

Lost memories and useless coins: Revisiting the absentminded driver*

Wolfgang Schwarz

16 March 2014

Abstract. The puzzle of the absentminded driver combines a version of the Sleeping Beauty problem with a Newcomb problem. Its analysis therefore depends on the choice between “halving” and “thirring” as well as that between “evidential” and “causal” decision theory. I show that the four combinations lead to interestingly different solutions, and draw some general lessons about the formulation of causal decision theory, the interpretation of mixed strategies and the connection between rational credence and objective chance.

1 Introduction

Sometimes it is controversial what rationality demands in a given situation. One-boxers and two-boxers disagree on the choice to make in Newcomb’s problem, halvers and thirders disagree on the beliefs to have in the Sleeping Beauty problem. Such disagreements often trace back to different general perspectives on rationality. At the heart of Newcomb’s problem lies the divide between causal and evidential decision theory. At the heart of the Sleeping Beauty problem arguably lies a tension between evidentialism and conservatism in epistemology. In this paper, I want to look at a case that raises both of these issues, as well as several others. The case was introduced in [Piccione and Rubinstein 1997], and goes as follows.

An absentminded driver has to take the second exit off the highway in order to get home. If she turns off at the first exit, she reaches a desolate area and has to spend the night in her car. If she continues at both exits, she has to stay at a motel at the end of the highway. Due to her absentmindedness, she cannot tell upon arriving at an exit whether it is the first or the second (unless, of course, she knows that she turns off at the first).

Our main question is what the driver ought to do. The answer varies between evidential and causal decision theory, and even between different formulations of the latter. In

* Thanks to Alma Barner, Rachael Briggs, Kenny Easwaran, Alan Hájek, Daniel Nolan, Mike Titelbaum, David Wiens and two anonymous referees for comments and discussion.

addition, what the driver ought to do depends on what she ought to believe, and this, too, turns out to be controversial: we will find essentially the same two options as in the Sleeping Beauty problem. We will also see that if the driver makes her choice by tossing a coin, then her degree of belief in the two possible outcomes (heads and tails) does not always match what she knows to be the objective chance. Consequently, several widespread ideas about the role of chance in game theory and decision theory threaten to break down.

The point of this paper is not to take sides in the debate between causal and evidential decision theory or between halving and thirding. In fact, I will argue that all four combinations give defensible answers to the puzzle if one keeps in mind the general perspective that motivates these combinations.

2 Absentmindedness and two types of expected utility

Before we begin, I should make some clarifications about the driver's predicament. The driver suffers from an unusual kind of absentmindedness. Her problem is not that she is likely to pass an exit without noticing it. On the contrary, she is certain to make a deliberate, rational choice at every exit she reaches. Her problem is that if she decides to stay on the highway at the first exit, then the monotony of the traffic will make her forget the whole event before she reaches the second exit, so that she arrives at that exit in the very same state of mind in which she arrived at the first exit. The two exits may look different, but the differences don't help the driver to figure out which is which. (There are, in particular, no street signs.) For some reason the driver also can't leave any marks on herself or her environment to counteract the memory loss brought on by the traffic. For example, she can't tie a knot in her handkerchief after continuing at the first exit and thus use the handkerchief to find out where she is. Throughout her journey, the driver is aware of all these facts.

Subject to the constraints of the scenario, we assume that the driver is ideally rational, and knows that she is ideally rational. We model her beliefs by a probability distribution P over possible states of affairs so that the probability assigned to a state represents the degree to which she believes that the state obtains. Similarly, the degree to which she desires the different states to obtain is represented by a utility function V . The driver would mostly like to get home, but also has a slight preference for staying at the motel over spending the night in her car. For concreteness, let's say that $V(Car) = 0$, $V(Home) = 4$ and $V(Motel) = 1$. We assume that due to the memory loss caused by the highway, her beliefs and desires are in all relevant respects the same whenever she gets to an exit.

It follows that if the driver's beliefs and desires determine a particular choice as uniquely rational, then that is what she is going to do at every exit. Let us assume for now that

this is indeed the case. So the driver can be confident that whatever she does at the present exit is also what she does at the other exit (if it is reached). She might then reason as follows.

“I can either leave the highway here or continue. If I leave, I must be at the first exit, for I know that I make the same choice at every exit, and I couldn’t be at the second exit after leaving at the first. So if I leave, I’ll end up spending the night in the car. Alternatively, if I continue on the highway, then it is clear that I’ll continue at both exits, so I’ll spend the night in the motel. That is the slightly better outcome, so I’d better continue.”

But is this an adequate way of evaluating the options? The driver is right that if she leaves the highway, she can infer that she is probably at the first exit and will spend the night in her car. Leaving the highway is bad news. Continuing is also bad news, since it equally entails that the driver won’t get home, although it is not quite as bad news as leaving. But arguably the driver’s aim is not to receive good news; it is to bring about good outcomes. The two aims often go together, but they come apart in situations in which a particular choice of action would be evidence for an undesirable state of affairs without having any influence over whether that state obtains. Famously, taking both boxes in Newcomb’s problem ([Nozick 1969]) is evidence that the opaque box is empty, but it doesn’t cause the box to be empty. Similarly, if our driver decides to leave the highway, this is strong evidence that she is at the first exit, but it doesn’t have any genuine influence over where she is. If, for example, she is actually at the second exit, then nothing she can do will bring it about that she is (right now) at the first exit. Again, if she decides to continue, this is evidence that she would continue at the other exit, but it doesn’t control the other (earlier or later) decision.

The different aims – receiving good news vs. bringing about good outcomes – show up in different formulations of decision theory. *Causal decision theory* advises agents to maximise expected utility; that is, to choose whichever option A maximises

$$EU(A) = \sum_{S \in W} P(S)V(S \& A),$$

where W is a suitable partition of states of affairs (which is here assumed to be finite). What is a “suitable” partition? The basic idea is that the states should be rich enough to distinguish relevantly different outcomes, while on the other hand the agent’s choice should have no causal influence over which of the states obtains.¹

Evidential decision theory instead says that agents should choose the option with the

¹ For more precise statements, see e.g. [Lewis 1981] and [Joyce 1999: ch.5]. While it is not essential to my discussion, I assume that states, acts and outcomes are all entities of the same kind (propositions), and that outcomes can be identified with conjunctions of states and acts; see [Joyce 1999: ch.2] for discussion.

highest conditional expected utility, defined by

$$CEU(A) = \sum_{S \in W} P(S/A)V(S \& A).$$

Conditional expected utilities are invariant under different choices of W , so the need to define suitability disappears.

Intuitively, the conditional expected utility of a proposition represents the extent to which the agent hopes, or desires, that the proposition is true. That’s because the degree to which an agent desires that a disjunction of incompatible propositions X and Y is true can plausibly be identified with the average of the degree to which she desires X and Y , weighted by their relative probability. That is, $V(X \vee Y) = P(X/X \vee Y) \cdot V(X) + P(Y/X \vee Y) \cdot V(Y)$. Since the act A is equivalent to the disjunction of $S \& A$, for all $S \in W$, it follows that $V(A) = CEU(A)$. The advice of evidential decision theory thus amounts to the advice to choose the option for which you have the strongest desire that it be chosen. The problem with this, from the perspective of causal decision theory, is that sometimes you may desire that you make a particular choice merely because that would be evidence for something good, without contributing at all to bringing it about: the conditional probability $P(S/A)$ does not distinguish causal correlation from merely evidential correlation. This is what went wrong with the driver’s argument that she ought to continue.²

If we analyze the driver’s problem in causal decision theory, we find that what the driver *should* do inversely depends on what she believes she *will* do. Suppose she is confident that she will leave the highway. She can then infer that she is probably at the first exit, in which case the best choice is to continue. On the other hand, if she is confident that she will continue, then she knows that her journey eventually brings her to both exits, in which case she should give equal credence to being at the first exit and being at the second. Leaving the highway then has an equal probability of bringing her to the desolate area and bringing her home. Since getting home is much better than the other two outcomes, the best choice is then to take the risk and turn off. Either way, if the driver is confident that she will do one thing, she is better off doing the other!

Let’s spell this out in a bit more detail. The outcome of the driver’s present choice depends first of all on whether she is at the first exit or at the second. If she is at the second exit, then her choice settles whether she reaches home or the motel. If she is

² There are other ways to understand desire which aren’t represented by conditional expected utilities. Roughly speaking, conditional expected utilities measure the extent to which the agent would be pleased to learn that the relevant proposition is true. Something else is in play when we say, for example, “I wish Oswald hadn’t killed Kennedy”, by which we don’t mean that it would be better if someone else had killed him. Formally, one might define *indicative utility* as $V^i(X) = \sum_{w \in X} P(w/X)V(w)$, and *subjunctive utility* as $V^s(X) = \sum_{w \in X} P^X(w)V(w)$, where P^X is P imaged on X , for a suitable imaging function. See [Byrne and Hájek 1997], [Etlin 2008: ch.2], [Williams 2010] for discussion.

at the first exit, the outcome also depends on what she is disposed to do later if she reaches the second exit. Thus whenever she reaches an exit, she faces a decision problem summarised by the following matrix.³

	<i>First & Continue₂</i>	<i>First & Leave₂</i>	<i>Second</i>
<i>Continue</i>	1	4	1
<i>Leave</i>	0	0	4

The two rows represent the driver’s present options; the columns represent the possible states of the world that matter to the outcome of her choice. *Continue₂* means that the driver *either* continues at the second exit *or* does not reach the second exit but would have continued if she had reached it. Recall our assumption that the driver’s state of mind when arriving at the second exit would be the same in all relevant respects as her state of mind when arriving at the first. So even if the driver doesn’t reach the second exit, we can still ask what would happen if she were in that situation in the given state of mind. *Continue₂* says that she would continue, *Leave₂* that she would exit. It is not entirely obvious that one of these must be true: if the driver’s actions are genuinely random, for example, there may be no fact of the matter about what would happen at the second exit. But let’s set this possibility aside for now. We know that the driver is perfectly rational; as long as rationality gives an answer to her decision problem, we can be sure that there is a fact of the matter about what she does whenever she faces her problem.

It is crucial for causal decision theory that the states (the columns in the matrix) are outside the agent’s present control. At the first exit, the driver can control *whether* she’ll face another decision problem at the second exit, but she doesn’t have any direct control over the choice she either will or would then make.

To compute the expected utilities, we need to know the probability of the three states. I will assume for the whole of this paper that conditional on reaching both exits, the driver gives equal credence to being at the first exit and being at the second:

$$P(\textit{First}/\textit{Continue}_1) = P(\textit{Second}/\textit{Continue}_1) = 1/2.$$

Call this the *symmetry* assumption. (*Continue₁* means that the driver continues at the first exit.) As indicated above, I will also assume – although not for the whole of the

³ Given the uniformity assumption, one might worry about the value 4 in the bottom-right cell: if the driver is absolutely certain that she does not continue at the first exit and leave at the second, how can she assign a utility to *Second & Leave*? The problem could be avoided by using subjunctive utility instead of indicative utility (see footnote 2), or more simply by replacing the value 1 in the uniformity assumption by “almost 1”, and sprinkling “almost” over all results based on this assumption.

paper – that the driver is confident that she makes the same choice at every exit; more specifically,

$$P(\textit{Continue}_1 / \textit{Continue}_2) = P(\textit{Continue}_2 / \textit{Continue}_1) = 1.$$

Call this the *uniformity* assumption. (The precise value 1, like 1/2 in the symmetry assumption, serves mainly to simplify the calculations.)

Now let c be the driver's degree of belief that she will continue. Notice that *Continue* can be defined as $(\textit{First} \ \& \ \textit{Continue}_1) \vee (\textit{Second} \ \& \ \textit{Continue}_2)$. Symmetry and uniformity then entail that⁴

$$P(\textit{First} \ \& \ \textit{Continue}_2) = c/2;$$

$$P(\textit{First} \ \& \ \textit{Leave}_2) = 1 - c;$$

$$P(\textit{Second}) = c/2.$$

Hence

$$EU(\textit{Continue}) = c/2 \cdot 1 + (1 - c) \cdot 4 + c/2 \cdot 1 = 4 - 3c;$$

$$EU(\textit{Leave}) = c/2 \cdot 0 + (1 - c) \cdot 0 + c/2 \cdot 4 = 2c.$$

The two are equal at $c = 4/5$. If the probability of continuing is greater than 4/5, leaving maximises expected utility, if it is less than 4/5, continuing is best. The grass is always greener on the other side.

In evidential decision theory, by comparison, we would replace the probability of the states by their probability conditional on the relevant choice. Conditional on *Leave*, all probability goes to the state *First* & *Leave*₂. Conditional on *Continue*, the driver's probability is evenly divided between *First* & *Continue*₁ and *Second*. Thus

$$CEU(\textit{Continue}) = 1/2 \cdot 1 + 0 \cdot 4 + 1/2 \cdot 1 = 1,$$

$$CEU(\textit{Leave}) = 0 \cdot 0 + 1 \cdot 0 + 0 \cdot 4 = 0.$$

So the driver ought to continue.

⁴ We first show that c is also the probability of *Continue*₁ and *Continue*₂. Observe that *Second* entails *Continue*₁, wherefore *Continue* entails *Continue*₁. More specifically, *Continue*₁ is the disjunction of *Continue* and *Second* & *Leave*₂. By uniformity, the latter has probability 0, as does *Continue*₁ & *Leave*₂. Hence $c = P(\textit{Continue}) = P(\textit{Continue}_1) = P(\textit{Continue}_1 \ \& \ \textit{Continue}_2)$. Since *Continue*₂ & *Leave*₁ also has probability 0, it follows that $P(\textit{Continue}_2) = P(\textit{Continue}_2 \ \& \ \textit{Continue}_1) = c$ as well.

Now *Continue*₂ divides into *First* & *Continue*₂ and *Second* & *Continue*₂, so $c = P(\textit{First} \ \& \ \textit{Continue}_2) + P(\textit{Second} \ \& \ \textit{Continue}_2)$. By uniformity, it follows that $c = P(\textit{First} \ \& \ \textit{Continue}_2) + P(\textit{Second} \ \& \ \textit{Continue}_1)$. Moreover, by symmetry, $P(\textit{First} \ \& \ \textit{Continue}_1) = P(\textit{Second} \ \& \ \textit{Continue}_1)$, and so $P(\textit{First} \ \& \ \textit{Continue}_2) = P(\textit{Second} \ \& \ \textit{Continue}_1)$ by uniformity. Hence $P(\textit{First} \ \& \ \textit{Continue}_2) = P(\textit{Second} \ \& \ \textit{Continue}_1) = c/2$. *Second* & *Continue*₁ is equivalent to *Second*. The remaining possibility *First* & *Leave*₂ must then have probability $1 - c$.

3 Rational indecision

In causal decision theory, the driver faces an *unstable decision problem*: a situation where any tendency to do one thing makes it advisable to do something else. In these situations, the advice to maximise expected utility is not very useful. Suppose the driver believes that she will continue, and thereby figures out that she ought to leave. Shouldn't this make her reconsider her belief that she will continue? How can she even arrive at a reasonable belief about what she is going to do? Knowing that she is rational, she knows that she will do whatever maximises expected utility. So to find out whether she ought to leave or continue, she first has to find out which of these maximises expected utility. She seems stuck in a loop.

Nevertheless, the driver's epistemic problem can be solved. We will solve it by drawing on Brian Skyrms's work on the dynamics of rational deliberation (see [Skyrms 1990]). Suppose that at the start of her deliberation, the driver is completely undecided about what she will do: $P(\textit{Continue}) = 1/2$. She can then figure out that *Continue* has a higher expected utility than *Leave*. Since she knows that she is rational, this should increase her degree of belief that she will continue. But as soon as $P(\textit{Continue})$ goes up, the expected utilities change, so she has to re-assess whether continuing is still the optimal choice. If $P(\textit{Continue})$ gets too high, *Leave* becomes the better option, in which case the driver's probability in *Continue* should decrease. Skyrms shows that if the details are filled in sensibly, this process always leads to an equilibrium: a point where the probabilities no longer change. For the driver, the deliberation equilibrium lies at a value of $P(\textit{Continue})$ at which both options have the same expected utility, so that the probability is then no longer pulled in either direction. There is only one such value, namely $P(\textit{Continue}) = 4/5$, which will be reached no matter in what state of uncertainty the driver begins her deliberation.⁵

Despite the apparent threat of circularity, we therefore know the driver's degree of belief that she will continue: it is $4/5$. Can we plug this into the expected utility calculation and conclude that choosing either option is OK, since they now have equal expected utility? That would suggest that the driver may rationally resolve to continue, or to leave. But as soon as she starts "resolving", as soon as she finds herself gravitating towards one of the options, her probabilities change and the other option becomes preferable. If she decides to leave, for example, she can be confident that she is at the first exit, and therefore that she really ought to continue. The driver's situation is not a simple case where the agent may rationally opt for either of two equally good options. It rather looks

⁵ Some authors have questioned whether decision makers can assign probabilities to their own actions. We could try to circumvent these worries by instead appealing to the driver's beliefs about what she either does or is disposed to do at the *other* exit. But I find the worries unconvincing in the first place; see [Rabinowicz 2002] and [Joyce 2002] for replies.

like rationality prohibits her from making any decision.

Considerations like these have tempted some authors to add a *ratifiability* constraint to decision theory, to the effect that a rational choice should maximise expected utility *conditional on being realised* (see [Harper 1984], [Sobel 1990], [Rabinowicz 1989]). This would mean that in the driver’s situation, no choice is rational. Others (notably [Arntzenius 2008]) have suggested that decision theory can’t tell our driver what to do, but only what to *believe*: she ought to be 4/5 confident that she will continue.

In my view, the best way to handle unstable decision problems is to assume that decision theory can sometimes recommend a *state of indecision*. The driver should be undecided between her options, but somewhat more inclined towards continuing. Indecision is a genuine psychological state that does not reduce to belief. To be sure, if the driver is rational, she will be ignorant about what she is going to do. But if she is genuinely undecided, she is not just epistemically, but also practically torn between her options. She doesn’t know *what to do*.

Formally, we can model a state of indecision as a probability measure over the agent’s options, with the probabilities specifying to what extent the agent is inclined to the relevant option. The limit case where all probability is concentrated on a single option represents a *decision*. Here the agent has fully resolved what to do. She may also have become certain that this is what she will do, but we should not identify decisions with beliefs, just as we shouldn’t identify “partial decisions” – non-trivial states of indecision – with partial beliefs.

On this model, the answer to a decision problem is always a state of indecision. For many problems, that state will be concentrated on a single option, in which case we can also say that this is what the agent ought to do. Not so in unstable decision problems. Here decision theory cannot tell the agent what to do. It can only tell her to what extent she should be inclined to her various options. By the calculations in the previous section, the rational state of indecision in the driver’s scenario is one where she is 4/5 inclined towards continuing and 1/5 towards leaving.

This view of rational choice has implications for the cognitive architecture of rational agents. If rationality sometimes requires indecision, we need a mechanism outside rational deliberation to break the ties. In the end, our driver must either leave or continue. She doesn’t have a third option – as it were, to park at the side of the road and remain forever undecided. So she needs a cognitive mechanism that turns her rational state of indecision into an action.

What would such a mechanism look like? Various implementations are conceivable. For example, one could have a tie-breaking mechanism that always chooses the option towards which the agent is most inclined. But this would be easy to discover through long-run observations, and then it would undermine the deliberational equilibria: being 4/5 inclined towards continuing, the driver could infer that she will continue, in which

case it would be better to leave. For parallel reasons, a good mechanism shouldn't be completely random, so that *Continue* could just as well be selected as *Leave*. From the driver's perspective, there should be a $4/5$ probability that the tie will be broken in favour of *Continue*. A natural idea is to bestow rational agents with a stochastic tie-breaker that uses the agent's degrees of inclination as weights for the individual options.

I am not here talking about something the agent herself is supposed to do. In the present setup, the driver doesn't have the option of choosing a tie-breaker. Her options are *Leave* and *Continue*. What she ought to do is be $4/5$ inclined towards *Continue*. The question is how the driver's neural wiring should translate this state of indecision into an action. What I suggest is that if an agent is in a state of rational indecision, then the eventual act must be selected by a process that at least to the agent looks stochastic, with the probabilities matching her degrees of inclination.

At this point, the uniformity assumption has become very implausible. The driver knows that she faces the same decision problem at every exit. But what she will end up doing depends on the outcome of the tie-breaking, which may well be different at different occurrences of the same problem. Let's quickly figure out the equilibrium state without the uniformity assumption – assuming instead that, from the driver's perspective, the outcomes of different tie-breaking events are independent. (A more thorough version of the following argument will be given in the next section.)

In equilibrium, the driver has not figured out whether she ought to continue, but she has figured out the extent c to which she ought to be inclined towards continuing, reflected in her present degree of belief that she will continue. She also knows that if she is at the first exit, then continuing would lead her to another instance of the same decision problem, where she'll reach the same state of indecision, so that the probability of continuing will again be c (because the tie-breakings are independent). So $P(\text{First} \ \& \ \text{Continue}_2) = P(\text{First}) \cdot c$. What is $P(\text{First})$? For a start, *First* divides into *First* & *Continue*₁ and *First* & *Leave*₁. By the symmetry assumption,

$$P(\text{First} \ \& \ \text{Continue}_1) = P(\text{Second} \ \& \ \text{Continue}_1) = P(\text{Second}) = 1 - P(\text{First}).$$

Moreover, the driver knows that no matter at which intersection she is, her inclination towards continuing is c ; so

$$P(\text{Continue}_1 / \text{First}) = P(\text{Continue}) = c.$$

Since $P(\text{First} \ \& \ \text{Leave}_1) = P(\text{Leave}_1 / \text{First})P(\text{First}) = (1 - P(\text{Continue}_1 / \text{First}))P(\text{First})$, it follows that $P(\text{First}) = 1 - P(\text{First}) + (1 - c)P(\text{First})$, which resolves to $P(\text{First}) = 1/(c + 1)$. Hence the equilibrium probabilities of the three states are

$$P(\text{First} \ \& \ \text{Continue}_2) = c/(c + 1);$$

$$P(\text{First} \ \& \ \text{Leave}_2) = (1 - c)/(c + 1);$$

$$P(\text{Second}) = c/(c + 1),$$

which yields

$$EU(Continue) = \frac{c}{c+1} \cdot 1 + \frac{1-c}{c+1} \cdot 4 + \frac{c}{c+1} \cdot 1 = \frac{4-2c}{c+1};$$

$$EU(Leave) = \frac{c}{c+1} \cdot 0 + \frac{1-c}{c+1} \cdot 0 + \frac{c}{c+1} \cdot 4 = \frac{4c}{c+1}.$$

In equilibrium, both options must have the same expected utility, which means that $c = 2/3$. So without the uniformity assumption, the driver's inclination towards continuing should be $2/3$, not $4/5$.

This is a rather satisfying result: if the driver always continues with probability $2/3$, then she gets home with probability $2/3 \cdot 1/3 = 2/9$, reaches the motel with probability $2/3 \cdot 2/3 = 4/9$, and the desolate area with probability $1/3$. In the long run, the average (or expected) payoff is $2/9 \cdot 4 + 4/9 \cdot 1 + 1/3 \cdot 0 = 4/3$. In general, if at every exit, c is the probability for continuing, then the expected payoff is $(1-c) \cdot 0 + c(1-c) \cdot 4 + c^2 \cdot 1 = 4c - 3c^2$. This function has its maximum at $c = 2/3$. Within the constraints of the scenario, the driver couldn't do any better.

Recall that evidential decision theory straightforwardly tells the driver to continue, for a guaranteed payoff of 1. Thus agents following the causal theory perform on average better than agents following the evidential theory. This is noteworthy because unstable situations like the driver's are often presented as intuitive *problems* for causal decision theory (see e.g. [Richter 1985], [Egan 2007]). In my view, the form of causal decision here presented gets such cases exactly right. Suppose the driver follows evidential decision theory and decides to continue. Passing an exit, she may wonder where it would lead. Since she gives equal credence to being at the first and being at the second exit, she is 50% certain that the exit she passes leads home. But she much prefers getting home to the other two outcomes – and note that it makes no difference on the evidential account whether the utility of getting home is 4 or 4 million. Shouldn't this make her reconsider her decision? Shouldn't she be tempted to turn around, take the exit and try her luck?⁶

Confronted with the better long-run performance of causal decision theory, proponents of evidential decision theory might complain that we have given our causal agent an unfair advantage by endowing her with a stochastic tie-breaker. That's not true. We can grant that the evidential agent, too, has a stochastic tie-breaker. But evidential decision theory doesn't recommend indecision, so the tie-breaker won't get used. What's true is that if we gave the evidential agent a further *option* to randomise her choice, then she

⁶ The fact that evidential and causal decision theory give different verdicts about the driver shows that the two accounts really do come apart. This is sometimes questioned, notably by Huw Price (see e.g. [Price 1986], [Price 1991]). Price's attempt to bridge the gap between the two theories implies that the driver actually does have causal influence over where she is. In particular, if the driver has chosen to continue at the first exit and now finds herself at the second, then by deciding to leave, she could make it the case that she is actually at the first exit and never reaches the second. That strikes me as absurd. Absentmindedness does not convey powers to undo the past!

might decide to randomise in such a way that continuing has probability $2/3$ (as we will see). But it is important to be clear that our causal agent hasn't been given any new options. She was deliberating only between *Continue* and *Leave*. There was no further possibility of delegating her choice to a stochastic mechanism, and this is not what she decided to do.⁷

Let's see what happens if we add randomisation as a genuine option.

4 Randomisation: a puzzle and an alleged solution

The scenario we are going to look at for the rest of the paper is the same as before, except that the driver now has the additional option of tossing a coin of any bias she likes. Let's say that heads means leave, tails continue.

In this setup, it makes sense to redefine the *uniformity* assumption to say that the driver is confident that she chooses the same bias at every exit she reaches. As we saw at the end of the previous section, the optimal coin to choose at every exit seems to have bias $2/3$ towards tails, as the expected payoff from choosing a coin with bias c at every exit is $4c - 3c^2$, which is maximal for $c = 2/3$. (Here and henceforth, 'bias' always means bias towards tails, i.e. towards continuing.) Given the redefined uniformity assumption, we might therefore expect that evidential decision theory recommends this choice. After all, conditional on choosing any particular bias, the driver is certain that this is her choice at every exit, so comparing conditional expected utilities amounts to considering which coin would be best given that it is tossed at every exit.

Oddly, this is not what we find – at least on a *prima facie* plausible way of modelling the situation. Instead, we find that *causal* decision theory recommends the coin with bias $2/3$, while evidential theory seems to recommend a sub-optimal coin with bias around 0.53.

The causal argument mirrors the argument at the end of the previous section, but I will spell it out a bit more carefully this time. To apply causal decision theory, we first have to find a suitable partition of states. This is not entirely straightforward, because the eventual outcome depends (among other things) on the result of the present coin toss, which is *partly* under the driver's control: by choosing a certain bias, she can make it more or less likely that she will continue, but she doesn't have any further control over how the coin actually lands. To model this, we will follow the advice of [Lewis 1981] and [Skyrms 1984] and use a partition of states which, combined with a choice of the driver, only determines an *objective chance* for the relevant outcomes. Thus being at the second exit is a complete state, because it determines, combined with any choice of a particular coin, the chances of getting home and of reaching the motel. In general, the cells in our

⁷ Pace [Lewis 1981: 29f.].

decision matrix are propositions assigning objective probabilities to the three ultimate outcomes. I will call such propositions *lotteries*.

One of the states represents the driver as being at the second exit. In the other states, she is at the first exit; combined with the choice of a bias, this determines an objective probability for reaching the second exit. That's not yet a complete lottery, because the objective probabilities of the eventual outcomes further depend on what coin the driver is disposed to choose at the second exit. In the previous section, we divided *First* into *First & Continue₂* and *First & Leave₂*. Similarly, we now divide *First* by all possible choices the driver could make at the second exit. Let '*Bias₂ = x*' be the proposition that the driver either chooses bias *x* at the second exit or (if she doesn't reach it) is disposed to choose bias *x* there.

If there are uncountably many possible values of *Bias₂*, we get uncountably many states, which leads to minor complications further down the line. To keep things simple, let's assume – as seems realistic anyway – that the driver has only a finite number of coins at her disposal. Let *B* be the set of available bias values. It doesn't matter exactly which values are in *B*; I assume it contains at least all ratios *n/m* for moderately sized *n* and *m* with *m* ≥ 1 and *n* ≤ *m*.

Now we have finitely many states, each of which, combined with one of the driver's options, determines a lottery: an assignment of objective chance to the three eventual outcomes. The utility of such a lottery is plausibly the average of the utility of the outcomes, weighted by the chances. For example, the utility of tossing the 2/3 coin at the second exit is 1/3 times the utility of getting home (4) plus 2/3 times the utility of reaching the motel (1). Let *Bias = b* be the proposition that the driver chooses bias *b* at the present exit. Then the utilities are, for all *b, c* ∈ *B*,

$$\begin{aligned} V(\text{First} \ \& \ \text{Bias}_2=c \ \& \ \text{Bias}=b) &= (1-b)0 + b(1-c)4 + bc1 = 4b - 3bc; \\ V(\text{Second} \ \& \ \text{Bias}=b) &= (1-b)4 + b1 = 4 - 3b. \end{aligned}$$

Next we need to know a few things about how the driver's degrees of belief are distributed over the states. By probability theory, $P(\text{First} \ \& \ \text{Bias}_2=c) = P(\text{Bias}_2=c)P(\text{First}/\text{Bias}_2=c)$. By the reasoning in the previous section, the driver's degree of belief in *First*, given that she continues with probability *c* at the first exit, should be $1/(c+1)$; i.e. $P(\text{First}/\text{Bias}_1=c) = 1/(c+1)$. Assuming uniformity, we can substitute *Bias₁* for *Bias₂*. It then follows that $P(\text{Second}/\text{Bias}_2=c) = 1 - P(\text{First}/\text{Bias}_2=c) = c/(c+1)$. Hence

$$\begin{aligned} P(\text{First} \ \& \ \text{Bias}_2=c) &= P(\text{Bias}_2=c)/(c+1); \\ P(\text{Second}) &= \sum_{c \in B} P(\text{Bias}_2=c)c/(c+1). \end{aligned}$$

Putting all this together, the expected utilities are

$$EU(Bias=b) = \sum_{c \in B} P(Bias_2=c) \frac{4b + 4c - 6bc}{c + 1}.$$

Once again, what the driver ought to do depends on what she believes she will do. For example, if she is certain that she will choose a coin with bias 1, so that $P(Bias_2=1) = 1$, the expected utility of choosing bias b is $(4b + 4 - 6b)/2$, which linearly decreases with b . This means that $Bias=1$ is not an equilibrium in the deliberation dynamics: the more the driver is inclined towards bias 1, the more all other options appear better. This time, however, there is an option which the driver can rationally choose: $Bias=2/3$. If the driver is certain that she'll choose bias $2/3$, the expected utility of bias b is $(4b + 8/3 - 4b)/(5/3) = 8/5$. This is constant, so the deliberation is no longer pulled anywhere else.⁸

From here, it is easy to compute the recommendation of evidential decision theory. Conditional on any $Bias=b$, the driver can be confident that $Bias_2=b$. The *conditional* expected utility of choosing bias b is therefore the “diagonal” of $(4b + 4c - 6bc)/(c + 1)$, with $b = c$, which works out to

$$CEU(Bias=b) = (8b - 6b^2)/(b + 1).$$

This has its maximum at $b = \sqrt{336}/12 - 1 \approx 0.53$, where the conditional expected utility is around $5/3$. So causal decision theory recommends choosing $2/3$ for the coin's bias, and evidential decision theory 0.53 .

This is puzzling. By using conditional expectations, evidential decision theory effectively represents the driver's choice as a choice of which bias to use not only at the present exit, but at every exit. Given that the optimal bias to use at every exit is $2/3$, one might therefore have expected evidential decision theory to recommend bias $2/3$. But now we find that *causal* decision theory recommends $2/3$, while evidential decision theory recommends something else. As a consequence, drivers who follow the causal advice score higher utility, on average, than drivers who follow the evidential advice. Is this the revenge for Newcomb's problem, where evidential decision makers are often said to outperform causal decision makers?

As Piccione and Rubinstein [1997] point out, drivers who follow the evidential advice also seem to undergo a curious change of mind. Suppose at the start of her journey, the driver already considered what she ought to do at every exit, and saw that $2/3$ is the optimal bias. (If the driver had to fix the bias of her coin once at the start of her

⁸ One might wonder how, in equilibrium, the driver could be certain that she'll choose bias $2/3$, given that all her options now seem equally good. The answer lies in the deliberation dynamics: as soon as she considers the possibility of choosing another bias, the expected utilities change and she'll be drawn back to $2/3$.

journey, both causal and evidential decision theory say she ought to choose bias $2/3$.) As soon as she reaches an exit, she would now change her mind and prefer a coin with bias 0.53 . What could justify this change of opinion? It doesn't seem to be prompted by her learning any interesting facts: if she had anticipated at the start of the journey everything she would observe at the exit, she still would have regarded the coin with bias $2/3$ to be optimal.

Causal decision theorists might be tempted to put the blame on evidential decision theory. After all, the causal theory gives the intuitively correct advice. This is basically the answer of [Aumann et al. 1997] to Piccione and Rubinstein's puzzle.⁹ However, to anyone who has followed the deadlocked philosophical debate over causal vs. evidential decision theory, such a simple diagnosis should raise suspicion.

Indeed, set aside the debate between causal and evidential decision theory. On either account it is plausible that conditional expected utilities measure the agent's degree of desire, or hope, that the relevant proposition is true. By the above argument, the driver should therefore be happy to discover that she uses a coin with bias 0.53 rather than $2/3$. But is this correct? It is certainly not what the driver would have thought at the start of her journey. There, she would have preferred to learn that she uses bias $2/3$. Whence her change of mind, in the absence of any relevant new information?

To understand what is going on here, we need to look at another puzzle that was also introduced in [Piccione and Rubinstein 1997] (as 'example 5') and has gained fame in philosophy by the exchange between [Elga 2000] and [Lewis 2001]: the Sleeping Beauty problem.

5 Halving and Thirring

The scenario is probably familiar. On Sunday night, while Sleeping Beauty is asleep, a fair coin is tossed. If it lands tails, Beauty will be awoken on Monday and again on Tuesday; however, before the second awakening, all her memories of Monday will be

⁹ Piccione and Rubinstein [1997] not only set $c = b$, but also overlook the fact that the probability of being at the first exit (evidentially) depends on the chosen bias b . Instead they use the observation that $P(\text{First}) = 1/(c + 1)$ to compute $P(\text{First}) = 3/5$ based on the driver's prior decision to use bias $2/3$. Setting $c = b$ then yields the payoff function $(6b - 3b^2 + 8)/5$, which has its maximum at $b = 1/3$. The corrected evidential formula $(8b - 6b^2)/(b + 1)$ is given in [Rabinowicz 2003: Appendix].

Aumann et al. [1997] object that by setting $c = b$, Piccione and Rubinstein erroneously represent the driver's present choice as controlling her choice at the other exit. They suggest that if the driver is certain that she chooses bias c , then the expected utility of $\text{Bias} = b$ is $(4b + 4c - 6bc)/(c + 1)$. According to Aumann et al., the optimal bias is the value of b which maximises that function when plugged in for c . In effect, this means that b is an equilibrium in the causal deliberation dynamics – although neither Piccione and Rubinstein nor Aumann et al. derive their solution from general decision-theoretic principles, and neither of them notes the connection to causal vs. evidential decision theory.

erased. If the coin lands heads, Beauty will be awoken only on Monday and made to sleep through Tuesday. Beauty knows all these facts.

The parallels to the absentminded driver should be obvious. Imagine the driver decides to use a fair coin, and focus on the outcome of the toss at the first exit. If the coin lands tails, the driver reaches both the first and the second exit (“Monday” and “Tuesday”), but will have lost all memories of the first by the time she reaches the second. If the coin lands heads, she only reaches the first exit (“Monday”).

The “Sleeping Beauty problem” is the question what Beauty should believe about the outcome of the coin toss when she wakes up on Monday morning. Analogously, we can ask what the driver should believe about the outcome of the first toss when she arrives at the first exit. *Halfers* say her credence in heads should be $1/2$; *thirders* say it should be $1/3$.

Numerous arguments have been given for either side, and this is not the place to recapitulate the whole debate. Broadly speaking, considerations of diachronic rationality tend to support halving, while considerations of evidential support tend to support thirding. For example, halving may be supported by the following principle of *doxastic conservatism*:

If an agent rationally assigns positive credence to a proposition A which is certain not to change its truth-value, then her credence in A should remain unchanged as long as she receives no evidence which, by the lights of her own previous beliefs, has any bearing on the truth of A .

Everyone agrees that on Sunday, Beauty’s credence in heads should be $1/2$. On Monday morning, she then learns various facts (including “centred” facts): that she is awake, that the sky is overcast, and so on. For each of these, we can ask whether Beauty’s Sunday credence in heads would have been any different if she had already known that she would learn the relevant fact when awakening. The answer is plausibly ‘no’. In this sense, whatever Beauty learns on Monday is, by the lights of her previous beliefs, neutral on the outcome of the coin toss. By the principle of doxastic conservatism, her new credence in heads should therefore still be $1/2$.¹⁰

A standard argument for thirding, by contrast, looks just at Beauty’s evidence on Monday morning. Regarding the outcome of the coin toss, Beauty’s evidence includes (i)

¹⁰ This basic line of thought occurs in [Elga 2000] and [Lewis 2001], and is further developed e.g. in [Bradley 2011], [Schwarz 2012a] and [Schwarz 2012b]. One complication here is that Beauty may be forced to violate doxastic conservatism if the coin lands tails: for example, her credence in the proposition that the sky is overcast on either Monday or Tuesday will be high on Monday evening, but low on Tuesday morning, although she doesn’t acquire any relevant information. (At least this should be so if we assume, as we did for the driver, that her belief state on Monday morning is identical to her belief state on Tuesday morning.) One might think that this threat of diachronic irrationality somehow undermines the force of doxastic conservatism at the earlier transition from Sunday to Monday. In [Schwarz 2012b] I argue that it doesn’t.

her knowledge of the general setup, (ii) her experience of being awake, and (iii) the fact that she has no memories from later than Sunday. By itself, (i) lends equal support to the four combinations *Heads & Monday*, *Heads & Tuesday*, *Tails & Monday*, *Tails & Tuesday*; (iii) rules out any further possibilities (such as *Heads & Wednesday*), and (ii) excludes *Heads & Tuesday*. The remaining possibilities should therefore have probability $1/3$ each – assuming that Beauty should believe these propositions to the degree to which they are supported by her present evidence.¹¹

These considerations apply with equal force to the absentminded driver. Of course, the driver plausibly doesn't use a fair coin, so we have to generalise the halfer and thirder positions. Suppose the driver knows that the coin she chooses at the first exit has bias c towards tails. The halfer position then suggests that her credence in tails, when she arrives at the first exit, should equal c . Thirder, on the other hand, suggests that her credence should equal $2c/(c+1)$. Let me run through the thirder argument once more to explain why. To bring out the similarities, imagine the driver knows that she reaches the first exit at a time called "Monday" and the second (if she continues) at "Tuesday". Now when she arrives at the first exit, her relevant evidence consists of (i) her general knowledge of the setup, including the fact that a coin with bias c is tossed at the first exit, and (ii) her observation that she is still on the highway. (i) supports the hypothesis that the coin at the first exit lands tails to degree c ; it is neutral on *Heads & Monday* vs. *Heads & Tuesday*, and on *Tails & Monday* vs. *Tails & Tuesday*; (ii) then rules out *Heads & Tuesday* as well as any further possibilities. This leaves the two tails possibilities with probability $c/(c+1)$ each, and the remaining heads possibility with $(1-c)/(c+1)$.¹²

Here I have assumed that the driver knows all along that the coin at the first exit has bias c . But the arguments plausibly carry over to the driver's conditional beliefs. Thus halving suggests that

$$P(Tails_1/Bias_1=c) = c.$$

Analogously, thirder suggests that

$$P(Tails_1/Bias_1=c) = 2c/(c+1).$$

I have added a subscript '1' to 'Tails' because there might be a second coin toss at the second exit, which doesn't exist in the Sleeping Beauty story.

It is worth teasing out a few consequences of the two positions. Combined with the symmetry assumption (that the driver assigns equal credence to being at the first

¹¹ This argument can be found, in different variations, e.g. in [Piccione and Rubinstein 1997], [Dorr 2002], [Arntzenius 2003], [Horgan 2004], [Horgan 2008] and [?].

¹² For another motivation of the thirder formula, observe that if the experiment were repeated indefinitely, the ratio of occasions where the driver finds herself at an exit and the coin lands tails to such occasions where the coin lands heads would converge to $2c : (1-c)$.

and being at the second exit conditional on reaching both), the halfer formula entails that $P(\text{First} \ \& \ \text{Tails}_1 / \text{Bias}_1 = c) = c/2$ and $P(\text{First} \ \& \ \text{Heads}_1 / \text{Bias}_1 = c) = 1 - c$, while the thirder formula says that $P(\text{First} \ \& \ \text{Tails}_1 / \text{Bias}_1 = c) = c/(c + 1)$ and $P(\text{First} \ \& \ \text{Heads}_1 / \text{Bias}_1 = c) = (1 - c)/(c + 1)$.¹³ Thus

$$P(\text{First} / \text{Bias}_1 = c) = \begin{cases} 1 - c/2 & \text{on halving} \\ 1/(c + 1) & \text{on thirding.} \end{cases}$$

Moreover, since *First* excludes one of the two *Tails*₁ possibilities,

$$P(\text{Tails}_1 / \text{First} \ \& \ \text{Bias}_1 = c) = \begin{cases} c/(2 - c) & \text{on halving} \\ c & \text{on thirding.} \end{cases}$$

Unlike the first toss, the second toss (if it comes about) has no influence on how many exits the driver will reach. From an epistemic perspective, it is an ordinary coin toss, so halvers and thirders should agree that

$$P(\text{Tails}_2 / \text{Tails}_1 \ \& \ \text{Bias}_2 = c) = P(\text{Tails}_2 / \text{Tails}_1 \ \& \ \text{First} \ \& \ \text{Bias}_2 = c) = c.$$

I will call this the *neutral* assumption. (Conditioning on *Tails*₁ here only serves to ensure that the second coin toss actually takes place.)

We can also consider the driver's attitude towards the proposition that the coin at the present exit, whichever it may be, lands tails. That is, define *Tails* as $(\text{First} \ \& \ \text{Tails}_1) \vee (\text{Second} \ \& \ \text{Tails}_2)$. Then

$$P(\text{Tails} / \text{Bias}_1 = \text{Bias}_2 = c) = \begin{cases} c/2 + c/2 \cdot c = (c + c^2)/2 & \text{on halving,} \\ c/(c + 1) + c/(c + 1) \cdot c = c & \text{on thirding.} \end{cases}$$

In the presence of the neutral and symmetry assumptions, all the above characterisations of halving and thirding are equivalent. To illustrate, here is how we can go back from $P(\text{Tails} / \text{Bias}_1 = \text{Bias}_2 = c) = c$ to the original thirder formula on p.16: *Tails*₁ is equivalent to $\text{Tails} \vee (\text{Second} \ \& \ \text{Heads}_2)$; since $P(\text{Tails} / \text{Bias}_1 = \text{Bias}_2 = c) = c$, it follows that $P(\text{Tails}_1 / \text{Bias}_1 = c) = c + c/(c + 1) \cdot (1 - c) = 2c/(1 - c)$.

¹³ In the case of Sleeping Beauty, the symmetry assumption is sometimes motivated by a stipulation that Beauty's two tails awakenings are "subjectively indistinguishable". Symmetry is then supposed to follow from a general principle of self-locating indifference. Halfers sometimes object to symmetry, suggesting that Beauty ought to be certain on Monday that it is Monday and on Tuesday that it is Tuesday; see e.g. [Meacham 2010], [Schwarz 2012a]. This response gets less attractive if it is stipulated, as I have, that the driver has the same beliefs at every exit. (Whether the two situations are otherwise indistinguishable is to my mind irrelevant, but feel free to suppose so if you think it matters.)

6 The puzzle resolved

In section 4, we calculated the optimal bias for the driver's coin, getting $2/3$ for causal decision theory and around 0.53 for evidential decision theory. If you look back at the calculations, you can see that we have unwittingly presupposed thirding. (To my knowledge, this presupposition is universally shared in the literature on the absentminded driver.) In particular, we assumed that $P(\textit{First}/\textit{Bias}_1=c) = 1/(c+1)$. This is in line with thirding, but not with halving. On the halfer account, $P(\textit{First}/\textit{Bias}_1=c) = 1 - c/2$: intuitively, if the probability of reaching both exits is c , and the conditional probability of *Second* given that both exits are reached is $1/2$, then *Second* has probability $c/2$ and so *First* has $1 - c/2$.

Another point where we presupposed thirding is in the calculation of the utilities. We identified the utility of a lottery with the expectation of the utility of the outcomes relative to their objective probability. For example, we assumed that

$$V(\textit{First} \ \& \ \textit{Bias}_2=c \ \& \ \textit{Bias}_1=b) = 4b - 3bc.$$

The reasoning was that conditional on *First*, $\textit{Bias}_1=b$ and $\textit{Bias}_2=c$, the driver's subjective probability for reaching the desolate area should be $1-b$, for getting home $b \cdot (1-c)$, and for getting to the motel $b \cdot c$; then $(1-b) \cdot 0 + b(1-c) \cdot 4 + bc \cdot 1 = 4b - 3bc$. But this presupposes that the driver's probability for *Tails*₁, conditional on $\textit{First} \ \& \ \textit{Bias}_1=b \ \& \ \textit{Bias}_2=c$, equals b . This happens to be correct according to thirding. On the halfer account, $P(\textit{Tails}_1/\textit{First} \ \& \ \textit{Bias}_1=b) = b/(2-b)$. If the driver obeys halving, her probability for reaching the desolate area conditional on $\textit{First} \ \& \ \textit{Bias}_1=b \ \& \ \textit{Bias}_2=c$ is therefore not $1-b$, but $1 - b/(2-b)$. Similarly, her conditional probability for getting home is $b/(2-b) \cdot (1-c)$, and for reaching the motel $b/(2-b) \cdot c$. The desirability of the lottery is therefore $b/(2-b) \cdot (1-c) \cdot 4 + b/(2-b) \cdot c \cdot 1 = (4b - 3bc)/(2-b)$.

For the second exit, we assumed that

$$V(\textit{Second} \ \& \ \textit{Bias}=b) = (1-b)4 + b1 = 4 - 3b.$$

This is correct on both halving and thirding, since on either account the probability of *Tails*, given $\textit{Second} \ \& \ \textit{Bias}=b$, is b .

Now let's reconsider the conditional expected utilities. Conditional on $\textit{Bias}=b$, the uniformity assumption makes it certain that $\textit{Bias}_1=\textit{Bias}_2=b$. So

$$\begin{aligned} CEU(\textit{Bias}=b) &= P(\textit{First}/\textit{Bias}_1=b)V(\textit{First} \ \& \ \textit{Bias}_1=\textit{Bias}_2=b) \\ &\quad + P(\textit{Second}/\textit{Bias}_1=b)V(\textit{Second} \ \& \ \textit{Bias}_2=b) \\ &= \begin{cases} (1-b/2) \cdot \frac{4b-3b^2}{2-b} + b/2 \cdot (4-3b) = 4b - 3b^2 & \text{on halving,} \\ \frac{1}{b+1} \cdot (4b - 3b^2) + \frac{b}{b+1} \cdot (4 - 3b) = \frac{8b-6b^2}{b+1} & \text{on thirding.} \end{cases} \end{aligned}$$

The thirder function, with its maximum at around 0.53, is what we found in section 4; the halfer function is what we expected to find: it coincides exactly with our reasoning that the expected payoff from choosing bias b at every exit is $(1-b) \cdot 0 + b \cdot (1-b) \cdot 4 + b \cdot b \cdot 1 = 4b - 3b^2$, with maximum at $b = 2/3$. If the driver obeys halving, her preferences therefore don't change between the start of the journey and the first exit. Similarly, evidential decision theory no longer recommends her to use the sub-optimal bias 0.53. The puzzle disappears.

Above, I suggested that halving and thirring may be regarded as consequences of two more general attitudes towards rational belief: halving is motivated by the idea that rational agents should only revise their attitude towards a proposition if they receive evidence they regard as relevant to that proposition (at least if the proposition is certain not to change its truth-value over time); thirring, on the other hand, is motivated by the idea that an agent's degree of belief in a proposition should match the extent to which the proposition is supported by their present evidence. In cases like the absentminded driver, the two constraints pull in opposite directions.

To illustrate, suppose the driver is certain that she tosses a coin with bias $2/3$. At the start of her journey, her degree of belief that she will get home was then $2/3 \cdot 1/3 = 2/9$. When she reaches the first exit, she does not receive any information that would allow her to rule out previously open possibilities in which she doesn't get home. If she had known at the start of the journey exactly what she would later learn at the first exit, her credence in getting home would still have been $2/9$. This is why halving entails that her belief should remain unchanged. On the other hand, consider the driver's evidence when she has reached the first exit. The evidence tells her that she is at one of the two exits, at each of which the chance of continuing is $2/3$. Given this information, how likely is it that the driver will get home? If she is at the first exit, the chance of getting home is $2/3 \cdot 1/3 = 2/9$; if she is at the second, the chance is $1/3$. Arguably, the evidential probability for getting home is therefore a mixture of $2/9$ and $2/3$. Thirring says that this should also be the driver's degree of belief. More specifically, it says that if the driver is certain that she tosses a coin with bias b , then $P(\text{Home}) = P(\text{Home}/\text{First})P(\text{First}) + P(\text{Home}/\text{Second})P(\text{Second}) = b(1-b) \cdot 1/(b+1) + (1-b) \cdot b/(b+1) = 2(b-b^2)/(b+1)$, which is $4/15$ for $b = 2/3$. Relative to her previous beliefs, the driver did not gain any information that would lend further support to getting home. Nevertheless, her new evidence, by itself, more strongly suggests that she will get home than her evidence at the start of her journey.

If the driver's beliefs change in accordance with thirring, it is also understandable why she would suddenly be happy to learn that she uses a coin with bias 0.53 rather than $2/3$. This evidence makes it slightly more probable that she is at the first exit and that she will reach the desolate area, but it also makes it more probable that she will get home if she is at the second exit. The latter effect outweighs the former, albeit not by very much.

It may help to imagine many repetitions of the driver's situation, with different choices of the bias. If we look at all drivers who arrive at the first exit, those with bias $2/3$ perform best. However, if we look at the drivers arriving at the second exit, those with a lower bias do better. Theorists who endorse both thirding and evidential decision theory therefore shouldn't accept that their recommendation has a worse average performance: that's true if we look at drivers at the first exit, but it's not true among drivers at the second exit – and for all the driver knows, she might be one of those.

Causal decision theory agrees that the optimal bias to choose at the second exit is 0. But it doesn't agree that $2/3$ is optimal at the first exit: since the driver's choice only controls her present action, the optimal bias at the first exit is 1. If the driver obeys thirding, then by the reasoning from section 4 the right compromise isn't 0.53, but $2/3$. As we will see, this is not true if the driver follows halving.

Before we turn to this, however, we should have another look at the coin-free scenario of section 2. The argument at the end of section 2, which meant to show that the driver should be $2/3$ inclined towards continuing, was completely analogous to the argument for choosing bias $2/3$ in section 4. Did we implicitly presuppose thirding already in section 2?

Recall that without coins, the driver could not decide whether to continue or leave. The more she was inclined towards one option, the more the other looked better. She then needed a tie-breaker to turn her state of indecision into an action. Ideally, the tie-breaker works stochastically, choosing each action with an objective probability that matches the driver's equilibrium degree of inclination, which is also her equilibrium probability that she will carry out the relevant action. For the sake of vividness, let's imagine that a little coin is tossed inside the driver's head to determine whether she will continue; the coin's bias c is set to equal the driver's equilibrium probability that she will continue, $P(\textit{Tails})$. Suppose the driver is aware of this fact, and also knows that she'll reach the same state of indecision at every exit. Plausibly, her equilibrium credence that she will continue at the second exit, given that it is reached, should equal her degree of belief c that she will continue at the present exit, whether it is the first or the second. (This is especially plausible now that the driver knows that c is the objective chance that she will continue at the second exit.) But now we have all the ingredients to show that the driver's equilibrium probability satisfies thirding – as per the argument at the very end of the previous section.

From a halfer perspective, this is odd. Why should the driver obey thirding about the coin in her head, but halving about the coin in her hand? Surely it makes no epistemic difference whether the coin is tossed inside or outside the driver's head!

But there really is an important difference. Consider the evolution of the driver's beliefs in the scenario where she has only two options. To strengthen the conservative case for halving, assume that the driver has already figured out at the start of her journey that the

deliberation equilibrium in the problem she will face at any exit lies at $P(\textit{Continue}) = 2/3$. Hence she knows that a coin with bias $2/3$ will be tossed at the first exit, and again at the second exit if the first one lands tails. Now she reaches the first exit. Obeying the principle of doxastic conservatism, her new credence in \textit{Tails}_1 is still $2/3$, as is her new credence in \textit{Tails}_2 conditional on \textit{Tails}_1 . Moreover, the latter credence is not sensitive to whether she is at the first exit or the second: $P(\textit{Tails}_2/\textit{Tails}_1) = P(\textit{Tails}_2/\textit{Tails}_1 \ \& \ \textit{First})$. This means that the driver fully obeys halving. Her credence in \textit{Tails} , given by the halfer equation $P(\textit{Tails}/\textit{Bias}_1 = \textit{Bias}_2 = c) = (c + c^2)/2$, is now $5/9$. But then her credence is not in equilibrium: at $P(\textit{Continue}) = 5/9$ and $P(\textit{Continue}_2) = 2/3$, continuing has a higher expected utility than leaving. In the process of deliberation, the driver's probability for $\textit{Continue}$ will therefore increase. Under plausible assumptions about the deliberation dynamics – for example, that the symmetry assumption remains true throughout the deliberation –, this process converges to the thirder distribution with $P(\textit{Continue}) = 2/3$. Thus the coin with bias $2/3$ will indeed be tossed, just as the driver anticipated at the start of her journey.

The crucial point is that in the scenario from section 2 the coin's bias is determined by the driver's state of indecision, and therefore affected by the dynamics of deliberation. Compare the situation where the driver can actually decide to toss a coin. Suppose again that she has figured out in advance that she will decide to use the coin with bias $2/3$; that is, she has figured out that the equilibrium in her decision problem lies at $P(\textit{Bias}=2/3) = 1$. As before, upon reaching an exit, the conservative belief update will result in a halfer distribution, with $P(\textit{Continue}) = 5/9$. But $\textit{Continue}$ here doesn't stand for one of the driver's options. Her options are given by the propositions $\textit{Bias}=b$. To be sure, one of her options, $\textit{Bias}=0$, can be identified with the option to continue, but since the driver is certain that she chooses bias $2/3$, $P(\textit{Bias}=0)$ is 0, not $5/9$. In fact, both before and after the belief update, $P(\textit{Bias} = 2/3) = 1$. This, we assumed, is also her equilibrium state. So deliberation does not change her beliefs and she will retain the halfer distribution.

7 Useless coins

There are four ways of combining causal and evidential decision theory with halving and thirding. We know what three of these combinations say about the optimal bias in the driver's decision problem:

	<i>EDT</i>	<i>CDT</i>
<i>Thirding</i>	0.53	2/3
<i>Halving</i>	2/3	

What about the fourth combination, of causal decision theory and halving?

We already know the corrected utilities:

$$\begin{aligned} V(\text{First} \ \& \ Bias_2=c \ \& \ Bias_1=b) &= (4b - 3bc)/(2 - b); \\ V(\text{Second} \ \& \ Bias=b) &= 4 - 3b. \end{aligned}$$

As for the probabilities, we figured out that $P(\text{First}/Bias_1=c) = 1 - c/2$. By uniformity, we can substitute $Bias_2$ for $Bias_1$. Then $P(\text{First} \ \& \ Bias_2=c) = P(\text{First}/Bias_2=c)P(Bias_2=c) = P(Bias_2=c)(1 - c/2)$. So

$$\begin{aligned} P(\text{First} \ \& \ Bias_2=c) &= P(Bias_2=c)(1 - c/2); \\ P(\text{Second}) &= \sum_{c \in B} P(Bias_2=c)c/2. \end{aligned}$$

The resulting expected utilities are

$$EU(Bias=b) = \sum_{c \in B} P(Bias_2=c) \frac{3bc^2 + 3b^2c - 20bc + 8c + 8b}{4 - 2b}.$$

(The diagonal, with $b=c$, is again the *CEU* function $4b - 3b^2$.)

Does this still recommend the coin with bias $2/3$? If the driver is confident that $Bias_2 = 2/3$, the expected utility of $Bias = 2/3$ is $4/3$, but the expected utility of $Bias = 1$ is $5/3$. So the driver can't resolve to use bias $2/3$. In fact, for every available value b , if the driver knows that she chooses b , then choosing either bias 1 or bias 0 has higher expected utility. For the extreme values, of course, the right choice is always the opposite of what she thinks she will choose. There is no stable option: the driver should be undecided which coin to choose!

Without a stable solution, it is worth to once again drop the (revised) uniformity assumption, that the driver is certain to use the same bias at every exit. We can then follow the same basic line of thought as at the end of section 2. In equilibrium, $P(Bias_1=c/\text{First}) = P(Bias=c)$. By halving, $P(\text{Tails}_1/\text{First} \ \& \ Bias_1=c) = c/(2 - c)$. Hence $P(\text{Tails}_1/\text{First}) = \sum_{c \in B} P(Bias=c)c/(2 - c)$. Together with the symmetry assumption, it follows that $P(\text{First}) = 1/(e + 1)$, where $e = \sum_{c \in B} P(c)c/(2 - c)$. The equilibrium probabilities of the states are

$$\begin{aligned} P(\text{First} \ \& \ Bias_2=c) &= P(Bias=c)/(e + 1); \\ P(\text{Second}) &= e/(e + 1). \end{aligned}$$

Thus

$$EU(B=b) = \sum_{c \in B} \left(\frac{P(c)}{e + 1} \cdot \frac{4b - 3bc}{2 - b} \right) + \frac{e}{e + 1} \cdot (4 - 3b).$$

In equilibrium, all options with positive probability must have maximum expected utility. One such equilibrium (the only one, as far as I can tell) is given by the probability function that assigns $2/3$ to $Bias = 1$ and $1/3$ to $Bias = 0$.

Now, tossing a coin with bias 1 or 0 is evidently pointless: $Bias = 1$ means to continue, $Bias = 0$ to leave. The upshot is that the driver will find no use in her coins: in the end, she will be torn between continuing and leaving – exactly like in section 2 – and she will be certain that she won’t use any of the coins.¹⁴

Let me briefly mention two consequences of this result. First, there is a widely accepted result in game theory suggesting that every finite game has a Nash equilibrium if the players are allowed to randomise their choices. By essentially the same argument, it has been suggested that every unstable decision problem becomes stable if randomisation is allowed (see e.g. [Harper 1986]). The argument assumes that the utility of a randomised choice in a given state equals the expectation of the utility of the individual options relative to the chosen probabilities. In practice, this assumption may be false if, for example, there is a punishment on randomisation, or if one of the original options is a bet that the agent never resorts to randomisation. The absentminded driver illustrates a very different reason why the assumption may be false: it is false if the agent’s beliefs about the outcome of her randomised choice are not in line with the objective probabilities. If our driver decides to toss a coin with bias $2/3$, then (on the halfer account) her rational credence in *Tails* is not $2/3$, but $5/9$. This turns out to have a similar effect as adding extra costs to randomisation: no matter what the driver thinks she will do, one (or both) of the non-randomised choices always looks better than every randomised choice.¹⁵

Relatedly, the present observations count against formulating causal decision theory in terms of expected conditional chance. Several authors, most prominently Brian Skyrms [1984: ch.4] (see also [Williams 2010], [Joyce 1999: 166f.]) have suggested that agents should maximise

$$EU_C(A) = \sum_{S \in W} \sum_x x \cdot P(Ch(S/A)=x) \cdot V(S),$$

where W is a partition of possibilities whose members have uniform utility (meaning that for any $S \in W$ and any proposition B compatible with S , $V(S \& B) = V(S)$) and $Ch(S/A)=x$ is the proposition that the objective chance of S , conditional on A , equals

¹⁴ In the equilibrium state, choosing any coin has a strictly lower expected utility than continuing or exiting. The same is true if we stick with the uniformity assumption. Then deliberation leads to an equilibrium where $Bias=1$ has probability $4/5$ and $Bias=0$ has $4/5$ – again exactly like in section 2.

¹⁵ There are other interpretations of mixed (randomised) equilibria in game theory, on which they don’t reflect the players’ choice of a randomised strategy. [Aumann and Brandenburger 1995], for example, suggest interpreting Nash equilibria as equilibria in belief.

x .¹⁶ But this is only plausible if $P(S/A \ \& \ Ch(S/A) = x) = x$, which is false for the absentminded driver if she obeys halving.

Admittedly, it is somewhat surprising that the driver should assign credence 5/9 to tails if she knows that the objective chance is 2/3. The driver thereby seems to violate a form of the “Principal Principle” according to which one’s rational credence should match the known objective chances (compare [Lewis 1980]).¹⁷ However, some such violation is unavoidable: if the driver knows that the objective chance of tails at every exit (and hence also at the present exit, whichever it is) equals c , then $P(Tails_1)$ and $P(Tails)$ cannot both equal c as long as the driver is open-minded about her location and the outcome of the second toss. That’s because $Tails_1 = Tails \vee (Second \ \& \ Heads_2)$. If $P(Second \ \& \ Heads_2) > 0$, it follows that $Tails_1$ and $Tails$ cannot have the same probability. If one of them matches the known chance, the other one doesn’t.

8 Conclusions

Let me sum up the main observations of this paper. In section 2 I looked at the driver’s predicament assuming her only options are *Leave* and *Continue*. Here, evidential decision theory recommended *Continue*. In causal decision theory, the decision problem is unstable. I have argued that in this type of case, the agent should adopt a state of indecision. Specifically, the driver should be 2/3 inclined towards *Continue* and 1/3 towards *Leave*.

Turning to the second version of the puzzle, where the driver can choose to toss a coin, the outcome depends not only on evidential vs. causal decision theory, but also on halving vs. thirding. Two of the four combinations – thirding with causal decision theory and halving with evidential decision theory – recommend tossing a coin with bias 2/3. Given that this is *prima facie* the optimal choice, this mirrors arguments in [Arntzenius 2002] and [Briggs 2010] to the effect that thirders should be causal decision theorists and halfers evidential decision theorists. However, I have argued that the other two combinations are perfectly acceptable as well. The combination of halving with causal decision theory recommends a state of indecision with the same long-run payoff as choosing the 2/3 coin. Thirding with evidential decision theory recommends the coin with bias 0.53. In general, this is the coin which, according to thirding, the driver should desire to toss, no matter which decision theory she follows. I have argued that it is indeed the most desirable choice given the ahistorical, “time-slice” perspective on rationality that motivates thirding.

16 In Skyrms’s formulation, $EU_C(A)$ is expressed as $\sum_K P(K) \sum_C Ch_K(C/A) V(C)$, where K ranges over chance hypotheses, C over “consequences” and $Ch_K(C/A)$ is the conditional chance of C given A according to K .

17 The driver does in fact not violate the Principle formulated in [Lewis 1980], since her credences aren’t “initial”. One might also question whether there even is an objective chance for the centred proposition *Tails*, as opposed to the uncentred *Tails*₁ and *Tails*₂.

Along the way, we have encountered some other interesting facts. One was that on any account, the driver’s rational degrees of belief about her coins cannot match the known objective chances. This casts doubt on some popular ideas in game theory and decision theory. We have also seen that if agents are equipped with a stochastic tie-breaker to resolve states of rational indecision, then their attitudes towards the outcomes of this mechanism can diverge from the attitudes they would have towards the outcomes of an explicitly chosen randomisation device.

References

- Frank Arntzenius [2002]: “Reflections on Sleeping Beauty”. *Analysis*, 62: 53–62
- [2003]: “Some problems for conditionalization and reflection”. *Journal of Philosophy*, 100: 356–370
- [2008]: “No Regrets, or: Edith Piaf Revamps Decision Theory”. *Erkenntnis*, 68: 277–297
- Robert Aumann and Adam Brandenburger [1995]: “Epistemic Conditions for Nash Equilibrium”. *Econometrica*, 63: 1161–1180
- Robert Aumann, Sergiu Hart and Motty Perry [1997]: “The Absent-Minded Driver”. *Games and Economic Behavior*, 20: 102–116
- Darren Bradley [2011]: “Confirmation in a Branching World: The Everett Interpretation and Sleeping Beauty”. *British Journal for the Philosophy of Science*, 62: 323–342
- Rachael Briggs [2010]: “Putting a Value on Beauty”. In T. Szabo Gendler and J. Hawthorne (Eds.) *Oxford Studies in Epistemology*, vol Vol. 3. Oxford: Oxford University Press
- Alex Byrne and Alan Hájek [1997]: “David Hume, David Lewis, and Decision Theory”. *Mind*, 106: 411–728
- Cian Dorr [2002]: “Sleeping Beauty: In defence of Elga”. *Analysis*, 62: 292–296
- Andy Egan [2007]: “Some Counterexamples to Causal Decision Theory”. *Philosophical Review*, 116: 93–114
- Adam Elga [2000]: “Self-locating belief and the Sleeping Beauty problem”. *Analysis*, 60: 143–147
- David Etlin [2008]: “Desire, Belief, and Conditional Belief”. PhD Dissertation

- William Harper [1984]: “Ratifiability and Causal Decision Theory: Comments on Eells and Seidenfeld”. *PSA*, 2: 213–228
- [1986]: “Mixed Strategies and Ratifiability in Causal Decision Theory”. *Erkenntnis*, 24: 25–36
- Terry Horgan [2004]: “Sleeping Beauty Awakened: New Odds at the Dawn of the New Day”. *Analysis*, 64: 10–21
- [2008]: “Synchronic Bayesian Updating and the Sleeping Beauty Problem: Reply to Pust”. *Synthese*, 160: 155–159
- James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press
- [2002]: “Levi on Causal Decision Theory and the Possibility of Predicting One’s Own Actions”. *Philosophical Studies*, 110: 69–102
- David Lewis [1980]: “A Subjectivist’s Guide to Objective Chance”. In Richard Jeffrey (Ed.), *Studies in Inductive Logic and Probability* Vol. 2, University of California Press.
- [1981]: “Causal Decision Theory”. *Australasian Journal of Philosophy*, 59: 5–30
- [2001]: “Sleeping Beauty: Reply to Elga”. *Analysis*, 61: 171–176
- Christopher Meacham [2010]: “Unravelling the Tangled Web: Continuity, Internalism, Non-Uniqueness and Self-Locating Beliefs”. In Tamar Szabo Gendler and John Hawthorne (Eds.) *Oxford Studies in Epistemology, Volume 3*, Oxford University Press, 86–125
- Robert Nozick [1969]: “Newcomb’s problem and two principles of choice”. In N. Rescher (Ed.) *Essays in honor of Carl G. Hempel*, Dordrecht: Reidel, 114–146
- Michele Piccione and Ariel Rubinstein [1997]: “On the Interpretation of Decision Problems with Imperfect Recall”. *Games and Economic Behavior*, 20: 3–24
- Huw Price [1986]: “Against Causal Decision Theory”. *Synthese*, 67: 195–212
- [1991]: “Agency and Probabilistic Causality”. *British Journal for the Philosophy of Science*, 42: 157–176
- Wlodek Rabinowicz [1989]: “Stable and Retrievable Options”. *Philosophy of Science*, 56: 624–641
- [2002]: “Does Practical Deliberation Crowd Out Self-Prediction?” *Erkenntnis*, 57: 91–122

- [2003]: “Remarks on the Absentminded Driver”. *Studia Logica*, 73: 241–256
- Reed Richter [1985]: “Rationality, Group Choice and Expected Utility”. *Synthese*, 63: 203–232
- Wolfgang Schwarz [2012a]: “Changing Minds in a Changing World”. *Philosophical Studies*, 159: 219–239
- [2012b]: “Diachronic rationality”. Manuscript
- Brian Skyrms [1984]: *Pragmatics and Empiricism*. Yale: Yale University Press
- [1990]: *The Dynamics of Rational Deliberation*. Cambridge (Mass.): Harvard University Press
- Jordan Howard Sobel [1990]: “Maximization, Stability of Decision, and Actions in Accordance with Reason”. *Philosophy of Science*, 57: 60–77
- J. Robert G. Williams [2010]: “Counterfactual Desire as Belief”. Unpublished Manuscript