Two-Dimensional Deference

J. DMITRI GALLOW*

Abstract

Principles of expert deference say that you should align your credences with those of an expert. This expert could be your doctor, the objective chances, or your future self, after you've learnt something new. These kinds of principles face difficulties in cases in which you are uncertain of the truth-conditions of the thoughts in which you invest credence, as well as cases in which the thoughts have different truth-conditions for you and the expert. For instance, you shouldn't defer to your doctor by aligning your credence in the *de se* thought 'I am sick' with the doctor's credence in that same *de se* thought. Nor should you defer to the objective chances by setting your credence in the thought 'The actual winner wins' equal to the objective chance that the actual winner wins. Here, I generalise principles of expert deference to handle these kinds of problem cases. This generalisation has surprising consequences for Adam Elga's *Sleeping Beauty* puzzle.

[This paper is long, but the casual reader need only read §§1–2, pp. 1–12]

1 | INTRODUCTION

I have opinions about how likely various things are. I think that it's unlikely to snow in Dubai, that a flipped coin is just as likely to land heads as tails, and that global leadership is unlikely to take serious steps to address climate change. Others have these kinds of opinions, too. For instance, weather reporters have opinions about how likely it is to snow in Dubai; the objective chances have opinions about how likely it is that a flipped coin will land heads;[1] and my future, better informed, selves have opinions about how likely global leadership is to take serious steps to address climate change. Principles of expert deference say that I should treat the opinions of these experts—the weather reporters, the objective chances, or my future, better informed self—as a particularly strong kind of evidence. Roughly, given that one of these experts is n% sure of 'p', I too should be n% sure of 'p'—that is to say, my own subjective degree of confidence in 'p', or my *credence* in 'p', should be n%.[2]

Draft of February 16, 2021. Please do not cite without permission.
Comments appreciated. ✉: dmitri.gallow@gmail.com

* Thanks to Daniel Drucker, Barry Loewer, James Shaw, and Juhani Yli-Vakkuri for helpful conversations on this material. Thanks also to audiences at the Rutgers Foundations of Probability Seminar, the Dianoia Institute of Philosophy, and the University of Helsinki Formal Epistemology Workshop.

1. I speak about the objective chances as though they were people with probabilistic opinions. This is a satisfying metaphor, though I hope it's clear that I do not take the metaphor at all seriously.

2. I will enclose a sentence in single quotation marks to form a name for the thought which corresponds to that sentence. When using a schematic variable, p, I will use "p" for the result of writing ' ', followed by the substituend of 'p', followed by ' '. (That is: I'll sloppily use single quotation marks for quasi-quotation.)

Principles like these are plausible, but their plausibility depends upon an implicit restriction on the thoughts which we are permitted to substitute for '*p*'. It is plausible that, when it comes to matters of my health, I should defer to my doctor. Let '*p*' be the thought 'I am sick'. Presumably, '*p*' concerns my health. But, just because my doctor is *n*% confident in 'I am sick', this doesn't give me a reason to be *n*% confident in 'I am sick'. After all, when my doctor entertains the thought 'I am sick', that thought will be true just in case *she* is sick, and when I entertain the thought 'I am sick', that thought will be true just in case *I* am sick. Or let '*p*' be 'It is now Monday'. On Sunday, I can know that one of my future, better informed selves will be nearly certain of 'It is now Monday'. But it doesn't follow that, on Sunday, I should be nearly 100% sure of 'It is now Monday'.

These substituends for '*p*' are about matters *de se et nunc*—that is, they concern who I am, and where I am located in space and time. But there are also concerns for principles of expert deference when '*p*' has nothing to do with self-location. Consider the following example, from Hawthorne & Lasonen-Aarnio (2009): tomorrow, we will draw one of 100 names from an urn, and the person whose name is drawn will win a prize. Before the draw takes place, we introduce the name 'Lucky' for the person whose name is actually drawn. We don't yet know the truth-conditional content of 'Lucky wins'. If Sabeen actually wins, then the truth-conditional content of 'Lucky wins' is that Sabeen wins. If Evîn actually wins, then the truth-conditional content of 'Lucky wins' is that Evîn wins. Even though we don't know what the truth-conditions of 'Lucky wins' are, we know for sure that those truth-conditions have a 1% objective chance of being satisfied. *Whoever* Lucky is, they have a 1% chance of winning the prize—same as everyone else. So it appears that objective chance is 1% confident in 'Lucky wins'. But surely *we* should think 'Lucky wins' is nearly 100% likely. It is, after all, *a priori* knowable that, if anybody wins, then Lucky does.[3]

What thoughts like 'I am sick' and 'Lucky wins' have in common is that their truth-conditions depend upon who entertains them, when and where they entertain them, or what the world is like when they entertain them. If 'I am sick' is entertained by me, then it has the truth-conditional content that [Author] is sick. If it is entertained by my doctor, then it has the truth-conditional content that *she* is sick. If 'Lucky wins' is entertained in a world in which Sabeen wins, then it has the truth-conditional content that Sabeen wins. If it is entertained in a world in which Evîn wins, then it has the truth-conditional content that Evîn wins.

Several authors have modeled thoughts like 'I am sick' and 'Lucky wins' with a *two-dimesional* semantics.[4] As I'll be understanding it here, the first dimension models your *a priori* ignorance of who you are, when and where you are, and what the world is like. The second dimension models how the truth-conditional contents of your thoughts depend upon who you are, when and where you are, and what the world is like. If we confine our attention to thoughts like 'Somebody is sick on August 16, 2020', this second dimension will be uninteresting. Thoughts like this have the same truth-conditions, no matter who, when, or where you are, and no matter what the

3. Similar cases are discussed in Schulz (2011), Titelbaum (2012), Nolan (2016), and Salmón (2019).

4. See, in particular, Kaplan (1978, 1989), Stalnaker (1978), Evans (1979), Davies & Humberstone (1980), Jackson (1998), and Chalmers (2006a,b).

world is like. Principles of expert deference work well when we only concern ourselves with thoughts like these. They face difficulties when we begin to entertain thoughts like 'I am sick' and 'Lucky wins'—thoughts whose truth-conditions vary depending upon who, when, and where you are, or what the world is like.

Here, I will introduce and explore an emendation of principles of expert deference which allows them to deal with these more interesting thoughts. This emendation comes in two parts. Standard principles of expert deference say that you should align your opinions about '*p*' with the expert's opinions about '*p*'. However, when it comes to thoughts like 'I am sick', I should not align my opinion with my doctor's opinions about 'I am sick', but instead with her opinion about some other thought—a *surrogate* for '*p*'. The first part of the emendation concerns this surrogate.

The second part of the emendation will be required to deal with expert deference in cases in which uncertainty about who, when, or where you are has led to uncertainty about the truth-conditions of your thoughts. Roughly, I'll deal with these cases by suggesting that you should only align your opinions with those of the expert *conditional* on who you both are, and when and where you both are located in space and time.

In a slogan, my proposal is this: you should defer to the expert about whether your thoughts are true, given the location at which you entertain them. This proposal will reduce to the familiar principles of expert deference when you only have opinions about boring thoughts like 'Somebody is sick on August 16, 2020'. But, once you have *some* opinions about interesting thoughts like 'I am sick', the emendation will disagree with the familiar principles, even when it comes to the boring thoughts. I'll illustrate this with a discussion of Elga (2000)'s *Sleeping Beauty* puzzle. Lewis (2001) took his principle of chance deference to militate against Elga's 'thirder' solution to that puzzle. However, my proposed emendment of his principle of chance deference will be perfectly consistent with Elga's 'thirder' solution. In contrast, it will be inconsistent with Lewis's own 'halfer' solution.

2 | OVERVIEW: THE CASE OF CHANCE

In this section, I will provide an overview of my proposal by developing a principle of deference for a particular kind of expert: the objective chances. Once I've introduced and motivated the proposal for the case of chance, I will go on in subsequent sections to further clarify and develop the proposal, and to consider experts other than chance. Causal readers will be able to stop reading at the end of this section while still appreciating the paper's central innovations.

2.1 Lewis's Principle of Chance Deference

Lewis (1980) thought that you should defer to the objective chances by adhering to the following principle.[5]

> LEWIS'S PRINCIPLE OF CHANCE DEFERENCE
> For any thought '*p*', any number *n*%, and any time *t*, your credence in '*p*',

5. LCD isn't the same as Lewis's *Principal Principle*, though it follows from the Principal Principle given the updating rule of conditionalisation, which Lewis accepted (see his 1999, e.g.).

given that the time t chance of 'p' is n%, should be n%,

(LCD) $$C(p \mid \mathcal{C}h_t(p) = n\%) \stackrel{!}{=} n\%$$

(so long as you lack any time t inadmissible information)

Let me offer a few comments on this principle. Firstly, on notation: '$C(p)$' is your credence function. You hand it a thought, 'p', and it hands you back a number between 0% and 100% '$C(p \mid q)$' is your *conditional* credence function. You hand it a pair of thoughts, 'p' and 'q', and it hands you back a number between 0% and 100% which indicates how confident you are that p, on the supposition that 'q' is true. I will take it for granted that your conditional and unconditional credences are related in the following way, for any 'p' and 'q': $C(p \mid q) \cdot C(q) = C(p \wedge q)$.[6] '$\mathcal{C}h_t$' is the definite description "the time t objective chance function". Thus, '$\mathcal{C}h_t(p) = n\%$' says that the time t objective chance of 'p' is n%. I place an exclamation mark over an equals sign to indicate that the equality *ought* to hold, and not that it *does* hold. Thus, LCD says what your credence *ought* to be like; it doesn't say anything about what they *are* like. Secondly, some comments on terminology. For now, I stipulatively reserve the word 'thought' for whatever the arguments of your credence function happen to be. (I will say more about my use of the term 'thought' in §3 below.) For LEWIS (1980), information counts as time t *inadmissible* if and only if (iff) the information is *about* times after t. So long as you are at a time before t, the only way LEWIS (1980) thought you could come by inadmissible information was by way of time travellers, crystal balls, oracles, and the like. So long as there's no funny business like that, and so long as it's before the time t, LEWIS's criterion of inadmissibility will allow us to ignore the parenthetical proviso.

I'll argue here that Lewis's principle LCD faces two kinds of problems. In the first place, it faces problems with *a priori* knowable contingencies. (This problem has been noted and discussed by HAWTHORNE & LASONEN-AARNIO (2009), SCHULZ (2011), NOLAN (2016), and SALMÓN (2019), among others.) In the second place, it faces problems in cases where you've lost track of the time. (To my knowledge, this second problem has not been recognised before.)

To illustrate the first problem, suppose that we are about to flip a coin at time t, and, at some point before t, I introduce the name "Uppy" by saying "Let's call whichever side of the coin actually lands up 'Uppy'". Let 'u' be the thought that the coin lands with Uppy facing up. Then, if we set p equal to u and we set n equal to 50%, LCD tells us that your credence in 'u', given that the objective chance of 'u' is 50%, should be 50%.

$$C(u \mid \mathcal{C}h_t(u) = 50\%) \stackrel{!}{=} 50\%$$

6. I will assume throughout that your unconditional credences satisfy the following two rationality constraints: (i) if it's *a priori* knowable that 'p' is true, then $C(p) = 100\%$; and (ii) if it's *a priori* knowable that no two of 'p_1', 'p_2', 'p_3', … are both true, then $C(p_1 \vee p_2 \vee p_3 \vee \ldots) = C(p_1) + C(p_2) + C(p_3) + \ldots$. I will also take for granted that your *conditional* credences satisfy the following rationality constraint, known as *conglomerability*: for any thoughts 'p' and 'q', and any set of thoughts $\mathbf{r}$ which *partitions* 'q', your conditional credence $C(p \mid q)$ lies in the range of the conditional credences $C(p \mid r)$, for the $r \in \mathbf{r}$—that is: $\inf_{r \in \mathbf{r}} C(p \mid r) \leqslant C(p \mid q) \leqslant \sup_{r \in \mathbf{r}} C(p \mid r)$. (I say that the set of thoughts $\mathbf{r}$ *partitions* 'q' just in case (a) the disjunction of every 'r' $\in \mathbf{r}$ is *a priori* equivalent to 'q', (b) it is *a priori* knowable that no two thoughts in $\mathbf{r}$ are both true at once, and (c) for no 'r' $\in \mathbf{r}$ is it *a priori* knowable that 'r' is false.

But you know for sure that the objective chance of 'u' is 50%. For you know for sure that Uppy is either heads or tails. If Uppy is heads, then the chance of the coin landing on Uppy is the chance of the coin landing on heads, which is 50%. And if Uppy is tails, then the chance of the coin landing of Uppy is the chance of the coin landing on tails, which is 50%. So, either way, the chance of the coin landing on Uppy is 50%. If C is a probability—and I'll suppose for the nonce that it is—and you know something for sure, then you may ignore it whenever it shows up on the right-hand-side of the '|'. That is: if $C(q) = 100\%$, then $C(p \mid q) = C(p)$. So LCD says that your credence in 'u' should be 50%,

$$C(u) \stackrel{!}{=} 50\%$$

This looks like bad advice. After all, it is *a priori* knowable that the coin lands on Uppy (so long as it lands on anything at all). So it looks like your credence in 'u' should be close to 100%, and not down around 50%. (The reader may not yet be convinced that our first problem *is* a problem for LCD. I'll consider some objections they may have in §4.1 below.)

To illustrate the second problem, suppose that you don't know whether it's Monday or Tuesday, but you think it's equally likely to be either. That is: you're 50% sure that today is Monday, and 50% sure that today is Tuesday. And, while you don't know what day it is, you know for sure that *today*'s chance of Secretariat winning the race ('w') is 75% and that *yesterday*'s chance of Secretariat winning the race was 25%. Then, if we set 'p' equal to 'w', t equal to Monday (*mon*), and n% equal to 25% and 75%, respectively, LCD tells us that

$$C(w \mid \mathcal{C}h_{mon}(w) = 25\%) \stackrel{!}{=} 25\%$$
$$\text{and} \qquad C(w \mid \mathcal{C}h_{mon}(w) = 75\%) \stackrel{!}{=} 75\%$$

You know for sure that $\mathcal{C}h_{mon}(w) = 25\%$ iff it is Tuesday ('*tuesday*'), and you know for sure that $\mathcal{C}h_{mon}(w) = 75\%$ iff it is Monday ('*monday*'). If you know for sure that $q \leftrightarrow r$, then $C(p \mid q) = C(p \mid r)$. So this implies:

$$C(w \mid \mathit{tuesday}) \stackrel{!}{=} 25\%$$
$$\text{and} \qquad C(w \mid \mathit{monday}) \stackrel{!}{=} 75\%$$

Since you are 50% sure that it is Monday and 50% sure that it is Tuesday, this implies (*via* the law of total probability) that

$$\begin{aligned} C(w) &= C(w \mid \mathit{monday}) \cdot C(\mathit{monday}) + C(w \mid \mathit{tuesday}) \cdot C(\mathit{tuesday}) \\ &\stackrel{!}{=} 75\% \cdot 50\% + 25\% \cdot 50\% \\ &= 50\% \end{aligned}$$

But this looks like bad advice. After all, you know for sure that *today*'s chance of Secretariat winning is 75%. Given that, it seems that you should be 75% sure that Secretariat wins, and not merely 50% sure. (Again, the reader may not yet be convinced that this second problem is a problem for LCD. I'll consider objections in §4.2 below.)

There have been proposed emendations of LEWIS's principle which successfully

deal with the first problem; but they do not solve the second problem. See §4 for further discussion. The principle I will propose below will allow us to solve both problems.

2.2 A Two-Dimensional Principle of Chance Deference

In general, principles of expert deference face difficulties when it comes to *de se* thoughts—thoughts which are in part about who you are and where you are located in space and time. For instance, let the relevant expert be Beyoncé's doctor. A naïve principle of doctor deference would tell Beyoncé that, given that her doctor's credence in 'p' is n%, her credence in 'p' should be n%, too. That is, if 'C' is Beyoncé's credence function, then:

$$C(p \mid \mathcal{D} = D) \stackrel{!}{=} D(p)$$

where '$\mathcal{D}$' is the definite description 'Beyoncé's doctor's credence function', and 'D' is any particular probability function. (The reader will notice differences between the form of this principle of doctor deference and the form of Lewis's principle of chance deference. These differences are negligible; the curious reader should consult §6 and [author].)

Set 'p' equal to the *de se* thought 'I am sick' ('s'). Then, this principle of doctor deference will tell Beyoncé: "given that your doctor is confident in 'I am sick', your should be confident in 'I am sick', too". But this is terrible advice. When Beyoncé's doctor entertains the thought 'I am sick', they entertain a thought with the truth-conditional content that *they* are sick. When Beyoncé entertains that same thought, she entertains a thought with the truth-conditional content that *she* is sick. Since there's no connection between Beyoncé's health and her doctor's health, she should not see her doctor's high credence in 'I am sick' as imposing any rational constraint on her own credence in 'I am sick'. (The reader may suspect that the doctor's thought 'I am sick' is not the same as Beyoncé's thought 'I am sick'. These kinds of worries are addressed in §3 below.)

Beyoncé should not defer to her doctor by setting her credence in 'I am sick' equal to the doctor's credence in *that same de se thought*. Instead, she should defer to them by setting her credence in 'I am sick' equal to their credence in some appropriately chosen *de dicto surrogate* of that *de se* thought. In this case, the appropriate surrogate is 'Beyoncé is sick'. Below, I will provide a *general* surrogate which you should use whenever you are deferring to an expert. But to lay the groundwork for that general surrogate, let me begin by introducing the notion of a *location*.

2.2.1 Locations and *De Dicto* Surrogates. Say that a thought is *purely de se* iff it only says something about who you are, or when and where you are located in time and space, and it doesn't additionally tell you anything about what the world is like—that is, it doesn't provide you with any *de dicto* information. What I will call a *location* is a thought which is strong enough to settle the truth-value of all of your purely *de se* thoughts—and no stronger. In other words: a *location* tells you who you are, where you are, and what time it is in as rich a detail as your thoughts will permit—and it doesn't tell you anything more than this. (I give a more careful definition of 'location' in §5.4 below.) A location is just another thought; however, as a notational convention, I'll use lowercase Greek letters like 'λ' to indicate that a thought is a location.

Now, take any thought 'p', and any location, 'λ'. Then, the *de dicto* λ-surrogate of 'p'—which I will write 'p_λ'—is a thought which is true so long as 'p' is true when entertained at the location λ. That is: 'p_λ' says that the thought 'p' expresses a truth

when it is entertained at λ. (I give a more careful definition of 'p_λ' in §5.5 below.) For instance, if 'β' is Beyoncé's location, and 's' is the *de se* thought 'I am sick', then the *de dicto* β-surrogate of 's'—'s_β'—says that 'I am sick' is true when entertained at Beyoncé's location. That is: 's_β' says that Beyoncé is sick.

Introducing *de dicto* surrogates is enough to solve our first problem for Lewis's principle LCD (the problem with *a priori* knowable contingencies). Recall: 'u' says that the coin lands on Uppy, where 'Uppy' is a name for whichever side the coin actually lands on. Let 'λ' your (known) location. Then, the proposal is that you shouldn't defer to chance about *whether the coin lands on Uppy*. Instead, you should defer to chance about whether *your thought 'u'* expresses a truth. That is: you should satisfy:

$$C(u \mid \mathcal{C}h_t = Ch) \stackrel{!}{=} Ch(u_\lambda)$$

'u_λ' says that the thought 'u' expresses a truth when it is entertained at your location, λ. And, even though there's only a 50% chance than the coin lands on Uppy, there is a *100%* chance that your thought 'the coin lands on Uppy' will express a truth. If the coin lands heads, then your thought 'the coin lands on Uppy' will say that the coin lands on heads, and this will be true. On the other hand, if the coin lands tails, then your thought 'the coin lands on Uppy' will say that the coin lands on tails, and this will be true. So your thought will be true no matter how the coin lands. So the principle will tell you that

$$C(u \mid \mathcal{C}h_t = Ch) \stackrel{!}{=} 100\%$$

And since it will say this for *every* potential chance function Ch, the principle will imply that your unconditional credence in 'u' should be 100%.[7]

This works well so long as you know for sure what your location is. But what if you are uncertain about your location? Suppose, for instance, that you don't know whether you are Kelly or Beyoncé. In that case, I think we should say this: given that you are Beyoncé and your doctor is n% sure that Beyoncé is sick, you should be n% sure in 'I am sick'. And, given that you are Kelly and your doctor is n% sure that Kelly is sick, you should be n% sure in 'I am sick'. That is, if 'β' is Beyoncé's location, 'κ' is Kelly's location, '$\mathcal{D}$' is the definite description 'my doctor's credence function', and 'D' is any probability function, your credence function, C, should satisfy:

$$C(s \mid \mathcal{D} = D \wedge \beta) \stackrel{!}{=} D(s_\beta)$$

$$\text{and} \quad C(s \mid \mathcal{D} = D \wedge \kappa) \stackrel{!}{=} D(s_\kappa)$$

More generally, you should defer to an expert, $\mathcal{E}$, as described below:

7. Here, I appeal to the principle of *conglomerability* (the third rationality constraint from fn 6). For conglomerability implies that, if you have a set of thoughts $\mathbf{r}$ which *partitions* the thought 'q', and $C(p \mid r) = n\%$ for each 'r' $\in \mathbf{r}$, then $C(p \mid q) = n\%$, too. In the special case in which the set of thoughts $\mathbf{r}$ is countable, and each 'r' $\in \mathbf{r}$ has positive credence, this follows from countably additivity (the second rationality constraint from fn 6); however, if $\mathbf{r}$ is an uncountably infinite partition of 'q', then the principle is independent of countable additivity. Nonetheless, I take credences which violate the principle to be irrational. This means that, when $C(q) = 0$, I will have to understand a conditional credence like $C(p \mid q)$ as being relativised to a partition (*i.e.*, a set of thoughts which partitions a tautology). See Easwaran (2013, 2019) for more.

TWO-DIMENSIONAL DEFERENCE

Given that the expert $\mathcal{E}$'s probability function is E, and given that you are located at λ, your credence in 'p' should be E's probability in the *de dicto* λ-surrogate of 'p', 'p_λ'.

$$C(p \mid \mathcal{E} = E \wedge \lambda) \stackrel{!}{=} E(p_\lambda)$$

In a slogan: you should defer to the expert about whether your thoughts are true, given the location at which you are entertaining them. (In §6.1 below, I'll give some reasons why this principle should be generalised further, but those generalisations won't be relevant when the expert is the objective chances, so I'll ignore them for now.)

2.2.2 Two-Dimensional Chance Deference. Applying this general principle to the expert of chance, then, we should say this:

> So long as you lack any time t inadmissible information, your credence in 'p', given that the time t objective chance function is Ch, and given that you are located at λ, should be equal to $Ch(p_\lambda)$.

$$C(p \mid \mathcal{C}h_t = Ch \wedge \lambda) \stackrel{!}{=} Ch(p_\lambda)$$

In a slogan: you should defer to chance about whether your thoughts are true, given the location at which you are entertaining them—so long, that is, as you lack any inadmissible information.

This principle only applies in cases where you lack inadmissible information. However, if we combine this principle with *ur-prior conditionalisation*, it tells us exactly what your credences should be, even if you have inadmissible information. The principle of ur-prior conditionalisation I have in mind says that there should be some ur-prior credence function C_0 (a credence function which it would be rational to hold in the absence of any evidence) such that, for any 'e', when your total evidence is 'e', your credence in 'p', conditional on 'q', should be $C_0(p \mid q \wedge e)$.[8] Now, if we set 'q' in this principle equal to '$\mathcal{C}h_t = Ch \wedge \lambda$', it tells us that

$$\begin{aligned} C(p \mid \mathcal{C}h_t = Ch \wedge \lambda) &\stackrel{!}{=} C_0(p \mid \mathcal{C}h_t = Ch \wedge \lambda \wedge e) \\ &= \frac{C_0(p \wedge e \mid \mathcal{C}h_t = Ch \wedge \lambda)}{C_0(e \mid \mathcal{C}h_t = Ch \wedge \lambda)} \end{aligned}$$

Now, we may apply our principle of chance deference to both the numerator and the denominator of the fraction above. After all, the ur-prior credence function C_0 doesn't have any inadmissible evidence—it doesn't have any evidence at all! So:[9]

$$C(p \mid \mathcal{C}h_t = Ch \wedge \lambda) \stackrel{!}{=} \frac{Ch(p_\lambda \wedge e_\lambda)}{Ch(e_\lambda)} = Ch(p_\lambda \mid e_\lambda)$$

8. Cf. Meacham (2016).

9. Here, I am assuming that '$(p \wedge e)_\lambda$' = '$p_\lambda \wedge e_\lambda$'. This follows from the careful definition of '*de dicto* λ-surrogate' I provide in §5.5 below.

If you lack any time t inadmissible information, then our principle tells us that $C(p \mid \mathcal{C}h_t = Ch \wedge \lambda)$ should *also* be $Ch(p_\lambda)$. Therefore, if the evidence 'e' is time t admissible, it should be that $Ch(p_\lambda \mid e_\lambda) = Ch(p_\lambda)$, for any thought '$p$', any potential chance function Ch, and any potential location, 'λ'. Set 'p' equal to 'e', and this implies that it should be that $Ch(e_\lambda) = Ch(e_\lambda \mid e_\lambda) = 100\%$, for any Ch and 'λ'. So the principle of ur-prior conditionalisation has provided us with a sufficient condition for inadmissible information. If $Ch(e_\lambda) < 100\%$ for any potential time t chance function and location, Ch and 'λ', then 'e' is time t inadmissible.

I propose we strengthen this sufficient condition for inadmissibility into a necessary and sufficient condition. That is, I propose the following criterion of inadmissibility:

> INADMISSIBLE INFORMATION
> 'e' is time t *inadmissible* iff, for some potential location and time t chance function, 'λ' and Ch,
>
> $$Ch(e_\lambda) < 100\%$$

In this definition, a location 'λ' and a time t chance function Ch are *potential* iff your evidence is consistent with λ being your location and Ch being the time t objective chance function. That is: λ and Ch are a *potential* pair of location and time t chance function iff your total evidence doesn't entail that either $\mathcal{C}h_t \neq Ch$ or $\neg\lambda$. In a slogan, the criterion tells us that 'e' is inadmissible iff it might be news to the objective chances.

Given this criterion for inadmissibility, we may provide a fully general principle of chance deference which applies even in cases where you have inadmissible information.

> TWO-DIMENSIONAL CHANCE DEFERENCE
> If 'e' is your time t inadmissible information, then your credence in 'p', given that the time t objective chance function is Ch, and given that you are located at λ, should be equal to $Ch(p_\lambda \mid e_\lambda)$.

(2DCD) $$C(p \mid \mathcal{C}h_t = Ch \wedge \lambda) \stackrel{!}{=} Ch(p_\lambda \mid e_\lambda)$$

We've already seen how this principle solves our first problem (the problem with *a priori* knowable contingencies). It also solves the second problem (the problem with losing track of the time). Recall, in the problem case, you are 50% sure that today is Monday, 50% sure that today is Tuesday, and you know for sure that today, the chance of Secretariat winning the race ('w') is 75%, and that yesterday, the chance of 'w' was 25%. Let 'μ' be any Monday location, and let 'τ' be any Tuesday location. Then, notice that the information that today's chance of 'w' is 75%—'$\mathcal{C}h_{today}(w) = 75\%$'—is Monday inadmissible. The reason is that this information might be news to the Monday chance function. For τ is a potential location, and if you are at the location τ, then the *de dicto* τ-surrogate of '$\mathcal{C}h_{today}(w) = 75\%$' (namely: the *Tuesday* chance of 'w' is 75%) is news to the Monday chance function.

In this case, there are two relevant kinds of potential Monday chance functions: those according to which the chance of 'w' is 75% and those according to which the chance of 'w' is 25%. Take an arbitrary function of the first kind and call it '$Ch^{75\%}$'. Take an arbitrary function of the second kind and call it '$Ch^{25\%}$'. You know for sure that

$\mathcal{C}h_{mon} = Ch^{75\%}$ only if today is Monday, and you know for sure that $\mathcal{C}h_{mon} = Ch^{25\%}$ only if today is Tuesday. Now, since '$\mathcal{C}h_{today}(w) = 75\%$' is your total inadmissible information, 2DCD implies that

$$\begin{aligned} C(w \mid \mathcal{C}h_{mon} = Ch^{75\%} \wedge \mu) &\stackrel{!}{=} Ch^{75\%}(w \mid \mathcal{C}h_{today}(w) = 75\%_{\mu}) \\ &= Ch^{75\%}(w \mid \mathcal{C}h_{mon}(w) = 75\%) \\ \text{and} \quad C(w \mid \mathcal{C}h_{mon} = Ch^{25\%} \wedge \tau) &\stackrel{!}{=} Ch^{25\%}(w \mid \mathcal{C}h_{today}(w) = 75\%_{\tau}) \\ &= Ch^{25\%}(w \mid \mathcal{C}h_{tues}(w) = 75\%) \end{aligned}$$

Assuming that the chance function knows its own values for sure, $Ch^{75\%}(\mathcal{C}h_{mon}(w) = 75\%) = 100\%$, so the first constraint above implies that

$$C(w \mid \mathcal{C}h_{mon} = Ch^{75\%} \wedge \mu) \stackrel{!}{=} Ch^{75\%}(w) = 75\%$$

And, assuming that the objective chances defer to their future selves,[10] $Ch^{25\%}(w \mid \mathcal{C}h_{tues}(w) = 75\%) = 75\%$, so the second constraint above implies that

$$C(m \mid \mathcal{C}h_{mon} = Ch^{25\%} \wedge \tau) \stackrel{!}{=} 75\%$$

Since $\mu, \tau, Ch^{75\%}$, and $Ch^{25\%}$ were arbitrary, the same will hold for *any* potential Monday location, *any* potential Tuesday location, *any* potential Monday chance function which gives a 75% probability to 'w', and *any* potential Monday chance function which gives a 25% probability to 'w'. Since these are the only kinds of potential locations and Monday chance functions, we will have in general:[11]

$$C(w) \stackrel{!}{=} 75\%$$

So the second problem is resolved.

2.3 Sleeping Beauty

The principle of chance deference I've developed here has a surprising consequence for Elga (2000)'s *Sleeping Beauty* puzzle. In this puzzle, we suppose that, on Sunday evening, you are informed of the following: you will be put to sleep with a powerful sedative and awoken on Monday morning. On Monday evening, you will be put back to sleep and a fair coin will be flipped. If this coin lands heads, then you will be kept asleep throughout Tuesday, and you will not be woken again until Wednesday. If, on the other hand, the coin lands tails, then your memories of Monday will be erased, and you will be awoken again on Tuesday. Also, just by the way: you are beautiful.

When you awake on Monday morning, you will know for sure that, if it is Tuesday, then the coin flip on Monday landed tails. However, you won't know for sure whether it is Monday or Tuesday. For all you know for sure, it is Tuesday and your memories of

10. More carefully, I am assuming that, for any times t, t^* such that $t < t^*$, and any p in the domain of the chance function, $Ch_t(p \mid Ch_{t^*}(p) = x) = x$. Cf. van Fraassen (1984, 1995).

11. Here, I again appeal to the principle of *conglomerability*. See footnote 7.

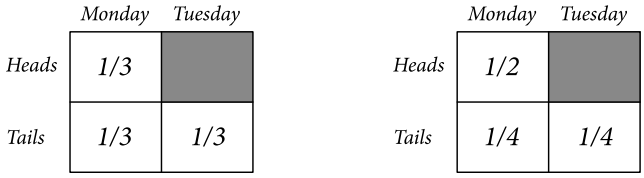

FIGURE 1: The *thirder* thinks you should have the credence distribution in figure 1a, whereas the *halfer* thinks you should have the credence distribution in figure 1b.

being awoken on Monday have been erased. The central debate over *Sleeping Beauty* concerns how confident you should be that Monday's flip landed heads, 'h'. So-called *thirders* say that your credence in 'h' should be one third. They advocate the credence distribution shown in figure 1a.[12] So-called *halfers* are generally unhappy with this distribution, in part because it means that your credence in 'h' departs from the known Monday *chance* of 'h'. They say instead that your credence in 'h' should be one half. They advocate the credence distribution shown in figure 1b.[13,14]

Let's use 'μ' for any arbitrary Monday location, and 'τ' for any arbitrary Tuesday location. Let 'Ch' be any potential Monday chance function. And let 'a' be the thought 'I am awake'. Importantly, 'a' is information you have when you wake up on Monday—this is the information which allows you to rule out that it is Tuesday and Monday's flip landed heads. Moreover, given our criterion for inadmissibility, this information will count as Monday inadmissible. For τ is a potential location, and the *de dicto* τ-surrogate of 'a', 'a_τ'—which says that you are awake on Tuesday—is news to the Monday chances. Because you're awake on Tuesday iff the coin lands tails, the Monday chances think that there's only a 50% probability that you'll be awake on Tuesday, $Ch(a_\tau) = 50\%$. (Of course, 'a_μ' is *not* news to the Monday chances—the Monday chances know for sure that you are awake on Monday.)

Then, 2DCD implies that

$$C(h \mid Ch_{mon} = Ch \wedge \mu) \stackrel{!}{=} Ch(h_\mu \mid a_\mu)$$

But the Monday chances are already certain that a_μ, and the *de dicto* surrogate of 'h_μ' is just 'h', so this reduce to

$$C(h \mid Ch_{mon} = Ch \wedge \mu) \stackrel{!}{=} Ch(h) = 50\%$$

Moreover, since this holds for *any* potential Monday chance function Ch, this implies

12. See, for instance, ELGA (2000, 2004), DORR (2002), ARNTZENIUS (2003), HITCHCOCK (2004), HORGAN (2004), and WEINTRAUB (2004).

13. See, for instance, LEWIS (2001), HALPERN (2004), BOSTROM (2007), and MEACHAM (2008).

14. Of course, the *thirder* and *halfer* positions are not exhaustive. For one alternative, see the 'imprecise' suggestion discussed in MONTON (2002) and defended in SINGER (2014).

that[15]

$$C(h \mid \mu) \stackrel{!}{=} 50\%$$

And since this holds for *any* potential Monday location μ, this in turn implies that

$$C(h \mid \mathit{Monday}) \stackrel{!}{=} 50\%$$

(where '*Monday*' is the *de se* thought that today is Monday.)

This, it turns out, is a powerful constraint. It is incompatible with the halfer's favoured distribution, and compatible with the thirder's. So, surprisingly, if we accept the principle of chance deference which I've developed here (for quite independent reasons), then it will be the *thirder*, and not the halfer, who properly defers to the known chances. It is of course true that the thirder's credence in '*h*' is not *equal* to the known chance of '*h*'. But, if we accept my proposed criterion of inadmissibility, then the thirder has a ready excuse: their credence in '*h*' departs from the known chance of heads because they have the inadmissible information that they are awake.[16] This is not information which is *about* times after Monday, so it will not count as inadmissible according to Lewis (1980)'s criterion—it is, after all, Monday, and there are, after all, no time travellers, oracles, crystal balls, nor any other form of divination or prognostication. Nonetheless, it is information which might be news to the Monday chances—for it might be Tuesday, and if it is Tuesday, then your being awake is news to the Monday chances. So it counts as inadmissible information given our criterion. And, given that they have this inadmissible information, the thirder is correctly showing deference to the objective chances.

* * *

This concludes the big-picture overview of the paper's proposal. The casual reader may feel free to stop reading at this point. In the following sections, I will further develop and defend the approach to deference that I've outlined here. In particular, in §3 below, I will say more about my use of the term 'thoughts', and how it relates to familiar debates about the objects of belief. In §4, I will return to the two objections I raised for Lewis's principle of chance deference, LCD. In §4.1, I'll consider ways you might attempt to defend LCD from the first problem—the one about *a priori* knowable contingencies—as well as some alternative treatments of the problem from the literature. In §4.2, I'll consider some ways you might want to defend LCD from the second problem—the one about losing track of the time. In §5, I will develop a two-dimensional framework for thinking about how the truth-conditions of your thoughts vary, depending upon who, when, and where you are, and what the world is like. This framework will bear some similarities to other two-dimensional frameworks, though there will also be substantive differences. In §§5.4 and 5.5, I use this framework to give a careful definition of the notion of a *location*, and of the *de dicto* λ-surrogate of a thought. In §6, I provide a fully general two-dimensional principle of deference, for an arbitrary expert. Finally, in §7, I use this general principle to explain how, on

15. This and the next inference rely upon the principle of conglomerability. See fn 7.

16. Cf. Horgan (2004) and Weintraub (2004).

my view, you should show deference to your future, better-informed self. There, I 'two-dimensionalise' VAN FRAASSEN (1984)'s principle of *reflection*, and show how the resulting principle escapes some objections from ARNTZENIUS (2003).

If you're not interested, you can easily skip past §3 and §4. Most of §5 may also be skipped, though you should read §5.5 to understand what I mean by a location being *occupied* before going on to §6. And while much of §6 may be skipped, you'll want to read the introductory paragraphs (up to the start of §6.1) before going on to §7.

3 | THOUGHTS

I use 'thoughts' as a technical term for the arguments of your credence function. They are those things to which you assign degrees of confidence. I'll suppose that, in many cases, at least, one and the same thought may be entertained by different people and at different times, that thoughts can be true or false, and that it makes sense to talk about the *negation* of a thought, p, which I write '$\sim p$', as well as the *disjunction* and *conjunction* of two thoughts, p and q, which I write '$p \vee q$' and '$p \wedge q$', respectively. I use English sentences to give names to thoughts. Intuitively, a sentence names the thought to which a thinker would express commitment by asserting that sentence. This works well enough for my purposes here; though it's worth noting that, in some cases, a single sentence will correspond to multiple thoughts. For instance, for someone who falsely thinks that there are two famous Polish people with the name 'Pederewski', there will be two thoughts corresponding to the one sentence 'Paderewski is a musician'.[17] In cases like these, we would have to reach for more detailed sentences like 'Paderewski the politician is a musician' or 'Paderewski the musician is a musician' to name these thoughts.

I assume that a thought, together with the information of who entertains the thought, when and where, and in what possible world, is *a priori* sufficient to determine the truth-conditions of that thought (that is: the set of possible worlds in which the thought is true)—more on this in §5 below. Importantly, I do not hold that thoughts are just truth-conditional contents, or sets of metaphysically possible worlds. Nor should you. Your credence in 'Mark Twain is a gifted humorist' can differ from your credence in 'Samuel Clemens is a gifted humorist', though both of these are true in exactly the same metaphysically possible worlds (because Twain is identical to Clemens). So thoughts should be in some respects more fine-grained than truth-conditional contents. And in other respects, they should be more coarse-grained. When Beyoncé and Kelly both entertain the thought 'I am sick', they thereby become related to different truth-conditional contents. Beyoncé becomes related to the truth-conditional content that Beyoncé is sick, whereas Kelly becomes related to the truth-conditional content that Kelly is sick. However, Kelly might be confused about whether she is Kelly or Beyoncé. She shouldn't thereby be confused about which thought she's investing credence in when I she is 50% sure of 'I am sick'. That is: when you are 50% sure of 'I am sick', you can know which thought you're 50% sure of, even if you don't know who you are. So we should say that Kelly and Beyoncé entertain the same thought, even though, by so doing, they become related to two different truth-conditional contents. In this

17. Cf. KRIPKE (1979)

respect, thoughts should be individuated a bit more coarsely than truth-conditional contents.

Beyond these assumptions, I hope to remain officially neutral on what a thought is. I believe that my use of the term 'thought' is broad and ecumenical enough that, no matter your views about the objects of belief, you will be able to find something to play the role of my thoughts, and which can serve as the arguments of your credence function. For everyone should accept that there's an important difference between the belief state that I would report by saying 'I think that Twain is clever, and Clemens is not' and the belief state I would report by saying 'I think that Twain is clever and that Twain is not'. The first belief state could be rational, while the second could not. So there is something to distinguish those belief states. Whatever that something is, thoughts should be partly individuated with respect to it, so that 'Twain is clever' and 'Clemens is clever' can be distinct thoughts. And everyone should accept that, when Kelly and Beyoncé each have the belief they would express with the sentence 'I am sick', there is something that that they thereby have in common. Whatever that something is, thoughts should be partly individuated with respect to it, so that when Kelly and Beyoncé entertain the thought they would each express with the sentence 'I am sick', they are entertaining one and the same thought. In the remainder of this section, I will say more about how to individuate thoughts, given some popular views about the objects of belief.

First, some terminology: as I'll use the term here, a *proposition* is the referent of a 'that'-clause in an attitude ascription. So the referent of 'that Fermat didn't have a proof of Fermat's last theorem' in an attitude ascription like 'John suspects that Fermat didn't have a proof of Fermat's last theorem' is a proposition. Some hold that propositions are *fine-grained*, in the following sense: the referent of 'that Twain is gifted humorist' is a different proposition than the referent of 'that Clemens is a gifted humorist'. Others want to identify these two propositions. Call the first group *fine-grainers*, and the second, *coarse-grainers*. Coarse-grainers say that, if you believe that Twain is a gifted humorist and yet disbelieve that Clemens is a gifted humorist, then you both believe and disbelieve one and the same proposition. Nonetheless, coarse-grainers will want to allow that you may still be rational (after all, you may not know that Twain is Clemens). They will want to distinguish your rational belief state from an irrational state of believing that Twain both was and was not a gifted humorist. To do this, they will appeal to the notion of a *guise*. A guise is a way of being acquainted with a proposition. For coarse-grainers, when you bear an attitude to a proposition, you do so under some guise or other. The reason you can rationally believe and disbelieve one and the same proposition is that the guise under which you believe it is distinct from the guise under which you disbelieve it.

A coarse-grainer should distinguish thoughts from propositions. For I have supposed that your credences are a function from thoughts to real numbers. It follows that you cannot give one and the same thought two different credences: if $C(p) \neq C(q)$, then $p \neq q$. But, according to the coarse-grainer, you *can* give one and the same *proposition* two different credences. For instance, you can be confident that Twain is a gifted humorist but not very confident that Clemens is. So, if you are a coarse-grainer, you should individuate thoughts by something other—or something more—than their propositional content.

I see two natural suggestions for the coarse-grainer: thoughts could be guises, or

they could be guise-proposition pairs (where the paired proposition is the one you entertain *via* that guise).[18] BRAUN (2016) opts for the second option, though from my perspective, the first is more attractive. On a coarse-grained view of propositions, rational credence has everything to do with the guises under which propositions are entertained, and nothing to do with the propositions thereby entertained. Just to illustrate the point, consider a coarse-grained view on which propositions are individuated by their truth-conditions—that is, if it is necessary that the propositions P and Q have the same truth-value, then $P = Q$. Take any thought, 'p', and let P be the proposition which your thought 'p' expresses. So long as 'p' is true, there is a guise such that you may be rationally certain in P, under that guise. For, if 'p' is true, then 'p' has the very same truth-conditions as '$p \leftrightarrow @p$'. Here, '$@p$' says that p is *actually* true. (That is: if 'p' is actually true, then '$@p$' is necessarily true; and if 'p' is actually false, then '$@p$' is necessarily false.) We've supposed that 'p' is true; that means that '$@p$' is necessarily true. So the biconditional '$p \leftrightarrow @p$' will have a necessarily true right-hand-side, and so it will be true iff its left-hand-side, 'p', is true. So 'p' and '$p \leftrightarrow @p$' have the very same truth-conditions. On the coarse-grained view we are considering, then, they correspond to precisely the same propositions. Now, '$p \leftrightarrow @p$' is *a priori* knowable (it is *a priori* knowable that any thought of your is true if and only if it is *actually* true.) If a thought is *a priori* knowable, then you may be rationality certain of it. So you can be rationally certain of a thought with the same truth-conditional content as 'p', for any true 'p' whatsoever. (I owe this observation to GIBBARD, 2012, appendix 1 and YLI-VAKKURI & HAWTHORNE (forthcoming); Gibbard additionally shows that similar results hold for other kinds of coarse-grainers.) So, from my perspective, when it comes to rational credence, it's most natural for a coarse-grainer to think that propositions are an idle wheel, and so to identify thoughts with guises, not guise-proposition pairs. (Of course, propositions have other roles to play for the coarse-grainer. They are needed; they're just not needed in the domain of your credence function.)

Whether a coarse-grainer identifies thoughts with guises or guise-proposition pairs may make a difference to the book-keeping in this paper. Consider the guise associated with my belief that I am sick now. Suppose, for the sake of illustration, that both my future self and my doctor are also capable of entertaining a proposition under this guise—though that guise will determine different propositions for me, my future self, and my doctor. For me, the guise determines the proposition that [Author] is sick on August 16, 2020; whereas, for my future self, it determines the proposition that [Author] is sick on August 17, 2020, and, for my doctor, the guise determines the proposition that *they* are sick on August 16, 2020. If the thought 'I am sick now' is just the guise, then both me, my future self, and my doctor can have a credence in this one thought. On the other hand, if the thought 'I am sick now' is a pair of a guise and a proposition, then my doctor is not capable of entertaining my thought 'I am sick'. I doubt that there is any substantive issue here. Suppose that thoughts are identified with guise-proposition pairs. Then, we may say that two thoughts are *equivalent* iff they have a guise-component in common. Then, even if my doctor cannot entertain my thought 'I am sick', they can entertain a thought which is equivalent to it. And this will be enough for my purposes. In what follows, I'll opt for the first form of book-keeping, supposing

18. See the proposals discussed in CHALMERS (2011), BRAUN (2016), and FITTS (2014).

that me and my doctor can both have credences in the thought I'd express with 'I am sick', but the reader should feel free to keep their own books differently, supposing instead that my and my doctor's thoughts are merely equivalent, and not identical. If you keep your books this way, then whenever I say "the thought'*p*'", you should read this as "the equivalence class of '*p*'". So far as I can see, nothing substantive will change.

Turning now to fine-grainers: Insofar as they are happy to say that the proposition I express when I say 'I am sick' is the same as the proposition you expresses when you say 'I am sick', fine-grainers may identify thoughts with propositions. If they distinguish between these propositions, then their propositions are finer than my thoughts. Just as with the coarse-grainer, I don't think there is any substantive issue here. Even if you and I believe different propositions when we each believe the propositions we'd express with 'I am sick', there is nonetheless something that our belief states have in common. Perhaps both of our beliefs are mediated by the same sentence in the language of thought, perhaps they are mediated by the same guise, or belief *state*. In any case, if you are this kind of fine-grainer, you may understand my talk about thoughts as talk about equivalence classes of belief states which have that feature in common. When I take two thoughts to be the same, you can understand me as talking about thoughts which are *equivalent*, though strictly speaking distinct. Again, so far as I can see, nothing substantive will change.

4 | PROBLEMS FOR LEWIS'S PRINCIPLE OF CHANCE DEFERENCE

In §2.1, I raised two problems for LEWIS (1980)'s principle of chance deference, LCD. In this section, I will consider some alternative responses to those problem cases.

4.1 *A Priori* Knowable Contingencies

Recall that the first problem concerned *a priori* knowable contingencies. Before we flip a fair coin, we name whichever side actually lands up 'Uppy'. Then, it is *a priori* knowable that the coin will land on Uppy, though we also know for sure that the objective chance of the coin landing on Uppy is 50%. So, while it appears that we should be nearly 100% sure that the coin lands on Uppy, LCD tells us to be only 50% sure that the coin lands on Uppy.

One reaction to this kind of case is to suggest that the naming ceremony in which 'Uppy' was introduced has provided you with some kind of inadmissible information. This kind of flat-footed appeal to inadmissibility often shows up in conversation about cases like these, and more careful versions of the reaction show up in SCHWARZ (2014) and SPENCER (2020). While the flatfooted version of the response faces some serious problems, the more careful approach of SCHWARZ and SPENCER is able to successfully deal with this problem. However, the principles advocated by SCHWARZ and SPENCER still face the second difficulty from §2.1 above. That is: these principles still give bad advice when you've lost track of the time.

Let me make two points about the flat-footed version of this response. Firstly, if the dubbing ceremony provides you with inadmissible evidence, then inadmissible evidence is much easier to come by than Lewis thought. As I mentioned in §2.1 above, so long as they are sitting around before the time t, Lewis thought that ordinary humans left to their own devices would only have time t admissible evidence—it is only with *via*

time travel or prognostication that ordinary humans could come to possess inadmissible information. But ordinary humans left to their own devices are perfectly capable of introducing names like 'Uppy' without the assistance of crystal balls, oracles, or time machines.[19] Secondly—and more importantly—we can generate this problem for LCD without any dubbing ceremony or the introduction of any *name* at all. All we need is the rigidified definite description 'the side of the coin which actually lands up'. You should be certain, or nearly certain, that the side of the coin which actually lands up lands up, but you are also certain, or nearly certain, that the chance of this happening is 50%. We can even create this kind of trouble for LCD with just demonstratives like 'this coin'. So it seems that any solution which appeals to the kind of knowledge gained in dubbing ceremonies is going to solve the problem in general.

I take it that what drives this kind of response to the problem is the idea that you simply shouldn't be deferring to chance about thoughts like 'the coin lands on Uppy'. This is a natural thought, though it's worth pointing out that it's difficult for LEWIS to say anything like this. The reason is that, for LEWIS (1980), the arguments of your credence function are truth-conditional contents, or sets of metaphysically possible worlds. In §3, I argue that thoughts should not be identified with truth-conditional contents. But, even putting those concerns aside, if we hold that thoughts are sets of metaphysically possible worlds, and we hold that dubbing ceremonies give us inadmissible evidence, then we will *always* be in a position to escape the rational requirement imposed by LCD. Suppose you know that the chance of Secretariat winning the race is 5%, but you wish nonetheless to be very confident than Secretariat will win. Then, you can simply introduce the verbs "flerm" and "glurg" with the following speech: "If Secretariat actually wins, then 'flerm' means 'win' and 'glurg' means 'lose', and if Secretariat actually loses, then 'flerm' means 'lose' and 'glurg' means 'win'." Now, if you think you shouldn't defer to chance about thoughts like 'the coin lands on Uppy', then you should also think that you should not defer to chance about either 'Secretariat flerms' or 'Secretariat glurgs', and you know that 'Secretariat wins' has the same truth-conditional content as one of those thoughts. If this linguistic ceremony is enough for you to no longer defer to chance about either 'Secretariat flerms' or 'Secretariat glurgs', then you are now free to adopt a high credence that Secretariat wins while still abiding by the principle of chance deference.

The bounds of rationality are not so easily slipped. So it's unsurprising that, when SCHWARZ (2014, §3) deals with puzzles like these, he assumes that the objects of credence are more fine-grained than truth-conditional contents.[20] He then offers the

19. At least, I will suppose that we are capable of introducing names in this way, though some have disagreed. I won't be engaging with that position here, since it's increasingly unpopular, and it doesn't ultimately get us out of our puzzle, as we can raise the same troubles with definite descriptions and demonstratives (see the discussion in the body above).

20. He assumes (with LEWIS, 1979) that the arguments of your credence function are sets of centred possible worlds, and that, when we attribute a property to an individual, we do so with what Schwarz calls an *identifier*: a relation we bear to the individual, and by which we pick the individual out. For instance, we can identify the heads side of the coin in two ways: either as the side with George Washington's face on it, or as the side which actually lands up. These two different identifiers make for two different objects of credence, so they make for two different thoughts. In the terminology from §3, an 'identifier' like this is associated with a *guise*—it gives us a way of being related to a proposition (*i.e.*, a truth-conditional content). Schwarz, then, takes the objects of credence to be guise-proposition pairs, and not guises on their own. Since nothing I have to say in this section depends upon the identity conditions for your thoughts, this complication won't

following emendation of LCD (the difference is in the parenthetical proviso):

> SCHWARZ'S PRINCIPLE OF CHANCE DEFERENCE
> For any thought 'p', any number $n\%$, and any time t, your credence in 'p', given that the time t chance of p is $n\%$, should be $n\%$,
>
> (SCD) $$C(p \mid \mathcal{C}h_t(p) = n\%) \stackrel{!}{=} n\%$$
>
> (so long as you don't have any inadmissible information, *and so long as the thought 'p' is apt for deference at t*).

In this principle, '$\mathcal{C}h_t(p) = n\%$' is also a thought, so it should also be distinguished from its truth-conditional content. Just as 'the coin lands on Uppy' has the same truth-conditional content as 'the coin lands on heads' (since the coin actually lands on heads), 'the chance that the coin lands on Uppy is 50%' has the same truth-conditional content as 'the chance that the coin lands on heads is 50%'. As I'll understand it here, Schwarz's principle SCD requires the embedded thought 'p' on the right-hand-side of the '|' to be the same as the unembedded 'p' on the left-hand-side of the '|'.[21] So long as the thought 'the coin lands on Uppy' is not apt for deference, SCD won't fall prey to the counterexample which beset LCD.

What SCD is capable of telling us depends upon how many thoughts are apt for deference at t; and, before it tells us anything at all, we must have an account of which thoughts are apt for deference at t and which are not. A natural first suggestion is this: a thought is not apt for deference at t whenever the truth-conditional content of that thought depends upon matters which are chancy at t. That is: 'p' is apt for deference at t iff, for some truth-conditional content P, there's a positive chance at t that P is the truth-conditional content of 'p', and there is a positive chance at t that P is *not* the truth-conditional content of 'p'. This first pass suggestion might require further Chisholming, but it will do as a rough-and-ready characterisation of which thoughts are apt for deference at t.

SPENCER (2020, fn 20) floats a superficially different way of responding to these kinds of worries. Where 'p' and 'q' are any thoughts, $n\%$ is any number, and 'Ex('p', q)' says that 'p' expresses the truth-conditional content that q, SPENCER suggests modifying LCD to this:

$$C(p \mid \mathcal{C}h_t(q) = n\% \wedge \mathrm{Ex}(\text{`}p\text{'}, q)) \stackrel{!}{=} n\%$$

so long as you don't have any time t inadmissible information, and *so long as* 'Ex('p', q)' *is only about matters before t*.[22] If 'p' is 'the coin lands Beatrice up', and 'q' says that the

make any difference to my discussion here. One final caveat: Schwarz intends his principle to apply to deterministic chances as well as tychistic chances; for this reason, his explicit presentation of the principle has some additional bells and whistles which mine lacks. In the present context, I'm only concerned with tychistic chance (for this reason, the examples involving coin flips should really involve quantum measurement instead, but I've gone with coin flips for ease of exposition).

21. Cf. CHALMERS (2011, p. 630). The same qualification applies to my own principle of chance deference 2DCD.

22. SPENCER formulates his principle in terms of an *initial* credence function; but, given conditionalisation, his principle will entail the one in the body. Instead of using the thought 'Ex('p', q)', he uses 'Of(g, q)', where g is the guise of the thought 'p', and 'Of(g, q)' says that g is a guise of the truth-conditional content that q. I've used 'Ex('p', q)' instead in an attempt to translate Spencer's principle into into my idiom. To be

coin lands heads up, then '$\text{Ex}(\text{`}p\text{'}, q)$' will be partly about matters after t. So Spencer's proposed principle won't fall prey to the counterexample which beset LCD.

Whatever truth-conditional content 'p' expresses, this content could be entertained *via* the thought 'p'. Then, if 'p' in fact expresses the truth-conditional content that q, then one way of entertaining the truth-conditional content of '$\text{Ex}(\text{`}p\text{'}, q)$' ("$p$' expresses the truth-conditional content that $q$') is with the thought '$\text{Ex}(\text{`}p\text{'}, p)$' ("$p$' expresses the truth-conditional content that p'). And this thought is knowable *a priori*. Thus, it is certain to be true. If a thought is certain to be true, then we can ignore it whenever it shows up on the right-hand-side of the '|'. So if we set 'q' to 'p' in SPENCER's proposed principle, then it reduces to the following:

$$C(p \mid \mathcal{C}h_t(p) = n\%) \overset{!}{=} n\%$$

so long as you don't have any time t inadmissible information, and *so long as* '$\text{Ex}(\text{`}p\text{'}, p)$' *is only about matters before* $t$. And this is just SCD with a more explicit specification of which thoughts are apt for deference at $t$. According to SPENCER's principle, a thought is apt for deference at $t$ iff '$\text{Ex}(\text{`}p\text{'}, p)$' is about matters prior to t. This is in the same ballpark as the rough-and-ready characterisation of 'apt for deference' that I offered SCHWARZ above. There is some additional strength in SPENCER's principle, due to the fact that we don't have to co-ordinate the thought 'q' embedded in '$\mathcal{C}h_t(q) = n\%$' on the right-hand-side of the '|' with the thought 'p' on the right-hand-side. However, this additional strength won't be relevant to my discussion below. In §4.2 below, I'll focus on SCHWARZ's principle, but everything I have to say about it there applies to SPENCER's principle as well.

4.2 Losing Track of the Time

Principles like SCD are able to deal with *a priori* knowable contingencies like 'the coin will land Beatrice up'. However, just like LCD, they face difficulties in cases where you've lost track of the time.

Recall the case from §2.1: you're 50% sure that it's Monday and 50% sure that it's Tuesday, and you know for sure that *today's* chance of Secretariat winning ('w') is 75%, and *yesterday's* chance of Secretariat winning was 25%. In this case, SCD, just like LCD, says that, conditional on $\mathcal{C}h_{mon}(w) = 25\%$, your credence in '$w$' should be 25%, and conditional on $\mathcal{C}h_{mon}(w) = 75\%$, your credence in '$w$' should be 75%. Just as with LCD, these two constraints imply that your credence in 'w' should be 50%. (See §2.1.)

A thought like 'Secretariat wins' should count as apt for deference on either Monday or Tuesday. The race won't be run until Wednesday (let's say), and there doesn't seem to be any funny business with naming. Moreover, 'Secretariat wins' will count as apt for deference according to the rough-and-ready characterisation I offered SCHWARZ above. On Monday, it is not a matter of chance which truth-conditional content the thought 'Secretariat wins' expresses. Likewise with SPENCER's implicit characterisation. The thought that 'Secretariat wins' expresses the truth-conditional content that Secretariat wins is not about times after Monday. So on either criterion, 'Secretariat

explicit: if you are a guise theorist, then my thoughts are your guises, and the thought 'p' will express the truth-conditional content that q just in case 'p' is a guise of that truth-conditional content (for you). See §3.

wins' will be apt for deference.

But there is another way out: SCD (and LCD, for that matter) will only imply that your credence that Secretariat wins should be 50% *if* we assume that you don't have any Monday-inadmissible information. And you might suspect that, in this case, you *do* have some Monday-inadmissible information. After all, for all you're in a position to know, today is Tuesday. And if it is Tuesday, then your knowledge that *today's* chance of 'w' is 75% is Monday-inadmissible information.

It's true that, *if* today is Tuesday, then your information that today's chance of 'w' is 75% will count be about times after Monday, and so will count as Monday-inadmissible, according to LEWIS's criterion. But nothing about the case requires us to suppose that today is Tuesday. Suppose that, unbeknowst to you, today is in fact Monday. If that's the case, then your information that today's chance of 'w' is 75% will not be about times after Monday, and so will not count as inadmissible, given LEWIS's criterion. More broadly, if today is in fact Monday, then—given that there are no crystal balls or oracles about—it's difficult to see how you could have come by any information about times after Monday. And so it's difficult to see how you could have acquired any information which is inadmissible, given LEWIS (1980)'s criterion.

But perhaps we should revise LEWIS's criterion of inadmissibility. Perhaps we should say that information is time t inadmissible iff, *for all you know*, it is about times after t. Then, we could say that losing track of the time has given you the Monday-inadmissible information that the Tuesday chance of 'w' *might* be 75%.

Let me make three observations about this reply. Firstly, while it's true that 'the Tuesday chance of 'w' is 75%' is Monday-inadmissible information, we should be cautious about the inference from "e' is inadmissible' to "it might be that e' is inadmissible'. After all, 'Secretariat wins' is Monday-inadmissible, and you know that Secretariat might win. But surely knowing that Secretariat *might* win isn't a sufficient reason to stop deferring to the known chance that Secretariat wins.

Secondly, suppose that I have been keeping track of the time, and I inform you that today is Monday. At that point, you know for sure that it is Monday and that the Monday chance of 'w' is 75%. Once I've told you the day, there is no difference between your epistemic situation with respect to 'w' and my own. Since I don't have any information which is Monday-inadmissible, you shouldn't have any Monday-inadmissible information, either. But, in learning that it is Monday, you didn't *lose* any information. You only *gained* information. Since you don't end up with any inadmissible information, you must not have had any inadmissible information to begin with. I can foresee some denying this by claiming that, whenever you gain the information that p, you will thereby lose the information that it might be that $\sim p$. So gaining the information that it is Monday means *losing* the putatively inadmissible information that the Tuesday chance of 'w' might be 75%. It seems to me that, if you learn something new and you haven't forgotten anything, then you shouldn't count as losing information in any interesting sense of the word 'information'. But I don't have much to say to anyone who disagrees with me on this point. More importantly, however, it simply isn't true in this case that you've lost the information that the Tuesday chance of 'w' might be 75%. That's something you continue to know even after you're informed that today is Monday.

Thirdly: suppose that there is a countable infinity of times, $t_1, t_2, t_3, \ldots$, which might, for all you know, be the current time. As before, while you don't know the

time, you do know that the *current* chance of Secretariat winning is 75%. In this kind of case, a principle of chance deference should tell you to set your credence in '*w*' to 75%. However, if all that it takes for you to have information about times after t_i is for you to know something about what the t_{i+1} chances *might* be, then you will have information about times after t_i, for *every* time t_i. And, in that case, SCD (or LCD, for that matter) wouldn't tell you to defer to the chances at *any* time.

In sum, I don't think it's plausible to suggest that losing track of the time gives you information about the future. However, let me emphasise that I think that there is *something* deeply right about this response. In particular, I think that you have sufficient reason to have a credence of 75% in '*w*', in spite of the fact that your expectation of the Monday chance of '*w*' is 50%. Moreover, I think that this is true precisely *because* you think that it might be Tuesday. However, I don't see how to understand this in terms of your possessing information which is *about* times after Monday. And if you don't have any information about times after Monday, then SCD (and LCD) will require your credence in '*w*' to be 50%. This seems bad enough, but it's worth emphasizing that SCD (and LCD) will *also* require your credence in '*w*' to be 75%. After all, your expectation of the *Tuesday* chance of '*w*' will be 75%.[23] So, if you don't have any inadmissible information, then SCD (and LCD) will require your credence in '*w*' to be *both* 50% *and* 75%.

5 | A TWO-DIMENSIONAL FRAMEWORK

In §3, I survey several options for what a thought may be. On any of them, it will turn out that you can be uncertain about the truth-conditions of your thoughts. When you think 'I am sick', you may not know whether you think a thought which has the same truth-conditions as 'Beyoncé is sick'. For, if you suffer from amnesia, you may not know whether you are Beyoncé; and you know that 'I am sick' has the same truth-conditions as 'Beyoncé is sick' iff you are Beyoncé. So your uncertainty about whether you are Beyoncé translates into uncertainty about the truth-conditions of your thought 'I am sick'.

In this section I will introduce a *two-dimensional* framework to model the ways that the truth-conditions of your thoughts can vary depending upon what the world is like, who you are, and where you are located. This framework is similar to the two-dimensional semantics of KAPLAN (1989), STALNAKER (1978), EVANS (1979), DAVIES & HUMBERSTONE (1980), JACKSON (1998), and CHALMERS (2006a,b), though there are several formal and interpretational differences. §§5.1–5.3 lay out the framework. §5.1 introduces what I call an *index*—an index is a set of thoughts which, for all you know *a priori*, might completely describe a possibility in as rich a detail as your thoughts permit. To fix ideas: indices play a role in my framework similar to the role played by LEWIS (1979)'s *centred* possible worlds. §5.2 defines a relation of relative metaphysical possibility over these indices. And §5.3 defines a two-dimensional valuation function for thoughts. This function maps a thought, '*p*', together with a *pair* of indices (i, j), to

23. If today is Tuesday, then you know for sure that the Tuesday chance of '*w*' is 75%, and if today is Monday, then you know for sure that the Monday chance of '*w*'s is 75%, and so you'll expect the Tuesday chance of '*w*' to be 75% as well. By the law of iterated expectations, your current expectation of the Tuesday chance of '*w*' is 75%.

a truth-value—which represents whether the truth-conditional content which 'p' has when entertained at the index i is true or false at the index j. In §5.4, I formulate a careful definition of a *world*—which, intuitively, completely describes a possibility in as rich a detail as your *de dicto* thoughts permit. To fix ideas: worlds in my framework play a role similar to possible worlds in Lewis's framework. Also in §5.4, I formulate a careful definition of what I called in §2.2.1 a *location*, 'λ'. Finally, in §5.5, I define the *de dicto* λ-surrogate of a thought 'p', 'p_λ'. In this section, I also explain what it is for a location to be *occupied*. If you're primarily interested in the material in §6 and §7, you can skip most of this section, but you should pause for a while on §5.5.

5.1 Indices

The framework begins with a domain of indices, generated from an underlying set of thoughts.[24] An *index*, as I will be understanding it, is a maximally consistent set of thoughts. A consistent set of thoughts is *maximal* just in case no proper superset of it is also consistent. Let me offer three clarificatory comments about 'consistency'. Firstly, consistency is not metaphysical compossibility. The thoughts in the set {'Twain is gifted', 'Clemens is not gifted'} are not metaphysically compossible; but they are consistent, as I use the term here. Secondly, even if a set of thoughts can be *a priori* known to contain a falsehood, it can nonetheless be consistent. For instance, any set containing both the thoughts 'q' and '$\sim@q$' can be *a priori* known to contain a falsehood, as can any set containing the thought 'I am not here now'.[25] Nonetheless, these sets may still count as consistent. Although it is *a priori* knowable that their members do not describe *actuality*, it is not *a priori* knowable that their members do not describe a *possibility*. Consistency concerns the latter notion, not the former. Thirdly, if you have thoughts *de se et nunc*—thoughts about who you are and where and when you are located in space and time—then an index will include this kind of information. For instance, one index might include the thought 'I am Beyoncé'. Another index might include the thought 'I am not Beyoncé'. In some sense, these two thoughts are consistent. Beyoncé could truly think the first while Kelly truly thinks the second. However, they are not consistent in the sense that I'm using the term here. For, when you entertain these two thoughts, here and now, they explicitly contradict each other. Summing up, we may think of consistency like this: can you *a priori* determine that it is not possible for each thought in the set to be true (given the truth-conditional content they have for you, here and now)? (As the examples of {'q', '$\sim@q$'} and {'I am not here now'} demonstrate, determining that it's not *possible* for each thought in a set to be true is more demanding than determining that it's not *actually* the case that each thought in the set is true.) If the answer is 'yes', then the set is inconsistent; if the answer is 'no', then the set is consistent.

Exactly one index will contain all and only the true thoughts (as actually entertained by you, here and now). Call that 'the actual index'. Call the other indices 'non-actual'. Some indices may be known, *a priori*, to be non-actual. For other indices,

24. I'll assume that this set of thoughts is closed under negation, disjunction, and conjunction.

25. Some may disagree on the grounds that the question of whether you exist is *a posteriori*. I don't think that's the right way to think about *a priority*, but it doesn't matter. If you think 'I am not here now' cannot be known *a priori* to be false for this reason, feel free to substitute 'I exist and I am not here now'.

their non-actuality will not be *a priori* knowable. Call the latter kinds of indices *epistemically possible*. We can lift your credence distribution over thoughts to a credence function over sets of epistemically possible indices, as follows. For every thought, there will be the corresponding set of epistemically possible indices at which that thought is true. Set your credence in that set equal to your credence in the thought. If your credences over thoughts were probabilities, then this new credence function over sets of epistemically possible indices will be a probability function, as well.[26]

5.2 Metaphysical Possibility

Consider an index which contains both 'Twain is clever' and 'Clemens is not clever'. This is not the actual index. It does not describe the way things actually are—nor does it even describe a way they (metaphysically) *might* have been, given the metaphysical necessity of identity, which I will assume here.[27] Nonetheless, it is not *a priori* knowable that the index does not describe a way things might have been. If the actual index turns out to contain the thought 'Twain ≠ Clemens', then this index needn't describe a metaphysically impossibility. Similarly, if the actual index contains 'I am not Beyoncé' and 'today is Monday', then any index which contains 'I am Beyoncé' or 'today is Tuesday' will describe an impossibility. So which indices are metaphysically possible varies depending upon which index is actual.

To model this, let me introduce a binary accessibility relation, R, defined over the set of indices. R will model relative metaphysical possibility between indices. If iRj, then I'll say that j is metaphysically possible from i. I'll work my way up to the relation R by first defining a smaller relation, R^-, which I'll then extend to R. Say that an epistemically possible index, i, bears the relation R^- to another index, j, iR^-j, iff, if i is actual, then j is metaphysically possible. The relation R^- tells us which indices are metaphysically possible from each *epistemically* possible index. I assume that the correct logic of metaphysical possibility is S5; so, every index should be metaphysically possible from itself (*i.e.*, R should be reflexive), and, if both j and k are metaphysically possible from i, then k should be metaphysically possible from j (*i.e.*, R should be Euclidean). When we restrict attention to the epistemically possible indices, R^- will already be reflexive and Euclidean; but this won't hold for the epistemically impossible indices. So we should extend R^- by taking its reflexive and Euclidean closure. R is the relation which results.

For an illustration, let's suppose that you only have opinions about the following two thoughts (and Boolean combinations thereof): 'The coin lands on heads'—which we can dub 'h'—and 'The coin lands on Uppy'—which we can dub 'u'. (Recall, the name 'Uppy' was introduced for whatever side of the coin actually lands up.) Then, there are four indices, which we can name 'i_{hu}', '$i_{h\overline{u}}$', '$i_{\overline{h}u}$', and '$i_{\overline{hu}}$'. At i_{hu}, the coin

26. More carefully, I mean this: suppose that your credences satisfy the following constraints: 1) your credence in each thought is no lower than zero; 2) if it is *a priori* knowable that t is true, then your credence in t is 100%; and 3) if it is *a priori* knowable that at least one of p and q is false, then your credence in the disjunction $p \vee q$ is equal to the sum of your credence in p and your credence in q. Then, when we lift your credences to a function over sets of epistemically possible indices, the lifted function will satisfy the following constraints: 1*) your credence in every set is non-negative; 2*) your credence in the set of all epistemically possible indices is 100%; and 3*) if P and Q are disjoint sets, then your credence in their union is equal to the sum of your credence in P and your credence in Q.

27. The metaphysical necessity of identity says that, if $x = y$, then it is metaphysically necessary that $x = y$.

row R col	i_{hu}	$i_{\overline{h}u}$	$i_{h\overline{u}}$	$i_{\overline{hu}}$
i_{hu}	✓	×	×	✓
$i_{\overline{h}u}$	×	✓	✓	×
$i_{h\overline{u}}$	×	✓	✓	×
$i_{\overline{hu}}$	✓	×	×	✓

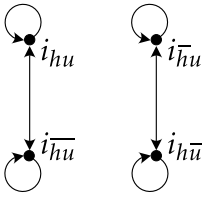

FIGURE 2: The matrix on the left shows the relation of relative metaphysical possibility, R. (A '✓' means that the row index bears R to the column index; an '×' means that it does not.) The same relation is displayed graphically on the right. i_{hu} and $i_{\overline{h}u}$ are epistemically possible, and $i_{h\overline{u}}$ and $i_{\overline{hu}}$ are not epistemically possible.

lands on heads, and it lands on Uppy. At $i_{h\overline{u}}$, the coin lands on heads, but not on Uppy. At $i_{\overline{h}u}$, the coin lands on tails and on Uppy. And at $i_{\overline{hu}}$, the coin lands on tails, but not on Uppy. The second and fourth of these indices, $i_{h\overline{u}}$ and $i_{\overline{hu}}$, are not epistemically possible. Given the way 'Uppy' was introduced, it is *a priori* knowable that the coin lands on Uppy. So it is *a priori* knowable that $i_{h\overline{u}}$ and $i_{\overline{hu}}$ contain at least one falsehood. All four, however, are consistent. Take $i_{\overline{hu}}$. If i_{hu} is the actual index, then 'Uppy' refers to heads. Since it's possible that the coin not land on heads, it is possible that the coin not not land on Uppy. So $i_{\overline{hu}}$ is metaphysically possible from i_{hu}. And, if $i_{\overline{hu}}$ is metaphysically possible from i_{hu}, then i_{hu} must be metaphysically possible from $i_{\overline{hu}}$, and $i_{\overline{hu}}$ must be metaphysically possible from itself. Figure 2 shows which indices are metaphysically possible from which other indices in this case.

Notice that, when we ask about whether an index j is metaphysically possible from i, we give the thoughts in j the truth-conditions they have *when entertained at* i, and not the truth-conditions they have when entertained at j. Were the thought '$\sim u$' entertained at $i_{\overline{hu}}$, it would have the truth-conditional content that the coin didn't land on *tails*. It is because these truth-conditions conflict with 'the coin didn't land on heads' that $i_{\overline{hu}}$ is epistemically impossible.[28] However, at i_{hu}, '$\sim u$' is true iff the coin didn't land on *heads*. And this clearly is consistent with '$\sim h$'. That's why, from i_{hu}, the index $i_{\overline{hu}}$ is metaphysically possible (though it is known *a priori* to be non-actual).

This highlights an important lesson which we will need in the next subsection: an index represents the world as being a certain way—in particular, it represents the world as meeting the truth-conditions of the thoughts it contains. However, it represents the world as meeting the *actual* truth-conditions of those thoughts, and *not* the truth-conditions those thoughts would have, were they entertained in a possibility described by that index. The index $i_{\overline{hu}}$ describes a possibility at which the coin doesn't land on Uppy. However, it *also* describes a possibility at which the thought 'the coin lands on Uppy' would express a truth. Suppose the coin actually lands heads, and consider a possibility which *we* would accurately describe with the thought '$\sim h \wedge \sim u$'. The inhabitants of this possibility would *not* accurately describe their world with the thought '$\sim h \wedge \sim u$'. Instead, *they* would accurately describe their world with the thought '$\sim h \wedge u$'. That's because, even though *our* thought 'u' is true iff the coin lands

28. I'm treating 'the coin landed on tails' as the negation of 'the coin landed on heads'. So the coin must either land on heads or tails—if the coin lands on its edge or doesn't land, that counts as a tails landing.

$$
\begin{array}{c}
[\![h]\!]^{row,\,col} \\
\begin{array}{c} i_{hu} \\ i_{\overline{h}u} \\ i_{h\overline{u}} \\ i_{\overline{hu}} \end{array}
\end{array}
\begin{array}{c}
\begin{array}{cccc} i_{hu} & i_{\overline{h}u} & i_{h\overline{u}} & i_{\overline{hu}} \end{array} \\
\left[\begin{array}{cccc} T & * & * & F \\ * & F & T & * \\ * & F & T & * \\ T & * & * & F \end{array}\right]
\end{array}
$$

(a)

$$
\begin{array}{c}
[\![u]\!]^{row,\,col} \\
\begin{array}{c} i_{hu} \\ i_{\overline{h}u} \\ i_{h\overline{u}} \\ i_{\overline{hu}} \end{array}
\end{array}
\begin{array}{c}
\begin{array}{cccc} i_{hu} & i_{\overline{h}u} & i_{h\overline{u}} & i_{\overline{hu}} \end{array} \\
\left[\begin{array}{cccc} T & * & * & F \\ * & T & F & * \\ * & F & T & * \\ F & * & * & T \end{array}\right]
\end{array}
$$

(b)

FIGURE 3: Two-dimesional valuations for the thoughts h = 'the coin lands on heads' (in figure 3a) and u = 'the coin lands on Uppy' (in figure 3b). In both matrices, the first index comes from the row, and the second index comes from the column.

on heads, *their* thought 'u' is true iff the coin lands on tails.

5.3 Two-Dimensional Valuations

With this lesson appreciated, we may introduce a *two-dimensional* valuation function, $[\![\,]\!]$. Hand this valuation function a thought, 'p', and it hands you back a two-place function, $[\![p]\!]^{i,j}$, where i and j are indices. The output of the function $[\![p]\!]^{i,j}$ depends upon, firstly, whether j is metaphysically possible from i, and, secondly, whether j satisfies the truth-conditions which the thought 'p' has when it is entertained at i. If j is not metaphysically possible from i, then $[\![p]\!]^{i,j}$ will be undefined. In that case, I'll write '$[\![p]\!]^{i,j} = *$'. If j is metaphysically possible from i, then, if j satisfies the truth-conditions which 'p' has when entertained at i, then $[\![p]\!]^{i,j} = T$. If, on the other hand, j is metaphysically possible from i, but j *doesn't* satisfy the truth-conditions 'p' has when entertained at i, then $[\![p]\!]^{i,j} = F$. ('T' is the truth-value *true*, and 'F' is the truth-value *false*.) For instance, figure 3 shows the outputs of the two-place functions $[\![h]\!]$ (in figure 3a) and $[\![u]\!]$ (in figure 3b).

In these matrices, the truth-values along each row correspond to the truth-conditional contents which 'h' and 'u' express, if that row's index is actual. Look at the first and fourth rows in figure 3a. If i_{hu} is actual, then there are two metaphysically possible indices: i_{hu} and $i_{\overline{hu}}$. We know *a priori* that we are not at the index $i_{\overline{hu}}$; but, even so, we can ask which truth-conditions 'h' would have, when entertained at i_{hu} or $i_{\overline{hu}}$. And the answer is: its truth-conditions are exactly the same, no matter which of these indices it is entertained at: the thought is true at i_{hu} and false at $i_{\overline{hu}}$ no matter which index it is entertained at. Likewise, if $i_{\overline{h}u}$ is actual, then there are two metaphysically possible indices: $i_{\overline{h}u}$ and $i_{h\overline{u}}$. And, again, the thought 'h' has the same truth-conditions, whether it is entertained at $i_{\overline{h}u}$ or $i_{h\overline{u}}$. No matter which of these indices it is entertained at, 'h' is false at $i_{\overline{h}u}$ and true at $i_{h\overline{u}}$.

In general, if a thought has the same truth-conditions, no matter which index it is entertained at, say that it is a *boring* thought. (Or, more carefully, a thought 'p' is boring iff, for every index i, and every index j metaphysically possible from i, the truth-conditions 'p' has when entertained at i is the same as the truth-conditions 'p' has when entertained at j.) Figure 3b demonstrates that 'u' is not boring. Start with the first and fourth rows. If 'u' is entertained at i_{hu}, it expresses a truth-conditional content which is true at i_{hu} and false at $i_{\overline{hu}}$. On the other hand, if 'u' is entertained at $i_{\overline{hu}}$, then 'u' expresses a truth-conditional content which is true at $i_{\overline{hu}}$ and false at i_{hu}. Similarly, if 'u' is entertained at $i_{\overline{h}u}$, it expresses a truth-conditional content which is

true at $i_{\overline{h}u}$ and false at $i_{h\overline{u}}$; whereas, if 'u' is entertained at $i_{h\overline{u}}$, it expresses a content which is false at $i_{\overline{h}u}$ and true at $i_{h\overline{u}}$. If a thought isn't boring, say that it's *interesting*.

There are two important asymmetries between the boring 'h' and the interesting 'u'. The first asymmetry is that, while the actual truth-value of 'h' determines the truth-conditional content of 'u', the actual truth-value of 'u' underdetermines the truth-conditional content of 'h'. Note that we could coarse-grain by ignoring the thought 'u', partitioning our indices into cells by grouping together any indices which agree on thoughts not involving 'u'. Let 'H' be the cell in which 'h' is true, $\{i_{hu}, i_{h\overline{u}}\}$, and let '$\neg H$' be the cell in which '$h$' is false, $\{i_{\overline{h}u}, i_{\overline{hu}}\}$. Then, we may show how the truth-conditional content of 'u' depends upon the actual truth of 'h' with a two-dimensional valuation function like the one shown below.

$$\begin{array}{cc} [\![u]\!]^{row,\,col} & \begin{array}{cc} H & \neg H \end{array} \\ \begin{array}{c} H \\ \neg H \end{array} & \left[\begin{array}{cc} T & F \\ F & T \end{array}\right] \end{array}$$

That is: the truth-conditional content of 'u' only depends upon whether the coin actually landed on heads or tails. The same cannot be said in reverse. If we coarse-grain by ignoring the thought 'h', and grouping together any indices which agree on thoughts not involving 'h', then we will not have enough information to determine the truth-conditional content of 'h'. Following Chalmers (2012), we may call $\{H, \neg H\}$ a *scrutability base* for 'u', since 'u"s truth-conditional content is *a priori* implied by each cell of this partition. Chalmers (2006a,b, 2012) argues that there is an appropriate scrutability base which allows us to provide a two-dimensional array like this for every thought not in the base. I'm sympathetic to Chalmers's position, but I won't require this additional thesis for my purposes here.[29]

The second asymmetry concerns the *diagonal* entries in the matrices shown in figures 3a and 3b. (These are the entries where the row index is the same as the column index.) When we set both the first and second indices to either $i_{\overline{h}u}$ or $i_{\overline{hu}}$, the thought 'h' is evaluated as false, $[\![h]\!]^{i_{\overline{h}u}, i_{\overline{h}u}} = [\![h]\!]^{i_{\overline{hu}}, i_{\overline{hu}}} = F$. In contrast, the thought 'u' is true whenever the first and second indices are the same. We may define a *diagonal* valuation function as follows. Given a thought, 'p', and an index, i, let $[p]^i \stackrel{\text{def}}{=} [\![p]\!]^{i,i}$. Since every index is metaphysically possible from itself, $[p]^i$ will always be defined. Call the function $[p]^i$ the *diagonal* content of the thought 'p'. In these terms, the second asymmetry is that the diagonal content of 'h' is contingent, whereas the diagonal content of 'u' is necessary.[30] Notice that, by construction, if a thought's diagonal content is

29. I won't require any of this additional apparatus myself, but just to dot my 'i's and cross my 't's: I've extended the two-dimensional valuation function so that it takes as inputs *sets* of indices, and not just single indices. Given any two sets of indices, I and J, the output of the function $[\![p]\!]^{I,J}$ will depend upon, firstly, whether any $j \in J$ is metaphysically possible from any $i \in I$, and secondly, whether 'p' is true or false at every $j \in J$ which is metaphysically possible from some $i \in I$. If no $j \in J$ is metaphysically possible from any $i \in I$, then $[\![p]\!]^{I,J}$ is undefined. If 'p' is true at some $j \in J$ which is metaphysically possible from some $i \in I$, and false at some other $j \in J$ which is metaphysically possible from some $i \in I$, then $[\![p]\!]^{I,J}$ is *multiply* defined. If, however, there's just one truth-value which 'p' takes on at every $j \in J$ which is metaphysically possible from any $i \in I$, then $[\![p]\!]^{I,J}$ is defined to be that truth-value.

30. I say that 'p"s diagonal content is *necessary* iff $[p]^i = T$ for every index i; and 'p"s diagonal content is *contingent* iff $[p]^i = T$ for some index i, and $[p]^j = F$ for some index j.

necessary, then it will be *a priori*. The converse thesis—that, if a thought is *a priori*, then its diagonal content will be necessary—does not follow from what I've said so far. However, if we confine our attention to just the epistemically possible indices, we may give a characterisation of which thoughts are *a priori* knowable and those which are only *a posteriori* knowable in terms of a kind of necessity of the diagonal content. Say that 'p' is an *epistemic diagonal necessity* iff, for every epistemically possible index i, $[p]^i = T$. And say that 'p' is an *epistemic diagonal contingency* iff, for some epistemically possible indices i and j, $[p]^i = T$ while $[p]^j = F$. Then, it follows by construction that 'p' is *a priori* knowable iff it is an epistemic diagonal necessity; and 'p' is only *a posteriori* knowable iff it is an epistemic diagonal contingency.

An important point for my purposes is this: while truth-conditional contents live along the rows—*i.e.*, while the truth-conditional content of a thought is given by the row of the actual index—*rational credence* lives along the diagonal. Earlier, I lifted your credence distribution over thoughts to a credence distribution over sets of epistemically possible indices by setting your credence in the set of epistemically possible indices containing 'p' to your credence that p. But the epistemically possible indices containing 'p' are precisely the epistemically possible indices i such that $[p]^i = T$. So we should sharply distinguish a rational credence in 'p' from a rational credence in 'p"s truth-conditional content. We can illustrate the distinction with the example of 'the coin lands on Uppy', or 'u'. Suppose that is rational to have a credence of 50% that i_{hu} is the actual index, and a credence of 50% that $i_{\overline{hu}}$ is the actual index. And suppose the coin actually lands on heads. Then, the truth-conditional content of 'u' is: true at i_{hu} and false at $i_{\overline{hu}}$. Since your credence in i_{hu} is 50% and your credence in $i_{\overline{hu}}$ is 0%, your credence in the truth-conditional content of 'u' is 50%. In contrast, the diagonal content of 'u' is true at both i_{hu} and $i_{\overline{hu}}$. So your credence in the diagonal content of 'u' is 100%. So rational credence goes with diagonal contents, and not with truth-conditional contents.

5.4 Worlds and Locations

Some of the thoughts in an index only concern what the world is like, and other thoughts concern who you are, or when and where you are located in the world. Following Lewis (1979), I call the former kinds of thoughts *de dicto*, and the latter *de se*. Note that a *de se* thought may tell you *more* than just when, where, or who you are; it may also tell you something about what the world is like. For instance, the thought 'I am Beyoncé and grass is green' provides both the *de se* information that you are Beyoncé and the *de dicto* information that grass is green. In this section, I will want to pull apart these two kinds of information. I'll do that by first defining what I will call a *world*, and then going on to define what I will call a *location*. Intuitively, a world will tell you everything about what the world is like *except* who, when, and where you are located, in as rich a detail as your thoughts will permit. And a location will tell you who, when, and where you are, in as rich a detail as your thoughts will permit.

Let's start with 'worlds'. Take any two indices, i and j. I'll say that j is a *world-mate* of i's iff there's no *de dicto* thought included in i whose negation is included in j. World-mates do not disagree about any *de dicto* matters, though they may disagree about *de se* matters. Being world-mates is an equivalence relation, so it partitions the

indices into equivalence classes.[31] I'll call a cell of this partition a 'world', and I'll write 'W_i' for the cell of this partition to which the index i belongs. Intuitively, W_i is the world to which the index i belongs. In this notation, then, i and j are world-mates iff $W_i = W_j$; that is, i and j are world-mates iff they belong to the same world. I'll call the world to which the actual index belongs 'the actual world'.

On to 'locations': the guiding idea is that a location is a kind of thought which tells you everything there is to tell about who, when, and where you are, but doesn't tell you anything else about the world. I will build up to my definition of 'location' by starting with the notion of a *purely de se* thought. A purely *de se* thought is a thought which does not have any *de dicto* consequences. Thus, if 'p' is purely *de se*, then, for any world W_i, there will be some index $j \in W_i$ which contains 'p'. Notice that this has the consequence that 'I am Beyoncé' does not count as purely *de se*, since it has the *de dicto* consequence that Beyoncé exists. However, there is a nearby thought which *is* purely *de se*—namely, the thought 'I am Beyoncé, *if she exists*'.[32] Now, take any two indices, i and j. I'll say that j is a *location-mate* of i iff there's no purely *de se* thought included in i whose negation is included in j, nor any purely *de se* thought included in j whose negation is included in i.[33] Location mates do not disagree about any purely *de se* matters. Being location-mates is an equivalence relation, so it partitions the indices into equivalence classes.[34] Take any cell of this partition, Λ. If a thought, 'λ', is included in all and only the indices in Λ, then I will call 'λ' a *location*. That is: a location provides exactly the information that you are within one of these equivalence classes.

It's important to note that, even though locations are defined in terms of purely *de se* thoughts, they need not be purely *de se* themselves. To illustrate, notice that both 't' = 'I am Twain, if he exists' and 'c' = 'I am Clemens, if he exists' are purely *de se*. While neither of these *individually* has any *de dicto* consequences, *together* they have the *de dicto* consequence that, if both Twain and Clemens exist, then Twain is Clemens. If you have opinions about the thoughts 't' and 'c', then any index which includes both 't' and 'c' will only be location-mates with other indices which include both of these thoughts. And so any equivalence class of indices including both 't' and 'c' will only include indices at which Twain is Clemens, unless one of them doesn't exist. And

31. To be explicit: an equivalence relation on the set of indices, $\mathbf{I}$, is any relation $R \subseteq \mathbf{I} \times \mathbf{I}$ which is reflexive and Euclidean. (A relation R is reflexive iff for every $i \in \mathbf{I}$, iRi; and R is Euclidean iff, for every $i, j, k \in \mathbf{I}$, if iRj and iRk, then jRk.) And given any $i \in \mathbf{I}$, the *equivalence class*—which I'm here writing 'W_i'—is the set of indices to which i bears R: $W_i =_{df} \{j \in \mathbf{I} \mid iRj\}$. It is easy to show that, if R is an equivalence relation, then the set of equivalence classes, $\{W_i \mid i \in \mathbf{I}\}$ will form a partition of $\mathbf{I}$ (that is: each equivalence class will be a non-empty subset of $\mathbf{I}$ and every $i \in \mathbf{I}$ will belong to exactly one of the equivalence classes).

    To show that being world-mates is an equivalence relation, we must show that it is reflexive and Euclidean. The relation is trivially reflexive. To see that it is is Euclidean, suppose (for *reductio*) that j and k are both world-mates of i, but that k is not a world-mate of j. Then, there's some *de dicto* thought 'p' such that 'p' $\in j$ and '$\sim p$' $\in k$. Now, by the maximality of indices, either 'p' or '$\sim p$' is included in i. If 'p' $\in i$, then k is not a world-mate of i's. If '$\sim p$' $\in i$, then j is not a world-mate of i's. Either way, we have a contradiction.

32. Notice that there are similar complications with distinctness claims like 'I am not Beyoncé'. This entails the *de dicto* claim that Beyoncé is not the *only* person that exists. Here, too, there is a nearby purely *de se* claim—namely, 'if there is someone distinct from Beyoncé, then I am distinct from Beyoncé'.

33. Notice that the negation of a purely *de se* thought needn't be a purely *de se* thought itself. For instance, the negation of 'I am Beyoncé, if she exists' entails that Beyoncé exists.

34. The proof that being location-mates is an equivalence relation is almost exactly the same as the proof that being world-mates is an equivalence relation, with 'world' exchanged for 'location', and '*de dicto*' exchanged for 'purely *de se*'.

row R col	$i_{\beta bs}$	$i_{\beta\overline{bs}}$	$i_{\overline{\beta}bs}$	$i_{\overline{\beta}b\overline{s}}$	$i_{\overline{\beta bs}}$	$i_{\overline{\beta bs}}$
$i_{\beta bs}$	✓	✓	×	×	×	×
$i_{\beta\overline{bs}}$	✓	✓	×	×	×	×
$i_{\overline{\beta}bs}$	×	×	✓	✓	✓	✓
$i_{\overline{\beta}b\overline{s}}$	×	×	✓	✓	✓	✓
$i_{\overline{\beta bs}}$	×	×	✓	✓	✓	✓
$i_{\overline{\beta bs}}$	×	×	✓	✓	✓	✓

FIGURE 4: The matrix shows the relation of relative metaphysical possibility, R. (A '✓' means that the row index bears R to the column index; an '×' means that it does not.)

therefore, any *location* which implies both 't' and 'c' will imply the *de dicto* information that Twain is Clemens, unless one of them doesn't exist. And this information is not be purely *de se*.

For illustration, consider the following three thoughts: 'I am sick' ('s'), 'Beyoncé is sick' ('b'), and 'I am Beyoncé' ('β'). If we begin with just these three thoughts, then there will be six indices, which we can call '$i_{\beta bs}$', '$i_{\beta\overline{bs}}$', '$i_{\overline{\beta}bs}$', '$i_{\overline{\beta}b\overline{s}}$', '$i_{\overline{\beta bs}}$', and '$i_{\overline{\beta bs}}$'. At $i_{\beta bs}$, you are Beyoncé, Beyoncé is sick, and (therefore) you are sick. At $i_{\overline{\beta}b\overline{s}}$, you are not Beyoncé, Beyoncé is sick, but you are not. And so on for the other indices, in the natural way. The relation of relative metaphysical possibility for these indices is shown in figure 4. If you are Beyoncé, then it's not possible that you're not Beyoncé; likewise, if you're not Beyoncé, then it's not possible that you are Beyoncé.

If we begin with just this set of thoughts (and we close the set under negation and conjunction), 'b' and '$\sim b$' will be the only *de dicto* thoughts.[35] Since the indices $i_{\beta bs}$, $i_{\overline{\beta}bs}$, and $i_{\overline{\beta}b\overline{s}}$ all share the *de dicto* thought, 'b', they are world-mates—intuitively, they all belong to the world in which Beyoncé is sick. Likewise, since $i_{\beta\overline{bs}}$, $i_{\overline{\beta bs}}$, and $i_{\overline{\beta bs}}$ all share the *de dicto* thought '$\sim b$', they are all world-mates. They each belong to the world in which Beyoncé is not sick. The relation of 'being world-mates' thus partitions the indices into the equivalence classes $W_b = \{i_{\beta bs}, i_{\overline{\beta}bs}, i_{\overline{\beta}b\overline{s}}\}$ and $W_{\overline{b}} = \{i_{\beta\overline{bs}}, i_{\overline{\beta bs}}, i_{\overline{\beta bs}}\}$. W_b is the world at which Beyoncé is sick, and $W_{\overline{b}}$ is the world at which she's not sick.

Relative to this set of indices, the thought 'β' will count as purely *de se*, since it does not imply any *de dicto* information. For every world, there is an index included in that world which contains 'β'. For $i_{\beta bs}$ is in the world at which Beyoncé is sick, and it contains 'β', and $i_{\beta\overline{bs}}$ is in the world at which Beyoncé is not sick, and it contains 'β'. (Of course, if you had opinions about the *de dicto* thought 'Beyoncé exists', then 'β' would not count as purely *de se*; but given that we're beginning with the expressively impoverished set of thoughts {'β', 'b', 's'}, 'β' will not entail any *de dicto* thoughts about which you have opinions.) Likewise, '$\sim\beta$' is purely *de se*. For it is included in the index $i_{\overline{\beta}bs}$, which is in the world at which Beyoncé is sick, and it is included in the index $i_{\overline{\beta bs}}$, which is in the world at which Beyoncé is not sick. And these are the only two purely *de se* thoughts.[36] So the indices $i_{\beta bs}$ and $i_{\beta\overline{bs}}$ are location-mates, as are the indices $i_{\overline{\beta}bs}$,

35. That is to say: all the *de dicto* thoughts you have opinions about will be *a priori* equivalent to either 'b' or '$\sim b$'.

36. That is: any purely *de se* thought you have opinions about is *a priori* equivalent to either 'β' or '$\sim\beta$'.

$i_{\overline{\beta}b\overline{s}}$, $i_{\overline{\beta}bs}$, and $i_{\overline{\beta}\overline{bs}}$. The relation of being location-mates thus partitions indices into the cells $\Lambda_\beta = \{i_{\beta bs}, i_{\beta\overline{bs}}\}$ and $\Lambda_{\overline{\beta}} = \{i_{\overline{\beta}bs}, i_{\overline{\beta}b\overline{s}}, i_{\overline{\beta}bs}, i_{\overline{\beta}\overline{bs}}\}$. '$\beta$' is included in all and only the indices in Λ_β, so it is a location. Likewise, '$\sim\beta$' is included in all and only the indices in $\Lambda_{\overline{\beta}}$, so it, too, is a location. (There is of course more to be said about who, when, and where you are than the location '$\sim\beta$' tells us, but the thoughts β, b, and s are not expressive enough to allow you to say those things.) These are the only two locations; all other locations are *a priori* equivalent to one of them.

5.5 *De Dicto* Surrogates

Take any thought 'p', and any location 'λ'. With these, I wish to find a *de dicto* surrogate for 'p' which is true anywhere in a world so long as the thought 'p' is true when entertained at the location λ in that world. I'll call this 'the *de dicto* λ-surrogate of 'p'', and I'll write it 'p_λ'. For instance, take the thought 'I am sick' and the location 'I am Beyoncé' ('β'). Then, the *de dicto* β-surrogate of 'I am sick' is a thought which says that 'I am sick' expresses a truth when entertained at the location 'β'. But 'I am sick' expresses a truth at this location iff Beyoncé is sick. So the *de dicto* β-surrogate of 'I am sick' is a thought which is *a priori* equivalent to 'Beyoncé is sick'. In general, in terms of the diagonal valuation function from §5.3, we may say that 'p_λ' is true at an index i iff there is some index j which is a world-mate of i ($W_j = W_i$), which is at the location λ ($[\lambda]^i = T$), and at which 'p' is true ($[p]^i = T$). That is,

$$(p_\lambda) \qquad [p_\lambda]^i = \begin{cases} T & \text{if } \exists j \,:\, W_i = W_j \wedge [p]^j = [\lambda]^j = T \\ F & \text{else} \end{cases}$$

A thought will count as a *de dicto* λ-surrogate of 'p' iff it satisfies the constraint (p_λ).[37]

Now that we have this definition of a *de dicto* λ-surrogate, we may use it to formulate the thought that a location is *occupied*. Take, for instance, a location which implies both 'I am Twain, if he exists', and 'I am Clemens, if he exists'. This location will not be occupied at every world. In particular, it will not be occupied at any world at which Twain and Clemens both exist, but Twain is not Clemens. In general, if 'λ' is a location, then the thought that 'λ' is occupied is just the *de dicto* λ-surrogate of 'λ'. That is, 'λ_λ' is the thought that 'λ' is occupied. For 'λ_λ' is true at an index i iff there is some index, j, which is a world-mate of i's, and at which both 'λ' and 'λ' are true—that is to say, iff there is some index, j, which is a world-mate of i's, and at which 'λ' is true. And this will be so iff 'λ' is occupied at the world to which i belongs.

6 | TWO-DIMENSIONAL DEFERENCE

In this section, I'm going to explain how, in general, you should show deference to an expert once you or the expert has opinions about interesting thoughts like 'the coin will land on Uppy', 'I am sick', and 'Today is Monday'. (See §5.3 to appreciate what I

37. The constraint (p_λ) singles out a class of *a priori* equivalent thoughts, any of which counts as a *de dicto* λ-surrogate of 'p'. For instance, 'b' and '$b \vee b$' are both *de dicto* β-surrogates of 'I am sick'. Since any probability function will assign the same probability to any two *a priori* equivalent thoughts, your credence function shouldn't distinguish between any two *de dicto* λ-surrogates of 'p'.

mean when I call a thought *interesting*.) To work my way up to that, though, let me begin by introducing some more standard principles of expert deference.

If you're going to attempt to align your credences with those of an expert, then you must have some views about what the expert's opinions are. So all of the principles of expert deference I'm going to discuss take it for granted that you have credences in thoughts of the form '$\mathcal{E} = E$', where the script '$\mathcal{E}$' stands for the definite description 'the expert's credence function', and 'E' is a particular credence function. When I'm speaking about a generic expert, I'll use '$\mathcal{E}$'. When I'm talking about the time t chances, I'll use '$\mathcal{C}h_t$'. And when I'm talking about your time t self, I'll use '$\mathcal{C}_t$'. Thus, '$\mathcal{C}h_t = Ch$' says that the time t chance function is given by Ch, and '$\mathcal{C}_t = C_t$' says that, at time t, your credence function is C_t. Given thoughts of the form '$\mathcal{E} = E$', we can construct thoughts of the form '$\mathcal{E}(p) = n\%$'. To get this latter thought, we take every potential expert function E which is such that $E(p) = n\%$ and disjoin them all. That is: the thought $\mathcal{E}(p) = n\%$ is the disjunction $\bigvee_{E\,:\,E(p)=n\%} \mathcal{E} = E$.

Several authors[38] suggest the following general recipe for deference: conditional on the expert thinking that 'p' is $n\%$ likely, you too should think that 'p' is $n\%$ likely. That is, for every p and every $n\%$,

(D1) $$C(p \mid \mathcal{E}(p) = n\%) \stackrel{!}{=} n\%$$

Other authors suggest the following: conditional on E being the expert's probability function, your credence in 'p' should be $E(p)$. That is, for every p and every E,

(D2) $$C(p \mid \mathcal{E} = E) \stackrel{!}{=} E(p)$$

D2 is *nearly* equivalent to D1. If you satisfy D2, then you will satisfy D1, as well; and, if you satisfy D1—and you are not in an *incredibly* singular and special case—then you will satisfy D2, as well. The two constraints are not *precisely* equivalent, but they are equivalent for all philosophical purposes.[39] So we should not worry ourselves too much about the differences between them. (See [redacted for blind review] for more discussion.)

Finally, some authors suggest that your credence in 'p' should be your *expectation* of the expert's credence in 'p'.[40] That is, your credences should satisfy D3, for every 'p',

(D3) $$C(p) \stackrel{!}{=} \sum_E E(p) \cdot C(\mathcal{E} = E)$$

D3 is strictly weaker than D1 and D2; both D1 and D2 entail D3, but D3 does not entail either D1 or D2. Unlike the differences between D1 and D2, the difference between D3 and the other two principles is not negligible. By the way, given the difference between D1 and D2, you might expect there to be a difference between D3 and a similar principle

38. See, *e.g.*, Skyrms (1980) and Gaifman (1988).

39. Gaifman (1988) provides a case in which you satisfy D1 but not D2. In [redacted for blind review], I show that Gaifman's case is, in a good sense, the *only* kind of case in which D1 and D2 come apart.

40. See, for instance, Ismael (2008, 2015).

which says that, for every 'p',

$$(\text{D3}^*) \qquad C(p) \stackrel{!}{=} \sum_{n\%} n\% \cdot C(\mathcal{E}(p) = n\%)$$

But there is no difference between D3 and D3*. You will satisfy the one iff you satisfy the other.

There are other forms which principles of expert deference take, but those other forms are introduced to deal with complications orthogonal to my purposes here. For instance, D2 entails that the expert knows for sure what its own probabilities are. In some applications, this seems implausible, and several authors have suggested the remedy of replacing the right-hand-side of D2 with '$E(p \mid \mathcal{E} = E)$'.[41] For the substituends for 'p' I'll be looking at here, it is safe to suppose that whatever uncertainty the expert may have about their own probabilities won't make a significant difference to their probability in 'p', so this additional complication won't be relevant to anything I have to say here. And precisely the same remedy is available for the principle 2DD (to be introduced below): just condition the expert's probability function on '$\mathcal{E} = E$' on the right-hand-side.[42]

As I see things, the central source of difficulty for the traditional principles is that *your* thoughts 'I am sick' and 'Today is Monday' could be true without *the expert's* thoughts 'I am sick' and 'Today is Monday' being true. More generally, the truth-conditions of *your* thought 'p' could differ from the truth-conditions of *the expert's* thought 'p'. So you shouldn't want to align your credence in 'p' with the expert's credence in that same thought, as the traditional principles of expert deference assume. To deal with this problem, I will suggest that, instead of deferring to the expert by aligning your credence in 'p' with the expert's credence in 'p', you should defer by aligning your credence in 'p' with the expert's credence in an appropriate *surrogate* of 'p'. We want to find a thought which, when entertained by the expert, will have the same truth-conditions that 'p' does, when it is entertained by you.

What makes interesting thoughts interesting is that their truth-conditions depend upon matters about which you may be uncertain. For instance, you may not know who you are, in which case you may not know which truth-conditional content your thought 'I am sick' expresses. This will complicate our search for the expert's surrogate for your thoughts. For instance, if you are Beyoncé, then your doctor's surrogate for 'I am sick' should presumably be a thought like 'Beyoncé is sick'; and, if you are Kelly, then your doctor's surrogate for 'I am sick' should be a thought like 'Kelly is sick'. But what if you're not certain whether you are Beyoncé or Kelly? In that case, it doesn't look like we will be able to find any *one* surrogate for your thought 'I am sick'. Instead, I'll suggest that, conditional on you being Beyoncé, you should defer to your doctor's opinion about 'Beyoncé is sick'; and, conditional on you being Kelly, you should defer to your doctor's opinion about 'Kelly is sick'.

In general, our principles of expert deference must control for any uncertainty you

41. See, *e.g.*, HALL (1994), LEWIS (1994), and ELGA (2013).

42. DORST (forthcoming) defends a different principle of expert deference called 'Trust'. DORST et al. (ms) generalise this principle to what they call 'Total Trust'. I believe that these principles can also be 'two-dimensionalised' in the ways I'm proposing, though I won't explore this any further here.

or the expert may have about your location. I'll also suggest that these principles must control for any uncertainty either you or the expert may have about *their* location. Then, in rough outline, the principle I'll propose says that you should defer to the expert by setting your credence in 'p' to the expert's credence in an appropriate *surrogate* of 'p', when both your and the expert's credence functions have been conditioned on both your and the expert's locations. (And not just your actual locations, but any locations which you and the expert might occupy.)

That's the rough outline. More carefully, here's the modification of D2 I'll be proposing: for any thought, 'p', any locations λ and ϵ, and any potential expert function E, your credence in 'p', given that the expert's credence function is E, you are located at λ, and ϵ is occupied, should be equal to E's credence in the *de dicto* λ-surrogate, 'p_λ', given that λ is occupied, and given that they are located at ϵ.

$$C(p \mid \mathcal{E} = E \wedge \lambda \wedge \epsilon_\epsilon) \stackrel{!}{=} E(p_\lambda \mid \lambda_\lambda \wedge \epsilon) \tag{2DD}$$

(I explain what it is for a location to be *occupied* in §5.5 above.) A similar modification could replace D1; I'll leave the details in this footnote.[43] I will discuss the natural generalisation of D3 in §7.2 below. In the remainder of this section, I'll explain the principle 2DD by walking through some examples to motivate successive revisions to D2, until we eventually arrive at the principle 2DD. Once this is done, I'll explain how the principle of expert deference 2DD relates to the principle of *chance* deference 2DCD from §2.2.2.

6.1 The Principle Explained

Suppose you know for sure that you are Beyoncé, and you wish to set your credence in 'I am sick' by deferring to the opinion of your doctor. To do so, you should not set your credence in 'I am sick' equal to your doctor's credence in that same thought. Instead, you must find some *surrogate* for your thought 'I am sick'. This should be a thought which will be true when entertained by the doctor iff 'I am sick' is true when entertained by you.

You and your doctor may faultlessly disagree about the *de se*—for instance, you may truly think 'I am sick' while she truly thinks 'I am not sick'. But you and your doctor may not faultlessly disagree about the *de dicto*. For this reason, coordination of opinion between you and your doctor should go by way of the *de dicto*. If your doctor's opinions are going to constrain your credence in the thought 'I am sick', then we must find a *de dicto* surrogate for this thought. In many case, a suitable *de dicto* surrogate will be clear. Suppose, for instance, that both you and your doctor know that you are Beyoncé. Then, the thought 'Beyoncé is sick' will fit the bill. But this particular solution won't offer any guidance in other cases. It would be preferable to

43. For any thought 'p', any locations λ and ϵ, and any number $n\%$, your credence in 'p', given that you are located at λ, ϵ is occupied, and given that the expert's credence in 'p_λ' is $n\%$, conditional on λ being occupied and them being at ϵ, should be $n\%$:

$$C(p \mid \lambda \wedge \epsilon_\epsilon \wedge \mathcal{E}(p \mid \lambda_\lambda \wedge \epsilon) = n\%) \stackrel{!}{=} n\%$$

have a *general* solution, one which will allow us to find a suitable *de dicto* surrogate for *any* thought of yours.

The question to ask yourself is this: 'how confident is the doctor that *my thought* 'I am sick' is true?' That is: 'how confident is she that, when *I* entertain the thought 'I am sick', it expresses a truth?' However confident she is of that is how confident you should be in 'I am sick'. In §5.5, I defined the thought 'p_λ', which is true at any location in a world so long as 'p' is true at location λ within that world. For this reason, 'p_λ' says that the thought 'p', entertained at λ, expresses a truth. It is a general surrogate for your thought 'p', given that you are at location λ. Return to the simple example from §5.4. In that example, 'β' was the location 'I am Beyoncé', 's' was the *de se* thought 'I am sick', and 'b' was the *de dicto* thought 'Beyoncé is sick'. Then, it is *a priori* knowable that 's_β' is true if and only if b is true. If '$\mathcal{D}$' is the definite description 'the doctor's credence function', and 'D' is any credence function, then it looks like, if you and your doctor both know that you are Beyoncé, then you should defer to your doctor's opinion by setting your credence in s, given that the doctor's credence function is D, equal to D's credence in the surrogate s_β, which will just be D's credence in b. And this motivates emending D2 to say that your credence in any thought 'p', given that the expert $\mathcal{E}$'s probability function is E, should be E's credence in p_λ:

(D2*) $$C(p \mid \mathcal{E} = E) \stackrel{!}{=} E(p_\lambda)$$

This works well in the present case, but in constructing the surrogate 's_β', we took for granted that you were Beyoncé. This is something about which you may uncertain. Suppose you don't know whether you're Beyoncé or Kelly, and you know that your doctor thinks Beyoncé is very likely sick and Kelly is very likely not sick. In that case, you shouldn't be very confident in 'I am sick'. Instead, you should proportion your confidence in 'I am sick' to your confidence that you are Beyoncé. That is, for each potential credence function D, and each location λ, you should defer to the doctor's opinion by setting your credence in s, given that $\mathcal{D} = D$ *and you are at location* λ, equal to D's credence in the surrogate 's_λ'—that is, $C(s \mid \mathcal{D} = D \wedge \lambda)$ should be $D(s_\lambda)$. If 'λ' is a location at which you are Beyoncé, then 's_λ' will imply the *de dicto* thought 'Beyoncé is sick'. And if 'λ' is a location at which you are Kelly, then 's_λ' will imply the *de dicto* thought 'Kelly is sick'. This motivates revising D2* to say that your credence in 'p', given that the expert $\mathcal{E}$'s probability function is E *and given that you are at location* λ, should be E's credence in p_λ:

(D2**) $$C(p \mid \mathcal{E} = E \wedge \lambda) \stackrel{!}{=} E(p_\lambda)$$

Just as you can be uncertain about which locations you occupy, your doctor could be uncertain about which locations *are* occupied. Suppose you know for sure that you are Jekyll, and you are 50% confident that you are *also* Hyde. However, your doctor is very confident that Jekyll and Hyde are two different people. She has one medical record filed under 'Jekyll' and another filed under 'Hyde'. There are positive test results in the 'Hyde' file, and no test results in the 'Jekyll' file. The disease is very rare, so she thinks 'Jekyll is sick' is very unlikely; but the test is reliable enough that she thinks 'Hyde is sick' is very likely. However, were she to learn that Jekyll and Hyde are one

and the same person, she would think it's very likely that that person is sick.[44]

Let 'η' be a location at which you are Jekyll and you are Hyde.[45] Then, if we abide by D2**, we'll say that your credence in 'I am sick' ('s'), given that you are are the location η, should be the doctor's credence in the *de dicto* thought 's_η'. But the thought 's_η' is only true in worlds at which the location η is occupied. At any world at which Jekyll and Hyde are different people, 's_η' will be false.[46] Since your doctor is very confident that Jekyll and Hyde are different people, she will think that 's_η' is very *unlikely*. So the principle D2** would tell you to not be very confident that you are sick, given that you are Hyde. This is the wrong verdict. Given that the doctor is very confident that Hyde is sick, you should be very confident that you are sick, given that you are Hyde.

The trouble with the principle D2** is that, on the left-hand-side, you have conditioned on some information—the information that you occupy location λ—which the expert may lack. If the expert doesn't have this information, then before deferring to them, you should first bring them up to speed by conditioning the function E on this information. Of course, the location 'λ' is a *de se* thought; so you don't want to bring the function E up to speed by conditioning it on *this* information. Instead, you want to bring it up to speed by conditioning it on an appropriate *de dicto* surrogate of this information. That is, you want to condition it on the *de dicto* information that location λ is *occupied*: 'λ_λ'. This thought, 'λ_λ', is true in any world at which 'λ' is true *somewhere*. In the case of the location 'η', 'η_η' is a *de dicto* thought which implies that Jekyll and Hyde are the same person. Thus, your credence in 's', given that your doctor's credence function is D, and given that you are at the location 'η', should be equal to D's credence in the *de dicto* η-surrogate 's_η', *given that* η *is occupied*:

$$C(s \mid \mathcal{D} = D \wedge \eta) \stackrel{!}{=} D(s_\eta \mid \eta_\eta)$$

Since you know that your doctor is very confident that Jekyll/Hyde is sick ('s_η'), conditional on the information that Jekyll and Hyde are the same person, this principle will tell you, correctly, to be very confident that you are sick, given that you are Hyde.

These considerations suggest that we should modify D2** even further, by requiring that your credence in a thought 'p', given that the expert $\mathcal{E}$'s probability function is E and you are at location λ, should be E's credence in the *de dicto* λ-surrogate 'p_λ', given that location λ is occupied, 'λ_λ':

$$\text{(D2***)} \qquad C(p \mid \mathcal{E} = E \wedge \lambda) \stackrel{!}{=} E(p_\lambda \mid \lambda_\lambda)$$

But wait—what if the expert is uncertain about *their* location? In some cases, this may not matter. If her location is irrelevant to the question of whether you are sick,

44. Cf. Chalmers (2011).

45. Or rather, it implies that you are Jekyll and Hyde, so long as both Jekyll and Hyde exist (see §5.4). In the present example, we can take for granted that both you and the doctor know for sure that Jekyll and Hyde both exist, so I'll ignore this caveat.

46. Recall from §5.5 that 's_η' is true at an index i iff there is some index j which is a world-mate of i, and at which *both* 's' and 'η' is true. If Jekyll and Hyde are different people at an index i, then at any index which is a world-mate of i, 'η' will be false. So 's_η' will be false at any index at which Jekyll and Hyde are different people.

then you may defer to your doctor about whether you are sick without either of you knowing her location. However, in some cases, her location *may* be relevant to whether you are sick. Suppose, for instance, that both you and your doctor suffer from amnesia. However, you both know the following four things: 1) you are either Alfred or Cyril, and she is either Blanche or Dinah; 2) Alfred and Blanche are fraternal twins, as are Cyril and Dinah; 3) the disease is congenital, so if one fraternal twin is sick, then the other one is sick, too; and 4) the doctor is very confident that she is sick.

In this kind of situation, you should only defer to your doctor conditional on a hypothesis about where you and she are both located. Let α be a location which implies (given your and the doctor's background information) 'I am Alfred and I am the patient', let 'β' be a location which implies 'I am Blanche and I am the doctor', and let 'δ' be a location which implies 'I am Dinah and I am the doctor'. Then, conditional on you being Alfred and the doctor being Blanche, your credence in 'I am sick' ('s') should be the doctor's credence in the *de dicto* α-surrogate 's_α', given that you are Alfred and she is Blanche:

$$C(s \mid \mathcal{D} = D \wedge \alpha \wedge \beta_\beta) \stackrel{!}{=} D(s_\alpha \mid \alpha_\alpha \wedge \beta)$$

Since the doctor's credence in s_α is high, given that you are Alfred and she is Blanche, you should have a high credence in s, given that you are Alfred and she is Blanche. Likewise, conditional on you being Alfred and her being Dinah, your credence in 'I am sick' should be her credence in the surrogate 's_α', given that you are Alfred and she is Dinah:

$$C(s \mid \mathcal{D} = D \wedge \alpha \wedge \delta_\delta) \stackrel{!}{=} D(s_\alpha \mid \alpha_\alpha \wedge \delta)$$

Since the doctor's credence in s_α is low, given that you are Alfred and she is Dinah, you should have a low credence in s, given that you are Alfred and she is Dinah.

Cases like this motivate replacing D2*** with the following principle: for any thought 'p', any locations λ and ϵ, and any potential expert function E, your credence in 'p', given that the expert function is E, you are at λ and the expert is at location ϵ, should be equal to E's credence in the *de dicto* λ-surrogate 'p_λ', given that you are at λ and they are at ϵ:

(2DD) $$C(p \mid \mathcal{E} = E \wedge \lambda \wedge \epsilon_\epsilon) \stackrel{!}{=} E(p_\lambda \mid \lambda_\lambda \wedge \epsilon)$$

In certain applications, we may be able to ignore some of 'λ', 'ϵ', 'λ_λ', and 'ϵ_ϵ' because either you know your or the expert's location for sure, or else the expert knows their or your location for sure. For instance, the objective chance function at t, $\mathcal{C}h_t$, does not have a spatial location. It's only location is temporal, and that location is both necessary and *a priori* knowable. If the objective chances at t were at t^*, they wouldn't be the objective chances *at* t.[47] So, when we are considering a principle of *chance* deference, 2DD reduces to the principle 2DCD from §2.2.2.

Parenthetically: I've motivated the principle 2DD by considering cases where you wish to defer to another human, but I believe similar reasoning carries over when we consider an expert like the objective chances. In that application, by the way, we will

47. Of course, you may wish to defer, not to the *time* t chances, but instead the *current* chances. Then, even though the current chances know their temporal location, you may not (if you've lost track of the time). In that application, the current chance's temporal location cannot be ignored.

have to understand my talk of 'the expert's thoughts' slightly differently. When I talk about 'the expert's thoughts', I am still using the term 'thought' stipulatively to refer to refer to whatever the arguments of the expert's probability function happen to be. So, if the expert is the objective chances, then their thoughts are just truth-conditional contents.[48] In that case, the difficulty isn't that the truth-conditions of your thoughts differ from the truth-conditions of chance's thoughts. The difficulty is that (in the terms from §5.3 above), your credences live along the diagonal, while chance's probabilities live along the rows. For that reason, your thought could be epistemically necessary even while the corresponding truth-conditional content is contingent. This is what led to the problem with *a priori* knowable contingencies in §2.1 and §4.1 above. Just as with human experts, before you defer to chance, you need to find some common content on which you and chance can coordinate. And in my view, just as with the human experts, this common content is the diagonalised *de dicto* λ-surrogate 'p_λ'.

Returning to the principle 2DD: notice that, if 'p' is a boring thought, and both your and $\mathcal{E}$'s credence in 'p' is independent of your and the expert's location, then this proposed principle entails the more familiar principle of expert deference D2. For, if your credence in 'p' is independent of your and the expert's location (conditional on $\mathcal{E} = E$) then the left-hand-side of 2DD equals $C(p \mid \mathcal{E} = E)$. And, if '$p$' is a boring thought, then the right-hand-side of 2DD will be equal to $E(p \mid \lambda_\lambda \wedge \epsilon)$. If E's credence in 'p' is independent of λ_λ and ϵ, then this is equal to $E(p)$. So the principle reduces to $C(p \mid \mathcal{E} = E) \stackrel{!}{=} E(p)$.

7 | TWO-DIMENSIONAL DEFERENCE TO YOUR FUTURE SELF

7.1 The Principle of Reflection

VAN FRAASSEN (1984, 1995) proposes that your credences should satisfy the following constraint: for any future time t, any thought 'p', and any function C_t which might be your credence function at time t,

$$C(p \mid \mathcal{C}_t = C_t) \stackrel{!}{=} C_t(p)$$

That is: conditional on your time t credence function being C_t, your credence that p should be $C_t(p)$. From this principle, it follows that your current credence that p should equal your *expectation* of your time t credence that p, for any thought 'p' and any future time t. That is, for any 'p' and any future t:[49]

$$C(p) \stackrel{!}{=} \sum_{C_t} C_t(p) \cdot C(\mathcal{C}_t = C_t)$$

48. At least, I will be taking for granted that the arguments of the objective chance function are truth-conditional contents. There are some who have responded to the kinds of puzzles I considered in §2.1 and §4.1 above by suggesting that the arguments of chance are more fine-grained than this. See, for instance, NOLAN (2016) and SALMÓN (2019).

49. Here, we are summing over credence functions, C_t, which might be your time t credence function. I assume in the body that there are at most countably many such functions; if there are more, then the principle should be emended to say that $C(p)$ ought to be your expectation of your time t credence in p, $Exp[C_t(p)]$.

There are a host of counterexamples to VAN FRAASSEN's principle. Many of these counterexamples are cases in which you expect your future self to be irrational, or to have strictly less information that you have now. For instance, TALBOTT (1991) notes that, in ten years' time, you will have little to no idea what you ate for lunch today; but this is no reason to drop your *current* credence that you had spaghetti for lunch today. Likewise, suppose you are about to be brainwashed to believe that Elmo from Sesame Street is the Antichrist. That's unfortunate, but it's no reason for you to *now* think that Elmo is the Antichrist.[50]

These kinds of counterexamples can seem most damning when we think of VAN FRAASSEN's principle as a constraint on your *current* credences. But the principle is more plausible, and more defensible, if we think of it as a constraint on your *learning dispositions*. To set the stage for this understanding, suppose that, just before the time t, you will learn exactly one of the thoughts in $\mathbf{E} = \{e_1, e_2, \ldots, e_N\}$ (and no more). Let 'D_e' be the credence function which you are *d*isposed to adopt if you learn 'e'. Then, the version of VAN FRAASSEN's principle which I will focus on here says that the credence in 'p' which you are disposed to adopt after learning shouldn't be expected to be any higher or lower than your current credence in 'p'. That is, for any 'p', your learning dispositions should satisfy the following constraint, which I'll call 'REFLECTION'.[51]

(REFLECTION) $$\sum_{e \in \mathbf{E}} D_e(p) \cdot C(\mathcal{C}_t = D_e) \stackrel{!}{=} C(p)$$

As I understand REFLECTION, it imposes the constraint that your learning dispositions not be *biased*. If, after learning, your credence that p is expected to be higher than your current credence that p, then your learning dispositions are biased in favor of 'p'. On the other hand, if your credence that p is expected to be lower than $C(p)$ after learning, then your learning dispositions are biased against 'p'. So, if your learning dispositions are not biased in favor or of against 'p', then they will satisfy REFLECTION.[52] In a slogan, REFLECTION says that your learning dispositions should warrant deference, in the sense that, after learning, you should be disposed to become somebody worthy of being deferred to.[53]

TALBOTT's counterexamples involve irrationality and forgetting. Take forgetting first. I contend that, at least in mundane cases, if you're currently disposed to forget something you now know upon learning, that is a rational defect in your learning dispositions. These kinds of rational defects may be widespread and forgivable, but they are nonetheless departures from ideal epistemic rationality. Likewise, in cases where you are disposed to adopt irrational credences after learning, this is a rational defect in

50. For more on these kinds of objections to VAN FRAASSEN's principle, see HITCHCOCK & GREEN (1994), ELGA (2007), and BRIGGS (2009).

51. Any name or definite description which you know for sure to denote a unique time is an acceptable substituend for 't'. So, for instance, in the right circumstances, '5:55 Tuesday morning', 'five minutes from now', and 'the time immediately after I learn one of the thoughts in **E**' could all be acceptable. (The same goes for the subscripted 't' in the principle of chance deference 2DCD.)

52. See SALOW (2018) for more discussion.

53. More specifically, you should be disposed to become somebody worthy of being deferred to *in expectation*. For reasons I don't have the space to discuss here, I don't think that your learning dispositions should satisfy $C(p \mid \mathcal{C}_t = D_e) = D_e(p)$, for every e. See [redacted for blind review] for more discussion.

your learning dispositions. Note that, even if your learning dispositions are irrational, REFLECTION does not put any pressure on you to adjust your *current* credences. This may be a way of getting your learning dispositions to satisfy REFLECTION, but it is not a way of meeting the demands of rationality. Compare: a legal goal requires getting the ball between the goal posts; but it doesn't follow that, if the ball is off-course, moving the goal posts is a way of getting a legal goal. Moving the goal posts is itself illegal. Likewise, adjusting your current credences to forget something you now know is itself irrational.

ARNTZENIUS (2003) presents some more pressing counterexamples to REFLECTION. These counterexamples all involve thoughts *de se et nunc*—thoughts about when and where you are located in time and space. Thoughts like these can give rise to some straightforward counterexamples to REFLECTION. For instance: on Sunday before going to sleep, I am nearly certain to awake, and I am disposed to be nearly certain that it is Monday upon awaking. However, I am now nearly certain that it is not Monday. So my expectation of my credence in the thought 'It is Monday' after learning is much greater than my current credence in the thought 'It is Monday', in violation of REFLECTION. One reaction to counterexamples like these is to limit REFLECTION to apply only to *de dicto* thoughts. ARNTZENIUS's counterexamples show that REFLECTION is not defensible, even after it is restricted to *de dicto* thoughts.

Let me introduce two of these counterexamples. Both involve your credence that a fair coin flip landed heads. This is is a *de dicto* thought; moreover, in the terminology from §5.3, it is a *boring* thought. First counterexample: you know that it is now 8:30, and you know that, if the coin landed heads, then the lights in your room will turn off at 12:00. However, you don't currently have a clock. As you sit in your room, you will learn that time has passed, although you know that you won't keep perfect track of the time. So, by the time 11:30 rolls around, the most you'll know for sure is that between two and four hours have passed, so that it's between 10:30 and 12:30. Let's suppose that, at 11:30, your credence that it's earlier than 11:00 will be 25%, your credence that it's later than 12:00 will be 25%, and your credence that it's between 11:00 and 12:00 will be 50%. But at 11:30, the lights in your room will certainly be on. So at 11:30, after learning that somewhere between two and four hours have passed, you'll think it's 75% likely that the lights don't tell you anything about whether the coin landed heads, and you'll think it's 25% likely that the lights tell you that the coin landed tails. So your credence that the coin landed heads will be 37.5%, or 3/8. So at 11:30, after learning that between two and four hours has passed, you expect your credence that the coin landed heads to be lower than it currently is, in violation of REFLECTION. Since it doesn't appear that these dispositions to update your credences after learning that between two and four hours have passed are irrational, it seems that REFLECTION is making unreasonable demands.

The second counterexample is ELGA (2000)'s *Sleeping Beauty* puzzle, which we've already encountered in §2.3. To explain why it is a counterexample, I'll need to introduce an additional detail to the case. Recall: on Monday morning, you are awoken; and, in the evening, you are put back to sleep. The new detail: before being put to sleep on Monday evening, you will be told that it is Monday. From there, the case is the same: a fair coin will then be flipped. If the coin lands heads, then you will be kept asleep all through Tuesday. If, on the other hand, the coin lands tails, then your memories of Monday will be erased, and you will be awoken again on Tuesday. Because your memories will have been erased, if you awake on Tuesday, you will not know for

sure whether it is Monday or Tuesday.

Back in §2.3, I introduced the *thirders* and the *halfers*. With this new addition to the case, halfers sub-divide into two camps: *Lewisian halfers* and *double halfers*. Both Lewisian and double halfers agree that, upon waking Monday morning, you should be disposed to be one half confident that Monday's coin flip lands heads ('*h*'). What differentiates the Lewisian and the double halfers is their answer to this question: how confident should you be in '*h*' *after* you're told that it is Monday? Lewisian halfers answer: 'two thirds'.[54] Double halfers, on the other hand, continue to answer: 'one half'.[55] (By the way, when it comes to this question, the thirders side with the double halfers. While they think you should be one third confident in '*h*' upon waking Monday morning, they think that you should be one half confident in '*h*' after being told that it is Monday on Monday evening.) These three positions are of course not exhaustive, but the overwhelming majority of authors who have discussed ELGA's *Sleeping Beauty* puzzle have fallen into one of these three camps. And all three of them violate REFLECTION.

Start with the thirder. Before going to sleep on Sunday, the thirder is disposed to lower their credence in '*h*' from 1/2 down to 1/3 upon learning that they've awoken. On Sunday evening, before going to sleep, they know for sure that they will learn that they've awoken Monday morning. So they know for sure that after awaking on Monday morning, their credence in '*h*' will be lower than their current credence in '*h*', in violation of REFLECTION. Similarly, before going to sleep on Sunday, the Lewisian halfer is disposed to raise their credence in '*h*' to 2/3 upon learning that they've awoken and that it is Monday on Monday evening. So, on Sunday evening, they know for sure that Monday evening, their credence in '*h*' will be higher than their current credence in '*h*', in violation of REFLECTION. Finally, consider the double halfer on Monday morning, before learning what day it is. They know that they will either learn that it is Monday or that it is Tuesday. If they learn that it is Monday, then they are disposed to keep their credence in '*h*' fixed at 1/2. If they learn that it is Tuesday, they are disposed to lower their credence in '*h*' to zero. They think that it's 75% likely to be Monday,[56] so their expectation of their credence in '*h*' after learning is 3/8,[57] which is less than their current credence in '*h*', in violation of REFLECTION. So, no matter which of the dominant three positions we take on *Sleeping Beauty*, we will violate the principle of REFLECTION.

In the following subsection, I will show how the two-dimensional principle of deference I developed in §6 can be used to emend REFLECTION. This emended principle will allow us to escape both of ARNTZENIUS's counterexamples—though, in order to escape the second counterexample, it will require us to be thirders.

54. See LEWIS (2001)

55. See, for instance, HALPERN (2004), BOSTROM (2007), and MEACHAM (2008).

56. I am supposing that the double halfer thinks that, conditional on the coin landing tails, it's just as likely to be Monday as Tuesday (that is, I am supposing that they have the distribution shown in figure 1b). This doesn't ultimately matter, however. So long as their credence that it's Tuesday is greater than zero, their expectation of their credence in '*h*' after learning will be lower than 1/2.

57. Their expectation of their updated credence in heads is given by $D_{monday}(h) \cdot C(monday) + D_{tuesday}(h) \cdot C(tuesday) = 1/2 \cdot 3/4 + 0 \cdot 1/4 = 3/8$, where '$D_{monday}$' and '$D_{tuesday}$' are the credence functions they're disposed to adopt upon learning that it is Monday or Tuesday, respectively.

7.2 Two-Dimensional Reflection

Return to the general two-dimensional principle of deference 2DD from §6. Let the relevant expert be the credences you are disposed to adopt after you learn. Then, the principle tells us that, for any thought 'p', any locations 'λ' and 'ϵ', and any potential updated credence function D_e (where 'D_e' is the credence function you're disposed to adopt at t after you learn e), your credence in 'p', given that you are at λ, your future credences are given by D_e, and your future self is at ϵ, should be equal to D_e's credence in 'p_λ', given that you are at λ and they are at ϵ.

$$C(p \mid \mathcal{C}_t = D_e \wedge \lambda \wedge \epsilon_\epsilon) \stackrel{!}{=} D_e(p_\lambda \mid \lambda_\lambda \wedge \epsilon)$$

From this principle, it follows that, given you and your future self's locations, your current credence in 'p' should equal your conditional expectation of your updated credence that p, given your and your future self's location (where the conditional expectation is conditional on you and your future self's location).

$$\sum_{e \in \mathbf{E}} D_e(p_\lambda \mid \lambda_\lambda \wedge \epsilon) \cdot C(\mathcal{C}_t = D_e \mid \lambda \wedge \epsilon_\epsilon) \stackrel{!}{=} C(p \mid \lambda \wedge \epsilon_\epsilon)$$

That's a bit of a mouthful, but the principle simply says that, once any confusion about you and your future self's locations have been cleared up, your learning dispositions should be *unbiased*—in the sense that you shouldn't expect your updated credence in the *de dicto* surrogate of 'p' to be any higher or lower than your current credence in 'p'.

Fortunately, for my discussion here, we can simplify the principle somewhat. In all the cases I'll consider here, neither you nor your future self will suffer from any uncertainty about your *current* self's location, and you will be certain about which location your future self will occupy (though your future self may be uncertain about what location they occupy). In these special circumstances, the above equation entails:

(2D REFLECTION) $$\sum_{e \in \mathbf{E}} D_e(p_\lambda \mid \epsilon) \cdot C(\mathcal{D} = D_e) \stackrel{!}{=} C(p)$$

Recall the straightforward counterexample to REFLECTION: on Sunday, before going to sleep, I know for sure that it is Sunday. However, I also know for sure that tomorrow morning I'll learn that I've awoken, and upon learning that I've awoken, I'm disposed to be certain that 'It is Sunday' is *false*. These learning dispositions violate REFLECTION, but they do not violate 2D REFLECTION. Let λ be a potential location of your Sunday self—a location to which your Sunday self assigns positive credence. Assuming that your Monday morning self suffers from no ignorance about their location, 2D REFLECTION tells us that your current credence in 'It is Sunday' should equal your expectation of your Monday morning credence—*not* in 'It is Sunday'—but instead in the *de dicto* λ*-surrogate* of 'It is Sunday'. This surrogate says that the thought 'It is Sunday' expresses a truth when it is entertained by you at λ. Since λ implies that it is Sunday, it is *a priori* knowable that 'It is Sunday' expresses a truth at λ. So your Monday morning self will know for sure that the *de dicto* λ-surrogate of 'It is Sunday' is true. So your learning dispositions will not violate 2D REFLECTION in this case.

Next, consider the case from ARNTZENIUS in which you sit in a room without a clock, knowing that the lights will go off at 12:00 iff a fair coin landed heads. At 8:30,

when you know the time, your credence that the coin lands heads is 50%, or 1/2. Suppose that, at 11:30, after learning that somewhere between 2 and 4 hours have passed, you are disposed to be 37.5%, or 3/8, confident that the coin landed heads. In this case, we may suppose that, at 8:30, you know precisely what your location is. After learning, however, you will be uncertain about where you are located in time. Let ϵ be your current location, shifted forward to 11:30. Then, at 8:30, you know that, at 11:30, after you've learnt that between 2 and 4 hours have passed, ϵ will be true, but you won't be disposed to be certain that ϵ is true. And even though your 11:30 credence in '*h*' is 37.5%, your 11:30 credence in '*h*' *conditional on it being 11:30*, is 50%. So, even though you are disposed to lower your credence in '*h*' upon learning that between 2 and 4 hours have passed, you are *not* disposed to lower your credence in '*h*', *given your location*. And so your learning dispositions will not violate 2D REFLECTION in this case.

Finally, let us consider *Sleeping Beauty*. As we saw above, the thirder, the Lewisian halfer, and the double halfer all violate REFLECTION in this case. However, the thirder will satisfy 2D REFLECTION. To simplify things, let's suppose that, on Sunday, you know your location for sure. Then, the only *de se* uncertainty you'll suffer from upon awaking on Monday is uncertainty about what day it is. Let 'μ' be your Sunday location, shifted forward to Monday. You know for sure what your credences will be on Monday morning, since you know for sure that, on Monday morning, you will learn that you've awoken, *a*, and nothing else. So, if D_a is the credence function you're disposed to adopt upon learning that you've awoken, then 2D REFLECTION says that

$$D_a(h_\sigma \mid \mu) \stackrel{!}{=} C(h)$$

(where 'σ' is your (known) location on Sunday.) Since '*h*' is a *de dicto* thought, 'h_σ' = '*h*', and we can ignore the subscripted 'σ'. If you are a thirder, then you are disposed to adopt a credence function upon learning that *a* on Monday morning which is such that $D_a(h \mid \mu) = 1/2$. And on Sunday, you think the coin is just as likely to land heads as tails, $C(h) = 1/2$. So, when it comes to your Monday morning credences, the thirder will satisfy the principle 2D REFLECTION. (Since both the Lewisian and the double halfer sets their Monday morning credence in '*h*', conditional on it being Monday, to two thirds, $D_a(h \mid \mu) = 2/3$, neither the Lewisian nor the double halfer will satisfy 2D REFLECTION.)

Your Tuesday morning credences are more complicated, and more interesting. Before discussing them, however, let me consider a different case.[58] Colonel Aureliano Buendía faces the firing squad, the ranks of which are filled with men who once served under his command. He does not know whether these men are still loyal or not. However, in the minute after the order to fire is given, one of two things will be the case: either Colonel Buendía will be dead, in which case he won't have any credences about whether his men are loyal, or he will survive (perhaps because the men do not fire, perhaps because they all miss). If he survives, he will be nearly certain that the men are loyal—and rationally so. Learning that all of the men have either refrained from firing

58. I borrow this case from LESLIE (1989), who introduces it in his discussion of the 'fine-tuning' argument. My presentation of the case differs in superficial ways from Leslie's.

or have missed is strong evidence that those men are loyal.[59] This case might be seen as a problem for a principle like 2D REFLECTION, since Buendía has only one potential future credence function, and it is much more confident in the *de dicto* thought 'the men are loyal' than Buendía is currently. However, I think that, properly undestood, 2D REFLECTION does not apply in this case. For 2D REFLECTION simply presupposes that, no matter what happens, at the relevant future time, your learning dispositions will manifest themselves with the adoption of some credence function or other. Against the backdrop of this presupposition, it goes on to assert that, in expectation, these learning dispositions shouldn't be biased for or against any thought. However, if we are taking seriously the possibility of you not having any credences at all at the relevant time, then this presupposition is false. In that case, there will be potential outcomes in which your future credence is undefined. And if a quantity is undefined in some possibilities with positive credence, then you cannot take an expectation of that quantity. Nonetheless, you *can* take a *conditional* expectation of the quantity—you can ask what value you expect it to take on, conditional on it taking on a value at all. So we can ask about what credence in 'p' you expect to have after learning, conditional on your credences being defined. And we can say that this conditional expectation should match your current credence in 'p', conditional on your credences being defined.

Let '$\mathcal{C}_t \neq *$' say that your future credences are defined. Then, we should modify 2D REFLECTION so that it says that your expectation of your future credence that p, conditional on your future credences being defined, should equal your current credence that p, conditional on your future credences being defined. That is, if λ is your (known) location and ϵ is your future self's location, then for any thought 'p',

(2D REFLECTION*) $$\sum_{e \in \mathbf{E}} D_e(p_\lambda \mid \epsilon) \cdot C(\mathcal{C}_t = D_e \mid \mathcal{C}_t \neq *) \stackrel{!}{=} C(p \mid \mathcal{C}_t \neq *)$$

Now, let us return to the case of *Sleeping Beauty*, and consider again the credence you are disposed to adopt in 'Monday's coin flip lands heads' ('h') on Tuesday after learning that you've awoken. If the coin flip on Monday lands heads, then you will be asleep on Tuesday, and you will not be around to have any credences about the day or about whether the coin flip on Monday lands heads or tails. Let 'τ' be your (known) Sunday location shifted forward to Tuesday, and let '$\mathcal{C}_t$' be the definite description 'your Tuesday morning credences'. Then, you know for sure that, if $\mathcal{C}_t$ is defined, then you will have learnt that you are awake, a. So, if $\mathcal{C}_t$ is defined, there is only one potential credence function, D_a. And 2D REFLECTION* says that it should satisfy the following equality:

$$D_a(h \mid \tau) = C(h \mid \mathcal{C}_t \neq *)$$

Upon waking on Tuesday, your credence that the coin lands heads, given that it is Tuesday, will be 0%. And, on Sunday, your credence that the coin lands heads, given that you're around to have credences on Tuesday, will be 0%. So you will satisfy the extended principle 2D REFLECTION*.

59. At least, I'm inclined to take this as a datum, but there's been some dispute about this in the literature. See, for instance, SOBER (2005); but note that, in his 2009, Sober ends up changing his mind and agreeing that Buendía's survival is evidence of loyalty.

REFERENCES

Arntzenius, Frank. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy*, vol. 100 (7): 356–370. [11], [13], [39], [40], [41]

Bostrom, Nick. 2007. "Sleeping beauty and self-location: A hybrid model." *Synthese*, vol. 157: 59–78. [11], [40]

Braun, David. 2016. "The Objects of Belief and Credence." *Mind*, vol. 125 (498): 469–497. [15]

Briggs, R. A. 2009. "Distorted Reflection." *The Philosophical Review*, vol. 118 (1): 59–85. [38]

Chalmers, David J. 2006a. "The Foundations of Two-Dimensional Semantics." In *Two-Dimensional Semantics: Foundations and Applications*, M. Garcia-Carpintero & J. Macia, editors. Oxford University Press, Oxford. [2], [21], [26]

—. 2006b. "Two-Dimensional Semantics." In *Oxford Handbook of the Philosophy of Language*, E. Lepore & B. Smith, editors. Oxford University Press, Oxford. [2], [21], [26]

—. 2011. "Frege's Puzzle and the Objects of Credence." *Mind*, vol. 120 (479): 587–635. [15], [18], [35]

—. 2012. *Constructing the World*. Oxford University Press, Oxford. [26]

Davies, Martin & Lloyd Humberstone. 1980. "Two Notions of Necessity." *Philosophical Studies*, vol. 38 (1): 1–30. [2], [21]

Dorr, Cian. 2002. "Sleeping Beauty: In Defense of Elga." *Analysis*, vol. 62 (276): 292–296. [11]

Dorst, Kevin. forthcoming. "Evidence: A Guide for the Uncertain." *Philosophy and Phenomenological Research*. [32]

Dorst, Kevin, Benjamin A. Levinstein, Bernhard Salow, Brooke E. Husic & Branden Fitelson. ms. "Deference Done Better." [32]

Easwaran, Kenny. 2013. "Expected Accuracy Supports Conditionalization—and Conglomerability and Reflection." *Philosophy of Science*, vol. 80: 119–142. [7]

—. 2019. "Conditional Probabilities." In *Open Handbook of Formal Epistemology*, Richard Pettigrew & Jonathan Weisberg, editors, chap. 4, 131–198. [7]

Elga, Adam. 2000. "Self-locating belief and the Sleeping Beauty problem." *Analysis*, vol. 60 (2): 143–147. [3], [10], [11], [39], [40]

—. 2004. "Defeating Dr. Evil with Self-Locating Belief." *Philosophy and Phenomenological Research*, vol. 69 (2): 383–396. [11]

—. 2007. "Reflection and Disagreement." *Noûs*, vol. 41 (3): 478–502. [38]

—. 2013. "The puzzle of the unmarked clock and the new rational reflection principle." *Philosophical Studies*, vol. 164: 127–139. [32]

Evans, Gareth. 1979. "Reference and Contingency." *The Monist*, vol. 62 (2): 161–189. [2], [21]

Fitts, Jesse. 2014. "Chalmers on the Objects of Credence." *Philosophical Studies*, vol. 170 (2): 343–358. [15]

Gaifman, Haim. 1988. "A Theory of Higher Order Probabilities." In *Causation, Chance, and Credence: Proceedings of the Irvine Conference on Probability and Causation*, Brian Skyrms & William L. Harper, editors, vol. 1, 191–220. Kluwer Academic Publishers, Dordrecht. [31]

Gibbard, Allan. 2012. *Meaning and Normativity*. Oxford University Press, Oxford. [15]

Hall, Ned. 1994. "Correcting the Guide to Objective Chance." *Mind*, vol. 103 (412): 505–517. [32]

Halpern, Joseph Y. 2004. "Sleeping Beauty Reconsidered: Conditioning and reflection in asynchronous systems." In *Proceedings of the Twentieth Conference on Uncertainty in AI*, 226–234. [11], [40]

Hawthorne, John & Maria Lasonen-Aarnio. 2009. "Knowledge and Objective Chance." In *Williamson on Knowledge*, Patrick Greenough & Duncan Pritchard, editors, 92–108. Oxford University Press, Oxford. [2], [4]

Hitchcock, Christopher. 2004. "Beauty and the Bets." *Synthese*, vol. 139: 405–420. [11]

Hitchcock, Christopher & Mitchell S. Green. 1994. "Reflections on Reflection: van Fraassen on Belief." *Synthese*, vol. 98 (2): 297–324. [38]

Horgan, Terrence. 2004. "Sleeping Beauty awakened: New odds at the dawn of the new day." *Analysis*, vol. 64: 10–24. [11], [12]

Ismael, Jeann. 2008. "Raid! Dissolving the Big, Bad Bug." *Noûs*, vol. 42 (2): 292–307. [31]

—. 2015. "In Defense of IP: A Response to Pettigrew." *Noûs*, vol. 49 (1): 197–200. [31]

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press, Oxford. [2], [21]

Kaplan, David. 1978. "Dthat." In *Syntax and Semantics*, Peter Cole, editor, 221–243. Academic Press, New York. [2]

—. 1989. "Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and other Indexicals." In *Themes From Kaplan*, John Perry Joseph Almog & Howard Wettstein, editors, 481–563. Oxford University Press, Oxford. [2], [21]

Kripke, Saul. 1979. "A Puzzle About Belief." In *Meaning and Use*, A. Margalit, editor, vol. 3, 239–283. Springer, Dordrecht. [13]

Leslie, John. 1989. *Universes*. Routledge, London. [42]

Lewis, David K. 1979. "Attitudes De Dicto and De Se." *The Philosophical Review*, vol. 88 (4): 513–543. [17], [21], [22], [27]

—. 1980. "A Subjectivist's Guide to Objective Chance." In *Studies in Inductive Logic and Probability*, Richard C. Jeffrey, editor, vol. II, 263–293. University of California Press, Berkeley. [3], [4], [5], [6], [7], [12], [16], [17], [20]

—. 1994. "Humean Supervenience Debugged." *Mind*, vol. 103 (412): 473–490. [32]

—. 1999. "Why Conditionalize?" In *Papers in Metaphysics and Epistemology*, vol. 2, chap. 23, 403–407. Cambridge University Press, Cambridge. [3]

—. 2001. "Sleeping Beauty: reply to Elga." *Analysis*, vol. 61 (3): 171–176. [3], [11], [40]

Meacham, Christopher J. G. 2008. "Sleeping Beauty and the Dynamics of De Se Belief." *Philosophical Studies*, vol. 138 (2): 245–269. [11], [40]

—. 2016. "Ur-Priors, Conditionalization, and Ur-Prior Conditionalization." *Ergo*, vol. 3 (17). [8]

Monton, Bradley. 2002. "Sleeping Beauty and the Forgetful Bayesian." *Analysis*, vol. 62 (1): 47–53. [11]

Nolan, Daniel. 2016. "Chance and Necessity." *Philosophical Perspectives*, vol. 30 (1): 294–308. [2], [4], [37]

Salmón, Nathan. 2019. "Impossible Odds." *Philosophy and Phenomenological Research*, vol. 99 (3): 644–662. [2], [4], [37]

Salow, Bernhard. 2018. "The Externalist's Guide to Fishing for Compliments." *Mind*, vol. 127 (507): 691–728. [38]

Schulz, Moritz. 2011. "Chance and Actuality." *Philosophical Quarterly*, vol. 61 (242): 105–129. [2], [4]

Schwarz, Wolfgang. 2014. "Proving the Principal Principle." In *Chance and Temporal Asymmetry*, Alistair Wilson, editor, 81–99. Oxford University Press. [16], [17], [19]

Singer, Daniel Jeremy. 2014. "Sleeping beauty should be imprecise." *Synthese*, vol. 191 (14): 3159–3172. [11]

Skyrms, Brian. 1980. "Higher Order Degrees of Belief." In *Prospects for Pragmatism*, D. H. Mellor, editor, chap. 6, 109–137. Cambridge University Press. [31]

Sober, Elliot. 2005. "The Design Argument." In *Blackwell Guide to the Philosophy of Religion*, William E. Mann, editor, chap. 6, 117–147. Blackwell Publishing, Malden, MA. [43]

—. 2009. "Absence of evidence and evidence of absence: evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads." *Philosophical Studies*, vol. 143 (1): 63–90. [43]

Spencer, Jack. 2020. "No Crystal Balls." *Noûs*, vol. 54 (1): 105–125. [16], [18], [19]

Stalnaker, Robert C. 1978. "Assertion." In *Syntax and Semantics 9: Pragmatics*, P. Cole, editor, 315–332. Academic Press, New York. [2], [21]

Talbott, W. J. 1991. "Two Principles of Bayesian Epistemology." *Philosophical Studies*, vol. 62: 135–150. [38]

Titelbaum, Michael G. 2012. "An Embarrassment for Double-Halfers." *Thought*, vol. 1 (2): 146–151. [2]

van Fraassen, Bas C. 1984. "Belief and the Will." *The Journal of Philosophy*, vol. 81 (5): 235–256. [10], [13], [37], [38]

—. 1995. "Belief and the Problem of Ulysses and the Sirens." *Philosophical Studies*, vol. 77: 7–37. [10], [37]

Weintraub, Ruth. 2004. "Sleeping Beauty: the Simple Solution." *Analysis*, vol. 64 (1): 8–10. [11], [12]

Yli-Vakkuri, Juhani & John Hawthorne. forthcoming. "Intensionalism and Propositional Attitudes." *Oxford Studies in the Philosophy of Mind*. [15]