Synthese (2012) 187:123–145
DOI 10.1007/s11229-011-0022-6

Regret and instability in causal decision theory

James M. Joyce

Received: 30 September 2011 / Accepted: 30 September 2011 / Published online: 27 October 2011

Abstract Andy Egan has recently produced a set of alleged counterexamples to causal decision theory (CDT) in which agents are forced to decide among *causally unratifiable* options, thereby making choices they know they will regret. I show that, far from being counterexamples, CDT gets Egan's cases exactly right. Egan thinks otherwise because he has misapplied CDT by requiring agents to make binding choices before they have processed all available information about the causal consequences of their acts. I elucidate CDT in a way that makes it clear where Egan goes wrong, and which explains why his examples pose no threat to the theory. My approach has similarities to a modification of CDT proposed by Frank Arntzenius, but it differs in the significance that it assigns to potential regrets. I maintain, contrary to Arntzenius, that an agent facing Egan's decisions can rationally choose actions that she knows she will later regret. All rationality demands of agents it that they maximize unconditional causal expected utility from an epistemic perspective that accurately reflects all the available evidence about what their acts are likely to cause. This yields correct answers even in outlandish cases in which one is sure to regret whatever one does.

Keywords Expected utility · Ratifiability · Causal decision theory · Regret · Decision instability · Reflection principle · Dynamics of deliberation

For help with this paper, I would like to thank Gordon Belot, Richard Bradley, Aaron Bronfman, Andy Egan, Dmitri Gallow, Allan Gibbard, Alan Hájek, Bill Harper, Wlodek Rabinowicz, Jan-Willem Romeijn, Teddy Seidenfeld, Dan Singer, Paul Weirich, and audiences at the London School of Economics, University of Kent, University of Waterloo, University of Missouri, and the 2009 Progic Conference.

J. M. Joyce (✉)
Department of Philosophy, University of Michigan, Ann Arbor, MI, USA
e-mail: jjoyce@umich.edu

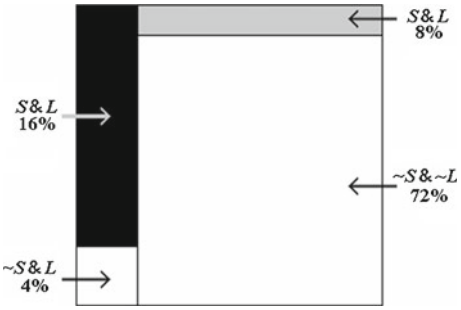

Fig. 1 In the *black region*, where you shoot (S) and have the lesion (L), you obtain the worst outcome $u\,(S, L) = -30$. In the *grey region*, where you shoot but lack the lesion, the best outcome $u\,(S, \sim L) = 10$ is achieved. In the white regions, where you cannot bring yourself to shoot, the status quo is preserved $u\,(\sim S, L) = u\,(\sim S, \sim L) = 0$

Egan (2007) has offered a series of purported counterexamples to causal decision theory (CDT) in which the choice of any act provides evidence about its own causal consequences, and this evidence undermines the act's rationale. Here is such a case:

> *Murder Lesion*. Life in your country would be better if you killed the despot Alfred. You have a gun aimed at his head and are deciding whether to shoot. You have no moral qualms about killing; your sole concern is whether shooting Alfred will leave your fellow citizens better off. Of course, not everyone has the nerve to pull the trigger, and even those who do sometimes miss. By shooting and missing you would anger Alfred and cause him to make life in your country much worse. But, if you shoot and aim true the Crown Prince will ascend to the throne and life in your country will improve. Your situation is complicated by the fact that you are a random member of a population in which 20% of people have a brain lesion that both fortifies their nerve and causes their hands to tremble when they shoot. Eight in ten people who have the lesion can bring themselves to shoot, but they invariably miss. Those who lack the lesion shoot only one time in 10, but always hit their targets. So, assuming for definiteness that the utility of killing Alfred has four times the magnitude of the disutility of shooting and missing,[1] your decision looks like Fig. 1.

Should you shoot?

The answer is not obvious. Since you know only the information given, your initial subjective probability estimates are $prob_0(S\,\&\,L) = 0.16$, $prob_0(S\,\&\sim L) = 0.04$, $prob_0(S\,\&\sim L) = 0.08$ and $prob_0(\sim S\,\&\sim L) = 0.72$. Thus, you initially see yourself as 20% likely to have the lesion and 24% likely to shoot.[2] Moreover, since $prob_0\,(S \mid L) = 0.8$ and $prob_0\,(S \mid \sim L) = 0.1$ you recognize a strong correlation between the presence/absence of the lesion and your tendencies toward/against

[1] This entails that you are entirely indifferent between the status quo and an arrangement in which a coin biased 3:1 in favor of tails is tossed and the bad/good outcome results from heads/tails. If one of these options seems better or worse to you, then you are operating with different utilities. The arguments of this paper go through just as well for different utility assignments.

[2] I assume throughout that it makes sense for agents to assign subjective probabilities to their potential actions in the course of their deliberations about what to do. There are decision theorists who disagree with this, most notably Levi (2000) and Spohn (1977). For defenses of act probabilities see Joyce (2002) and Rabinowicz (2002).

shooting. Likewise, since $prob_0(L \mid S) = 0.666$ and $prob_0(L \mid \sim S) = 0.055$, you also see the outcome of the decision to shoot/refrain as powerful evidence for/against the hypothesis that you have the lesion. This is odd. By deciding to shoot you will give yourself reason to think that you will miss, which makes shooting a bad choice. By deciding to refrain from shooting you give yourself reason to think that you would kill Alfred if you shot, which makes refraining a bad choice. So, neither act seems straightforwardly choiceworthy.

This sort of example is not original with Egan. It is an asymmetric version of the "Death in Damascus" case discussed in Allan Gibbard and William Harper's famous defense of CDT (1978), and it differs little from examples discussed in Gibbard (1992), Weirich (1985) and Pearl (2010). The defining feature of such examples is that they lack *causally ratifiable* acts. By choosing any act the agent provides herself with evidence for thinking that some alternative will be more effective at causally promoting desirable results.

Egan maintains that: (a) CDT recommends shooting as the only rational choice; (b) most people have a strong intuition that refraining is the only rational choice, even though it is not causally ratifiable[3]; (c) this intuition is correct. According to Egan, then, Murder Lesion falsifies both CDT and the idea that decisions should be ratifiabile.

I dispute (a) and (c). First, CDT does *not* recommend shooting as the uniquely rational act. Egan thinks otherwise only because he calculates expected utilities using probabilities that ignore causally relevant information. Second, while I agree with Egan that it would be wrong to shoot straightaway, I also think it would be wrong to refrain straightaway. Relative to the initial beliefs described in the problem, *neither* act can be rationally chosen. Even so, as I will show, agents who think their way through Murder Lesion from the perspective of CDT will wind up being correctly *indifferent* between shooting and refraining *once they have taken all their causally relevant information into account*. CDT, consistently applied, gets Murder Lesion exactly right, while those who follow their intuitions and choose to refrain straightaway are guilty of deciding before they have taken all their relevant information into account.

1 Why CDT does not advocate shooting straightaway

Egan believes that CDT requires you to shoot straightaway because it "enjoins [you] to *do whatever has the best expected outcome, holding fixed [your] initial views about the likely causal structure of the world*." (p. 96, emphasis Egan) According to CDT, the degree to which an act A promotes desirable results is given by its *causal expected utility*. In idealized cases where agents have subjective probabilities and utilities, one finds this quantity by first identifying an appropriate partition of "states of the world" $\{k_1, \ldots, k_N\}$ that offer alternative accounts of how outcomes causally dependon acts, and then calculating A's causal expected utility $\mathcal{U}(A) = \sum_n prob(k_n) \cdot u(A, k_n)$, where $prob(k_n)$ is the agent's subjective probability for k_n and $u(A, k_n)$ is the utility

[3] Actually, Egan recognizes that opinions are divided about Murder Lesion, but offers another puzzle—"The Psychopath Button"—about which people agree more frequently.

of the outcome that A would cause if k_n were to obtain. Applying this to Murder Lesion with $\{L, \sim L\}$ as the state partition we obtain:

$$\mathcal{U}_0(S) = prob_0(L) \cdot u(S, L) + prob_0(\sim L) \cdot u(S, \sim L) = 2$$
$$\mathcal{U}_0(\sim S) = prob_0(L) \cdot u(\sim S, L) + prob_0(\sim L) \cdot u(\sim S, \sim L) = 0$$

Since an act should be chosen only if its causal expected utility is at least as high as that of any alternative, CDT seems to recommend shooting right off the bat.

The weak link here is the initial probability assignment. Egan maintains that CDT requires using $prob_0(L) = 0.2$ to compute expected utilities. More generally, his view seems to be that CDT is committed to this:

> *Current Opinion Fixes Action*. If $prob_t$ characterizes an agent's beliefs at time t, then at t she is rationally obliged to perform an act A that maximizes her time t causal expected utility: $\mathcal{U}_t(A) = \sum_n prob_t(k_n) \cdot u(A, k_n)$.

If this were correct, then CDT would recommend shooting. But, causal decision theorists will resist the idea that current opinions should always determine actions. They will say, instead, that current opinions should only decide action when those beliefs reflect all the available evidence about what acts are likely to cause. This evidential completeness is precisely what is lacking in your initial beliefs in Murder Lesion.

Imagine a Blackjack player who has seen her top card (a seven) and the dealer's top card (an eight), but who has yet to peek at her hole card, which she can do cost-free. The player knows that she should stand pat if her cards total 17 or more, and that she should ask to be "hit" with another card if they total 16 or fewer. Suppose she calculates her chance of having at least 17 *without* looking at her hole card, and finds it to be 0.4, so that the expected payoff of taking a hit exceeds that of standing pat. While this is fine as an academic exercise, if the player took a hit on this basis we would think her daft. Even though she can assess probabilities and utilities without factoring in the hole card, she clearly should not *act* on such assessments. Rather, before deciding whether or not to take a hit the player faces a prior decision about whether or not to gather information about the likely effects of her acts. Since the costs of peeking at her hole card are negligible when compared to the costs of winning or losing the hand, she should not take a hit until she has ensured that her decision is based on beliefs that reflect all the freely available evidence about what her acts might cause.

In the same way, before deciding whether or not to shoot in Murder Lesion you should be sure to gather all the information about the causal consequences of your acts that is freely available to you. As a general matter, CDT is committed to two principles that jointly entail that initial opinions should fix actions most of the time, but *not* in decisions like Murder Lesion. The first is this:

> *Current Evaluation*. If $prob_t$ characterizes your beliefs at t, then at t you should *evaluate* each act by its causal expected utility computed using $prob_t$.

This says nothing about what you should *do*; it concerns only how you evaluate acts at t given your beliefs and desires at t. It is consistent with this that your time-t

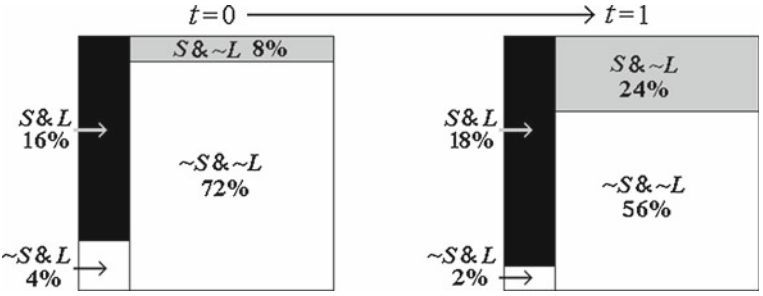

Fig. 2 $prob_0(S \,\&\, L \mid \mathcal{U}_0(S) = 2) + prob_0(\sim S \,\&\, L \mid \mathcal{U}_0(S) = 2) = 0.2$

evaluations should not be acted upon unless they meet some additional constraint. As the Blackjack example illustrates, the right constraint is:

> *Full Information*. You should act on your time-t utility assessments only if those assessments are based on beliefs that incorporate all the evidence that is both freely available to you at t and relevant to the question about what your acts are likely to cause.

The combination of this principle and the previous one tells you to act on your time-t causal expected utilities just when your time-t subjective probabilities reflect all available information about what your acts are likely to cause.

Maximizing expected utility relative to your initial opinions in Murder Lesion violates Full Information. It is just like taking a hit in Blackjack without looking at your hole card. You have easy access to your time-0 utilities, but you have yet to factor them into your thinking about the causal consequences of your acts. But, in any version of Murder Lesion that can pose problems for CDT it will be true that $prob_0(L \mid \mathcal{U}_0(S) = 2) \neq prob_0(L)$, which makes S's utility evidentially relevant to your beliefs about the lesion.[4]

To see why CDT is safe when learning time-0 utilities has no effect on L's probability, suppose that information about act utilities says nothing about the lesion, so that updating on $\mathcal{U}_0(S) = x$ leaves L's probability fixed at 0.2. In a case where S's probability rises from 0.24 to 0.42 the resulting change might look like Fig. 2. Such a picture would make sense if, say, you have a chance device implanted in your brain that, at the moment of choice, flips into the lesion state with frequency 0.2. Evidence about current utilities that raises or lowers S's probability will then leave L's probability intact since the fact that you judge shooting to be the best way to promote desirable results at $t = 0$ says nothing about how the chance process will resolve but, if your level of confidence in L remains the same after taking S's utility into account, then CDT does tell you to shoot, which is fine since shooting *is* the rational act in that situation! If $prob_0(L \mid \mathcal{U}_0(S) = x) = prob_0(L)$, then shooting continues to have a utility of 2 after you update, $\mathcal{U}_1(S) = \mathcal{U}_0(S \mid \mathcal{U}_0(S) = x) = 2$. Given this insensitivity of S's utility to information that alters its probability, the decision you face is no different

[4] You also know $\mathcal{U}_0(\sim S) = 0$, but since this is true whatever your beliefs about the lesion, I will consistently fail mention it. So, $\mathcal{U}_0(S) = 2$ really means $(\mathcal{U}_0(S) = 2 \,\&\, \mathcal{U}_0(\sim S) = 0)$.

from what it would be if you knew for sure that you lacked the lesion but learned that your gun misfires 20% of the time. Either way you are deciding whether to accept or reject a bet that offers a one-in-five chance of −30 utiles and a four in-five chance of 10 utiles. That choice is easy: you take the bet! So, the Murder Lesion of Fig. 2 poses no threat to CDT.

Accordingly, we will focus on versions of the problem in which the data about act utilities provides evidence about the likely consequences of your acts, i.e., those which obey the requirement

$$(\alpha) \quad \text{If } x \neq 0, \text{ then } prob_t(L \mid \mathcal{U}_t(S) = x) \neq prob_t(L).$$

In short, coming to recognize that you prefer shooting (or refraining) will lead you to reassess your views about how likely you are to have the lesion.

The data that $\mathcal{U}_0(S) = 2$ should also reinforce your confidence that you will shoot. The reasons for this have to do with your conception of yourself as a free agent with the power to perform any act you want. Your initial situation one in which $\mathcal{U}_0(S) > \mathcal{U}_0(\sim S)$ and $prob_0(S) < prob_0(\sim S)$. Assuming that you are a causal expected utility maximizer, this places you in the peculiar position of regarding shooting as your best option and yet of seeing yourself as *un*likely to shoot. No rational agent who takes herself to be free in the matter of shooting will let this stand. Once you believe that shooting is your best option you will become more inclined to shoot and, since you see yourself as controlling your acts, you will become more confident that you will shoot. In general, a causal expected utility maximizer who takes herself to have a free choice in the matter of A versus $\sim A$, will treat the fact that A has a higher current $\mathcal{U}$-value than $\sim A$ as a reason to increase her confidence in A. This is because: (i) the higher the $\mathcal{U}$-value of an act at t the more favorable the agent is toward that act at t, and (ii) since she regards herself as free to do what she wants, the more favorable she is toward an act at t, the more confident she is at t that she will ultimately perform that act. We will explore the relationship between knowledge of utilities and beliefs about acts more fully in the next section, but for now it suffices to understand that any causal expected utility maximizer who sees herself as having a free choice about A versus $\sim A$ at t will satisfy[5]:

$$(\beta) \quad \text{If } prob_t(A) < 1 \text{ and } x > y, \text{ then } prob_t(A \mid \mathcal{U}_t(A) = x \,\&\, \mathcal{U}_t(\sim A) = y) > prob_t(A).$$

[5] Some might object that a perfectly rational agent will already know $\mathcal{U}(A) = x$ since this is a logical consequences of facts about the initial probabilities and utilities. This is right: if perfect rationality requires logical omniscience, then my argument fails for perfectly rational agents. However, this will be cold comfort to Egan since a perfectly rational agent who takes herself to be free in the matter of A will never be in a position where $prob(A) < prob(\sim A)$ and $\mathcal{U}(A) > \mathcal{U}(\sim A)$, since this would contradict her belief in her own freedom of action. Applied to Murder Lesion, this means that our initial probability assignment is *not* one that any perfectly rational free agent can hold. Egan's counterexample is thus scotched at the start. As we will see, perfectly rational agents always hold beliefs that are in deliberational equilibrium. Less than perfect agents, who do not immediately see the consequences of their beliefs and desires, must reason their way into these equilibria by updating on what they learn about the utilities of their actions. In this connection, it is also important to note that the condition in α should be read so that $\mathcal{U}(A)$ and $\mathcal{U}(\sim A)$ are *non-rigid* designators for the causal expected utilities of A and $\sim A$.

In short, coming to recognize that A has a higher causal expected utility than $\sim A$ should make an agent more confident that she will do A (unless she is already certain of this). In Murder Lesion, someone with the initial probability assignments who comes to recognize that $\boldsymbol{\mathcal{U}}_0(S) = 2$ must revise her beliefs so that $prob_0(S \mid \boldsymbol{\mathcal{U}}_0(S) = 2) > prob_0(S)$.

Now, in light of α and β, it should be clear why Egan was wrong to think that CDT "enjoins [you] to do whatever has the best expected outcome, holding fixed [your] initial views about the likely causal structure of the world." As we have seen, CDT only advises you to act on your initial views when these views reflect all easily available information about what your actions might cause. In any version of Murder Lesion that can pose problems for CDT some of this information is missing since your initial views overlook the fact that shooting seems like your best option. This affects your estimate of the probability that you will shoot, and influences your estimate of how likely you are to have the lesion. So, maximizing causal expected utility relative to your initial opinions really is like taking a hit in Blackjack without peeking at your hole card. You see the value of $\boldsymbol{\mathcal{U}}_0(S)$ as evidentially relevant to the causal consequences of your acts, and, since you can easily know $\boldsymbol{\mathcal{U}}_0(S) = 2$, Full Information prevents you from coming to any final decision until you have taken this fact into account. CDT thus definitively *prohibits* you from acting on the basis of your initial probabilities in Murder Lesion, when the theory is properly understood as requiring Full Information and α and β.

It is, unfortunately, not surprising that Egan though otherwise since causal decision theorists often misleadingly speak as if only initial probabilities matter (Lewis 1981, pp. 12–13). The reasons for this can be traced to the historical accident that CDT was developed—by Robert Stalnaker, Allan Gibbard, William Harper, Brian Skyrms, Nancy Cartwright, and others—in response to a famous article by Robert Nozick (1969) about *Newcomb problems*. In these odd decisions a *dominated* act indicates a very desirable outcome that it does nothing to cause, while a *dominating* act causes a mildly desirable result but non-causally indicates a very undesirable outcome. We can "Newcombize" Murder Lesion by altering the utilities so that shooting dominates. It might be that if you shoot and miss then, instead of increasing the harshness of his rule, a chastised Alfred will be slightly less repressive for a day before going back to his old ways, in which case $u(\text{shoot \& hit}) = 10 > u(\text{shoot \& miss}) = 1 > u(\text{status quo}) = 0$. Here it is fine to act on the basis of initial probabilities, but this is *not* because they underwrite *correct* expected utility evaluations from which to act. Rather, it is because they are irrelevant to how you should act. At any time t, CDT assesses the merits of shooting in Newcombized Murder Lesion as $\boldsymbol{\mathcal{U}}_t(S) = prob_t(L) \cdot 1 + prob_t(\sim L) \cdot 10 = 10 - 9 \cdot prob_t(L)$, which exceeds $\boldsymbol{\mathcal{U}}_t(\sim S)$ for *any* value of $prob_t(L)$. The dominance structure of Newcomb problems thus renders moot the distinction between acting on less-than-fully-informed probabilities and acting on probabilities that reflect all available evidence about what your acts might cause, a distinction which is crucial in non-Newcombized Murder Lesion, where dominance considerations are not in play.

To put it differently, focusing exclusively on Newcomb problems can mislead one into thinking that CDT categorically prohibits one from using information about what one is inclined or likely to do as evidence for anything. In particular, it can seem that

one is required to ignore anything that one's attitudes toward acts might indicate about the causal structure of the world. If this were so, then probabilities of the form *prob*(act) or *prob*(state | act) would always be irrelevant to decision making; only *prob*(state) values would matter. This wrongly makes it seem as if the fundamental mistake in evidential decision theory lies in calculating expected utilities using *prob*(state | act) rather than *prob*(state). Properly understood, however, CDT *requires* rational agents to take account of the evidential import of their attitudes toward their own acts *insofar as these attitudes bear on questions about what those acts are likely to cause*. While CDT does tell agents to ignore what their acts might indicate about aspects of the world they cannot causally influence, it requires them to attend to all information, even information about what they are inclined or likely to do, that pertains to the causal powers of their acts.

When seen this way, the deep flaw in evidential decision theory shows up only *after* all causally relevant information is taken into account. Suppose you have processed all such information in the Newcombized Murder Lesion. It is at *this* point, when you know everything you can know about what your acts will cause, that the evidentialists go wrong. Instead of telling you to assess the utilities of all your actions on the same basis using your fully informed opinions, they tell you to assess shooting from an epistemic perspective which takes S to be certainly true, but also to assess refraining from a contrary epistemic perspective which takes S to be certainly false! At least one of these perspectives (and maybe both) will conflict with your fully informed views about S's probability. In situations like this, where $prob(\text{state}) \neq prob(\text{state} \mid \text{act})$ even *after* all causally relevant information has been factored in, disparities between causal and evidential expected utilities reflect real distinctions between the values of acts as causes and their values as mere indicators of outcomes the lie beyond the agent's influence. The mistake in evidential decision theory is not that it pays attention to what acts indicate, but that it pays attention to what acts indicate *about aspects of the world that the agent knows she cannot change*.

2 Ratifiability and deliberation

We have seen that CDT does not advocate shooting on the basis of your initial subjective probabilities. Even so, Egan's objection would still have force if the theory required you to shoot on the basis of whatever probabilities you come to have once you take all easily available causally relevant information into account. Fortunately, this is not the case. For CDT to uniquely recommend shooting there would have to be an epistemic state, reflected in a subjective probability $prob_t$ and associated expected utility $\mathcal{U}_t$, such that:

(i) So as not to run afoul of Full Information, $prob_t$ incorporates all available relevant information about what S and $\sim S$ might cause. In particular, it incorporates the utilities of S and $\sim S$, so that $prob_t(L \mid \mathcal{U}_t(S) = x) = prob_t(L)$.
(ii) $\mathcal{U}_t(S) > \mathcal{U}_t(\sim S)$, so that shooting is definitely preferred.
(iii) $prob_t$ assigns S a probability close to 1, so as not to run afoul of β.

To see what's wrong with such an assignment, we must explore the methods by which you might revise your subjective probabilities in light of information about current utilities. The details of this process will depend on your views about Murder Lesion's causal structure. You might think that the lesion influences your choice only by affecting your time-t beliefs and desires (perhaps as a consequence of its affect on your initial beliefs and desires), in which case S and L will be independent conditional on $\boldsymbol{\mathcal{U}}_t(S) = 2$. Or, you might see lesion's presence or absence as directly causing your action without mediation from desires, in which case L and $\sim L$ will screen off S from its utility, so that $prob_t(S \mid L \,\&\, \boldsymbol{\mathcal{U}}_t(S) = x) = prob_t(S \mid L)$ and $prob_t(S \mid \sim L \,\&\, \boldsymbol{\mathcal{U}}_t(S) = x) = prob_t(S \mid \sim L)$.[6] Or, you might think that your desires directly causes your act, but that there are non-causal correlations between having the lesion and freely shooting and between lacking the lesion and freely refraining. Here S and $\sim S$ will screen off L from $\boldsymbol{\mathcal{U}}_t(S) = 2$. Finally, you might see the lesion as continually influencing your beliefs and desires, while remaining convinced that you will perform whatever act you ultimately deem to be best (a judgment that may be influenced by the lesion). L will then be evidentially relevant to S given knowledge of S's utility, $prob_t(S \mid L \,\&\, \boldsymbol{\mathcal{U}}_t(S) = x) \neq prob_t(S \mid \boldsymbol{\mathcal{U}}_t(S) = x)$, but L will not be the whole story since you will see S's time-t utility as an additional source of information about your act, so that $prob_t(S \mid L \,\&\, \boldsymbol{\mathcal{U}}_t(S) = x) \neq prob_t(S \mid L)$.

For present purposes it does not matter which of these models is in play. What does matter is that data about S's utility should not undercut the basic character of Egan's counterexample. If, at any time t, updating on $\boldsymbol{\mathcal{U}}_t(S) = x \neq 0$, leaves L's probability unchanged, then the example poses no threat to CDT. When $x > 0$ updating on $\boldsymbol{\mathcal{U}}_t(S) = x$ increases the probability of shooting (as required by β), but leaves its utility at x, so that $\boldsymbol{\mathcal{U}}_t(S \mid \boldsymbol{\mathcal{U}}_t(S) = x) = x$. Conditioning on this new utility information will again raise S's probability and leave L's probability (and S's utility) intact. Iterating this cycle of belief revision, as Full Information requires, will lead you inexorably to a state in which you want to shoot, you are certain you will shoot, and where L's probability retains its initial value of $prob_t(L) < 0.25$. So, CDT tells you to shoot when $x > 0$ and $prob_t(L \mid \boldsymbol{\mathcal{U}}_t(S) = x) = prob_t(L)$. But, as we saw for $t = 0$ case, this poses no threat since shooting is the uniquely rational act in the situation. By symmetrical reasoning, if $x < 0$ and $prob_t(L \mid \boldsymbol{\mathcal{U}}_t(S) = x) = prob_t(L) > 0.25$, then CDT tells you to refrain from shooting, and this is the best thing to do as well. The moral is that if we are interested in versions of Murder Lesion that might pose a threat to CDT, we must supplement α and β with:

[6] This conception of the problem is not really compatible with the idea that you are a free agent. While the difficulty does not show up in Murder Lesion itself, it becomes obvious in a modified (Newcombized) version of the problem that has the same underlying causal structure but with utilities adjusted so that $u(S \,\&\, \sim L) = 10 > u(S \,\&\, L) = 1 > u$ (status quo) $= 0$, thereby making S the dominating act. In this case, α requires that reflecting on S's capacity to produce good results should *always* make you more confident that you will shoot (unless you already know you will). So, when you have taken all causally relevant information about your acts' utilities into account, you should wind up *certain* that you will shoot. But, if $prob(S \mid L)$ and $prob(S \mid \sim L)$ are held fixed, then S's probability must remain between 0.1 and 0.8, and you are ultimately left in the vexed position of knowing that shooting is your best option and yet of thinking that you are at least 20% likely *not* to do it. This is not compatible with seeing yourself as being fully free in the matter of shooting.

(χ) At each time t, $prob_t(L \mid \boldsymbol{\mathcal{U}}_t(S) = x) \neq prob_t(L)$ when $x \neq 0$.

This requirement is consistent with any of the update policies described above, and it seems essential to preserving the character of Egan's example.

It should be clear that no probability function satisfies (i), (ii) and χ. (i) requires $prob_t(L \mid \boldsymbol{\mathcal{U}}_t(S) = x) = prob_t(L)$. In light of χ this can only be true if $x = 0$, which requires $prob_t(L) = 0.25$. But, (ii) requires $prob(L) < 0.25$. So, CDT does not categorically advocate shooting in any version of Murder Lesion in which shooting is not the right choice. To complicate matters, CDT does not categorically advocate refraining either, unless it is the right choice, since no probability satisfies (i) together with (ii*) $\boldsymbol{\mathcal{U}}(S) < \boldsymbol{\mathcal{U}}(\sim S)$. Again, (i) requires that $prob_t(L) = 0.25$, but (ii*) requires $prob(L) > 0.25$. Thus, when properly understood CDT does *not* advocate either shooting or refraining as the uniquely permissible act except in cases where they are correct because χ fails.

This odd state of affairs is explained by the fact that (i) and χ and mandate that $\boldsymbol{\mathcal{U}}(S) = \boldsymbol{\mathcal{U}}(\sim S)$. For any value of x other than zero, χ requires that $\boldsymbol{\mathcal{U}}(S) = x$ is evidentially relevant to L. Hence, the *only* way to satisfy (i) is by having $\boldsymbol{\mathcal{U}}(S) = \boldsymbol{\mathcal{U}}(\sim S)$, so that $prob(L) = 0.25$. This state is precisely the one in which all your available causally relevant information has been taken into account.

To see this, it is useful to think of the $prob(L) = 0.25$ state as the limiting point of an idealized process of rational deliberation. In his (1990), Brian Skyrms developed a formal model of deliberation which incorporates the idea that rational agents believe they will maximize causal expected utility. On Skyrms' picture, deliberation starts with the agent in an initial belief/desire state $(prob_0, \boldsymbol{\mathcal{U}}_0)$, and proceeds through a sequence of stages $(prob_t, \boldsymbol{\mathcal{U}}_t)$, $t \leq 1$, to a final equilibrium $(prob_e, \boldsymbol{\mathcal{U}}_e)$. At every stage, each act A has an expected utility $\boldsymbol{\mathcal{U}}_t(A)$, which gives the agent's time-t assessment of A's value, and a probability $prob_t(A)$, which is her estimate at t of how likely she is to perform A once her deliberations end. The agent's total wellbeing at t, the "status quo", is given by $\boldsymbol{\mathcal{U}}_t(\mathrm{T}) = \sum_A prob_t(A) \cdot \boldsymbol{\mathcal{U}}_t(A)$. Acts with expected utilities that exceed the status quo are seen as potential improvements. Those that fall short are potential pitfalls.

In Skyrms' model, the deliberative process iterates through two steps, which I will describe for the case in which the act partition $\{A_m : m = 1, 2, \ldots, M\}$ is finite.

> *Step-1*: The agent assesses the utilities of acts in light of her time-t beliefs about the state of the world.

Here the agent learns a proposition $\boldsymbol{\mathcal{U}}_t(A_1) = x_1 \,\&\, \ldots \,\&\, \boldsymbol{\mathcal{U}}_t(A_M) = x_M$, where the value of x_m is acquired by computing $\boldsymbol{\mathcal{U}}_t(A_m) = \sum_n prob_t(k_n) \cdot u(A_m, k_n)$ using time-t probabilities for the states K_n that are assumed to be known (see *Step-2*).

> *Step-2*: The agent alters her probabilities for acts and states in light of utilities using an *update rule* that "seeks the good" by increasing probabilities of acts with utilities above the status quo, decreasing probabilities of acts with utilities below the status quo, and leaving probabilities of acts at the status quo unchanged.

This is a manifestation of the idea underlying β. One can think of the deliberator as learning the values of $\mathcal{U}_t(A_m)$, and then updating on what she learns in a way that reflects both (a) the fact that the $\mathcal{U}_t$-values are her best estimates, in light of her evidence at t, of the tendencies of her acts to cause desirable results, and (b) the fact that she is confident that she is free to choose the act that, in her estimation, will cause the best results.

For our purposes, the details of the update rule matter little. In addition to seeking the good, we only ask that the rule not change act probabilities so abruptly that acts with below average expected utilities are summarily assigned probability zero, and so removed from consideration. While acts that begin with positive probability can end up having zero probability in the equilibrium limit, their probability should not vanish along the way unless they achieve the minimum expected utility.[7]

For an example of such an update rule, consider *Bayesian dynamics in discrete time*. Here learning occurs at instants $t_0, t_1, \ldots$, governed by the rule $prob_{t+1}(A) = prob_t(A) \cdot [\mathcal{U}_t(A)/\mathcal{U}_t(\mathbf{T})]$, where the utility scale is chosen to have its minimum at zero. In keeping with (a) and (b) above, we think of this rule as expressing the agent's views about the evidential impact of learning the utilities of her acts at t, so that

$$\begin{aligned} prob_{t+1}(A_i) &= prob_t(A \mid \mathcal{U}_t(A_1) = x_1 \,\&\, \ldots \,\&\, \mathcal{U}_t(A_M) = x_M) \\ &= prob_t(A_i) \cdot \mathcal{U}_t(A_i) \Big/ \left[\sum_m prob_t(A_m) \cdot \mathcal{U}_t(A_m)\right].^{8} \end{aligned}$$

Once Step-2 is completed, the agent returns to Step-1 and repeats until she reaches an equilibrium at which $\mathcal{U}_{t+1}(A) = \mathcal{U}_t(A)$ for all acts A. (This equilibrium is sure to exist and be unique in decisions like Murder Lesion.) Broadly speaking, the agent's reasoning is a kind of feedback loop in which she revises her beliefs in light of varying assessments of the causal efficacy of acts, and then revises her assessments of the causal efficacy of acts in light of her varying beliefs. At any stage, a difference between $\mathcal{U}_{t+1}(A)$ and $\mathcal{U}_t(A)$ indicates that the time-t causal expected utilities provide the agent with relevant information about what her acts are likely to cause. An equilibrium state is reached only after all such information has been taken into account, which means that the agent can act on the basis of her equilibrium causal expected utility assessments without running afoul of Full Information.

In addition to the requirement to seek the good, Skyrms model imposes no further constraints on the updating process. In light of χ, however, we must ask that $prob_{t+1}(L) = prob_t(L \mid \mathcal{U}_t(S) = x) \neq prob_t(L)$ when $x \neq 0$, so that learning that you are definitely inclined toward or against shooting always provides you with evidence about the presence of the lesion. There are any number of update rules that can have this feature, but the crucial point is that on any of them it is only possible to obtain an equilibrium when $\mathcal{U}_t(S) = 0$ and $prob_t(L) = 0.25$. This reflects the fact that the

[7] This restriction makes sense given that the deliberative process should be sensitive to the fact that acts can become more or less desirable as they become more or less likely, as in Murder Lesion.

[8] In Murder Lesion, this yields $prob_{t+1}(S) = prob_t(S) \cdot (x + 30) / prob_t(S) \cdot x + 30$.

agent retains her views about the causal structure of the decision problem throughout her deliberations.

Deliberation often ends with a single action of maximum expected utility being assigned probability one and the rest being assigned probability zero, in which case the probability one act is chosen. But, in some decision problems, Murder Lesion being one such, deliberation can end in a mixed state in which $prob_e(A) > 0$ for more than one act, and $\mathcal{U}_e(A) = \mathcal{U}_e(B)$ for all such acts. Here the agent is torn among *equally desirable* actions given the beliefs she has after processing all available data about the likely causal consequences of her acts. In such a circumstance, CDT says that any act with positive probability, or any probabilistic mixture of such acts, may be rationally chosen since all have the same causal expected utility.

This has consequences for Murder Lesion whose equilibrium is attained when $\mathcal{U}_{e+1}(S) = \mathcal{U}_e(S) = 0$ and $prob_e(L) = 0.25$, the only point at which $prob_{e+1}(L) = prob_e(L)$. This is the unique stable point of deliberation, the only point at which all available information about the causal properties of acts has been taken into account. It is also the only epistemic state consistent with β and χ. Given Full Information, this is *the* state that you should use when assessing causal expected utilities for purposes of action. So, you must base your choice in Murder Lesion on the assessments of causal expected utility characterized by $\mathcal{U}(S) = \mathcal{U}(\sim S)$. It follows that CDT, rather than recommending either action alone, tells you to be entirely indifferent between shooting and not shooting. Once you have processed all the available information about what your acts might cause, you can rationally choose to shoot, to refrain from shooting, or to perform any "mixed act" that leads to shooting with probability p and to refraining with probability $1 - p$. All these choices are on a par with respect their ability to cause desirable results since each maximizes causal expected utility given full information about the causal properties of your acts. You can't go wrong in Murder Lesion, whatever you do!

3 Is it irrational to choose unratifiable acts?

Some will recoil from this conclusion. They will see the fact that CDT permits shooting at all as reason enough to jettison the theory. No plausible decision theory, they will say, should ever endorse shooting, even if it endorses refraining as well. There are, I think, two tempting reasons for holding such a view.

First, it just might just seem intuitively clear that shooting cannot be permitted. Indeed, Egan notes that many people have this intuition, and he takes this as a reason to regard shooting as irrational. I am less inclined to give such intuitions weight. Many philosophers seem to accept something like this: a strong and commonly held intuition that a given action is rational/irrational is powerful evidence for concluding that the act is rational/irrational, especially when the intuition persists in intelligent people under reflection. I think this is wrong. We have more than fifty years of research by cognitive psychologists showing that people make a wide range of predictable and systematic *errors* when evaluating acts, and that these errors often persist under reflection. Here is a partial list [see Shafir and Tversky (1995) and Gilovich et al. (2002) for more]: people are more concerned with gains and losses, seen as changes in some perceived status

quo, than with total well-being; people have inconsistent attitudes toward known risk, eschewing it when pursuing gains, but seeking it when avoiding losses; people pursue projects into which they have "sunk" costs even when doing so produces no benefits; people sometimes reverse their preferences when options are described differently or when they are offered an inferior choice whose availability says nothing about the values of other choices. In all these cases strong, stable, widely shared intuitions are *mistakes* which provide no reason for doubting normative theories that recommend the "unintuitive" acts.

Perhaps the common intuitions in Murder Lesion merit the same treatment.[9] The preference against shooting might simply be a manifestation of a misassessment of risks. After all, refraining seems like the "status quo" while shooting is a "risky" attempt to improve things, just the sort of situation in which people tend to be more risk averse than their beliefs and desires warrant. This interpretation is buttressed by the fact people feel less comfortable refraining when the problem is framed so that shooting is portrayed as a way to prevent losses. Suppose, for example, that life is pretty good in your country (utility 0), but that Alfred, newly crowned, is on his way to murder ten innocent people to celebrate the beginning of his rule (utility −10). You can stop him by killing him, but if you shoot and miss he will murder the ten and also enact measures that will make life terrible (utility −40). Are you comfortable letting the ten die (keeping in mind that a full third of all shooters lack the lesion)? Some people, at least, are not.

Consider also the fact that when one first encounters Murder Lesion it is hard to know what unconditional probability to assign to having the lesion. Part of the difficulty, of course, is that one tends to run together the question of one's chances of having the lesion with the question of one's chances of having the lesion given that one shoots, and one has no clear idea whether one will shoot. Whatever the reason, the initial probability of 20% seems to strike people as low. Indeed, if people really believed that their chance of having the lesion was 20% it would be hard to explain their reticence toward shooting. But, while people instinctively recognize that 20% is on the low side, they cannot tell, without giving the problem a lot more thought, what the right probability for L should be. In such cases of uncertainty, where probabilities are hard to estimate, people often choose the status quo ($\sim S$) or decide to max-the-min (again, $\sim S$) as a knee-jerk way out of the problem. These irrational reactions both involve ignoring relevant information about what acts are likely to cause. The bottom line is this: intuitions about what one might do in Murder Lesion are not worth much until one makes the effort to reason oneself into the equilibrium state in which all causally relevant information is taken into account. Once one is there, the acts seem to be on a par.

This brings us to a second, more formidable, reason for denying that any plausible decision theory can permit shooting. Even if one grants that rational agents should make decisions in light of full information about what their acts are likely to cause, and even if one recognizes that S and $\sim S$ have the same unconditional causal expected utility on this basis, one might still deny that either act is can be rationally chosen since

[9] My remarks here are informed by Wlodek Rabinowicz, whose forthcoming "Subjective Probabilities and Bets" makes many similar points.

neither is *causally ratifiable* in the most plausible versions of Murder Lesion. Though I have not yet said anything that constrains the equilibrium conditional probabilities $prob_e(L \mid S)$ and $prob_e(L \mid \sim S)$, it is natural, and entirely in the spirit of Egan's approach, to assume that L and S remain correlated in equilibrium, at least to the minimal extent that $prob_e(L \mid S) > prob_e(L \mid \sim S)$. Strange things happen when we make this assumption.

If, as you believe, your choice is likely to determine your act, then deciding to shoot is a good reason for thinking that you have the nerve to shoot, which is a good reason for thinking that you have the lesion, which is a good reason for thinking that you will miss, which is a good reason to refrain from shooting. Likewise, deciding to refrain is a good reason for thinking that you lack the nerve to shoot, which is a good reason for thinking that you lack the lesion, which is a good reason for thinking that you will hit your target, which is a good reason to shoot. In sum, deciding to shoot provides you with evidence for thinking that the causal consequences of shooting will be worse than those of refraining, while deciding to refrain provides you with evidence for thinking that the causal consequences of refraining will be worse than those of shooting. This, some will say, is why you cannot rationally choose either S or $\sim S$: once you settle on one of them, the other looks better.

Formally, an act A is causally ratifiable when it maximizes causal expected utility on the supposition that it is (irrevocably) decided upon. If we denote the decision to do A by δA, then A is causally ratifiable exactly when $\mathcal{U}(A \mid \delta A) \geq \mathcal{U}(B \mid \delta A)$ for all alternatives B, where $\mathcal{U}(\bullet \mid \delta A) = \sum_n prob(k_n \mid \delta A) \cdot u(\bullet, k_n)$. While there are subtleties involved in calculating expected utilities conditional on decisions, it is clear how things work Murder Lesion[10]:

$$\mathcal{U}(S \mid \delta S) = -30 \cdot prob(L \mid S) + 10 \cdot prob(\sim L \mid S) = 10 - 40 \cdot prob(L \mid S)$$
$$\mathcal{U}(S \mid \delta \sim S) = -30 \cdot prob(L \mid \sim S) + 10 \cdot prob(\sim L \mid \sim S) = 10 - 40 \cdot prob(L \mid \sim S)$$
$$\mathcal{U}(\sim S \mid \delta S) = \mathcal{U}(\sim S \mid \delta \sim S) = 0$$

Shooting is only ratifiable if your probability for having the lesion given that you shoot is not above 0.25, and refraining is only ratifiable if your probability for having the lesion given that you refrain is not below 0.25. This results in both acts being unratifiable in any equilibrium where $prob(L \mid S) > prob(L \mid \sim S)$! We already know $prob(L) = 0.25$ in equilibrium. It follows directly that $prob(L \mid S) > 0.25 > prob(L \mid \sim S)$, which entails $\mathcal{U}(S \mid \delta S) < \mathcal{U}(\sim S \mid \delta S)$ and $\mathcal{U}(S \mid \delta \sim S) > \mathcal{U}(\sim S \mid \delta \sim S)$. So, neither act is ratifiable.

Unratifiable acts do seem defective: by choosing them one puts oneself in an epistemic position from which they seem suboptimal as causes of desirable results. Indeed, one knows that one will *regret* choosing an unratifiable act as soon as one has (irrevocably) chosen it. But, it seems irrational, at least on the face of things, to choose acts that one knows one will regret when one has more information. So, it seems, at least on the face of things, that rational agents cannot choose unratifiable acts. Those who

[10] Here I assume that in Murder lesion the decision to perform an act is as reliable indicator of the state of the world as the act itself, so that $prob(L \mid \delta S) = prob(L \mid S)$ and $prob(L \mid \delta \sim S) = prob(L \mid \sim S)$. Some decision problems lack this feature.

hold this position, like Harper (1986), Weirich (1985) and Sobel (1990), will endorse the following:

> *Maxim of Causal Ratifiability*. An act A can be rationally chosen only if it is causally ratifiable. So, given a subjective probability *prob* with associated causal expected utility $\mathcal{U}$, if there is some alternative B such that $\mathcal{U}(A \mid \delta A) < \mathcal{U}(B \mid \delta A)$, then A can be eliminated from list of options that an agent with attitudes described by *prob* and $\mathcal{U}$ can rationally choose.

If this is right, then neither S nor $\sim S$ is permissible in Murder Lesion (under our working assumption that $prob\,(L \mid S) > prob\,(L \mid \sim S)$ holds in equilibrium).

While there is something to the thought that an act's unratifiably provides a reason against choosing it, there is something wrong in this idea as well. I shall contend that in decisions lacking ratifiable "pure" options, like Murder Lesion, no act is ever definitively recommended as the single right thing to do or definitively prohibited as the categorically wrong thing to do. Indeed, once you have taken all available causally relevant information into account, it is permissible to shoot, permissible to refrain from shooting, or permissible to toss a coin of any bias and shoot exactly if it comes up heads. These options are all permissible, *even though you will regret them*. So, I maintain, the mere fact that an act is unratifiable does *not* make it impermissible, and this can be true even when ratifiable options are available.

Usually, of course, the fact that one will regret an action is a good reason for not choosing it. As in Arntzenius (2008), one might seek to explain this sort of "no regrets" intuition as stemming from the following general principle:

> *Weak Desire Reflection*. If one's desires at time $t = 2$ arise from one's desires at an earlier time $t = 1$ as a result of conditioning on evidence acquired between the two times, then one's desires at the earlier time should be one's expectation of the later desires, so that $\mathcal{U}_1(A) = \sum_x prob_t(\mathcal{U}_2(A) = x) \cdot x$.

This seems to suggest a "no regrets" maxim which, as Arntzenius puts it, requires that "a rational person should not be able to foresee that she will regret her decisions." (p. 277) At a first pass, this maxim might be formulated as follows:

> *No Regrets-I*. It is impermissible to choose act A at time $t = 1$ if one knows that, at some later time $t = 2$, one will both have more relevant evidence about A's causal consequences and will regard A as suboptimal, so that $\mathcal{U}_2(A) < \mathcal{U}_2(B)$ for some B.

This seems to imply that unratifiable acts are impermissible. After all, one will know in advance both that one will be better informed after choosing an unratifiable act and that one will then rank some alternative as having a higher expected utility.

Upon closer inspection, however, *NR-I* leaves it unclear whether the knowledge that $\mathcal{U}_2(A) < \mathcal{U}_2(B)$ makes A impermissible on its own, or whether it does so by affecting the $t = 1$ utilities so that $\mathcal{U}_1(A)$ ceases to be maximal. To see the point, suppose you have never eaten a *bhut jolokia* pepper and are thinking about popping one into your mouth. Just before you do, a hot pepper expert tells you that you will deeply regret your decision because, upon ingesting the *bhut jolokia,* you will receive

information you currently lack (that it is *insanely* hot) and this information will be of such a character as to bring you to realize (instantly and with great vivacity) that eating it was a horrible mistake. Learning this new fact will likely lead you decide against eating. The reason is obvious: when you learn that you will regret a decision to eat you thereby acquire information about your future beliefs and desires that makes your *current* expected utility for eating less than that for not eating. Upon learning $\mathcal{U}(\text{eat} \mid \text{eat}) < \mathcal{U}(\sim\text{eat} \mid \text{eat})$ you readjust your *un*conditional expected utilities so that $\mathcal{U}(\text{eat}) < \mathcal{U}(\sim\text{eat})$. This scenario is consistent with Weak Desire Reflection, and with the following weakened "no regrets" principle.

> *No Regrets-II*. It is impermissible to choose A at $t = 1$ when one knows that one will have more relevant evidence about A's consequences at $t = 2$, and that this evidence requires $\mathcal{U}_1(A) = \sum_x prob_1(\mathcal{U}_2(A) = x) \cdot x < \sum_x prob_1(\mathcal{U}_2(B) = x) \cdot x = \mathcal{U}_1(B)$ for some B.

In other words, knowing you will regret A at some future time makes it impermissible for you to choose A in just those cases where this knowledge requires you to *now* expect that A's consequences will be suboptimal, i.e., when learning $\mathcal{U}_2(A) < \mathcal{U}_2(B)$ entails $\mathcal{U}_1(A) < \mathcal{U}_1(B)$ for some B.

NR-II, which all causal decision theorists will endorse, is too weak to justify a ratifiability requirement. As we have seen, it is possible for agent to know that an act is unratifiable and yet assign it maximal expected utility. In the Murder Lesion equilibrium you know that you will regret shooting since refraining has a higher causal expected utility conditional on a decision to shoot. Even so, shooting has an unconditional expected utility that is as high as that of any other option. So, *NR-II* does not prohibit shooting, despite its lack of ratifiability. The same is true of refraining, and of any mixed act which leads to shooting and refraining with anything other than the equilibrium act probabilities of $prob_e(S)$ and $prob_e(\sim S)$, whatever these might be. The mixed act, call it M, which leads to shooting with probability $prob_e(S)$ and to refraining with probability $prob_e(\sim S)$ is ratifiable,[11] but *NR-II* does not require you to prefer it since it has the same unconditional expected utility as S or $\sim S$, namely zero. So, irrespective of ratifiability considerations, *NR-II* does not prohibit or require any of the acts, pure or mixed, that you might have at your disposal.

Proponents of ratifiability will seek to portray *NR-II*'s failure to recommend M as a reason to favor *NR-I*, which, they will say, is better able to take your potential regrets into account. This is wrong. In fact, once you have achieved the equilibrium in which all available information about the effects of your acts has been taken into account, your unconditional causal expected utilities incorporate all relevant ratifiability considerations! Specifically, the fact that you know that you will regret any act but M is reflected in the unconditional $\mathcal{U}$-values of your acts. To see this, suppose for simplicity

[11] To secure the ratifiability of Mwe must assume that learning δM or $\sim\delta M$ is irrelevant to the probability of having the lesion. Without this assumption, M fails to be ratifiable. This would not alter the gist of the argument, however, since some other mixed act would then be ratifiable (which one would depend on the equilibrium values of $prob(L \mid \delta M)$ and $prob(L \mid \sim\delta M)$), and the reasoning could proceed along the same lines.

that S, $\sim S$ and M are your only options. We can write their unconditional expected utilities as:

$$\begin{aligned}
\mathcal{U}(S) &= prob(\delta S)\cdot\mathcal{U}(S \mid \delta S) + prob(\delta M)\cdot\mathcal{U}(S \mid \delta M) \\
&\quad + prob(\delta{\sim}S)\cdot\mathcal{U}(S \mid \delta{\sim}S) \\
\mathcal{U}(M) &= prob(\delta S)\cdot\mathcal{U}(M \mid \delta S) + prob(\delta M)\cdot\mathcal{U}(M \mid \delta M) \\
&\quad + prob(\delta{\sim}S)\cdot\mathcal{U}(M \mid \delta{\sim}S) \\
\mathcal{U}(\sim S) &= prob(\delta S)\cdot\mathcal{U}(\sim S \mid \delta S) + prob(\delta M)\cdot\mathcal{U}(\sim S \mid \delta M) \\
&\quad + prob(\delta{\sim}S)\cdot\mathcal{U}(\sim S \mid \delta{\sim}S)
\end{aligned}$$

Focus first on the $prob(\delta S)\cdot\mathcal{U}(\bullet \mid \delta S)$ terms. These are the components of total utility that reflect the potential evidential ramifications of a decision to shoot. Given that $prob(L \mid S) > prob(L \mid \sim S)$, it turns out that $\mathcal{U}(\sim S \mid \delta S) = 0 > \mathcal{U}(M \mid \delta S) > \mathcal{U}(S \mid \delta S)$. So, in virtue of what you will learn if you decide to shoot, $\sim S$ has an unconditional expected utility advantage of a $-prob(\delta S)\cdot\mathcal{U}(M \mid \delta S) > 0$ over M and an unconditional expected utility advantage of $-prob(\delta h)\cdot\mathcal{U}(S \mid \delta S) > 0$ over S, while M has an unconditional expected utility advantage of $prob(\delta S)\cdot[\mathcal{U}(M \mid \delta S) - \mathcal{U}(S \mid \delta S)] > 0$ over S. In this way, your unconditional utilities both implicitly assesses the relative desirability of acts in light of what you will learn if you decide on S, and then discount the results by the improbability of that decision. The values of $\mathcal{U}(S)$, $\mathcal{U}(M)$ and $\mathcal{U}(\sim S)$ thus already reflect not only the bare fact that S is unratifiable, but also the *extent* of its unratifiability (as measured by $\mathcal{U}(\sim S \mid \delta S) - \mathcal{U}(S \mid \delta S)$ and $\mathcal{U}(\delta M \mid \delta S) - \mathcal{U}(S \mid \delta S)$).[12]

This poor "regret profile" does not eliminate S from consideration however. As long as your estimate of S's probability is at its equilibrium value[13] you will see S as choiceworthy, despite its unratifiability. This is because your indecision about whether to shoot translates directly into uncertainty about whether the regrets you will come to have upon choosing to shoot will be *warranted*. Regrets are warranted, in the relevant sense, exactly if they would remain appropriate even if all relevant facts about the world's state were known.[14] So, in Murder Lesion you are warranted in regretting a

[12] This perspective helps clarify an example, due to Anil Gupta, which Egan discusses. Gupta imagines an agent who has the option of smoking a cigar, a cigarette or nothing at all. It turns out that those inclined to choose cigars would be a little better off smoking cigarettes, and a *lot* worse off not smoking at all. Those inclined to choose cigarettes would be a little better off smoking cigars, and a *lot* worse off not smoking at all. Those inclined to refrain from smoking would be only *slightly* less well off if they smoked cigars or cigarettes. Now, suppose the agent finds herself with a strong urge to smoke, so *prob*(cigarette or cigar) is nearly one. It would clearly be crazy for her to refrain from smoking, an act she thinks will lead to horrible results, merely because it is ratifiable. It should be clear why. The regret profiles of three actions are such that refraining comes out very badly given the decisions that are most likely. As a result, the unconditional expected utility of refraining is much lower than that of smoking a cigar or a cigarette.

[13] This requires the probabilities for δS and δM to obey $prob(S) = prob(\delta S) + prob(S)\cdot prob(\delta M)$, on the assumption that you are sure you will do shoot if you decide to shoot and are sure that you will shoot with probability $prob(S)$ is you decide on M.

[14] I do not say "all relevant facts about *outcomes* were known." This sort of "outcome regret" (Weber 1998) is not pertinent to the questions being asked here. Intuitively, regret is warranted in my sense just in case the objective expected utility of the chosen act is exceeded by that of some alternative.

decision to shoot/refrain just in case you have/lack the lesion. In equilibrium, your estimate of the chance of warrantedly regretting a decision to shoot is 25%, while your estimate of the chance of warrantedly regretting a decision to refrain is 75%. In these circumstances, you will not see S's unratifiability as a decisive mark against it. Since you know that your decisions cannot influence the presence or absence of the lesion, and since you know that the regrets you will come to have upon choosing S will be based on the belief, caused by the decision, that your chance of having the lesion is $prob(L \mid S) > 0.25$, it follows that you do not currently fully trust the accuracy of the future beliefs on which your regrets about shooting will be based! As a result, you need to temper the negative implications of S's unratifiability by your views about the potentially unwarranted nature of your regrets upon choosing it. In general, when you compare $\mathcal{U}(S)$, $\mathcal{U}(M)$ and $\mathcal{U}(\sim S)$ you implicitly compare their $prob(\delta S) \cdot \mathcal{U}(\bullet \mid \delta S)$ terms, thereby incorporating all relevant facts about how much you would regret choosing S, and factoring in your best assessment of the degree to which these regrets would be warranted. There is no need to levy further demerits against S: all relevant ratifiability considerations are incorporated into the unconditional probabilities. Similar things can be said about $\sim S$.

What about your ratifiable option M? It too is properly assessed on the basis of its unconditional expected utility. Given its "no regret" profile ($0 = \mathcal{U}(S \mid \delta M) = \mathcal{U}(M \mid \delta M) = \mathcal{U}(\sim S \mid \delta M)$) no advantage accrues to any act as a result of what you might learn if you choose M. The bare fact that M is ratifiable counts for nothing unless accompanied by some advantage in utility over the other acts conditional upon its being chosen. No matter how confident or doubtful you are about settling on M, the truth of δM or $\sim \delta M$ is irrelevant to what you should do because, upon becoming certain of either proposition, you retain your belief that $prob(L) = 0.25$,[15] and so see all your acts as having the same potential to cause desirable results. There is thus no need to heap further merits on M for being ratifiabile: indeed, doing so amounts to "double counting" since all relevant considerations about the matter are already taken into account. In an equilibrium in which all available information about the effects of one's actions has been processed, choiceworthiness is entirely a function of unconditional causal expected utility. The ratifiability statuses of acts are irrelevant, except insofar as these statuses are reflected in unconditional utilities.

This point is obscured in Murder Lesion because it can seem that you should believe both "If I decide to shoot my subsequent regrets will be warranted " and "If I decide to refrain my subsequent regrets will be warranted." In fact, you should believe neither, at least not very strongly. You are certain of this: "There is some act S or $\sim S$, I know not which, such that if I decide on it my subsequent regrets will be warranted." But, you are also certain of this: "There is some act S or $\sim S$, I know not which, such that if I decide on it my subsequent regrets will not be warranted." The right story is that, in equilibrium, (i) you know you will come to regret the act you choose, whichever act that is, *after* you choose it; (ii) you think the act that warrants regretting is 25% likely to be S and 75% likely to be $\sim S$; but

[15] See Footnote 9.

(iii) you do not regard the fact that you will regret either act as a decisive reason against choosing it simply because you are not certain that these regrets will be warranted. Indeed, since $\mathcal{U}(S) = \mathcal{U}(M) = \mathcal{U}(\sim S)$ everything cancels out exactly when we compare the regret profiles of the acts, and weight them in proportion to their probability. Since all this information is taken into account in the unconditional utilities, the bare facts that S and $\sim S$ are unratifiable, and that M is ratifiable, are irrelevancies.

It remains true, of course, that *after* you irrevocably decide on some act you will have more information about your chances of having the lesion, and so more evidence about the degree to which your regrets are warranted. If you irrevocably decide to shoot, you will *then* have more reason than you have now for believing that refraining would have been better. If you irrevocably decide to refrain, you will *then* have more reason than you have now for believing that shooting would have been better. However, these judgments reflect information you cannot possess until *after* you have irrevocably fixed on a course of action. Before that, when S and $\sim S$ are still live options, the best you can do is to use the available evidence about the consequences of your acts to assess to degree to which your future regrets will be warranted, and to choose on that basis. This is what *NR-II* and Weak Desire Reflection require. It is an error, however, to go on to conclude, as *NR-I* demands, that you can neither permissibly shoot nor permissibly refrain because you know you will regret either choice. You should not dismiss either option since each has, in your estimation, some chance of being the one you should *not* regret.

As a last ditch, friends of ratifiability might reply that choosing S or $\sim S$ cannot be rational since you would renege on either choice, if you could. That's true, but irrelevant. Either you face an *irrevocable* choice or you don't. If you do, then the only information you have to go on is encoded in your equilibrium beliefs, which leave you unsure as to whether your future regrets, and so your future inclinations to renege, are warranted. As a result, these inclinations cannot be counted as decisive reason against choosing S or $\sim S$. On the other hand, if you can revoke decisions, then immediately upon "choosing" to shoot you will reassess your views about the lesion, increasing $prob(L)$, and this will cause you to lean toward refraining. The further you lean the more inviting shooting will seem. Oscillations will continue until you achieve an epistemic state in which the impetus toward refraining caused by your inclinations to shoot is precisely offset by the impetus toward shooting caused by your inclinations to refrain. Of course, this is precisely the equilibrium $prob(L) = 0.25$. Your revocable decisions are highly unstable, and when you try to revocably "choose" either S or $\sim S$ you are always led straight back to the equilibrium.

This instability might seem like a reason to favor the ratifiable act M since it seems like the one choice on which you will not want to renege. This would be the wrong way to go, for two reasons. First, when you are in the equilibrium epistemic the fact that you will not renege on M is irrelevant to your views about its potential for causing desirable results. From that perspective there no advantage in choosing M and then reneging as opposed to choosing M and then carrying through. Second, while you will have no reason to renege on the choice of M you also will have no reason to not

to renege either since all your acts have the same causal expected utility given δM. It might be that choosing and carrying through on M is a particularly salient way of "picking"[16] (so as not to be left forever in a limbo of indecision), but in terms of the efficacy of actions as causes of desirable results there is no more reason to pick M than to pick any other act.

In the end, the Maxim of Causal Ratifiability and *NR-I* must go. Insofar as they are germane to what agents should do, considerations of ratifiability are encoded in the values of unconditional expected utilities. Weak Desire Reflection and *NR-II* accurately capture the force of these considerations by recognizing that future regrets should matter to current decisions only to the extent that anticipating these regrets affects current expected utilities. The resulting decision rule is simple: act to maximize unconditional causal expected utility in light of all the information about the causal consequences of your acts that is available to you. If you do this, matters of ratifiability and unratifiability will take care of themselves. To sum up, even though you cannot ratify either a decision to shoot or a decision to refrain in Murder Lesion, if you handle the facts properly, by first (a) taking all available information about what you acts might cause into account and then (b) discounting the regret profile of each decision by the probability that you will make that decision, then it is rationally permissible for you to pick any action.

4 Aside: comparison with Arntzenius

Before closing, let me explain how the position defended here differs from a somewhat similar one found in Arntzenius's excellent (2008). While I agree with most of what Arntzenius says about Egan's counterexamples, our views diverge on three main points. First, he thinks that Egan's examples show that CDT is "unsatisfactory" because it allows decisions that are inherently unstable. Second, he thinks this forces proponents of the theory to introduce an *additional* principle which requires agents to make choices in a deliberational equilibrium in which, "a rational person is one such that at the end of his deliberation as to which action to perform his credences are in equilibrium" (p. 293). Third, Arntzenius interprets Weak Desire Reflection as *NR-I* by maintaining that "a rational person should not be able to foresee that she will regret her decisions" (p. 277).

I have misgivings about all three points. First, I deny that the mere existence of unstable decisions, in which no pure acts are ratifiable, is any mark against CDT. To the contrary, I say that CDT, properly interpreted, supplies the right answer in these cases. It says, correctly, that knowing one will come to regret a decision is not, by itself, a reason to discard that decision. If one is choosing an action that maximizes unconditional causal expected utility from the perspective of an epistemic state

[16] I use "pick" here as a term of art. When *picking* one selects an act from a set of equally desirable alternatives via a process whose outcome does not reflect anything about one's current reasons for doing the act. When A and B coincide in expected utility, tossing a coin and being irrevocably bound to do A if heads and B if tails is a paradigm of picking: the fact that the coin falls one way rather than the other does not reflect anything about the merits of doing one act rather than the other. One uses picking procedures to avoid the fate of Buridan's ass, who was forever caught between equally attractive haystacks.

incorporates all available evidence about the causal consequences of acts, then further considerations about what one will and will not later regret are immaterial. There is no doubt that Murder Lesion, Death in Damascus, and the like, are odd decisions, but it redounds to the credit of CDT that it answers them correctly. Rather making rational action impossible in such cases of decision instability, CDT rightly deems any act permissible.

Second, while I agree that deliberative equilibrium is central to rational decision making, I disagree with Arntzenius that CDT needs to be *amended* in any way to make it appropriately deliberational. In cases like Murder Lesion a deliberational perspective is forced on us by what CDT says. It says this: *A rational agent should base her decisions on her best information about the outcomes her acts are likely to causally promote, and she should ignore information about what her acts merely indicate*. In other words, as I have argued, the theory asks agents to conform to Full Information, which requires them to reason themselves into a state of equilibrium before they act. The deliberational perspective is thus already part of CDT.

The main misgiving I have about Arntzenius's approach, however, has to do with his endorsement of a "no regrets" principle, again seen as an addendum to CDT. As I have indicated, insofar as questions of regret or ratifiabilty matter to decision making, they are already fully accounted for in CDT. If there is ever any conflict between the requirement to maximize unconditional causal expected utility and the requirement to choose acts one will not regret, the first requirement should always win. This is not to say that facts about what one will or will not regret are beside the point in decision making, only that their relevance, insofar as it matters, is taken into account in the values of unconditional causal expected utilities.

5 A final Salvo: strong correlations in murder lesion

Some people will still be unsatisfied by any picture that permits shooting, even as one option among many. Experience with the Newcomb Problem literature suggests that a further assault can be anticipated. Suppose the correlations between shooting/refraining and having/lacking the lesion are *very* strong. Imagine, for example, that in equilibrium we have $prob_e(S \mid L) = 0.999$ and $prob_e(S \mid \sim L) = 0.001$, so that upon choosing to shoot you will be better than 99.9% confident that you have the lesion, and upon choosing to refrain you will be better than 99.9% confident that you lack the lesion. Even though $\mathcal{U}(S) = \mathcal{U}(\sim S)$ in this equilibrium an objector might wonder whether we *really* should say that S and $\sim S$ are equally permissible given that you are practically certain that the first will lead to the worst possible outcome and that the second will leave the status quo intact? What if $prob_e(S \mid L) = 0.99999999999$ and $prob_e(S \mid \sim L) = 0.00000000001$, so that the odds of shooting and killing Alfred are less than one in a billion? Isn't there going to be a point at which the correlation between shooting/refraining and having/lacking the lesion becomes so strong that it would be utterly foolish to shoot?

No. Even in these extreme cases, once one has processed all the causally relevant information it will become clear that the punishment one risks by shooting is precisely offset by potential loss of opportunity associated with refraining. It may be true that

shooters almost always miss, but it is also true that non-shooters almost always pass up a golden opportunity to kill Alfred. One's assessment of the relative risks will change depending on the values of $prob_0(S \mid L)$ and $prob_0(S \mid \sim L)$, but in any version of Murder Lesion that can pose a threat to CDT, these risks balance out so that a fully informed rational agent will always wind up being indifferent between shooting and refraining.

I suspect it seems otherwise mainly because people have difficulty conceiving of a decision with Murder Lesion's structure in which $prob_e(S \mid L) \approx 1$ and $prob_e(S \mid \sim L) \approx 0$, and where the agent sees herself as free the sense of β. Under these conditions it seems unlikely that shooting or refraining is really a matter of decision; acts seem determined by presence or absence of the lesion in a way that makes free choice impossible. If this is how one responds to "extreme" Murder Lesions, then it becomes difficult to say what one should do. Indeed, without β it is unclear whether one even has a "decision" anymore. What is called for is less a rationale for action than a description of what behaviors an agent might exhibit when stuck in such unfortunate circumstances.

While the gambit of rejecting β does undercut the rationale for the indifference between S and $\sim S$, it also negates Murder Lesion as an objection to CDT. What the theory claims is only that a rational agent in Murder Lesion who sees herself as having a free choice in the matter of shooting or not shooting, in the sense captured by β, and who takes the time to consider all available evidence about what her acts might cause, will end up indifferent between shooting and refraining. This claim is correct no matter how large or small the equilibrium values of $prob(S \mid L)$ and $prob(S \mid \sim L)$ might get.

6 Conclusion: CDT stands

To sum up, CDT has nothing to fear from Murder Lesion or examples of its ilk. Contrary to what Egan maintains, it is neither true that CDT uniquely requires shooting nor that it is rationally compulsory to refrain from shooting. Contrary to ratificationism, even though one surely will regret both a decision to shoot and a decision to refrain, it is not impermissible to choose either of these acts. Instead, someone who has processed all the available information about what shooting and refraining are likely to cause can pick either act, or any probabilistic mixture of the two, without risking irrationality since all her acts have the same causal expected utility when assessed from that epistemic perspective. All rationality requires is that one first place oneself in an epistemic state that accurately reflects all one's evidence about what one's acts are likely to cause, and then that one maximize causal expected utility on that basis. Those who do this will never go wrong, even in outlandish cases, like Murder Lesion, in which one is sure to regret whatever choice one makes.

References

Arntzenius, F. (1990). Physics and common causes. *Synthese, 82*, 77–96.
Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis, 68*, 277–297.

Aumann, R., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica, 63*, 1161–1180.

Eells, E. (1982). *Rational decision and causality*. Cambridge, MA: Cambridge University Press.

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

Gibbard, A. (1992). Weakly self-ratifying strategies: Comments on McClennen. *Philosophical Studies, 65*, 217–225.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In C. Hooker, J. Leach, & E. McClennen (Eds.), *Foundations and applications of decision theory* (pp. 125–162). Dordrecht: Reidel.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Harper, W. (1986). Mixed strategies and ratifiability in causal decision theory. *Erkenntnis, 24*, 25–36.

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago: The University of Chicago Press.

Jeffrey, R. (1993). Causality and the logic of decision. *Philosophical Topics, 21*, 139–151.

Joyce, J. M. (1999). *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.

Joyce, J. (2002). Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies, 110*, 69–102.

Joyce, J. M. (2007). Are Newcomb problems really decisions?. *Synthese, 156*, 537–562.

Levi, I. (2000). Review essay: The foundations of causal decision theory. *Journal of Philosophy, 97*, 387–402.

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

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel. Synthese library*. Dordrecht: Reidel.

Pearl, J. (2010). *The curse of free-will and the paradox of inevitable regret. UCLA Cognitive Systems Laboratory, Technical Report (R-375).*

Rabinowicz, W. (2002). Does practical deliberation crowd out self-prediction. *Erkenntnis, 57*, 91–122.

Shafir, E., & Tversky, A. (1995). Decision making. In E. E. Smith & D. N. Osherson (Eds.), *An invitation to cognitive science, 2nd ed. (Vol. 3: Thinking)* (pp. 77–100). Cambridge, MA: MIT Press.

Skyrms, B. (1990). *The dynamics of rational deliberation*. Cambridge, UK: Cambridge University Press.

Sobel, J. H. (1990). Maximization, stability of decision, and actions in accordance with reason. *Philosophy of Science, 57*, 60–77.

Spohn, W. (1977). Where Luce and Krantz do really generalize savage's decision model. *Erkenntnis, 11*, 113–134.

Weber, M. (1998). The resilience of the Allais paradox. *Ethics, 109*, 94–118.

Weirich, P. (1985). Decision instability. *Australasian Journal of Philosophy, 63*, 465–472.