De Se Belief and Rational Choice

James R. Shaw
Draft, please don't cite

1 The Waking Game

Several philosophers have argued that *de se* beliefs—one's thoughts about oneself in a characteristically 'first-personal' way—have special features that set them apart from other kinds of belief. Frege famously seemed to argue that everyone's thoughts about themselves are distinct, and unshareable.[1] Other philosophers have argued that some *de se* beliefs require a refinement of attitudinal content or severing antecedently plausible connections between the objects of belief and belief states.[2]

These claims raise further questions about whether the peculiarities of *de se* belief require special adjustments to theories in which such beliefs may play a role: for example, in the compositional semantics of attitude reports, accounts of assertoric content, and theories of rational belief change. The Sleeping Beauty puzzle raised in Elga (2000) has been used to argue that the latter theories of rational belief change do require such adjustments.

> **Sleeping Beauty**. In an experiment, Beauty is put to sleep Sunday night by scientists until Wednesday. She will be woken up for certain on Monday and administered a drug to forget that waking. If and only if a fair coin tossed during the experiment lands tails she will be woken up again on Tuesday.

Beauty's predicament raises the following question: given that she knows how the experiment works, what ought she think is the likelihood of the coin's landing heads after she wakes at some point during the experiment? Two camps, "halfers" and "thirders", take the answer to be $\frac{1}{2}$ and $\frac{1}{3}$ respectively. A guiding halfer intuition is that before the experiment, Beauty's belief that a fair coin lands heads should be $\frac{1}{2}$ and that upon waking, no new information of relevance to the toss has been obtained. A guiding thirder intuition is that in repeated iterations of the experiment only one third of the wakings would be wakings in which the coin landed heads. I won't rehearse the array of more sophisticated arguments for each side here.[3] What's important is that the ways we use to flesh out either of these answers tend to have dramatic implications for how we should adjust our frameworks for rational belief change, or the principles governing them, to accommodate *de se* beliefs.

In a similar spirit, I'm going to use some variants of Sleeping Beauty to argue that theories of rational choice, as supplied in decision theory, require special changes to accommodate *de se* beliefs over and above those required from our theories of rational belief change. Questions about rational choice are, of course, closely connected to questions about rational belief since what it is rational to do (at least on one construal) depends not simply on what one believes, but what one *ought* to believe. I'll eventually be exploiting this connection between rational belief and rational choice in examining how *de se* beliefs may complicate decision theoretic frameworks.

For now, though, I want to provisionally *assume* the applicability of some standard frameworks for rational choice because such frameworks hold the promise of supplying us with an easy method for determining answers to the question *originally* posed by Sleeping Beauty: how our beliefs should change in cases of *de se* ignorance. Decision theory tells us how to get from rational beliefs and values to rational choices. Consequently, if we can present a case where it is clear what values an agent has, and what choices they should make, we should be able to 'work backwards' to determine what beliefs they should hold.

The following embellishment of the Sleeping Beauty scenario is designed to allow us to do just this. We suppose that Beauty gets various payoffs depending on what actions she performs during her experiment, and adjust those payoffs so that different answers to the question "what ought Beauty believe" supply us, though decision theory, with different actions she should perform. If we have intuitions about what Beauty should do in these cases, these will privilege one of the competing views about what Beauty ought to believe. Here's a generic form such an elaboration of the case might take.

> **The Waking Game**. Scientists put Beauty in an empty white room on Sunday with two buttons labeled "Left" and "Right". Beauty is put to sleep Sunday night for n days. If the toss of a fair coin early in the experiment lands tails she will woken up for n days beginning Monday. Otherwise she will be woken up only Monday and put back to sleep for $n-1$ days. Each time she is woken, she is given the opportunity to push the left or right buttons, and then will be administered a drug to forget the waking before being put back to sleep. Beauty is given certain payoffs in dollars depending on which buttons she pushes at which times, and is broached of the payoff structure at the outset of the experiment.

In part to simplify the application of decision theoretic frameworks, and in part to get the case to draw apart halfer and thirder views, I want to add two features to the scenario Beauty now faces.

> *Randomizing Prohibited*: Mixed strategies—where, for example, Beauty flips a coin to decide whether to push Left or Right—will unduly complicate the cases I want to consider. So suppose that scientists have prohibited such randomizing. They will allow Beauty's

memories to return after the experiment at which point they will administer a polygraph and ask her if she has attempted to randomize her choices. If she fails the test scientists will kill Beauty's pet dog—a consequence with boundless negative utility. Beauty knows she is very likely to fail the test if she tries to randomize her choices.

Previous Runs: Suppose Beauty has seen the experiment performed innumerable times before on late night reality television. In virtually all trials, when the coin landed tails, test subjects pushed the *same* button every day. Scientists have hypothesized that this is the case because erasing people's memories of previous wakings ensures that they are in the same state relevant to the determination of their choice of button-pushings each day.

Note, *Randomizing Prohibited* does not forbid Beauty from simply "choosing arbitrarily." We can suppose such arbitrary choices to be stable across tails wakings for prior test subjects in *Previous Runs*.

Now, to fill out the details of the Waking Game, suppose the number of days Beauty will wake if the coin lands tails is four and payoffs are given as follows.

Version 1

Payoffs on Heads:	\$400	if Left
	\$200	if Right
Payoffs on Tails:	\$100	if Left every day
	\$200	if Right every day
	\$200	if Left on Monday and Right another day
	\$100	if Right on Monday and Left another day

The payoffs for heads are straightforward, while the payoffs on tails are a little more complex. Assuming the coin lands tails, then if Beauty pushes Left every day she makes \$100, and if she pushes Right every day she gets \$200. If, however, she ever changes which button she pushes, then the payoff is determined by her push on the first day. If she pushed Left on that day then she makes \$200, and if she pushed Right then she makes \$100.

First, let's ask a relatively simple question: what ought Beauty to plan to do Sunday? It should be clear given *Randomizing Prohibited* that Beauty has only two options. Since Beauty can't randomize and since all the wakings are indistinguishable, she can only plan to push Left upon waking, or plan to push Right upon waking.

It should be clear that if Beauty plans to push Left and will succeed in doing so upon waking, she can expect an average payoff of \$250: about half of the time she'll get \$400, and about half of the time she'll get \$100. On the other hand, if she plans to, and succeeds in, pushing Right when she wakes up, she

will always get $200, whether heads or tails. I take it to be uncontroversial that if Beauty can *reliably* plan to push either Left or Right, she ought to plan to push Left.[4]

But there is a distinct question about what Beauty should do: what should she *actually do upon waking*? It turns out that different views about what Beauty's *de se* beliefs ought to be, when combined with standard decision theoretic frameworks, yield different answers.

When Beauty wakes, there are five scenarios she might be in: either the coin landed heads and it is Monday, or the coin landed tails and it is either Monday, Tuesday, Wednesday, or Thursday. I'll note these options by $MonH$, $MonT$, $TueT$, $WedT$, and $ThuT$ respectively. Standard thirder arguments support the view that Beauty's credence should be distributed as follows:

$$MonH = MonT = TueT = WedT = ThuT = \tfrac{1}{5}$$

On the other hand, standard halfer arguments support the alternative credence distribution:

$$MonH = \tfrac{1}{2},\ MonT = TueT = WedT = ThuT = \tfrac{1}{8}$$

Thirders think that upon waking Beauty should believe that she is most likely in a tails scenario. But if that is the case, Beauty should probably push Right: she most often stands to gain $100 by doing so. Halfers think that Beauty should give equal credence to being in a Tails and in a Heads scenario. But then Beauty should push Left: she stands to gain $200 half of the time by doing so, instead of gaining $100 the other half of the time by going right.

This plays out slightly differently depending on which version of decision theory one actually endorses, but the outcome is the same: thirders should push Right, halfers should push Left. To see this, I'll go through the calculations for halfers and thirders in the context of both causal and evidential decision theory (CDT and EDT respectively).[5,6]

Causal decision theorists think it is important to separate the outcome of one's choice and its causal effects from states of the world which are evidentially related to, but not causal outcomes of one's choice. This stance is relevant to the version of the Waking Game I have set up. To see this, suppose upon waking Beauty is considering what to do *assuming* that it is Tuesday. This raises the question: how should she see the relationship between her choice today, her choice the day before, and her choice on the subsequent two days?

CDT stresses that that even if Beauty's choice, say, to push Left is *evidence* that she has and will push Left on other days, this information shouldn't be factored into her decision *as determined* by her choice to push Left unless she takes her choice on Tuesday to *cause* these outcomes. On the present description, that seems unlikely (especially for choices in the past). Consequently CDT instructs Beauty to *fix* her beliefs about her performance on other days, and decide what to do on the basis of those credences.

Beauty's belief about what she did do and will do on other days should be constrained by *Previous Runs*. She should assign a very high credence to the

claim that she did do and will do the same things on every day other than 'today.' So, roughly, she should distribute most of her credence between two possibilities: between being someone who chose and will choose Left on other days—a 'lefty'—on the one hand, and being someone who chose and will choose Right on other days—a 'righty'. How should she allot her credence between these two claims? It turns out not to matter, since there is a dominance argument for Beauty to push Right on waking. This can be seen from the following computations of expected value of choosing Right 'today' (noted R_T) and choosing Left today (L_T), on the assumption that Beauty is a righty or a lefty and her utilities are linear in dollars.

Version 1, Computations for Thirder + CDT

$$\begin{aligned}
\text{if Lefty:} \quad EV(L_T) &= \frac{1}{5}(400) + \frac{1}{5}(100) + \frac{3}{5}(100) = 160 \\
EV(R_T) &= \frac{1}{5}(200) + \frac{1}{5}(100) + \frac{3}{5}(200) = 180 \\
\text{if Righty:} \quad EV(L_T) &= \frac{1}{5}(400) + \frac{1}{5}(200) + \frac{3}{5}(100) = 180 \\
EV(R_T) &= \frac{1}{5}(200) + \frac{1}{5}(200) + \frac{3}{5}(200) = 200
\end{aligned}$$

Let me go through the second of the four computations in a little more detail to spell out the reasoning. Assuming one is a lefty and a thirder, then there is a $\frac{1}{5}$ chance that the coin has landed heads, in which case pushing Right will get Beauty \$200. There is a $\frac{1}{5}$ chance that it is Monday and the coin landed tails, in which case pushing Right today will get Beauty \$100 (since, being a lefty, she will push Left tomorrow). In the remaining $\frac{3}{5}$, it is Tuesday, Wednesday, or Thursday and Beauty (again, being a lefty) has already pushed Left on Monday, and so Beauty stands to get \$200 if she pushes Right.

Note that whether Beauty thinks she is a lefty or a righty, she stands to gain more by pushing Right than pushing Left. Thus, *regardless of how much credence* she assigns to being either, she ought to push Right.

A similar verdict is reached by EDT. EDT, unlike CDT, allows evidential relations between the outcome of one's choice and states of the world to factor into one's decision as to what to do. In the present circumstances, for example, EDT might allow Beauty, on the assumption that the coin landed tails, to conceive of her choice 'today' as effectively *settling* her choices on other days. So beauty can think of herself, 'today', as choosing whether she is a lefty or a righty. Thus we have the following.

Version 1, Computations for Thirder + EDT

$$\begin{aligned}
EV(L_T) &= \frac{1}{5}(400) + \frac{1}{5}(100) + \frac{3}{5}(100) = 160 \\
EV(R_T) &= \frac{1}{5}(200) + \frac{1}{5}(200) + \frac{3}{5}(200) = 200
\end{aligned}$$

Again, the thirder thinks Beauty should push Right.

As I said before, halfers give a different response, regardless of whether they are causal or evidential decision theorists. CDT yields the following values.

Version 1, Computations for Halfer + CDT

$$\begin{aligned}
\text{Lefty:} \quad EV(L_T) &= \frac{1}{2}(400) + \frac{1}{8}(100) + \frac{3}{8}(100) = 250 \\
EV(R_T) &= \frac{1}{2}(200) + \frac{1}{8}(100) + \frac{3}{8}(200) = 187.5 \\
\text{Righty:} \quad EV(L_T) &= \frac{1}{2}(400) + \frac{1}{8}(200) + \frac{3}{8}(100) = 262.5 \\
EV(R_T) &= \frac{1}{2}(200) + \frac{1}{8}(200) + \frac{3}{8}(200) = 200
\end{aligned}$$

Regardless of whether one is a righty or a lefty, one stands to gain by pushing Left. The EDT thirder gets the same result.

Version 1, Computations for Thirder + EDT

$$\begin{aligned}
EV(L_T) &= \frac{1}{2}(400) + \frac{1}{8}(100) + \frac{3}{8}(100) = 250 \\
EV(R_T) &= \frac{1}{2}(200) + \frac{1}{8}(200) + \frac{3}{8}(200) = 200
\end{aligned}$$

Recall that the original motivation for examining this version of the Waking Game was to 'work backwards' within decision theory from values and rational choices to credences. The idea was that if we had strong intuitions about what we ought to do in a case where thirders and halfers diverged in their recommendations, we could use our intuitions to arbitrate between those views. I think Version 1 of the Waking Game is a case with a fairly intuitive response: the rational thing to do upon waking in the game is to push Left, for the simple reason that it seems one stands to gain by doing so. Before defending this claim, let me set up the argument which reveals how the supposition that Beauty ought to push Left constrains our alternatives.

Consider the following three premises governing the first version of the Waking Game.

(P_1) The rational thing for Beauty to do in Version 1 of the Waking Game is to push Left upon waking.

(P_2) There is a particular credence distribution over $MonH$, $MonT$, ..., $ThuT$ that it is rational for Beauty to have in the Waking Game (and hence the case of Sleeping Beauty) upon waking.

(P_3) If it is rational in the Waking Game to have a particular credence distribution over $MonH$, $MonT$, ..., $ThuT$ and the rational thing to do in Version 1 of the Waking Game upon waking is to do A, then standard decision theory yields the verdict that one should do A given the credences it is rational for Beauty to have upon waking.

I'm using the term "standard decision theory" as a blanket term to cover views like CDT and EDT, and reasonable variants of them which yield results such as I've shown above. Given (P_1), (P_2), and (P_3), we can derive

(C_1) Beauty should not have degree of belief $\frac{1}{3}$ that the coin lands tails in the Waking Game (hence in Sleeping Beauty).

Indeed, if we consider very similar situations with the numbers slightly adjusted, it is easy to see we can strengthen this to

(C_2) Beauty should have degree of belief $\frac{1}{2}$ that the coin lands tails in the Waking Game (hence in Sleeping Beauty).

This is a result which arguably puts pressure on the thirder. The thirder accepts (P_2) and rejects (C_1). Consequently the thirder must deny that it is rational in the Waking game to push Left, or jettison standard decision theory.

Though I think each of the premises in the above argument are plausible, there are grounds for doubting each. Consequently I'd like to briefly examine some issues relevant to each premise in turn.

(P_1): Beauty thinks she ought to aim to be a lefty on Sunday. This is because if she succeeds she'll likely be wealthier for her efforts. But as we've seen she might change her mind when she wakes up since she may, if she is a thirder, think it very likely the coin toss landed tails, which favors pushing Right. To alter her plan, however, seems like a bad idea. The reason for this is not just that it seems rational for Beauty to plan on Sunday to push Left since, after all, plans to act and the acts themselves might diverge in terms of rationality, especially in cases involving *de se* ignorance.[7] Rather, the reason it is rational for Beauty to push Left is simply because pushing Left seems the most reliable way for Beauty to get the most of what she wants. To sharpen intuitions let me alter the case in three ways: by increasing the number of tails cases, changing payoffs, and by iterating the game.

Suppose that scientists have designed an extremely efficient robot to run the experiment which can do multiple wakings per day—up to 1 every few minutes. And suppose Beauty is going into the experiment for an extended period—say about two months. Then there is the possibility that Beauty will wake up to 10,000 times. Suppose this is the case.

Version 2
Number of days woken up if coin lands tails: 9,999

Payoffs on Heads: \$5000 if Left

	\$2	if Right
Payoffs on Tails:	\$1	if Left every day
	\$2	if Right every day
	\$2	if Left on Monday and Right another day
	\$1	if Right on Monday and Left another day

Suppose further, that Beauty will get to go through the experiment about 10 times, where between experiments her memories are restored (so that she always knows which run of the experiment she is in). Halfers who, upon waking, push Left will average about \$25,000. Thirders who, upon waking, push Right will almost always end up with \$20. Why would they push Right? Their calculations look roughly as follows.

Version 2, Computations for Thirder + CDT

$$\begin{aligned}
\text{Lefty:} \quad EV(L_T) &= \tfrac{1}{10{,}000}(5,000) + \tfrac{1}{10{,}000}(1) + \tfrac{9{,}998}{10{,}000}(1) && \approx 1.5 \\
EV(R_T) &= \tfrac{1}{10{,}000}(2) + \tfrac{1}{10{,}000}(1) + \tfrac{9{,}998}{10{,}000}(2) && \approx 2 \\
\text{Righty:} \quad EV(L_T) &= \tfrac{1}{10{,}000}(5,000) + \tfrac{1}{10{,}000}(2) + \tfrac{9{,}998}{10{,}000}(1) && \approx 1.5 \\
EV(R_T) &= \tfrac{1}{10{,}000}(2) + \tfrac{1}{10{,}000}(1) + \tfrac{9{,}998}{10{,}000}(2) && \approx 2
\end{aligned}$$

The computations for EDT yield a similar verdict: the choice the thirder thinks they face is essentially a choice between \$2 and \$1.50.[8] This seems untrue to the case. Following the halfer who pushes Right on these grounds would, I think, be highly irrational.

The main point of the foregoing discussion has *not* been to further contrast the views of the thirder and halfer. Rather, I am merely trying to give the case for thinking that (P_1) holds: that when Beauty wakes up in Version 1 of the Waking Game, *the rational thing for her to do* is push Left.

How could one argue that the intuitions I have been trying to draw out are illusory? One response comes from consideration of the contrasting verdicts of CDT and EDT as regards Newcomb's puzzle. I don't want to get into the details of this case, since this would take us too far afield. I just want to note that Newcomb's puzzle presents a case where endorsing CDT leaves one less well off than endorsing EDT. This prompts a kind of question addressed to causal decision theorists, a version of which I am raising here for detractors of (P_1). It was succinctly put by David Lewis: "If you're so smart, why ain'cha rich?"[9]

A standard answer on behalf of CDT is to claim that the Newcomb situation is one in which 'rationality is being punished.' It is a controversial matter whether this idea is fruitfully appealed to by a defender of CDT. But it should be uncontroversial that this strategy is of no use in defending a rejection of (P_1). The charge that rationality is being punished in Newcomb's puzzle is made plausible by considering that the payoffs in that choice situation are restructured

based on one's dispositions to act. This might open up the possibility that the way the payoffs are restructured systematically penalizes persons disposed to a particular type of choice—perhaps those disposed to make the rational choice among them. In the Waking Game there is no similar restructuring of options. Payoffs are fixed in the scenario with Beauty's full awareness of them. Since her choice won't alter the payoff structure, but only (by normal means) her payoffs themselves, it seems implausible to suppose that the losses Beauty suffers by pushing Right are the outcome of a circumstance which penalizes rational choices. Rational people, so the story goes, capitalize on payoff structures and probabilities to secure the most of what they want. If the payoffs aren't being restructured, it is hard to see how Beauty's failure to secure more money is purely due to an unfair or unfortunate structure of the game.

There are perhaps other ways to defend the rejection of (P_1), but the standard way of coping with the "why ain'cha rich?" objection seems particularly unmotivated here.

(P_2): Though it might be difficult to arbitrate between the halfer and thirder views, it can feel obvious that at least these parties are debating a genuine question with a unique answer. That is, it can seem that whatever the case is, there is *some* unique credence distribution that Beauty ought to have upon waking in the case of Sleeping Beauty and the Waking Games.[10]

A challenge to this assumption, however, is furnished by Arntzenius (2002). Arntzenius stresses that when the coin lands tails, Beauty will have her beliefs artificially reset on Tuesday to conform to those she had Monday and that this is a highly relevant kind of cognitive mishap, in that it ensures that Beauty violates Bayesian conditionalization. Beauty is aware that she is going to be the subject of such a cognitive malfunction. Consequently, the main question Beauty faces is not what she ought to believe, but how she ought to behave to minimize the negative effects brought on by that malfunction.

Some evidence for Arntzenius' position comes from considering situations in which Beauty, on the assumption that she is a thirder (say) and endorses a particular decision theory, ought to accept bets at odds which apparently violate her credences. Similar problems afflict the halfer view. Arntzenius claims we can explain these situations as the upshot of the view that Beauty's credences in particular propositions in Sleeping Beauty are irrevocably corrupted by the cognitive malfunction she knows she either has or may yet suffer. The best Beauty can do is to consider herself to be somewhere at some point during the experiment, and consider what someone in that situation stands to gain or lose by adopting various plans.

Arntzenius sums up his position as follows: "[For Beauty not] to have a definite degree of belief in heads might be strange, but it might be the best that she can do given the forced irrationality that is inflicted upon her... The main moral of [Sleeping Beauty] is that in the face of forced irrational changes in one's degrees of belief one might do best simply to jettison them altogether."[11] It is unclear whether the examples Arntzenius supplies are enough to establish his position, but the idea is certainly one that can look more appealing after

considering variants of the Waking Game. I'll return to consider Arntzenius' suggestion again in §3.

(P_3): The idea that standard decision theories ought to be abandoned is one that has been advocated recently by Egan (2007) in response to a raft of examples where the seemingly rational things to do are not systematically reflected in the exclusive application of either EDT or CDT. The problem raised by the Waking Game for standard decision theory, however, is of a very particular variety. Egan's puzzles, if his analyses are accepted, seem to show that in some cases EDT wins out while in others CDT does. This seems to point to something like a hybrid view, or at least something in the neighborhood of standard decision theory. The first version of the Waking Game, however, seems to show that if one is a thirder, both EDT and CDT, and anything suitably similar to them, will have to go by the board. Consequently, the way in which (P_3) fails, if it is rejected here, will arguably be in a more dramatic way than Egan proposes.

There is more to be said about these premises, but let me recapitulate what I take to be some morals so far. The thirder accepts (P_2) and rejects (C_1). Consequently, she must reject either (P_1) or (P_3). Barring an account which overturns intuitions about maximizing gains that I have drawn on, it is extremely difficult to reject (P_1). Thus, without such an account, the thirder should seriously consider rejecting (P_3), and hence abandoning standard decision theory as a completely general account of what it is rational to do given what one believes.

This might in turn seem to apply a great deal of pressure to the thirder view. After all, standard decision theory is not merely a theory with 'good fit' to the set of data given by other uncontroversially rational choices. It is also a theory whose structure seems intuitively tailored to track rational decision making. What's more, the discussion so far may lead one to believe that thirders *alone* are in a bind. I suspect this apperance is illusory. It turns out that the halfer, and indeed any theorist who claims Beauty ought to have a particular credence distribution in the case of Sleeping Beauty, may ultimately face a challenge similar to the thirder. To see this, I'll have to introduce some new complications into the Waking Game.

2 Subjectively Distinguishable States

It is an important assumption of Sleeping Beauty, and my original Waking Games, that Beauty's wakings are subjectively indistinguishable. If things were otherwise, Beauty's rational beliefs—*de se* and otherwise—might change in dramatic ways. Sometimes theorists consider cases in which Beauty is capable of distinguishing her wakings, and in which it might be more clear what she ought to believe. The idea is that these situations can be used to try to glean information about what Beauty ought to believe in the original Sleeping Beauty case by way of analogy. Something like this strategy is adopted, for example, in Titelbaum (2008).

The cases I'd presently like to examine are variants of this kind. The examples will get quite complicated, but I believe this might be necessary to relieve thirders of the burdens of §1.

> **Modified Waking Game**. As in the original Waking Game, except that each day Beauty will be placed in a different colored room from among n options. The colors of the rooms are very easy to distinguish (i.e. red, green, etc.). She'll be placed in one of the n rooms each day. At the time scientists flip the coin to decide how many times beauty will be awakened, they will also roll an $n!$-sided die. Each $0 < i \leq n!$ corresponds to a permutation of the rooms that Beauty may be placed in. Thus, on tails, beauty is sure to wake to each of the n rooms at least once, whereas on heads, there is only a $1/n$ chance of her waking in any given room. Beauty is broached of these details and the colors in advance.

Let the number of days and payoffs be as before.

> **Modified Waking Game, Version 1**
> Number of days woken up if coin lands tails: 4
> Room Colors: Red, White, Green, Blue
>
> | | | |
> |---|---|---|
> | Payoffs on Heads: | \$400 | if Left |
> | | \$200 | if Right |
> | Payoffs on Tails: | \$100 | if Left every day |
> | | \$200 | if Right every day |
> | | \$200 | if Left on Monday and Right another day |
> | | \$100 | if Right on Monday and Left another day |

What should Beauty plan to do on Sunday, again provided she cannot randomize and is sure to execute her plan? The availability of colors to coordinate her decisions now allow for five equivalence classes of plans (equivalent under the relation of equal expected payoff) based on how many colored rooms she chooses to push Left in upon waking. The optimal plan is to always push Left except on one color.

Plan	**Approx. Expected Payoffs**
Left on all colors	\$250
Left on 3 colors, Right on 1	\$263
Left on 2 colors, Right on 2	\$225
Left on 1 colors, Right on 3	\$188
Right on all colors	\$200

Thus, adding subjectively distinguishable states allows for more elaborate, coordinated strategies with higher payoffs than in cases with subjective indistinguishability.

I want to contrast two different scenarios where these more elaborate strategies *may* or *may not* be available to Beauty, not because she is unaware of what circumstance she is in, but because of facts about her own psychology. To bring out this contrast we'll need to prevent Beauty from forming plans on Sunday.

> *In-Game Explanations*: Beauty knows she is in some Waking Game on Sunday, but doesn't know the exact rules, number of wakings, payoffs, and so forth. She'll be told them every day that she wakes by a recording over a loud speaker right after she gets up.

Again, let's suppose that *Randomizing Prohibited* and something analogous to *Previous Runs* hold. Consider the following circumstance.

> **Case 1:** Beauty wakes and hears the rules of Version 1 of the Modified Waking Game (with the colors of the rooms specified). She opens her eyes to find herself in a red room. She reasons as follows: "It would be ideal if I could get myself in a position to push Left in all rooms but one. Unfortunately, if the coin landed tails I'll have to coordinate with myself in other wakings—but I can't. Perhaps I could effectively coordinate by just picking an arbitrary color to be the "Right pushing" room right now, executing the corresponding plan, and hoping that I will adopt the same plan on other days. I'd probably pick red if that was what I ought to do. But the problem is *I've already seen that I'm in a red room*, and I'm a very suggestible person. Seeing the red room will doubtless systematically influence my decision as to which "arbitrary" color I choose in detrimental ways: it will make me very likely to pick the color of the room I'm in. This makes it highly likely that if I pick a color now and execute the corresponding plan (and the coin landed tails) I won't coordinate with myself in the right way: I'll be liable to choose to push Right every day. And I can't force myself to choose a color other than red now—then I'll just think the same thing every other day and always end up pushing Left. No, I can't capitalize on the different colors of the room to pry apart my choices on different days. I'm better off just making a decision independently of color considerations."

In this case Beauty has some very sophisticated views about her own psychological states. She thinks facts about those states put her in a bad position to coordinate her choices in the right way by capitalizing on subjective distinguishability. We can suppose, for the sake of the example, that Beauty has good evidence for this, and is in fact *right*. Having woken up and seen the color of the room she is in, planning to push Right on some one color is not a decision which will generally lead to her having coordinated her choices in the right way. Though lamentable, it appears it is best for Beauty to push Left in such circumstances. Her psychology prevents her from capitalizing on the benefits of subjective distinguishability, effectively putting her in the circumstance of the unmodified Waking Game.

In a contrasting case, though, we can suppose things had gone ever so slightly differently.

> **Case 2:** Just as in Case 1, except that before Beauty opens her eyes she checks herself: "What would be ideal is if I could get myself in a position to push Left in all rooms but one. If I open my eyes now, though, I might be forgoing a great option: pick a single arbitrary color before opening my eyes to single out a room in which I'll push Right. Since I'm likely to keep my eyes closed and reason this way on other days, and since (by *Previous Runs*) I'm liable then to pick the same arbitrary color, it will be as if I was able to form a plan on Sunday to push Right only once. Great! I choose red." Beauty opens her eyes to find herself in the red room.

Beauty's reasoning appears sound. Now, it seems, Beauty is in a great position to push Right.

Let me articulate another argument, analogous in structure to the one I gave in §1, which is suggested by the foregoing examples. It begins by tugging on the same intuitions concerning what it is rational for Beauty to do.

(P'_1) The rational thing for Beauty to do in Case 1 is push Left, and the rational thing for her to do in Case 2 is push Right.

As with the corresponding premise of §1, we can support (P'_1) by changing payoffs, increasing the duration of the experiment, and iterating games (with Beauty becoming aware of which trail she is in by having her memories of preceding trials return). As before, detractors from (P'_1) will systematically face high losses with apparently no explanation for why this is compatible with their choices being rational.

A second premise concerns what sort of credence distribution Beauty ought to have upon waking. Unlike before, it will be helpful to constrain the credence distributions which seem reasonable. In addition to the propositions *MonH*, *MonT*, *TueT*, *WedT*, and *ThuT* it will be useful to consider two (*de se*) propositions stating that Beauty will, if she wakes on several other days during the experiment, choose the same thing each day. Let *Righty* be the proposition that Beauty chooses Right every other day (if she is given the option), and *Lefty* be the proposition that Beauty chooses Left every other day. Then we can define the following notion of a reasonable credence distribution for my variants of the Modified Waking Game.

> A credence distribution $\mathcal{C}$ is *respectable* if the following three conditions hold.
>
> (i) $\mathcal{C}(MonH) \in \{\frac{1}{2}, \frac{1}{5}\}$
>
> (ii) $\mathcal{C}(Righty) + \mathcal{C}(Lefty) \approx 1$.
>
> (iii) $\mathcal{C}(MonT) \approx \mathcal{C}(TueT) \approx \mathcal{C}(WedT) \approx \mathcal{C}(ThuT)$.

Respectable credences are the ones it seems rational for Beauty to have in my variants of the Modified Waking Game.

(P'_2) There is some respectable credence distribution which it is rational in Case 1 for Beauty to have. Likewise for Case 2.

Let's go over the reasoning for each clause. (i) should hold because something like halfer or thirder reasoning should apply in the cases. Note that this does not mean, for example, that halfers *are committed to the claim that Beauty ought to believe to degree* $\frac{1}{2}$ *that the coin landed tails in these cases*. Halfers may claim that the changes brought about by introducing subjectively distinguishable wakings are relevant to what Beauty ought to believe. Ditto for thirders. Nonetheless, if one is a halfer or a thirder, one will likely expect it be rational for Beauty to have *some* credence in heads in the cases given, and that $\frac{1}{2}$ and $\frac{1}{5}$ are the best options. It seems hard to motivate other values.

(ii) should hold in Case 1 because Beauty essentially takes herself to be in the unmodified Waking Game, where the same assumption seems rational. Even if Beauty decides to act based on an attempt to use colors to coordinate her choices, I have assumed that she knows she would end up either pushing Left every day or Right every day. In Case 2, Beauty again may either ignore the coloring, in which case she should have the same beliefs as the unmodified Waking Game, but it is more likely that she will follow through on her plan, in which case she will push Right, but *only* in the red room—i.e. only 'today'. Thus she will push Left every other day and (ii) holds.

(iii) could be slightly more controversial. It is an application of a kind of indifference principle, which says that Beauty doesn't think that it is much more likely, say, for it to be Monday while the coin landed Tails, than for it to be Tuesday while the coin landed Tails. General indifference principles are not always easy to defend, but this particular application seems to be justified on intuitive grounds.[12]

Not only should Beauty's credence distribution in Cases 1 and 2 be respectable, but it also seems they should be identical, perhaps admitting for differences in how she proportions her beliefs between *Lefty* and *Righty*, and differences in her beliefs about her psychology, say. I'll ignore the latter divergences since they are irrelevant to the argument to follow.

(P'_3) Beauty should have the *same* respectable credence distribution in both Case 1 and 2, up to divergent relative credences between $\mathcal{C}(\mathit{Righty})$ and $\mathcal{C}(\mathit{Lefty})$.

This is because although Beauty has different evidence in Case 1 and Case 2 about what she is presently doing, what plans will succeed, and perhaps what she has already or may yet do, none of these differences are plausibly pertinent to $\mathcal{C}(\mathit{MonH})$ or to any other claims about what day it is or whether the coin landed tails.

So the premises (P'_1)–(P'_3) seem just as plausible as (P_1)–(P_3). However, they yield the following surprising conclusion.

(C') Supplied with values and rational credences, standard decision theory will not always compute the rational thing to do.

The argument is simple. Suppose (P'_2), (P'_3), and (C') hold. Suppose further that $\mathcal{C}(\mathit{MonH})=\frac{1}{2}$ in both Case 1 and Case 2. If the former holds, then both

EDT and CDT yield the result that Beauty should push Left in Case 2. Indeed, choosing Left dominates choosing Right in expected value: there is a $\frac{1}{2}$ chance of Beauty getting \$400 over \$200 by pushing Left, but (in the 'best case' scenario) only $\frac{1}{2}$ chance of getting \$200 over \$100 if she pushes Right. The resulting verdict contradicts (P'_1).

Suppose, instead, that $\mathcal{C}(\mathit{MonH})=\frac{1}{5}$ in both Case 1 and Case 2. Then both EDT and CDT direct Beauty to push Right in Case 1. Again, the choice dominates: at least $\frac{3}{5}$ of the time she stands to get \$200 over \$100 by pushing Right and faces a mere $\frac{1}{5}$ chance of forgoing \$400 for \$200. Again this contradicts (P'_1).

On the assumption of the premises, standard decision theory systematically produces the wrong results.

My argument so far has depended on crucial alterations to the original Sleeping Beauty case which animated the halfer and thirder positions—in particular, it has depended on the introduction of subjectively distinguishable wakings. However, it is easy to see how the new argument is indirectly relevant to both of their positions. We have a pair of cases where, on plausible assumptions about how credences should be alloted in those cases, decision theory falters. This shows that the result of §1 is not really a *special* problem for the thirder.

3 Conclusion

Let me elaborate a little on what conclusions I think we should draw from the foregoing arguments. First, let's return to the tension between (P'_1)–(P'_3) and standard decision theories.

(P'_1) The rational thing for Beauty to do in Case 1 is push Left, and the rational thing for her to do in Case 2 is push Right.

(P'_2) There is some respectable credence distribution which it is rational in Case 1 for Beauty to have. Likewise for Case 2.

(P'_3) Beauty should have the *same* respectable credence distribution in both Case 1 and 2, up to divergent relative credences between $\mathcal{C}(\mathit{Righty})$ and $\mathcal{C}(\mathit{Lefty})$.

If (P'_2) is true, it is hard to imagine that (P'_3) could be false. So barring grounds for rejecting (P'_1), the examples of §2 point to a tension between the idea that Beauty ought to have specific credences in Sleeping Beauty-like cases and standard decision theory. It would, however, be very difficult to reject (P'_2) to rescue universal application of standard decision theories. The reason is that if we want to preserve the intuition—given in (P'_1)—that it is *the* rational thing to do for Beauty to perform certain actions, we will want standard decision theory to never direct Beauty to do otherwise. But if we reject (P'_2), it seems like it should be rationally *permissible* for Beauty to have *any* credence distribution over the relevant propositions. Thus Beauty will not be rationally criticizable

if she selects a credence distribution on which standard decision theory directs her to, say, push Right in Case 1. This violates (P'_1), at least on its strongest reading. Even if Beauty is not permitted to rationally select her credences, then she will have *no* credences. Thus standard decision theory will not fault her for making either choice in Cases 1 or 2. Again this violates (P'_1).

Consequently, the cases of §2, and in retrospect the cases of §1, most strongly suggest accepting (C')—that is, they suggest jettisoning standard decision theory as the theory which uniformly yields results about what it is most rational to do in cases of *de se* ignorance.[13] This conclusion is a conclusion *just* about the implications of integrating *de se* beliefs into our theories of rational choice. Exactly how our frameworks for rational choice should be adjusted is a complicated issue I'll say a little bit more about shortly. In the interim, this conclusion about models of rational choice has some important *additional* implications for theories of rational belief update. Let me say why.

I originally presented the motivation for examining the Waking Game as a way of indirectly getting at the beliefs Beauty ought to have in her various predicaments: we could work backward from intuitions about what Beauty ought to do, along with knowledge of what her values are *through* standard decision theory to figure out what she ought to believe. That strategy has now been shown unreliable, because we've seen that standard decision theory can direct Beauty to do intuitively irrational things in various choice situations, regardless of which reasonable credence distribution she had.

This is significant, because we would normally expect that what one ought to believe in certain scenarios is conceptually tied to what one ought to do in them, and hence exploring rational choice should be a sure-fire way of uncovering rational beliefs. Consider, for example, the strategy of constructing 'dutch book' arguments to undermine particular credence distributions by showing that betting along the lines of those credences may lead one to systematically lose money. Such a strategy is merely a *special case* of the strategy of trying to figure out what Beauty ought to believe by seeing what she ought to do. Indeed, it turns out to be a relatively special case given that it is possible that sometimes one ought not to bet along one's credences, as emphasized by Arntzenius (2002).

Thus the foregoing reflections should cast suspicion on *every member* of a large class of arguments, including those involving dutch books, about what credences Beauty should have which, apparently, are the most fruitful and decisive we could give. This does not eradicate all hope of finding grounds for Beauty to prefer one credence distribution over another. There is always analogical reasoning—examining cases similar to those in which Beauty finds herself where there is no contest as to what Beauty's credences should be and extrapolating from those results to the more problematic cases. There is also the strategy, originally adopted by Elga (2000), of trying to reason to a conclusion in the case of Sleeping Beauty from putatively uncontroversial principles governing rational belief revision. Though such arguments might be available, the Waking Games add to worries that conclusive results will be very hard won.

Moreover, even if we *do* find what credences it is rational for Beauty to have, the Waking Games show that we will still be left with a theoretically distinct

and perhaps more pressing question about what Beauty *ought to do*. Typically we are interested in what one ought to believe *because* this should play a guiding role in action, and not merely for the satisfaction of knowing our credences yield to evidence in just the right way. The Waking Games seem to show that even if Beauty ought to have a certain credences in 'narrow' *de se* propositions, what she ought to do in those cases is computed in a way that swings free of those credences. Is there a way of systematizing her choices in such scenarios? And if so, how?

The answer to the second of these questions is clearly the topic for another investigation, but I would like to briefly say some things to help suggest that an answer to the first question is 'yes'. If the intuitions motivating (P_1) and (P_1') are reliable, then they already point to a strategy for Beauty to adopt. If Beauty knows what kind of situation she is in, regardless of whether she knows exactly where or when she is in it, she can use that information to assess the relative values of various choice strategies which might be adopted by persons in her situation. It seems the most rational choice can be pinpointed as the strategy which, when adopted, tends to yield the highest utility. In this calculation, it is not 'narrow' *de se* propositions concerning the time, e.g., which guide Beauty's choice, but 'broad' *de se* propositions such as *that I am presently in such-and-such a Waking Game*.

We've already seen the basic idea here suggested in the discussion of Arntzenius (2002). When rejecting the claim that Beauty ought to have any particular credence in heads, Arntzenius writes: "... [Beauty's] epistemic state upon waking up is best described by saying that she believes she is in the situation described in the Sleeping Beauty story."[14] Whether or not Beauty actually ought to have beliefs in narrower *de se* propositions, it is the broader ones Arntzenius discusses which apparently should play the decisive role in her choices. If this is true, perhaps it matters less in cases of centered uncertainty exactly where and when you are, than the general nature of your predicament.

Notes

[1]Frege (1956)

[2]See Lewis (1979) and Perry (1979) respectively.

[3]For a sampling, see and Dorr (2002), Hitchcock (2004), and Titelbaum (2008) in addition to the original Elga (2000) for some thirder arguments and Lewis (2001), White (2006), and Meacham (2008) for halfer views.

[4]Of course, she may not be able to reliably do this. This might be, for example, because she anticipates that it would be irrational for her to carry out her plan upon waking. Also, I'm assuming in this case that Beauty has no particular aversion to risk.

[5]Both CDT and EDT have slight variants, so I'm fixing on a particular construal here. I make no claims to exhaustiveness, but I hope the examples chosen are representative.

[6]This claim, if correct, shows that *pace* Briggs (2010) it's not clear that one's choice of causal or evidential decision theory specially privileges a halfer or thirder view.

[7]How could a plan to perform an action in the future be rational, but the future action itself be irrational? Perhaps I *now* have information which I know I may, or will, lack at the time of acting. Given what I know now I should wish my future self to do A. But were I to reason on the basis of only the more limited amount of information had by my future self, it would be more prudent to wish my future self to perform some distinct action B. If

my intention to act in the future is something like a cause of my so-acting, I may plan to A knowing that it will be subjectively irrational to do A in the future due to a loss of information at that time. See Elga (2004) for an example of this very kind of phenomenon which arises as one 'loses' *de se* information over time, without obviously suffering any sort of cognitive malfunction.

[8]In fact, for *any* position but the halfer's one can construct such a game.

[9]See Lewis (1981b) for a discussion of the status of the objection and the standard reply on behalf of CDT.

[10]More carefully: there is some credence distribution or constrained range of distributions over the various conjunctive propositions governing the outcome of the coin toss and the 'present day' that it is irrational for Beauty to deviate from.

[11]Arntzenius (2002), p.61.

[12]In fact, the assumption of (iii) isn't actually needed for the argument to follow, since we can create alternative scenarios where other days than Monday play the special role of determining what payoffs Beauty will get provided she changes her choice on other days. Appealing to (iii), however, will simplify the argument tremendously.

[13]In fact, since the case of ignorance in §2 is arguably not *irreducibly de se*, the arguments may show more: that in complex cases of ignorance like Sleeping Beauty, standard decision theoretic frameworks fail, regardless of whether irreducibly *de se* ignorance is at issue.

[14]Arntzenius (2002) p.61.

References

F. Arntzenius (2002). 'Reflections on Sleeping Beauty'. *Analysis* **62**(1):53–61.

F. Arntzenius (2003). 'Some problems for conditionalization and reflection'. *Journal of Philosophy* **100**(7):356–371.

N. Bostrom (2007). 'Sleeping Beauty and Self-location: A Hybrid Model'. *Synthese* **157**(1):59–78.

D. Bradley (2003). 'Sleeping Beauty: a Note on Dorr's Argument for 1/3'. *Analysis* **63**(3):266–268.

R. Briggs (2010). 'Putting a Value on Beauty'. *Oxford Studies in Epistemology* **3**:3–34.

C. Dorr (2002). 'Sleeping Beauty: in Defense of Elga'. *Analysis* **62**(4):292–96.

A. Egan (2007). 'Some Counterexamples to Causal Decision Theory'. *The Philosophical Review* **116**(1):93–114.

A. Elga (2000). 'Self-Locating Belief and the Sleeping Beauty Problem'. *Analysis* **60**(2):143–147.

A. Elga (2004). 'Defeating Dr. Evil with Self-Locating Belief'. *Philosophy and Phenomenological Research* **69**(2):383–396.

G. Frege (1956). 'The Thought: A Logical Inquiry'. *Mind* **65**(259):289–311.

J. Halpern (2005). 'Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems'. *Oxford Studies in Epistemology* **1**:111–42.

C. Hitchcock (2004). 'Beauty and the Bets'. *Synthese* **139**(3):405–420.

D. Lewis (1979). 'Attitudes de dicto and de se'. *The Philosophical Review* **88**(4):513–45.

D. Lewis (1981a). 'Causal decision theory'. *Australasian Journal of Philosophy* **59**(1):5–30.

D. Lewis (1981b). 'Why Ain'cha Rich?'. *Noûs* **15**:377–380.

D. Lewis (2001). 'Sleeping Beauty: Reply to Elga'. *Analysis* **61**(3):171–176.

C. J. Meacham (2008). 'Sleeping Beauty and the Dynamics of De Se Beliefs'. *Philosophical Studies* **138**(2):245–269.

J. Perry (1979). 'The Problem of the Essential Indexical'. *Noûs* **13**:3–21.

M. Titelbaum (2008). 'The Relevance of Self-locating Beliefs'. *The Philosophical Review* **117**(4):555.

R. White (2006). 'The Generalized Sleeping Beauty Problem: a Challenge for Thirders'. *Analysis* **66**(2):114–119.