that he'll lose his next bet puts the causalist in a case of self-frustrating decision-dependence – if he starts off thinking he'll bet on the Yankees, the prediction gives him evidence that the Red Sox will win and hence gives him reason to bet on the Red Sox, and similarly, *mutatis mutandis* for the case where he starts off thinking he'll bet on the Red Sox. What this means is that if we drop the stipulation that the causalist we are considering is non-self-aware, then we cannot get Arntzenius' desired result that the evidentialist predictably does worse than the causalist. Of course, it remains the case that the evidentialist would do better by becoming a non-self-aware causalist. Indeed, in the dialogue between the evidentialist and the causalist, we noted that the evidentialist in the Three Crates case would do better by taking Box B, but that doing so would only by licensed by CDT if she maintained some credence that she would take Box A, i.e. if she followed CDT without being aware that she was doing so. But given that non-self-awareness is an epistemic defect, it seems a poor defense of causalism to say that the evidentialist does worse than the non-self-aware causalist.

and ‘Yankees’ lose, assigning 0.9 credence to the former, 0.1 credence to the latter. That gives betting on the Yankees higher causal expected utility than betting on the Red Sox. So causalists always bet on the Yankees, and predictably enough do well, since the Yankees win 90% of the time.

But as with the initial ‘Why Aincha Rich?’ objection in the Newcomb Problem, Arntzenius’ version of the objection, this time raised against evidentialism, is inconclusive. The evidentialist will insist that we need to compare relative performance on cases that are *relevantly similar*, which for the evidentialist means cases that are alike with respect to the agent’s doxastic or evidential state. In this case, then, we need to compare how they do in cases where the predictor has said ‘You will win your next bet,’ on the one hand, and cases where the predictor has said ‘You will lose your next bet,’ on the other. But in each sort of case, the evidentialist does better than the causalist. In the former sort of case, the evidentialist wins $2 from betting on the Red Sox, while the causalist wins $1 from betting on the Yankees. In the latter, the evidentialist loses $1 from betting on the Red Sox, while the causalist loses $2 from betting on the Yankees. In ‘relevantly similar’ situations evidentialists, predictably, do better than causalists.[26]

So both ‘Why Aincha Richa?’ arguments yield a stalemate. The evidentialist says that when we look at what *she* regards as relevantly similar cases, we see that, predictably, she does better than the causalist, while the causalist says that when we

[26] Is it nevertheless true that in cases that are alike with respect to factors beyond the agent’s control, the causalist does better? No. In cases where the Yankees win, it is true that the causalist does better (winning $1, while the evidentialist loses $1), but in cases where the Red Sox win, the evidentialist does better than the causalist (winning $2, while the causalist loses $2). The reason the evidentialist does worse overall than the causalist is that cases where the Yankees win are so much more frequent than cases where the Red Sox win.

look at what she regards as relevant similar cases, we see that, predictably, she does better than the evidentialist.

The objection that we raise against Causal Decision Theory (and all theories that endorse Decision Dependence), based on the Three Crates case, yields no such stalemate. Recall that case:

Three Crates

You know the demon behaved like this:

	in A	**in B**	**in C**
If she predicted A, then she put	$1,000,000	$1,001,000	$0
If she predicted B, then she put	$0	$0	$1,000
If she predicted C, then she put	$0	$0	$0

Let the evidentialist and the causalist each operate with her preferred conception of relevant similarity of cases. The evidentialist can say that in relevantly-similar-in-her-way cases (similar with respect to the doxastic state of the agent) evidentialists, predictably, do better than causalists. All Three Crates cases are relevantly similar in this way, and in these cases the evidentialist always gets $1,000,000 while the causalist always gets $0. But the causalist *cannot* say that in relevantly-similar-in-her-way cases (cases similar with respect to things outside of the agent's control) causalists (who always take box C), predictably, do better than evidentialists (who always take A). In cases in which the demon predicted A, evidentialists do much, much better than causalists. In cases which the demon predicted C, evidentialists and causalists do equally badly (as we emphasized in the previous section, this is partly what makes the causalist's behavior so odd – she takes Box C while recognizing that even by causalist lights, she has no more reason to take C than to take A or B; it is just that were she to take one of the other boxes, she would either fail to be self-knowing, or fail to be epistemically rational, or fail to obey causalism). Only in cases in which the demon predicted B do causalists do better than evidentialists – but both causalists and evidentialists know they are not in such a case.

No stalemate.

7. Decision Dependence and the Newcomb Problem

Let us end by facing the Newcomb problem head-on and looking at how our argument that the subjective *ought* is not decision-dependent bears upon it.

First, note that the Newcomb Case is not a case in which CDT yields Decision Dependence. For CDT favors two-boxing no matter what credences you have concerning what you will do. No matter what credences you have concerning what you will do, the causal expected utility of two-boxing is $1,000 greater than the causal expected utility of one-boxing.[27]

The problem with CDT is that, as we have seen, it yields Decision Dependence in *other* cases. As a result it says that a practically and epistemically rational person will take Crate C in the Three Crates case. That raises a question: Can we save the intuition that two-boxing is the thing to do in the Newcomb Problem by coming up with another theory of practical rationality, some variant of CDT as standardly

[27] A qualification: as many commentators have noticed, if your credences concerning what you will do are *undefined* then in this case your credences concerning dependency hypotheses are undefined and the causal expected utilities of the options open to you are undefined. The point is that so long as you have credences concerning what you will do, the causal expected utility of two-boxing is greater than the causal expected utility of one-boxing.

understood, that tells us that a rational person will two-box in the Newcomb Case but does not tell us that a rational person will take Crate C in the <u>Three Crates</u> case?

We think the answer is no. There is no good theory of practical rationality with both these features. Two-boxing and taking-Crate-C stand or fall together.

We will not try to survey all possible attempts to devise a good theory of practical rationality with both features. That would be an endless game of whack-a-mole. Instead, we will aim for a general argument.[28]

To illustrate the idea, consider a sample theory that recommends two-boxing but not decision-dependence. In the Newcomb case, though the causal expected values of the options vary with your credences concerning what you will do, two-boxing always has higher causal expected value than one-boxing. So a natural first suggestion might be:

> *Supervaluationist CDT*
>
> You ought not to take an option iff there is some other option such that, no matter what credences you might have about what you will do, it has higher causal expected utility.[29]

This theory gives no recommendation between options in all cases where standard causal decision theory yields decision dependence between those options. But it says you ought not to one-box in the Newcomb case.

[28] Egan (2007) surveys and rejects a number of modified decision theories designed to yield two-boxing without giving counterintuitive results in a number of other cases. But he does not give a general argument that there can be no such decision theory. Briggs (2010) gives an Arrow-style argument that there are certain plausible constraints on rational decisions (including the constraint of respecting dominance reasoning) such that no decision theory can satisfy all the constraints together.

[29] More precisely, consider all the credences functions obtainable from your present credences by Jeffrey Conditionalizing with different credence distributions over the relevant act-propositions. If, relative to all of these new credence functions, one act-proposition has higher causal expected utility than another, then you ought not make true the other.

The problem with this theory is that sometimes it fails to take into account relevant evidence. Consider a modified version of the Newcomb case in which you have not two, but three options: *one-boxing*, *two-boxing*, and *taking a pistol and shooting yourself in the knee*. And suppose that if the predictor predicted that you would shoot yourself, then she put a little demand notice for $1,001 underneath the visible $1,000 in the transparent box, making one-boxing ever so slightly more attractive than two-boxing.

Newcomb Plus Pain

	by 1-boxing	**by 2-boxing**	**by shooting**
If she predicted *1-box*, then you get	$1,000,000	$1,001,000	pain
If she predicted *2-box*, then you get	$0	$1,000	pain
If she predicted *shoot*, then you get	$1,00,000	$999,999	pain

In this case Supervaluationist CDT yields that you ought not to shoot yourself in the knee, but not that you ought to one-box and not that you ought to two-box. If you assign sufficiently low credence to the proposition that you will shoot yourself in the knee (below around 0.99999) then two-boxing has higher expected utility than one-boxing. If you assign sufficiently high credence to the proposition that you will shoot yourself in the (above around 0.99999) then one-boxing has higher expected utility than two-boxing.

This is a very bad result. You know that you are not going to shoot yourself in the knee. That would be crazy. You are not crazy. So you know that there is no demand notice in the transparent box. It follows, we say, that you should treat this like the standard Newcomb case. If you are convinced that you ought to two-box in the standard Newcomb case, you should likewise be convinced that you ought to two-box in this modified case. By ignoring your conviction that you will not choose to shoot yourself, Supervaluationist CDT is ignoring relevant evidence about the contents of the boxes, relevant evidence about the causal consequences of your actions.

But the reasoning we have employed here (if you are rational then you will know you are not going to shoot yourself, so you will be epistemically licensed to 'cross out' the far-right column and the bottom row in Newcomb Plus Pain, and once you have done that, you will see that two-boxing has higher causal expected utility than one-boxing) is just the same as the reasoning that yielded taking Crate C in Three Crates (if you are rational then you will know you are not going to take Crate A, so you will be epistemically licensed to 'cross out' out the associated row and column in the decision matrix, and from there you will infer that you will not take Crate B, and so be licensed to 'cross out' the row and column for Crate B as well). In each case, you have evidence about which actions you will or will not perform, and this evidence is relevant to the causal consequences of the other actions. Why does this evidence have no bearing on what you ought to do?

This does not just point to a problem for a particular theory, Supervaluationist CDT. It also points to a more general argument that two-boxing in the Newcomb Problem and taking Crate C in Three Crates stand or fall together: The root motivation behind two-boxing in the Newcomb Problem is that actions ought to be evaluated based on their anticipated causal consequences. And it is an important fact about rationality that you ought to take into account your total evidence in determining what to believe or do. But these two things – focusing on causal consequences and taking into account total evidence – are all that is required to yield that a practically and epistemically rational person will take Crate C in Three Crates. So you can't have two-boxing be rational in the Newcomb Problem without also having taking Crate C be rational in Three Crates. To the extent that you agree with us that taking Crate C is irrational, you must think that two-boxing is irrational too.

8. Wrapping Up

Decision-dependence is a curious feature of a number of theories of practical rationality. Causal Decision Theory is the most famous such theory, but there are others, like SAD, which exhorts you to defer to your anticipated future desires, and SICK, which tells you to do what you believe will be good for your children.

Our view about Decision-Dependence is this: We want our best theories of practical rationality to hook up with our best theories of epistemic rationality, so as to allow us to paint an attractive picture of what someone who is rational in all epistemic and practical respects believes, desires and does. But, if any of SAD, SICK or CDT is true, then sometimes people who are ideal in all epistemic and practical respects, by taking into account evidence about the likely outcome of their present deliberation, end up compelled to perform actions that are no-good-by-the-lights-of-anybody. This is bad news for SAD, SICK and CDT, bad news for Decision-Dependence more generally, and bad news for two-boxers in Newcomb Cases.

References

Arntzenius, Frank. 2008. "No Regrets; Or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68 (2).

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

Derose, Keith. 2010. "The Conditionals of Deliberation." *Mind* 119 (476).

Eells, Ellery. 1981. "Causality, Utility, and Decision." *Synthese* 48: 295-329.

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

Gibbard, A. and Harper, W. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, Volume 1, ed. C. Hooker, J. Leach and E. McClennen, 125-162. Dordrecht, Holland: D. Reidel.

Hare, Caspar. 2007. "Voices From Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?" *Ethics* 117 (3).

Harman, Elizabeth. 2009. "'I'll Be Glad I Did it' Reasoning and the Significance of Future Desires." *Philosophical Perspectives 23, Ethics*

Harper, William. 1985. "Ratifiability and Causal Decision Theory." In Asquith and Kitcher (eds), *PSA 1984*, vol 2. East Lansing, Philosophy of Science Association. 213-228.

Harper, William. 1986. "Mixed Strategies and Ratifiability in Causal Decision Theory." *Erkenntnis* 24, 25-36.

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

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

Lewis, David. 1999. "Why Ain'cha Rich?" In *Papers in Ethics and Social Philosophy*, Cambridge: Cambridge University Press.

Richter, Reed. 1984. "Rationality Revisited." *Australasian Journal of Philosophy* 62, 392-403.

Skyrms, Brian. 1986. "Deliberational Equilibria." *Topoi* 5 (1).

Skyrms, Brian. 1988. "Deliberational Dynamics and the Foundations of Game Theory." *Philosophical Perspectives* 2, 345-367.

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

Sobel, J. Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press.

Weirich, Paul. 1986. "Decision Instability." *Australasian Journal of Philosophy* 63, 465-472.

Weirich, Paul. 1988. "Hierarchical Maximization of Two Kinds of Expected Utility." *Philosophy of Science* 55, 560-582.