Lost memories and useless coins: Revisiting the absentminded driver*

Wolfgang Schwarz

Draft, 03 June 2012

1 Introduction

Sometimes it is controversial what rationality demands in a given situation. One-boxers and two-boxers disagree about what choice would be rational in Newcomb's problem, halfers and thirders disagree about what beliefs would be rational for Sleeping Beauty. 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 the 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.

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. Moreover, 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.

Following [Piccione and Rubinstein 1997], I will begin with a simple version of the scenario in which the driver's only two options are to turn off and to continue. In causal

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

decision theory, this generates an "unstable" decision problem, in which the driver ought to do whatever she believes she won't do. Drawing on work by Brian Skyrms, I suggest that in situations like this, decision theory should recommend a certain state of indecision. This means that the cognitive architecture of rational agents must include a tie-breaking mechanism to convert states of indecision into acts. A state of indecision that is resolved by a stochastic tie-breaker must be sharply distinguished from a choice of a "mixed strategy". This is brought out in the second half of the paper, where I assume that the driver is allowed to randomise her choice by tossing a (possibly biased) coin. Here we run into a version of the Sleeping Beauty problem when we ask about the driver's beliefs concerning the outcome of her coin toss. The answer that is universally presupposed in the literature on the absentminded driver corresponds to thirding, and generates a puzzle: it entails that the driver should desire to use a bias that seems clearly sub-optimal. This is not so if the driver obeys halfing. Surprisingly, it then also turns out that the driver should disregard the option to toss a coin and instead remain undecided between her original two options.

The point of this paper is not to take sides in the debate between causal and evidential decision theory or between halfing and thirding. Rather, I will explain how all four combinations give sensible answers to the puzzle if one keeps in mind the general perspective that motivates these combinations. Special attention will however be payed to the combination of causal decision theory with halfing, partly because that has the most interesting consequences, and partly because it happens to be the view I myself prefer.

2 Absentmindedness and two types of expected utility

Before we begin, I should make some clarifications about the driver's predicament. Our driver suffers from a somewhat 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 completely forget the whole event before she reaches the second exit, so that she arrives at the second exit in the very same state of mind she was in at the first. 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: the probability assigned to a state represents the degree to which she believes that this 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 $V(Motel) = 1$. Due to the memory loss caused by the highway, her beliefs and desires are guaranteed to be the same in all relevant respects whenever she gets to an exit.[1]

Since the driver is in the same relevant state at every exit, and the exits themselves are similar in all relevant ways, it is plausible that the driver will make the same choice whenever she reaches an exit. In particular, if her beliefs and desires determine a particular choice as uniquely rational, then this is what she is going to do on every occasion. Let's assume the driver herself is confident that whatever she does at the present exit is also what she does at the other exit (if it is reached). There are then only two possible outcomes: either the driver turns off at the first exit and spends the night in her car, or she continues at both exits and ends up in the motel. Since she prefers the second outcome over the first, one might think that the rational strategy for her is to continue.

But let's have a closer look at the driver's decision problem once she reaches an exit. To be sure, she might 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 there is something odd about this line of thought. The driver is right that if she leaves the highway, then 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, but it is not quite as bad as leaving. But 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. Most famously, taking both boxes in Newcomb's problem 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

1 The driver's predicament is a little far-fetched and raises distracting questions about memory loss, but structurally similar situations without memory loss can easily arise when several agents need to coordinate their actions. For example, suppose two perfectly rational agents have been randomly assigned to two rooms, each containing a button. If both agents push their button ("continue") they each receive payoff 1; if agent 1 pushes and agent 2 does not, they receive payoff 4; if agent 1 doesn't push, they get nothing.

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. Classical ("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). Standard treatments like [Savage 1954] are often not very explicit about what counts as 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.[2]

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

2 For more precise statements, see e.g. [Lewis 1981] and [Joyce 1999: ch.5]. Although 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. See also [Joyce 2002: sec.1] in defense of the label "classical".

correlation from merely evidential correlation. This is what went wrong with the driver's argument that she ought to continue.[3]

If we try to repair this mistake, 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 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.[4]

	First & Continue$_2$	*First & Leave*$_2$	*Second*
Continue	1	4	1
Leave	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. *Continue*$_2$ means that the driver *either* continues at the second exit *or* doesn't reach it but would have continued if she had reached it. Recall that by the workings of the highway, the driver's state of mind when arriving at the second exit is guaranteed to be the same as her state of mind when

3 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.

4 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 3), 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.

arriving at the first. So 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. $Continue_2$ says that she would continue, $Leave_2$ 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, so 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.[5]

It is crucial for classical 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:

$$P(First/Continue_1) = P(Second/Continue_1) = 1/2.$$

Call this the *symmetry* assumption. ($Continue_1$ means that the driver continues at the first exit.) As indicated above, I will also assume – although not for the whole of the paper – that the driver is confident that she makes the same choice at every exit; more specifically,

$$P(Continue_1/Continue_2) = P(Continue_2/Continue_1) = 1.$$

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 $(First \,\&\, Continue_1) \vee (Second \,\&\, Continue_2)$. Symmetry and uniformity

5 Multi-agent versions of the driver's scenario are more perspicuous here (see footnote 1): '$Continue_2$' can then simply mean that the second agent chooses "Continue", and it is clear why this is not under the causal influence of the first agent.

then entail that[6]

$$P(\textit{First} \,\&\, \textit{Continue}_2) = c/2;$$
$$P(\textit{First} \,\&\, \textit{Leave}_2) = 1 - c;$$
$$P(\textit{Second}) = c/2.$$

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* & $\textit{Leave}_2$. Conditional on *Continue*, the driver's probability is evenly divided between *First* & $\textit{Continue}_1$ 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.

3 Rational indecision

In classical (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 ought that she ought to leave.

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

Now $\textit{Continue}_2$ divides into *First* & $\textit{Continue}_2$ and *Second* & $\textit{Continue}_2$, 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* & $\textit{Continue}_1$ is equivalent to *Second*. The remaining possibility *First* & $\textit{Leave}_2$ must then have probability $1 - c$.

Shouldn't this make her reconsider her belief that she will continue? Moreover, 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: $P(\mathit{Continue}) = 1/2$. 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 $P(\mathit{Continue})$ goes up, the expected utilities change, so she has to re-assess whether continuing is still the optimal choice. If $P(\mathit{Continue})$ gets too high, *Leave* becomes the better option, in which case the driver's probability in *Continue* should decrease. Skyrms shows that this process, if the details are filled in sensibly, always leads to an equilibrium: a point where the probabilities no longer change. (That such a point exists follows from Brouwer's fixed point theorem.) For the driver, the deliberation equilibrium lies at a value of $P(\mathit{Continue})$ at which both options have the same expected utility. The probability is then no longer pulled in either direction. There is only one such value, namely $P(\mathit{Continue}) = 4/5$, which will be reached no matter how the driver began her deliberation.[7]

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 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 maximize expected utility *conditional on being realized* (see [Harper 1984], [Sobel 1990], [Rabinowicz 1989]). This would mean that in the driver's situation, no choice is rational. Others (notably [Arntze-

7 Some authors have questioned whether decision makers can assign probabilities to their own actions, mainly because these probabilities are hard to directly elucidate in terms of betting odds. 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.

nius 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.

I would like to go a little further and suggest that decision theory can sometimes recommend a *state of indecision*. The driver ought to be undecided between her two options, but somewhat more inclined towards continuing. A state of indecision is not just a state of ignorance. It's true that if the driver is rational, then she won't know what she is going to do. But she also won't know *what to do*. She will be practically, not just epistemically, torn between her options. A limit case of a state of indecision is 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 the decision with that epistemic state. Similarly, we should not identify a "partial decision", in which an agent only has certain degrees of inclination towards her options, with a corresponding state of partial belief.

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. One way to get into this state is by a Skyrmsian process of rational deliberation. But we needn't require that the driver actually goes through that process. What matters is that somehow or other she ends up in the equilibrium state.

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, 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 neuronal 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 the 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 would again be c (because the tie-breakings aren't linked). So $P(\textit{First} \,\&\, \textit{Continue}_2) = P(\textit{First}) \cdot c$. What is $P(\textit{First})$? For a start, *First* divides into $\textit{First} \,\&\, \textit{Continue}_1$ and $\textit{First} \,\&\, \textit{Leave}_1$. By the symmetry assumption,

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

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

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

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

$$\begin{aligned}
&P(\textit{First} \,\&\, \textit{Continue}_2) = c/(c+1);\\
&P(\textit{First} \,\&\, \textit{Leave}_2) = (1-c)/(c+1);\\
&P(\textit{Second}) = c/(c+1),
\end{aligned}$$

which yields

$$\begin{aligned}
EU(\textit{Continue}) &= \frac{c}{c+1} \cdot 1 + \frac{1-c}{c+1} \cdot 4 + \frac{c}{c+1} \cdot 1 = \frac{4-2c}{c+1};\\
EU(\textit{Leave}) &= \frac{c}{c+1} \cdot 0 + \frac{1-c}{c+1} \cdot 0 + \frac{c}{c+1} \cdot 4 = \frac{4c}{c+1}.
\end{aligned}$$

In equilibrium, both options must have the same expected utility, which means that $c = 2/3$. Without the uniformity assumption, then, 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 $(1-c) \cdot 0 + c(1-c) \cdot 4 + c^2 \cdot 1 = 4c - 3c^2$, which has its maximum at $c = 2/3$. 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 the exit, she may now 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 since 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?[8]

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. But 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 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.[9]

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

8 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]). If I understand him correctly, Price would suggest that the driver's choice between continuing and leaving actually does have causal influence over where she is. For example, 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. But surely absentmindedness does not convey powers to undo the past!

9 Pace [Lewis 1981: 29f.].

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.

As we saw at the end of the previous section, if the driver has to select the same coin at every exit, then the optimal choice seems to be a coin with bias 2/3 towards tails. Again, the expected payoff from choosing a coin with bias c at every exit is $4c - 3c^2$, which is maximal for $c = 2/3$. ("Bias" always means bias towards tails.)

Suppose the driver is confident that she chooses the same bias at every exit. Given this revised uniformity assumption, we might expect evidential decision theory to agree with the recommendation of bias 2/3. 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.

But once again, let's have a closer look. When she has reached an exit, the driver doesn't know whether it is the first or the second. She can now reason as follows. "If I'm at the first exit, the optimal bias is 2/3. On the other hand, if I'm at the second exit, then I'm better off with a lower value. For example, a coin with bias 1/100 would then give me a 99% chance of getting home. Since I don't know where I am, I should choose a compromise between the bias that would be best at the first exit and the one that would be best at the second – that is, some value in between 2/3 and 0."

This is Piccione and Rubinstein's "paradox of the absentminded driver", from [Piccione and Rubinstein 1997]. The paradox is that it seems wrong to choose a coin with bias below 2/3. Agents choosing such a lower bias score less average utility than agents who use bias 2/3 – and our driver knows this. Piccione and Rubinstein also imagine that the driver already considered at the start of her journey what she ought to do when she reaches an exit: knowing that she'll use the same coin every time, she would then have figured out that 2/3 is the optimal bias. As soon as she reaches an exit, she is now supposed to change her mind and prefer the coin with a lower bias. But it's not like the driver gained any new information that would make it reasonable to swap the coins: if the lower bias is indeed the best choice, she should have been able to figure that out at the beginning.

It looks like something must be wrong with the above line of reasoning. Indeed, it isn't hard to spot what went wrong: as [Aumann et al. 1997] first pointed out, the argument implicitly assumes that the driver's present choice also determines her choice at the other exit (if reached). But while the present choice is strong evidence for what happens at the other exit, it doesn't directly control the other choice. In short, the paradoxical line of thought relies on an evidential conception of rational choice. If we properly distinguish

causal from merely evidential correlations, it turns out that when the driver finds herself at an exit, 2/3 is still the optimal choice.

Let's see why this is so. The argument is essentially the same as the argument at the end of the previous section, but I will spell it out a bit more carefully this time. To apply classical (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. One way to deal with this is to pretend that for every coin and every exit, there is a fact of the matter about how that coin would land if it were tossed at that exit. A complete specification of a state could then include the outcome of every possible coin toss. One state, for example, would be that the driver is at the second exit, the coin with bias 2/3 would land heads if tossed, the coin with bias 1/2 would land tails, and so on. Combined with any choice of the driver, this would settle the utility of the eventual outcome. But, if the coin with bias 1/2 isn't actually tossed, is there really a fact of the matter about how it would have landed? To avoid this problem, we will content ourselves with a partition of states which, combined with a choice of the driver, only determines an *objective chance* for the relevant outcomes. (Here we follow the advice of [Lewis 1981].) 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 reaching the motel. In general, the cells in our decision matrix are propositions assigning objective probabilities to the three ultimate outcomes. I will call such propositions *lotteries*.

One of the states is that the driver is 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. But that's not a complete lottery yet, 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* & $Continue_2$ and *First* & $Leave_2$. Similarly, we now divide *First* into all possible choices the driver could make at the second exit. Let '$Bias_2 = x$' 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 $Bias_2$, 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 with moderately sized n and m ($m \geq 1$, $n \leq m$).

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 given probabilities. 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 $b, c \in B$,

$$V(\text{First } \& \ Bias_2{=}c \ \& \ Bias{=}b) = (1-b)0 + b(1-c)4 + bc1 = 4b - 3bc;$$
$$V(\text{Second } \& \ Bias{=}b) = (1-b)4 + b1 = 4 - 3b.$$

Next, we need to know a few things about how the driver's probabilities are distributed over the states. Observe that $P(\text{First } \& \ Bias_2{=}c) = P(Bias_2{=}c)P(\text{First}/Bias_2{=}c)$. 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. $P(\text{First}/Bias_1 = c) = 1/(c+1)$. Assuming uniformity for the driver's options, we can substitute $Bias_1$ for $Bias_2$. It then follows that $P(Second/Bias_2{=}c) = 1 - P(First/Bias_2{=}c) = c/(c+1)$. Hence

$$P(\text{First } \& \ Bias_2{=}c) = P(Bias_2{=}c)/(c+1);$$
$$P(\text{Second}) = \sum_{c \in B} P(Bias_2{=}c)c/(c+1).$$

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.

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.

This, in essence, is the solution of [Aumann et al. 1997] to Piccione & Rubinstein's puzzle.[10] We can also fill in the details in the mistaken evidential reasoning now. 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 $5/3$.[11]

But has the puzzle really been solved? Recall how we got here. We saw that, if the driver had to choose which coin to use at every exit, the optimal choice would be $2/3$. This suggested that, given the uniformity assumption, *evidential decision theory* should also recommend choosing bias $2/3$. Instead it seemed to recommend a value less than $2/3$ – which we have now confirmed to be 0.53. Pointing out that causal decision theory recommends $2/3$ hardly answers this puzzle.

Let me reiterate why the puzzle is puzzling. First of all, evidential decision theory, by comparing the utility of different options conditional on them being chosen, effectively represents the driver's choice as a choice of which bias to use at every exit. Not so for causal decision theory, which treats the choice at the other exit as independently given. But now we find that causal decision theory recommends the coin that is best if chosen at every exit, while evidential decision theory recommends something else.

A second puzzle is that evidential decision theory seems to recommend a plainly bad choice. Using bias 0.53 has a provably worse average performance than using $2/3$. This time, the evidential theorist can't even complain about our use of tie-breaking mechanisms, because the mechanisms weren't used. To anyone who has followed the deadlocked philosophical debate over causal vs. evidential decision theory, the suggestion that agents who obey the evidential norm do worse in such a simple situation should raise suspicion.

The puzzle goes further. Set aside the debate between causal and evidential decision theory. On either account it is plausible that conditional expected utilities measure the

10 In [Aumann et al. 1997], the solution is not derived from general decision-theoretic principles. Aumann et al. do present $(4b + 4c - 6bc)/(c + 1)$ as the expected utility of $Bias{=}b$, assuming that the driver is certain that she chooses bias c; they further suggest that a solution to the driver's problem would have to be a value of b which maximises this function when plugged in for c. Although Aumann et al. don't say it, this effectively means that b is an equilibrium in the deliberation dynamics.

11 [Piccione and Rubinstein 1997] reach a different result, because they not only set b equal to c, but also overlook the fact that the probability of being at the first exit evidentially depends on the chosen bias b. In effect, Piccione and Rubinstein use the observation that $P(First) = 1/(c + 1)$ to compute $P(First) = 3/5$, 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 $(8b - 6b^2)/(b + 1)$ appears in [Rabinowicz 2003: Appendix].

agent's degree of desire, or hope, that the relevant proposition is true. By the argument of the present section, 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 that 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 Halfing and Thirding

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, but before the second awakening, all her memories of Monday will be erased. If the coin lands heads, Beauty will be awoken only on Monday and sleep through all 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 halfing, while considerations of evidential support tend to support thirding. For example, halfing is plausibly entailed 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 information which (by her lights) has any bearing on the truth of A.

In the case of Sleeping Beauty, everyone agrees that on Sunday, her credence in heads should be 1/2. When she then awakens on Monday, she doesn't seem to gain any new information that would support one of the outcomes over the other: whatever Beauty learns on Monday – that she is alive, that the sky is overcast, etc. –, her Sunday credence in heads, conditional on the assumption that she will make all these discoveries, is 1/2. By the principle of doxastic conservatism, her new credence in heads should therefore still be 1/2.[12]

A standard argument for thirding, by contrast, looks just at Beauty's evidence on Monday morning. Regarding the outcome of the coin toss, her evidence includes (i) 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.[13]

These considerations apply with equal force to the absentminded driver. Of course, the driver 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. Thirding, 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 and (ii) her observation that she is still on the highway. (i) includes the information that a coin with bias c is tossed at the first exit, which supports the outcome tails to degree c; on the other hand, (i) 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

12 This line of thought already occurs in [Elga 2000] and [Lewis 2001], and is further developed e.g. in [Bradley 2011], [**?**] and [**?**]. 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 [**?**] I argue that it doesn't.

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

$c/(c+1)$ each, and the remaining heads possibility with $(1-c)/(c+1)$.[14]

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 halfing suggests that

$$P(\mathit{Tails}_1/\mathit{Bias}_1 = c) = c.$$

Analogously, thirding suggests that

$$P(\mathit{Tails}_1/\mathit{Bias}_1 = c) = 2c/(c+1).$$

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 the first and the second exit conditional on the hypothesis that she reaches both), the halfer principle entails that $P(\mathit{First}\ \&\ \mathit{Tails}_1/\mathit{Bias}_1 = c) = c/2$ and $P(\mathit{First}\ \&\ \mathit{Heads}_1/\mathit{Bias}_1 = c) = 1-c$, while the thirder principle says that $P(\mathit{First}\ \&\ \mathit{Tails}_1/\mathit{Bias}_1 = c) = c/(c+1)$ and $P(\mathit{First}\&\mathit{Heads}_1/\mathit{Bias}_1 = c) = (1-c)/(c+1)$.[15] Thus

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

Moreover, since *First* excludes one of the two Tails_1 possibilities,

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

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 halfers and thirders should agree that

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

14 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 $2c : (1-c)$.

15 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 2012]. 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.)

I will call this the *neutral* assumption. (Conditioning on $Tails_1$ here 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 $(First \,\&\, Tails_1) \vee (Second \,\&\, Tails_2)$. Then

$$P(Tails/Bias_1 = Bias_2 = c) = \begin{cases} c/2 + c/2 \cdot c = (c + c^2)/2 & \text{on halfing,} \\ c/(c+1) + c/(c+1) \cdot c = c & \text{on thirding.} \end{cases}$$

In the presence of the neutral and symmetry assumptions, all the above characterisations of halfing and thirding are equivalent. To illustrate, here is how we can go back from $P(Tails/Bias_1 = Bias_2 = c) = c$ to the first thirder principle on p.18: $Tails_1 = Tails \vee (Second \,\&\, Heads_2)$; since $P(Tails/Bias_1 = Bias_2 = c) = c$, it follows that $P(Tails_1/Bias_1 = c) = c + c/(c+1) \cdot (1-c) = 2c/(1-c)$.

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.[16] In particular, we assumed that $P(First/Bias_1 = c) = 1/(c+1)$. This is in line with thirding, but not with halfing. On the halfer account, $P(First/Bias_1 = c) = 1 - c/2$: 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

$$V(\text{First} \,\&\, Bias_2 = c \,\&\, Bias_1 = b) = 4b - 3bc.$$

The reasoning was that under the given circumstances, the driver's probability for reaching the desolate area is $1 - b$, 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 $First \,\&\, Bias_2 = c \,\&\, Bias = b$, equals b. This happens to be correct according to thirding. On the halfer account, $P(Tails_1/First \,\&\, Bias_1 = b) = b/(2-b)$. If the driver obeys halfing, her probability for reaching the desolate area on the given assumption is therefore not $1 - b$, but

16 To my knowledge, this presupposition is universally shared in the literature on the absentminded driver.

$1 - b/(2 - b)$. Similarly, her probability for getting home, given $Bias_1 = b$ and $Bias_1 = c$, is $b/(2 - b) \cdot (1 - c)$, for reaching the motel it is $b/(2 - b) \cdot c$. The utility 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

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

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

Now let's reconsider the conditional expected utilities. Conditional on $Bias = b$, the uniformity assumption makes it certain that $Bias_1 = Bias_2 = b$. So

$$\begin{aligned} CEU(Bias = b) &= P(First/Bias_1 = b)V(First \ \& \ Bias_1 = Bias_2 = b) \\ &\quad + P(Second/Bias_1 = b)V(Second \ \& \ Bias_2 = b) \\ &= \begin{cases} (1 - b/2) \cdot \frac{4b - 3b^2}{2 - b} + b/2 \cdot (4 - 3b) = 4b - 3b^2 & \text{on halfing,} \\ \frac{1}{b+1} \cdot (4b - 3b^2) + \frac{b}{b+1} \cdot (4 - 3b) = \frac{8b - 6b^2}{b+1} & \text{on thirding.} \end{cases} \end{aligned}$$

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 halfing, 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 halfing and thirding may be regarded as consequences of two more general perspectives on rational belief. From this perspective, halfing is motivated by doxastic conservatism: by the idea that rational agents should only revise their attitude towards a proposition if they receive information which is relevant to that proposition (at least if the proposition is certain not to change its truth-value over time). Thirding, on the other hand, is motivated by a simple form of evidentialism: 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.

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 halfing 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. Thirding 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 $P(Home) = P(Home/First)P(First) + P(Home/Second)P(Second) = b(1-b) \cdot 1/(b+1) + (1-b) \cdot b/(b+1) = 2(b-b^2)/(b+1)$, 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 thirding, 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 the 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 these.

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 halfing.

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 might work 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. We may imagine that a little coin is tossed inside the driver's head to determine whether she will continue, with the coin's bias c being set by the driver's equilibrium probability that she will continue: $P(\mathit{Tails}) = c$. 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 the perspective of epistemic conservatism, this is odd. Why should the driver obey thirding about the coin in her head, but halfing 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 halfing, 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 $P(\mathit{Continue}) = 2/3$. 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 epistemic conservatism, her new credence in Tails_1 is still 2/3, as is her new credence in Tails_2 conditional on Tails_1. Moreover, the latter credence is not sensitive to whether she is at the first exit or the second: $P(\mathit{Tails}_2/\mathit{Tails}_1) = P(\mathit{Tails}_2/\mathit{Tails}_1 \,\&\, \mathit{First})$. This means that the driver fully obeys halfing. Her credence in *Tails*, given by the halfer equation $P(\mathit{Tails}/\mathit{Bias}_1 = \mathit{Bias}_2 = c) = (c + c^2)/2$, is now 5/9. But then her credence is not in equilibrium: at $P(\mathit{Continue}) = 5/9$ and $P(\mathit{Continue}_2) = 2/3$, continuing has a higher expected utility than leaving. In the process of deliberation, the driver's probability for *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 $P(\mathit{Continue}) = 2/3$. And 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 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 $P(\mathit{Bias}{=}2/3) = 1$. As before, upon reaching an exit, the conservative belief update will result in a halfer distribution, with

$P(\textit{Continue}) = 5/9$. But *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.

Now let's turn to our final task: to figure out the equilibrium in the coin scenario if the driver obeys halfing.

7 Useless coins

If the driver is a halfer, what will she do with her coins? We already know the corrected utilities:

$$V(\text{First } \& \ \textit{Bias}_2{=}c \ \& \ \textit{Bias}_1{=}b) = (4b - 3bc)/(2 - b);$$
$$V(\text{Second } \& \ \textit{Bias}{=}b) = 4 - 3b.$$

As for the probabilities, we figured out that $P(\textit{First}/\textit{Bias}_1{=}c) = 1 - c/2$. By uniformity, we can substitute $\textit{Bias}_2$ for $\textit{Bias}_1$. Then $P(\textit{First} \ \& \ \textit{Bias}_2 = c) = P(\textit{First}/\textit{Bias}_2 = c)P(\textit{Bias}_2{=}c) = P(\textit{Bias}_2{=}c)(1 - c/2)$. So

$$P(\text{First } \& \ \textit{Bias}_2{=}c) = P(\textit{Bias}_2{=}c)(1 - c/2);$$
$$P(\text{Second}) = \sum_{c \in B} P(\textit{Bias}_2{=}c)c/2.$$

The resulting expected utilities are

$$EU(\textit{Bias}{=}b) = \sum_{c \in B} P(\textit{Bias}_2{=}c)\frac{3bc^2 + 3b^2c - 20bc + 8c + 8b}{4 - 2b}.$$

(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 $\textit{Bias}_2 = 2/3$, the expected utility of $\textit{Bias} = 2/3$ is 4/3, but the expected utility of $\textit{Bias} = 1$ 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,

$P(Bias_1 = c/First) = P(Bias = c)$. By halfing, $P(Tails_1/First \,\&\, Bias_1 = c) = c/(2-c)$. Hence $P(Tails_1/First) = \sum_{c\in B} P(Bias = c)c/(2-c)$. Together with the symmetry assumption, it follows that $P(First) = 1/(e+1)$, where $e = \sum_{c\in B} P(c)c/(2-c)$. The equilibrium probabilities of the states are

$$P(First \,\&\, Bias_2 = c) = P(Bias = c)/(e+1);$$
$$P(Second) = e/(e+1).$$

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.[17]

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 than 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.[18]

17 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.

18 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.

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

$$EU_C(A) = \sum_{S \in W} \sum_{x \in [0,1]} x \cdot P(Ch(S/A) = x) \cdot V(S),$$

where W is a partition of possibilities whose members have uniform utility (meaning that for any $S \in W$ and any proposition X compatible with S, $V(S \,\&\, X) = V(S)$), and $Ch(S/A) = x$ is the proposition that the objective chance of S, conditional on A, equals x. But this is only plausible if $P(S/A \,\&\, Ch(S/A) = x) = x$, which is false for the absentminded driver if she obeys halfing.

Admittedly, it is somewhat surprising that the driver should assign credence 5/9 to tails if she knows the objective chance is 2/3. The driver thereby seems to violate a form of the "Principal Principle" linking rational credence and objective chance (see [**?**]).[19] 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 $P(\mathit{Tails}_1)$ and $P(\mathit{Tails})$ 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 $\mathit{Tails}_1 = \mathit{Tails} \vee (\mathit{Second} \,\&\, \mathit{Heads}_2)$. So if $P(\mathit{Second} \,\&\, \mathit{Heads}_2) > 0$, then 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. I have investigated the driver's predicament from all four combinations of halfing and thirding with evidential and causal decision theory. Two combinations – thirding with causal decision theory and halfing with evidential decision theory – recommend the apparently optimal choice: the coin with bias 2/3. This result 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 halfing 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 optimal choice given the evidentialist, ahistorical perspective that motivates thirding.

19 In fact, the driver does not violate the Principle as formulated in [**?**], 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.

Another interesting fact about the absentminded driver is that the driver's rational degrees of belief cannot fully match the known objective chances. I have argued that this casts doubt on some popular ideas in game theory and decision theory. Finally, we have 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

Terry Horgan and Anna Mahtani [forthcoming]: "Generalized Conditionalization and the Sleeping Beauty Problem". *Erkenntnis*: –

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 [1981]: "Causal Decision Theory". *Australasian Journal of Philosophy*, 59: 5–30. Reprinted in [**?**]

— [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

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

Leonard Savage [1954]: *The Foundations of Statistics*. New York. Wiley

Wolfgang Schwarz [2012]: "Changing Minds in a Changing World". *Philosophical Studies*, 159: 219–239

Brian Skyrms [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