

WHEN ARE CONTEXT-SENSITIVE BELIEFS RELEVANT?

MICHAEL TITELBAUM

Draft of March 24, 2006.

Please do not circulate or cite without permission.

Abstract: I begin with the question of whether self-locating beliefs can be relevant to beliefs *de dicto*. Can it ever be rational for an agent to change her degree of belief in a belief *de dicto* as a result of gaining only self-locating information? To analyze this question, I develop a formal technique for modeling the ideally rational evolution of an agent's degrees of belief as she learns new information over time. This modeling technique improves upon traditional conditionalization-based techniques by being general enough to correctly model stories involving context-sensitive beliefs. It also makes explicit the assumptions that go into a model and leaves no question what that model says about a given story. These improvements are achieved by devoting special attention to the language over which a model is defined; the central innovation of the technique is a principle for determining whether the verdicts of a model change when the modeling language expands. As I develop the modeling technique, I apply it to a number of stories — including Arntzenius's "Shangri-La" story involving the threat of cognitive mishap — and show that it yields intuitive results in obvious cases. I then apply the model to the Sleeping Beauty Problem and show that David Lewis's solution to the problem is incorrect. This result also demonstrates that self-locating beliefs can be relevant to beliefs *de dicto*. In a class of cases I describe, it is rational for an agent who learns only self-locating beliefs to adjust her relative degrees of belief in a set of beliefs *de dicto* all of which are logically compatible with what she has learned.

Introduction. How much do I learn when I learn what time it is, or where I am, or who I am? Beliefs about one's spatio-temporal location in the universe and beliefs about one's identity are often called "self-locating beliefs," and it is uncontroversial that a self-locating belief can be relevant to other self-locating beliefs. When I learn that today is Monday, it is rational to respond by adjusting my degree of belief that tomorrow is Tuesday. But is it ever rational to adjust one's degree of belief in a belief *de dicto* in response to learning a self-locating belief?¹

¹Here I am adopting David Lewis's terminology in (Lewis 1979), while altering it at one point. Lewis takes all beliefs to be self-ascriptions of properties, and therefore describes all beliefs as "beliefs *de se*." Among beliefs *de se* are beliefs *de dicto*, which limit the class of possible worlds to which their subject might belong. Lewis describes beliefs *de dicto* as locating their subject within "logical space." On the other hand, there are "irreducibly *de se* beliefs," which indicate the

We may have an intuition that the answer is no. For instance, as I move through the world it is inevitable that my spatio-temporal location will change. Learning that I have moved from one spatio-temporal position to another within the actual world should not alter my opinions about which possible world is actual. I glance at my watch; I see it is noon; five minutes later I glance again and see it is 12:05. Time marches on; I knew this would happen; it is not rational for self-locating beliefs marking my personal spatio-temporal progress to affect my degrees of belief in beliefs *de dicto* about the wider world.

Such intuitions might lead us to endorse the

Strong Relevance-Limiting Thesis: It is never rational for an agent who learns a self-locating belief to respond by altering her degree of belief in a belief *de dicto*.

Yet it is clear that the wall between self-locating beliefs and beliefs *de dicto* suggested by the Strong Relevance-Limiting Thesis can break down. To take a melodramatic example, if I glance at my watch while in a bunker that has been under continuous artillery fire, it might be quite a surprise to see that I have made it to 12:05. At noon I may have entertained as real possibilities a number of worlds in which I never see 12:05. So it may be quite rational for me to alter my degrees of belief in a number of beliefs *de dicto* when I learn the self-locating belief that it is now 12:05.

We need a more subtle position about the relevance of self-locating beliefs to beliefs *de dicto* than the Strong Relevance-Limiting Thesis.² A good first step is to admit that self-locating beliefs can rule out possible worlds with which they are incompatible. There are questions that should be asked at this step, for instance about the precise notion of “compatibility” involved. But even if we assume an available notion of “compatibility” clear enough for rough-and-ready use, a further step seems required. Admittedly, learning a self-locating belief can eliminate certain beliefs *de dicto* from the running. But can it also alter rational degrees of belief in the beliefs *de dicto* that remain?

Suppose that at noon my sergeant flipped a fair coin to decide whether to let me or another private off duty at 12:05. At noon I don’t yet know the outcome of the coin flip, so I entertain three possibilities: I won’t make it to 12:05; I will make it and be relieved; I will make it and not be relieved. Assuming I know my own identity, beliefs in all three possibilities are beliefs *de dicto*. Presumably, if rational I assign equal degrees of belief to the latter two possibilities. When I learn it is 12:05, the first possibility is eliminated. My degrees of belief in each of the remaining two increase, but isn’t it rational for me to continue to consider each of them equally likely? While the effects of self-locating beliefs on degrees of belief in beliefs *de dicto* are greater than we first thought, perhaps they are still limited as

subject’s identity or location in time or space—that is, the subject’s location within the possible world in which she dwells. Lewis thus takes all beliefs to be “self-locating beliefs.” I, on the other hand, reserve the term for those beliefs Lewis describes as “irreducibly *de se*.”

²Given all the technical definitions that appear in this paper, one might wonder whether I’m going to define the concept of “relevance” that appears from time to time. For the purposes of this paper, I prefer to treat relevance as an intuitive, pre-theoretical notion that answering more specific questions might give us some purchase on. Roughly speaking, relevance has to do with whether one piece of information bears upon another in a particular situation. I hope that this suggestion, along with context, will make my uses of the concept sufficiently comprehensible in what follows.

follows: self-locating beliefs have the ability to rule out beliefs *de dicto* altogether, but they cannot affect rational *relative* degrees of belief among the beliefs *de dicto* that remain.

This suggests the

Weak Relevance-Limiting Thesis: It is never rational for an agent who learns a self-locating belief to respond by altering her relative degrees of belief in two beliefs *de dicto* compatible with the self-locating belief.

(Notice that the Weak Relevance-Limiting Thesis is strictly weaker than the Strong Relevance-Limiting Thesis.)³

How can we evaluate proposals (such as the Weak Relevance-Limiting Thesis) concerning the relevance of self-locating beliefs to other types of beliefs? For a few decades now, formal epistemologists have been developing techniques for modeling rational degrees of belief. Included among these techniques are a number of quite precise principles for modeling the rational response to gaining new evidence. When we are curious how specific types of evidence can affect rational degrees of belief, these principles ought to be our guides.

Unfortunately, the principles for rational belief revision traditionally accepted by formal epistemologists systematically fail when self-locating beliefs get involved. This is because self-locating beliefs may be context-sensitive in the situation under consideration. While the broader epistemology community has been concerned with context-sensitive beliefs for some time, context-sensitivity has become a major issue in formal epistemology only over the last few years. This development has been prompted in large part by puzzlement over a story called the Sleeping Beauty Problem.

In this paper I present a formal technique for modeling rational degrees of belief as they change over time. This technique is general enough to accurately model rational degrees of belief in both context-sensitive and context-insensitive beliefs. I begin by introducing the basic framework of my models and some principles for modeling synchronically rational agents within that framework. I then discuss the most popular principle for modeling diachronically rational agents, updating by conditionalization, and explain why it yields inaccurate results for situations involving context-sensitive beliefs. I offer new principles for modeling diachronic rationality to take conditionalization's place, then show that they give the obviously correct results in some simple examples, thereby giving intuitive reason to think the principles are correct. I also offer thoughts about why these principles work, and explain how to use them to analyze situations of various types.

Finally, I present the Sleeping Beauty Problem and analyze it using my modeling technique. I argue that the controversial part of the problem can be solved without appeal to frequencies, objective chances, epistemic indifference, utilities, or Dutch Books. The correct approach to modeling context-sensitive beliefs is sufficient by itself to demonstrate that upon awakening and learning some self-locating

³With the slight revision I'll propose in Section 1.8 below, the Weak Relevance-Limiting Thesis is a consequence of the belief dynamics Christopher Meacham proposes in (Meacham manuscript). Meacham notes that Joseph Halpern (2005) defends a dynamics very similar to his own. Halpern's view may also support the Weak Relevance-Limiting Thesis, though it's difficult to draw such a general moral from (Halpern 2005) because the discussion there is limited to cases in which the agent is uncertain of the time. It is also possible that Lewis at one point subscribed to something like the Weak Relevance-Limiting Thesis; I address this issue in a note in Section 2.1 below.

information, Beauty should decrease her degree of belief in a belief *de dicto* that a particular coin flip came up heads. This in turn shows that in certain situations it can be rational for an agent, upon learning a self-locating belief, to alter her relative degrees of belief among a set of beliefs *de dicto* all of which are compatible with what she has learned. Thus even though we will improve the theses in an important respect along the way,⁴ in the end I will conclude that both the Strong and the Weak Relevance-Limiting Theses are false.

1. THE MODELING TECHNIQUE

1.1. Basic Framework. We will investigate the relevance-limiting theses by analyzing hypothetical scenarios I call “stories.” A story describes an agent who starts off with a particular set of beliefs and then learns pieces of information at various times over the course of the story. Beliefs learned by agents in stories are then believed with absolute certainty. Each story will pose a question about one of the agent’s degrees of belief at some moment during the story, or about relationships between her degrees of belief. The goal is to determine what would be true about those degrees of belief were they ideally rational.

I say “*ideally* rational” because the standards we will be examining may be stronger than are required for ordinary rationality. Our intuitions about what is required for rationality will also apply to ideal rationality, but it may be that we would be willing to call an agent rational even though she didn’t meet the requirements of ideal rationality. Ideal rationality requires that the set of beliefs to which an agent assigns degrees of belief be closed under the truth-functional connectives, and that an agent be certain of any belief logically implied by other beliefs of which she is certain. (The latter requirement is sometimes called “logical omniscience.”) Yet we might be willing to call an agent rational whose doxastic state did not have these features.

I will not assume that evaluative statements about what it would be ideally rational to believe lead in any straightforward fashion to statements normative for agents. My argument will proceed in the opposite direction. The Weak Relevance-Limiting Thesis holds that it is never *rational* for an agent who learns a self-locating belief to respond by altering her relative degrees of belief in beliefs *de dicto* with which it is compatible. I assume that a defender of the Weak Relevance-Limiting Thesis would also hold that it is never *ideally rational* for an agent to do so. Thus when, in the latter half of this paper, we demonstrate of an agent in a particular story (the Sleeping Beauty Problem) that it would be ideally rational for her to alter her relative degrees of belief in some beliefs *de dicto* in response to learning self-locating information, I take it this refutes the Weak Relevance-Limiting Thesis (and therefore the Strong).⁵

⁴See Section 1.8 below.

⁵It might be objected that I am placing my agents in a situation that makes rationality — much less ideal rationality — impossible. A position called “Regularity,” for example, holds that it is never rational to assign absolute certainty to an empirical belief, and many of the beliefs agents become certain of in the stories we will examine are empirical. (For references and discussion on Regularity see Section 4.5.4 of (Hájek 2003).) If ideal rationality is to be strictly stronger than rationality, it might be objected that it could never be ideally rational to take these claims for certain. While I’m not sure we need to maintain Regularity when analyzing hypothetical stories constructed to test particular philosophical theses, my hope is that even those who want to maintain Regularity in these cases will be willing to see that issue as independent of the relevance

To determine the ideally rational degrees of belief for a story, we will create formal models of the agent’s degrees of belief.⁶ The models we create will issue in verdicts, which I take to be necessary conditions for ideal rationality. I do not presume in what follows to have captured all the requirements of ideal rationality; there may turn out to be belief states compatible with our verdicts that would not be ideally rational to maintain. Thus I understand these verdicts as necessary but not sufficient conditions for ideal rationality.

Our general modeling strategy will be to represent the agent’s degrees of belief at a particular time as numerical values of a “credence function.” (A numerical value of a credence function is called a “credence.”) In the formal epistemology literature, a credence function is usually defined over a set of propositions. Yet it can be ideally rational to have different degrees of belief in two beliefs that might be taken to express the same proposition. Our modeling technique should be flexible enough to represent such a situation. For example, many theories of propositions take “Princess Aurora lives in the forest with the fairies” and “Briar Rose lives in the forest with the fairies” to express the same proposition; in that case, a credence function defined over propositions would make it a requirement of ideal rationality that an agent assign these two the same degree of belief. Similarly, some theories of propositions would take “I am here now” and “Marty McFly is in Hill Valley in 1955” to express the same proposition in a specific context, even though it might be ideally rational to assign them different degrees of belief.

I designed the modeling technique presented here to be both flexible and compatible with as many positions in other areas of philosophy as possible. For example, it can be usefully employed regardless of one’s position on whether there are “centered propositions” and regardless of one’s position on whether the objects of an agent’s doxastic states are propositions, sentences, self-ascriptions of properties, or something else. To maintain neutrality I will from this point forward describe possible objects of an agent’s doxastic states as “claims” and assume that whatever else their status, the four sentences quoted in the previous paragraph represent distinct claims. Instead of defining our credence functions over sets of propositions, I will define them over sets of sentences, and take distinct claims (potential objects of an agent’s doxastic states) to be represented by distinct sentences. This allows us to model stories in which it is ideally rational for an agent to assign different degrees of belief to two claims despite those claims’ expressing the same proposition.⁷

Whenever one creates a model of a situation, one begins with certain assumptions about the situation that are taken as fixed background points against which the model is defined. The aim of the model is not to question these assumptions; it is to draw out their implications. If someone doubts one of the assumptions, this is not a criticism of the modeling technique used to draw out those assumptions’ implications. Rather, it is a claim that the modeling technique should have started from a different point and thus provided a different model. Such discussions

relations we are after. Even if our stories’ requiring agents to take learned claims for certain makes those agents irrational in one way, my hope is that those stories can nevertheless serve as reliable test cases for the rather different issues brought up by the relevance-limiting theses.

⁶Note that I am using the term “model” in the scientist’s sense of a system that represents given data and makes predictions about what is not given, as opposed to the logician’s or mathematician’s sense of a set of objects that interprets a formal system.

⁷For another, rather different benefit of defining credence functions over sets of sentences, see (Fitelson manuscript).

demonstrate the particular importance of being explicit about the assumptions one is making going into the process of constructing a model.

The purpose of our modeling technique is to help us determine ideally rational degrees of belief for agents to assign under conditions of uncertainty. I am going to presume that for a wide variety of claims related to a story, we can use the details of that story to determine *prior to applying the modeling technique* whether or not it is ideally rational for the agent to be completely certain of that claim at a particular time. Judgments made (perhaps with the help of some deductive logic) about whether it is ideally rational for an agent in a story to be certain of a particular claim at a particular time will supply assumptions against which our models will be defined. (There will be other assumptions as well, but these will figure prominently.) The goal of those models is to yield more specific verdicts concerning the claims about which the agent is uncertain. There may be disagreement about the verdicts of our models because of disagreement about whether it would be ideally rational for a particular agent to take a particular claim for certain at a particular time. Such cases demonstrate the usefulness of the modeling technique in highlighting the precise source of disagreement among parties about particular verdicts.⁸

Another decision we have to make before constructing a model to represent a story is the choice of a “modeling language.” In any story, there will be a wide variety of claims to which the agent might assign degrees of belief. Yet our model will concern only a subset of these claims. Before constructing a model, we will explicitly delimit a set of sentences (representing claims in the story) that will be assigned credence values on our model. These will of course include sentences representing the claims asked about in the question posed by the story. But we will typically represent other claims as well, especially claims the agent learns over the course of the story. By explicitly delimiting the set of claims represented in the model at the start, we run the risk of failing to take into account the effects of some claims relevant to the degrees of belief asked about in the story. We also run the risk of prejudging one of the questions we want to use our models to investigate: what sorts of claims can be relevant to what others. These issues concerning language choice and the effects of increased language complexity will be set aside for now, then taken up again in Section 1.6 below.⁹

Once a modeling language has been specified, there will be two kinds of further constraints on the models we use to represent stories. First, there will be what I call “systematic constraints.” These constraints are parts of the modeling technique itself. They are common to all models built with this technique, whatever story they represent. I take them to represent general constraints on ideally rational degrees of belief, constraints based on considerations of consistency and relevance. Second, there are “extrasystematic constraints.” These constraints are

⁸Mark Lance argues in (Lance 1995) that a Bayesian model — indeed, *any* type of explicit decision-theoretic model — of a story must always work within a structure of empirical claims the agent is assumed to accept prior to the application of the model.

⁹One preliminary note on the complexity of modeling languages: The modeling languages employed in our models will begin with a finite set of atomic sentences and then be built into countably infinite sets by recursively joining these atomic sentences via truth-functional connectives. If there are stories for which the requirements of ideal rationality cannot be represented by modeling languages of this type — for instance, if there are stories that must be represented by uncountably infinite models — such stories cannot be modeled with the modeling technique as it is presented here.

not a built-in part of the modeling technique, and they are applied on a story-by-story basis. Most of the extrasystematic constraints we apply below will represent assumptions about whether it is ideally rational for an agent to be certain of a particular claim at a particular time in a particular story. For every sentence in the modeling language and every moment under consideration in the story, there will be an extrasystematic constraint stating either that the agent is required to be certain of the claim represented by that sentence at that moment or stating that the agent is required not to be certain of that claim at that moment. We will also consider some extrasystematic constraints based on what might be considered general requirements on ideal rationality. For example, when we restrict a model of a particular story according to the Principal Principle or an Indifference Principle, we will do so by applying an extrasystematic constraint. This is not to suggest that such principles, if correct, are any less important or general than those represented by the systematic constraints; it is just that they are not built into this modeling technique.¹⁰

Even considered all together, the systematic and extrasystematic constraints will often be insufficient to require that an agent assign one precise degree of belief to a particular claim at a particular time. It may be that the constraints rule out degrees of belief outside a particular range, but are equally compatible with any degree of belief in that range. At the same time, it may happen that while the constraints do not require for ideal rationality that the agent have a particular precise degree of belief at a particular time, they require the agent's later degrees of belief be related to her degrees of belief at that time in a very precise fashion. For this reason, each model will contain a set of "histories," and the set will usually have more than one element. Each history will assign a precise credence to each sentence in the modeling language at every time represented by the model. The verdicts issued by the model will stipulate that it is a necessary requirement of ideal rationality that the agent not assign degrees of belief represented by credence functions outside the histories contained in the model. As far as the requirements represented by the model go, any history in the model represents an ideally rational way for the agent to develop her degrees of belief over time.¹¹

¹⁰As we will see below, the decision to leave these principles out of the systematic constraints is not unmotivated. It is important that the formal structures involved in this modeling technique be precisely defined, so that (for instance) there is never any question precisely what values are assigned to particular sentences by particular models. Since the systematic and extrasystematic constraints play a role in defining the models, those constraints must be articulated precisely within the formal system of the modeling technique. At the time of this writing, I do not consider the Principal Principle or any Indifference Principle to have been formulated precisely enough that they could be articulated formally as general systematic constraints within the modeling technique. As a result, I prefer to use such principles informally to supply one-off judgments in precisely defined, non-controversial applications, then implement those individual judgments formally as extrasystematic constraints.

¹¹The presence of a range of histories in a model should not be read as a positive requirement of ideal rationality that an agent adopt a *vague* degree of belief in a particular claim represented by credence values ranging over the entire range of credences assigned to the corresponding sentence by the credence functions appearing in the histories. For example, a model may contain histories with credence functions assigning a particular sentence credences in the range 0.2 through 0.5 at a particular time. However, there may be a further requirement of ideal rationality not represented by the model that restricts the permissible set of degrees of belief in the corresponding claim at that time to degrees represented by credences higher than 0.4. Recall that the verdicts of our models represent necessary but not sufficient requirements for ideal rationality.

1.2. Formal System and Notation. With a general understanding of the framework in place, we'll now delve into the gritty details of the formal system and its notation. A **model** will be defined as a formal structure with the following components:

Each model represents a particular story. A model is defined relative to a specific **modeling language**, which is an infinite set of sentences. We specify a modeling language by giving a non-empty, finite set of **atomic sentences**. A **sentence** belongs to the modeling language just in case it is either an atomic sentence of that language or is built from other sentences of the language using the connectives \vee , \sim , $\&$, \supset , or \equiv according to the usual recursive rules. (We will also use parentheses at times for increased readability.) A model is also defined relative to a non-empty, finite **time sequence**. The time sequence is a sequence of moments (t_1, t_2, \dots, t_n) during the story, arranged in temporal order.

Given a time sequence and a modeling language, we define a set of **histories**. Each history contains a sequence of **unconditional credence functions** $P_{t,h}(\cdot)$. The first index indicates a moment in the time sequence, while the second index indicates the history to which the function belongs. An unconditional credence function takes a sentence in the modeling language as its argument and outputs a real number. Each history also contains a sequence of **conditional credence functions**. The indices for these are the same, the argument is an ordered pair of sentences in the modeling language, and the output is a real number.

A model represents its story as follows: A sentence X in the modeling language represents a claim concerning some aspect of the story. The connectives in the modeling language represent truth-functional connectives applied to claims. For some given h , a sequence of values $P_{1,h}(X), P_{2,h}(Y), \dots$ in the model represents a sequence of degrees of belief the agent might have in the claims represented by X, Y, \dots , etc. at times t_1, t_2, \dots during the story. Sequences of conditional credence values have similar interpretations, except that a value $P_{k,h}(X|Y)$ represents a degree of belief the agent might have at time t_k in the claim represented by X *conditional* on the supposition of the claim represented by Y .

From this point forward, we will usually be uninterested in the differences between histories in a given model; instead, we will be interested in the features that all the histories in a given model share. For this reason, we will have little use for the second index in a credence function's subscript. Thus we introduce a conventional shorthand: Any statement written without the second index will be read as being universally quantified over the history values available in the model. For example, $P_1(X) > P_2(Y)$ is shorthand for $\forall(h)[P_{1,h}(X) > P_{2,h}(Y)]$. Similarly, if I write, " $P_4(X)$ must be non-zero," I mean "In all histories h in the model under consideration, $P_{4,h}$ must be non-zero."

An **arithmetic statement** is formed by taking an equation or inequality joining expressions composed by applying arithmetic functions to real numbers and/or specific credence values within the same history, then universally quantifying it over histories. For example,

$$(1) \quad \forall(h)[P_{1,h}(\text{I like cheese}) + 0.5 \leq P_{2,h}(\text{I like cheese})^2 + P_{3,h}(\text{I like meat})]$$

is an arithmetic statement, which would be abbreviated in our shorthand as

$$(2) \quad P_1(\text{I like cheese}) + 0.5 \leq P_2(\text{I like cheese})^2 + P_3(\text{I like meat})$$

An arithmetic statement can contain no variables other than the universally quantified h variable. All credence expressions that occur in an arithmetic statement must be indexed to the history variable universally quantified over in the sentence. An arithmetic statement can contain no truth-functional connectives outside the arguments of credence functions.

A **verdict** of a model is an arithmetic statement that is true of the histories in that model. A verdict of a model represents a requirement of ideal rationality on the degrees of belief of the agent in the story.

The credence functions in each history of each model are subject to systematic constraints which are common to all models. Formulating the systematic constraints will be the business of the remaining sections of this first half of the paper; ultimately there will be six systematic constraints. Each model also contains extrasystematic constraints, which are common to all histories in the model but vary from model to model. The extrasystematic constraints take the form of arithmetic statements. The extrasystematic constraints are therefore also verdicts of the model (though typically not very interesting ones).

The set of histories belonging to a model M is the set of all possible histories consistent with the systematic and extrasystematic constraints on M . Thus, given a modeling language and a time sequence, the systematic and extrasystematic constraints on M completely define the histories in M . With a bit of algebra we can derive a model's verdicts from the systematic and extrasystematic constraints on that model.

To summarize, with the systematic constraints common to all models assumed in the background, we will be able to completely define a model by specifying:

- The story.
- A modeling language \mathbf{L} .
- A time sequence (t_1, t_2, \dots, t_n) .
- A set of extrasystematic constraints in the form of arithmetic statements.

When I initially describe a model, I will state the story it represents, the atomic sentences of the modeling language (along with the claims they represent), the time sequence, and some of the extrasystematic constraints. I will not go through the task of stating all the extrasystematic constraints, even where that is possible; I will typically state only those extrasystematic constraints that will be pertinent to our analysis. Finally, while these are deducible from the other elements in the description and so technically redundant, I will also for ease of reference include in the initial description of a model certain “learned information sets” common to all its histories. “Learned information sets” will be defined in Section 1.3 below. At the end of that section we will also have enough pieces in place to give a relatively simple example of a model.

A note on notation: Throughout what follows, I will use **this** font when naming a model. I will use **bold** font to name a set of sentences, and to demarcate when I am defining a new technical term. Atomic sentences of a language will be named by capital letters. $P_{t,h}(\dots)$ will always be a credence function, usually represented in our shorthand by $P_t(\dots)$. Finally, when multiple models are under discussion at once, I will use superscripts to indicate what goes with what. For example, the model M^+ is defined relative to the modeling language \mathbf{L}^+ and contains histories with unconditional credence functions denoted $P_{t,h}^+(\cdot)$.

I will sometimes use logical terms to describe relations between sentences in a modeling language. For example, I may say that two sentences X and Y are “logically equivalent”. Since sentences are just formal elements used by a model, they have no content, and so cannot actually be logically equivalent. Thus what I mean when I say that two sentences are logically equivalent is that they bear a particular formal relation that can be defined by recursive rules involving the connectives that appear in the sentences. Since the connectives in the modeling language represent the truth-functional connectives, I will not bother to lay out those recursive rules here; they can be adapted in an obvious way from classical propositional logic.

Classical propositional logic will be the *only* source for statements made about logical features of sentences in our models. Thus statements about relations of logical equivalence, entailment, mutual exclusivity, etc. between sentences, and statements that certain sentences are tautologies or contradictions, will depend on classical propositional logic. (Note that by the definition of a modeling language, every modeling language contains both tautologies and contradictions.) It may be that two sentences X and Y represent *claims* that are logically equivalent by, say, first-order logic but not propositional logic. Nevertheless, the sentences X and Y will not be taken to be logically equivalent within our models.

When we said in Section 1.1 that ideal rationality requires an agent to be certain of every claim logically implied by claims she is certain of, this included logical implication relations beyond propositional implication. So when the claim represented by a sentence X is logically equivalent to a claim represented by a sentence Y by some form of logical implication beyond propositional logic, we will want to represent the results of that equivalence for ideally rational degrees of belief. We will do so with extrasystematic constraints on our models, for example extrasystematic constraints representing the agent’s certainty at all times that the claim represented by X is true if and only if the claim represented by Y is true.

Finally, one additional shorthand that will prove useful: If we have a set of sentences $\mathbf{I} \subseteq \mathbf{L}$ and a sentence $X \in \mathbf{L}$, we will sometimes write $P_k(X | \mathbf{I})$. Technically speaking, this is improper, as a conditional credence function can take only a sentence as its second argument. So we will understand $P_k(X | \mathbf{I})$ to be an abbreviation for the following: For any nonempty $\mathbf{I} \subseteq \mathbf{L}$, there exists a sentence $I^* \in \mathbf{L}$ such that I^* is logically equivalent to the conjunction of all the sentences in \mathbf{I} . When we write $P_k(X | \mathbf{I})$ for a nonempty \mathbf{I} , we really mean $P_k(X | I^*)$ for some such I^* . If \mathbf{I} is empty, then $P_k(X | \mathbf{I}) = P_k(X | \mathbf{T})$, where \mathbf{T} is some tautological sentence in \mathbf{L} .

1.3. Synchronic Constraints. The remainder of the first half of this paper concerns systematic constraints on our models designed to ensure that they reflect the requirements of ideal rationality. We begin with “synchronic constraints” representing the way various degrees of belief held by an agent at the same time are required to cohere with each other in a rational fashion.

The first three synchronic constraints, Kolmogorov’s axioms, guarantee that each unconditional $P_{t,h}(\cdot)$ is a **probability function**. Given a modeling language \mathbf{L} and a time sequence (t_1, t_2, \dots, t_n) , the constraints are:

Systematic Constraints (1)-(3), Kolmogorov Axioms:

- (1) For any $t_k \in \{t_1, t_2, \dots, t_n\}$ and any sentence $X \in \mathbf{L}$, $P_k(X) \geq 0$.
- (2) For any $t_k \in \{t_1, t_2, \dots, t_n\}$ and any tautological sentence $\mathbf{T} \in \mathbf{L}$, $P_k(\mathbf{T}) = 1$.

- (3) For any $t_k \in \{t_1, t_2, \dots, t_n\}$ and any mutually exclusive sentences $X, Y \in \mathbf{L}$,
- $$P_k(X \vee Y) = P_k(X) + P_k(Y)$$

These axioms have the effect of restricting all credence values to the range $[0, 1]$. A higher credence in sentence X than in sentence Y represents a higher degree of belief by the agent in the claim represented by X . A credence of 1 in X represents absolute certainty at a particular time in the claim represented by X , while a credence of 0 in X represents absolute certainty in the negation of the claim represented by X . The axioms have various intuitively rational consequences, for example that at any given time the same credence will be assigned to any two logically equivalent sentences.

The next synchronic constraint, what I call the traditional definition of conditional probability, relates conditional probabilities to unconditional probabilities. Given a modeling language \mathbf{L} and a time-sequence (t_1, t_2, \dots, t_n) , the constraint is:

Systematic Constraint (4),

Traditional Definition of Conditional Probability:

For any $X, Y \in \mathbf{L}$ and any $t_k \in \{t_1, t_2, \dots, t_n\}$,

$$P_k(X | Y) = \frac{P_k(X \& Y)}{P_k(Y)}$$

Combined with Kolmogorov's axioms, this constraint also has various intuitively rational consequences, for example that an agent will be completely certain of any claim on the supposition of that very claim's truth. (In symbols, for any $X \in \mathbf{L}$ and $t_k \in \{t_1, t_2, \dots, t_n\}$, $P_k(X | X) = 1$.)

The traditional definition of conditional probability completes our list of synchronic constraints. In the next section, we will begin to consider possible diachronic constraints on our models. First, however, I want to define one more concept so that we can examine a sample model.

With the synchronic systematic constraints in place, we can define a "learned information set." A learned information set is one way of representing what the agent in a story learns between two times.¹² As was mentioned in Section 1.1, our stories assume that an agent who learns a claim comes to believe it with absolute certainty. This is represented in a history by the agent's increasing her credence in the sentence representing the claim to 1. The definition of a learned information set treats this as not only a necessary but also a sufficient condition for the agent's having learned a claim between two times. We define a **learned information set** as follows:

Given a model defined over a modeling language \mathbf{L} and two times t_j and t_k in the time sequence of the model, the learned information set $\mathbf{I}_{j,k} \subseteq \mathbf{L}$ is the set of sentences $X \in \mathbf{L}$ such that $P_j(X) < 1$ and $P_k(X) = 1$.

¹²I say *one* way because I do not want too much "real-world" significance to be read out of the learned information sets appearing in our models. The flow of information between models and the stories they represent is mostly one-way: we take facts about the story and represent them in the model. The only information we take in the other direction are the verdicts described in Section 1.2 above. We should not draw conclusions about what exactly the agent in the story learns from the appearance of certain sentences in a learned information set. To do so would be a tacit endorsement of strong theses about a precise definition of "learning" to which I am not committed. Learned information sets should be viewed as a technical apparatus providing one way of representing what an agent learns.

Two comments on this definition: First, learned information sets are common to every history in a model. The extrasystematic constraints on a model are sufficient to determine for each sentence in the modeling language and each moment in the time sequence whether that sentence is assigned a credence of 1 or a credence less than 1 at that moment. Since the extrasystematic constraints are common to all histories in the model, and since membership in learned information sets depends only on whether a sentence is assigned a credence of 1 or a credence less than 1 at particular times, learned information sets are common to all histories in a model. This is why we can include a description of the learned information sets in our general description of the model.

Second, the definition determines the information learned between t_j and t_k by comparing the agent's doxastic states at t_j and t_k without any reference to what happens between those two times. The learned information set thus represents the net increase in claims taken for certain between t_j and t_k . It in no way takes into account the order in which claims were learned, nor whether a claim was learned, forgotten, and then learned again between t_j and t_k .

We now have enough pieces in place to give an illustrative example of a model. We begin with the story:

The Die: Marilyn walks into a room. She is told that in a few minutes a fair die will be thrown, and then a loudspeaker will announce whether the outcome was odd or even. A few minutes later, the loudspeaker announces that the die came up odd. Assuming Marilyn believes everything she has heard with certainty, what does ideal rationality require of the relationship between her post-announcement degree of belief that the die came up “3” and her pre-announcement degrees of belief?

The most straightforward model for this story is model D:

Model: D

Story: The Die

L: Built on these atomic sentences:

TH The die came up 3.

OD The die came up odd.

t-seq: Contains these times:

t_1 After Marilyn is told about the die but before she hears the announcement.

t_2 After Marilyn hears the announcement.

- ES: (1) $0 < P_1(TH) < 1$
 (2) $0 < P_2(TH) < 1$
 (3) $0 < P_1(OD) < 1$
 (4) $P_2(OD) = 1$
 (5) $P_1(TH \supset OD) = 1$
 (6) $P_2(TH \supset OD) = 1$

I_{1,2}: *OD*, sentences entailed by *OD*

Note that from the point of view of the formal system, the sentences in the modeling language are just strings of symbols to which the credence functions assign real values. So from that point of view it doesn't matter if we consider the two-letter abbreviations above to be the sentences, or if we use grammatical English strings. I have included the full-length versions to make clear which claims are meant to be

represented by which sentences. Also keep in mind that the list of extrasystematic constraints (“ES”) here is not exhaustive; I have listed those I consider pertinent or illustrative. For efficiency’s sake, I have specified that credences in particular sentences fall in the range $(0, 1)$ to indicate that it is ideally rational for the agent to take neither the claim indicated by the sentence nor its negation for certain. Finally, it is a consequence of the synchronic constraints and the definition of a learned information set that if such a set is nonempty, it will always be an infinite set of sentences. (If the learned information set contains the sentence X , it will also contain X conjoined with a tautology, X conjoined with that tautology twice, etc.) Yet the learned information set can be characterized completely by specifying a finite set of sentences and then describing the learned information set as the set containing those sentences along with all the sentences entailed by their conjunction. This is the strategy for characterizing learned information sets I have adopted in the description above and will continue to adopt below.

To answer the question posed in the story of The Die, we need to find an arithmetic statement that relates $P_{2,h}(TH)$ to $P_{1,h}$ credences and that holds in every history in D . This would be a verdict of the model D . Intuitively, the verdict we want is

$$(3) \quad P_2(TH) = P_1(TH \mid OD)$$

The idea here is that ideal rationality requires Marilyn’s degree of belief at t_2 that the die came up 3 to equal her degree of belief in 3 at t_1 on the supposition that the die came up odd. At t_2 , after she’s learned that the die actually did come up odd, the ideally rational degree of belief for Marilyn in 3 is her t_1 degree of belief on the supposition of the truth of this information.

Equation (3) is supported by other intuitions about this case as well. For example, combining Equation (3) with the extrasystematic constraint that $P_1(TH \supset OD) = 1$ and using our systematic constraints (1) through (4), we can derive that $P_2(TH) > P_1(TH)$. This verdict expresses what we intuitively take to be a requirement of ideal rationality: that Marilyn increase her degree of belief in 3 once she learns the die has come up odd.

We could get even more precise, intuitive information out of Equation (3) if we incorporated additional extrasystematic constraints into the model. Since Marilyn is certain at t_1 that the die is fair, we might think that ideal rationality requires her at t_1 to assign equal degrees of belief to any of the possible numbered outcomes of the die roll. (The principle that rational degrees of belief are attuned to known objective chances is called the Principal Principle; it will be discussed in Section 2.1 below.) With a bit of work, we can convert this thought into a set of extrasystematic constraints guaranteeing that $P_1(OD) = \frac{1}{2}$ and $P_1(TH \ \& \ OD) = \frac{1}{6}$. With these constraints in place (alongside our systematic constraints), Equation (3) gives us

$$(4) \quad P_2(TH) = P_1(TH \mid OD) = \frac{P_1(TH \ \& \ OD)}{P_1(OD)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

which captures what we intuitively think are ideally rational degrees of belief for this story.

Having convinced ourselves that Equation (3) is intuitively correct, the challenge is to find the right diachronic constraint guaranteeing that Equation (3) correctly describes all the histories in model D .

1.4. Conditionalization. Our next task is to find a “diachronic constraint” representing the way an agent’s degrees of belief are required to cohere in a rational fashion over time. The diachronic requirement of ideal rationality most often discussed in the formal epistemology literature is a principle called “updating by conditionalization,” which can be represented in our modeling technique by the following systematic constraint:

Conditionalization: Given a model with modeling language \mathbf{L} , any two times t_j and t_k in the time sequence of the model with $j < k$, learned information set $\mathbf{I}_{j,k}$, and any belief $X \in \mathbf{L}$,
 $P_k(X) = P_j(X | \mathbf{I}_{j,k})$.

To illustrate the use of Conditionalization, we can apply it to our model D for The Die. Adopting Conditionalization as our fifth systematic constraint would cause the following arithmetic statement to hold true of all the histories in D:

$$(5) \quad P_2(TH) = P_1(TH | OD)$$

This equation results because the conjunction of the sentences in the learned information set $\mathbf{I}_{1,2}$ of D is logically equivalent to OD , so by systematic constraints (1) through (4) a credence in any sentence X conditional on $\mathbf{I}_{1,2}$ is equal to the credence in that X conditional on OD .¹³ The result is precisely the relationship we wanted in Equation (3) above. Conditionalization formalizes the intuitive idea we had earlier that when Marilyn learns some information $\mathbf{I}_{j,k}$ between times t_j and t_k , ideal rationality requires her t_k degree of belief in a claim represented by X to equal her t_j degree of belief in that claim on the supposition of all the information she learned between the two times.

Note that the notion of “all the information an agent learns between the two times” is being appealed to here on a purely intuitive level. It is typical in the formal epistemology literature to frame Conditionalization without the concept of a learned information set, and simply require the agent’s credence in X at a later time to be her credence in X at an earlier time conditional on “all the information she learns between the two times.” However, this formulation introduces imprecision into what is supposed to be a formal constraint. I have modified Conditionalization by replacing the phrase “all the information the agent learns between the two times” with the notion of a learned information set in order to eliminate two sources of vagueness in that phrase:

First, it is unclear what should count as information an agent “learns.” If between t_1 and t_2 someone tells me A (and I believe him with certainty), and A logically implies B , have I learned only A between t_1 and t_2 , or both A and B ? Ideal rationality requires me to be certain of B once I am certain of A , but should we understand this as my “learning” B when I learn A , or as my “learning” just A and then “concluding” that B ? From the point of view of our formal modeling technique, the question is immaterial. Ideal rationality requires me to be certain of B at t_2 . The goal of our modeling technique is not to determine what an agent should be certain of, but what degrees of belief it is ideally rational for her to assign at particular times to claims she is not certain of. If we are conditionalizing on information learned, systematic constraints (1) through (4) guarantee that if A

¹³From this point forward when conditionalizing we will typically skip directly to the stage where the conditional probability is conditionalized on a conjunction of the sentences from which the relevant learned information set is constructed.

implies B we will obtain the same results whether we conditionalize on just A or on $A \& B$. Thus we may as well include B in the learned information set, conditionalize on that entire set, and thereby sidestep the question of which of the claims an agent becomes certain of between two times are actually those she “learns.”

Second, vagueness enters with the concept of “*all* the information” the agent learns. If we were to catalogue *all* the claims an agent learns between two times — even in a simplified, hypothetical philosophical story — even the set of atomic sentences required to represent that list might be infinite. (For example, if I learn A , does ideal rationality also require me to learn that I learn A ?) In practice, what we actually do when we apply Conditionalization is focus on those learned claims that we take to be relevant to the degrees of belief in which we are interested. This leaves us perpetually open to the charge that we have left out some learned claims and therefore failed to be true to an updating principle based on “all the information the agent learns.” The modeling technique I am presenting deals with the issue of which claims to consider by making such choices explicit at the point when our modeling language is defined. A learned information set is then defined relative to that modeling language; within the context of a given model, there is no question which sentences belong in a given learned information set.

Even with the concept of a learned information set included, Conditionalization has its limitations. We will discuss those in Section 1.5 below, then go on to suggest an alternative. Thus Conditionalization will ultimately not be our diachronic systematic constraint. But before introducing that constraint, I want to discuss some attractive features of Conditionalization that our modeling framework has been designed to retain.

First, Conditionalization allows us to construct later credence functions in their entirety from earlier ones. Conditionalization gives us a set of relations between P_j and P_k such that every t_k credence value can be expressed as a conditional t_j credence value. While it is therefore natural to use Conditionalization to derive information about t_k from information about t_j , we can also move in the opposite direction: if we know something about a t_k credence value, we can conclude something about a conditional t_j credence value.

Second, the relations between credence functions we get from Conditionalization are determined by values of those credence functions and nothing else. The learned information set for t_j and t_k is determined entirely by whether or not various sentences are assigned a credence of 1 at t_j and at t_k . Conditionalization takes the sentences in the learned information set and uses them to determine relations between credences. As we observed above, the contents of the learned information set for two times do not depend on anything about the agent’s degrees of belief between those two times. The same goes for Conditionalization: We need not know anything about credence functions that come temporally between P_j and P_k to relate P_k to P_j by Conditionalization.

Finally, the credence functions related by Conditionalization represent the *outcome* of an ideally rational inference process based on what the agent learns. In reality, an agent’s learning process may go something like this: shortly after t_j the agent learns the claim represented by X ; she then notices that this claim logically implies the claim represented by X' ; she therefore learns X' ; etc. The logical omnipotence requirement embedded in Kolmogorov’s axioms prevents our credence

functions from modeling agents who have not yet drawn out all the logical implications of claims of which they are certain. Thus the credence function P_k related to P_j by Conditionalization represents the outcome of *all* the adjustments ideal rationality requires the agent to make to her doxastic state as the result of what she learns.

1.5. Limited Conditionalization. Given that one is working with stories in which an agent becomes certain of everything she learns, I am aware of two broad categories of exceptions to Conditionalization. The first category involves cognitive mishaps by the agent such as forgetting, and the threat of such cognitive mishaps. I am going to set this category of exception aside for now, and return to it in Section 2.3. The second category of exception provides the focus for this section and our motivation for moving away from Conditionalization as a systematic diachronic constraint.¹⁴

Consider the following story:

Sleeping In: After a dark and stormy night translating Slavic poetry, Olga falls into a deep slumber. She awakens the next morning to find that her clock reads 9am. However, she suspects that the power may have gone out during the night, and so does not take this as convincing evidence that it is indeed the morning rather than the afternoon. (It is too overcast outside for her to get any clear idea of the time of day.) Still groggy, she falls asleep again, and when she reawakens the clock reads 2pm. While Olga is certain it is the same day and suspects it is now the afternoon, for all she knows it might still be morning.

Question: Assuming Olga believes the reports of her senses implicitly, what does ideal rationality require of the relationship between her second-awakening degree of belief that it is afternoon and her degrees of belief after first seeing the clock?

Here is a reasonable initial model SI for this story:

Model: SI

Story: Sleeping In

L: Built on these atomic sentences:

$N9$ The clock now reads 9am.

$N2$ The clock now reads 2pm.

NA It is now afternoon.

t -seq: Contains these times:

t_1 After Olga first awakens and sees the clock reading 9am.

t_2 After Olga awakens for the second time and sees the clock reading 2pm.

- ES: (1) $P_1(N9) = 1$
 (2) $P_2(N9) = 0$
 (3) $P_1(N2) = 0$
 (4) $P_2(N2) = 1$
 (5) $0 < P_1(NA) < 1$
 (6) $0 < P_2(NA) < 1$

¹⁴Though not as widely noted as the first, the second category of exception has been discussed by Arntzenius (2003), Halpern (2005), and Meacham (manuscript).

I_{1,2}: $N2, \sim N9$, sentences entailed by their conjunction

To answer the question in the story, we need to find an arithmetic statement relating $P_2(NA)$ to P_1 credences. If we adopt Conditionalization as a systematic constraint, the following arithmetic statement will hold true:

$$(6) \quad P_2(NA) = P_1(NA | N2 \& \sim N9)$$

I hope that with a little thought it is clear that this is preposterous. Equation (6) suggests that ideal rationality demands of Olga that her t_2 degree of belief that it is afternoon equal her t_1 degree of belief that it is afternoon (at t_1) on the supposition that at t_1 the clock reads 2pm and not 9am. But we were after Olga's ideally rational t_2 degree of belief that it is afternoon at t_2 , not her t_2 degree of belief that it was afternoon at t_1 . Moreover, at t_1 Olga is certain that the clock reads 9am! This fact becomes especially damning when we apply the traditional definition of conditional probability to Equation (6):

$$(7) \quad P_2(NA) = \frac{P_1(NA \& N2 \& \sim N9)}{P_1(N2 \& \sim N9)}$$

We can conclude from our extrasystematic constraints and Kolmogorov's axioms that the denominator in Equation (7) must equal zero. Thus Conditionalization is telling us that every history in **SI** has an undefined $P_2(NA)$ value!

I could go on for quite some time describing absurdities that result from using Conditionalization as a systematic constraint in this case, but let me instead suggest why Conditionalization runs into trouble here. Conditionalization was originally designed to model agents' gradual accumulation over time of beliefs *de dicto*. A belief *de dicto*, if true, is always true. Thus it is rational for an agent who becomes certain of a claim expressing a belief *de dicto* to remain certain of that claim from that time forward. This fits perfectly with the fact that, if a credence function in a history assigns a credence of 1 to a sentence, Conditionalization requires all future credence functions in that history to assign a credence of 1 to that sentence as well. By systematic constraints (1) through (4), if an unconditional credence function assigns a sentence X a credence of 1, the related conditional credence function assigns X a credence of 1 conditional on any sentence whose unconditional credence is not 0.

Conditionalization runs into trouble when applied to sentences that can be true one moment, then false the next. In such cases, it can be perfectly rational to be certain of a sentence at one time and then uncertain of it (or even certain of its negation) a little bit later. For example, if we look over the extrasystematic constraints in model **SI**, we see that Olga goes from being certain that "The clock now reads 9am" is true at t_1 to being certain that it is false at t_2 , and this change is perfectly rational. Conditionalization results in verdicts inconsistent with the requirements of ideal rationality when it is applied to context-sensitive claims, by which I mean claims that have different truth-values in different contexts.¹⁵ Conditionalization is very good at modeling changes in an agent's epistemic state with

¹⁵Here I am adapting John MacFarlane's use of the term "context-sensitive" in (MacFarlane 2005). MacFarlane differentiates between claims that have different *contents* in different contexts, which he calls "indexical," and claims that have different *truth-values* in different contexts, which he calls "context-sensitive." I hope that by focusing on the context-sensitivity of claims I maintain neutrality among positions in the philosophy of language that differ over whether various kinds of claims have different *contents* in different contexts.

respect to a set of claims whose truth-values remain fixed. But Conditionalization has trouble with moving targets: it cannot take into account truth-value changes that occur over time when an agent entertains context-sensitive claims.¹⁶

The presence of context-sensitive sentences in a modeling language can make it consistent with ideal rationality for some of those sentences to be assigned a credence of 1 at one moment and a credence less than 1 the next. In those cases, Conditionalization is violated; so unrestricted Conditionalization cannot play the role of a systematic constraint in a modeling framework used to represent context-sensitive claims. Nevertheless, Conditionalization works well in a variety of situations in which context-sensitive sentences do not have this effect. The natural response is to retain Conditionalization in a limited capacity — to continue to use it when retreats from certainty are not a concern. This is accomplished by adopting the following systematic diachronic constraint:

Systematic Constraint (5), Limited Conditionalization (LC):

Given a model M defined over modeling language L , and two times t_j and t_k in the time sequence of M ,
if there does not exist a sentence $X \in L$ such that $P_j(X) = 1$ and $P_k(X) < 1$,
then for any $Y \in L$, if $I_{j,k}$ is the learned information set in M for t_j and t_k , then $P_k(Y) = P_j(Y | I_{j,k})$.

Barring instances of cognitive mishap (which we will not consider again until Section 2.3), the only way it can be ideally rational for an agent to be certain of a claim at one time and then less-than-certain of it later is if ideal rationality requires the agent to believe at that later time that the truth-value of the claim at that time may be different than it was at the earlier time. In some such cases, ideal rationality will require the agent to be certain at the later time that the claim she previously took to be true is now *false*; in other cases, ideal rationality will merely require the agent not to be certain that the claim remains true. Either way, it is

¹⁶One might object that I have created this problem for Conditionalization myself by the way I have chosen to set up our modeling technique. By treating “The clock now reads 9am” as the same claim at different times in the story, and representing it by the same formal object (a sentence) throughout our formal system, I have opened up the possibility that ideally rational histories could assign that object an unconditional credence of 1 at an earlier time and less than 1 later on. It might be suggested that the formal system should contain (at least) *two* sentences representing “The clock now reads 9am,” one for that claim when it is held at t_1 and one for when it is held at t_2 . In that way, the problems I have indicated for Conditionalization could be avoided.

In designing a modeling technique, one makes a number of choices with the hopes that they will yield a fruitful and elegant system. Early on in the construction of my modeling technique, I considered approaches involving the kind of proliferation of sentences described above. While solving some problems with Conditionalization, these approaches led to a host of other issues. Just to name one, it is difficult on such an approach to see how to use Conditionalization to relate credences in t_1 -stage sentences to credences in t_2 -stage sentences at all. So I ultimately chose to keep to a simple set of sentences, limit Conditionalization, and supplement the resulting gaps as described in the rest of this paper. I would be happy to compare my modeling technique with a technique that further differentiated sentences, left Conditionalization intact, and successfully accounted for our intuitions about all the stories I consider here.

I should also note that it doesn’t help matters to choose an alternative diachronic constraint to Conditionalization such as Jeffrey Conditionalization or the Reflection Principle. Each of those alternatives reveals its own shortcomings when context-sensitive claims get involved. (For discussion and references on these alternate diachronic constraints, see (Talbot 2001).)

when the possibility of context-sensitivity arises that (LC) is barred from yielding verdicts relating degrees of belief at two times.¹⁷

(LC) therefore maintains the results we wanted for the story of The Die: Marilyn is certain throughout that the truth-values of claims about the outcome of the die have not changed, and ideal rationality does not require her to be certain of one such claim at t_1 and then less-than-certain of it at t_2 .¹⁸ Thus (LC) allows us to conditionalize just as before. (LC) also does a nice job of avoiding the kinds of troubles we ran into when we tried to conditionalize in Sleeping In. The modeling language of SI contains sentences representing claims that Olga is quite certain have changed their truth-values between t_1 and t_2 — namely the claims about the current time displayed by the clock. This comes out in the fact that Olga is required by ideal rationality to be certain of both $N9$ and $\sim N2$ at t_1 and certain of their negations at t_2 . Thus the antecedent of the conditional in (LC) is not met for this model and the absurd Equation (6) need not hold.

While I will not go through them here, applying (LC) to a variety of examples affirms that it does an excellent job of preventing conditionalization from applying to situations in which it would generate absurd sentences like Equation (6), while at the same time allowing it to apply to many situations in which it generates intuitive results.¹⁹ This indicates that our analysis of the flaw in Conditionalization that generated Equation (6) is correct: Conditionalization is unable to relate degrees of belief at two times when ideal rationality requires an agent to be certain of some context-sensitive claim at the earlier time and then less-than-certain of it at the later time.

One can concoct a wide variety of stories in which the involvement of self-locating beliefs causes Conditionalization to fail. Some authors have responded by proposing that self-locating beliefs are a special class, requiring updating rules distinct from those applying to beliefs *de dicto*.²⁰ However, our analysis indicates that this is an over-reaction. For example, if an entire story takes place over the course of one day, and the agent is certain throughout that the day has not changed, the claim “Today is Monday” can be conditionalized upon just like any belief *de dicto* whose truth-value is incapable of change. The troublesome capacity of self-locating beliefs, and context-sensitive beliefs in general, is their capacity to require certainty

¹⁷Note that what matters here is not whether claims have *actually* shifted their truth-values between two times, but whether ideal rationality requires the agent to *believe* that claims may have shifted their truth-values between those times. Our systematic constraints reflect requirements of ideal rationality on the *internal* coherence of an agent’s degrees of belief.

¹⁸There is a relatively quick way to check if any of the sentences in a modeling language are context-sensitive between two times: if none of the *atomic* sentences in the language changes its truth-value between those times, none of the other sentences built from those atomic sentences can change their truth-values either. So we need only consider for each of the atomic sentences whether the agent would be certain at the later time that it had the same truth-value as at the earlier time.

¹⁹One interesting fact that comes out in the study of further examples is that once we restrict conditionalization to cases in which no sentence in the modeling language goes from a P_j value of 1 to a P_k value less than 1, the principle yields intuitive results even when t_k is a moment in the time sequence *earlier* than t_j . This accounts for a subtle change between traditional Conditionalization and (LC): (LC) does not require $j < k$. For analysis of this change, some relevant examples, and further discussion see (Titelbaum manuscript). For the stories considered this paper, however, we will need to conditionalize only in the time-forward direction, so I will continue to assume that $j < k$ and continue to refer to t_j as the “earlier time” and t_k as the “later time.”

²⁰See for example (Meacham manuscript) and (Halpern 2005).

at one moment and something less at the next; they cause trouble only when there is the threat that that capacity has been realized.²¹

Of course, we set out to analyze cases in which that capacity *is* realized, and (LC) gets us only part of the way there. (LC) prevents us from generating the absurd Equation (6), but it also keeps the only diachronic principle we had from applying in model SI. Thus we are left without any way to relate Olga's t_2 degree of belief that it is afternoon to her t_1 degrees of belief. To solve that problem, we will need our sixth and final systematic constraint, a constraint that helps us analyze stories involving context-sensitive claims.

Before we get to that, however, we are going on a slight detour: we are going to take up the worries voiced in Section 1.1 above about the choice between alternative modeling languages.

1.6. Conservative Modeling Principle. When you set out to model a story using our modeling technique, the trickiest choice to make is the initial choice of a modeling language. Choose too simple a language, and you may fail to represent claims the agent learns that are relevant to the degrees of belief you're interested in. Choose too complex a modeling language, and your model may have too few constraints to yield verdicts on those degrees of belief.

This dilemma is not peculiar to the modeling technique presented here. Whenever one is modeling using formal epistemology techniques (or just modeling in general), one has to choose what information to include in one's model and what information to leave out. The modeling technique presented here at least has the advantage that it makes such choices explicit at the start: it is clear at every stage which claims are under consideration by the model at hand, and precisely what verdicts result from that choice. Nevertheless, judgment must be exercised by those building models as to what to include and what to leave out. The resulting model will always be open to the challenge that it leaves unrepresented some claim that is relevant and whose representation would lead to verdicts differing from those obtained so far.

Let's be a bit more precise about what such a challenge would look like. Suppose model M is defined over modeling language L , and model M^+ is defined over modeling language L^+ . We call M a **submodel** of M^+ just in case

- M and M^+ model the same story,
- M and M^+ share the same time sequence,
- $L \subseteq L^+$,
- and each sentence in L represents the same claim in M^+ as it does in M .

If M is a submodel of M^+ , we call M^+ a **supermodel** of M . Keep in mind that the submodel/supermodel relation concerns the relation between the *languages* of M and M^+ . It will typically not be the case that the set of histories of M is a subset of

²¹It may even not be quite right even to say that beliefs *de dicto* are incapable of undergoing truth-value change. General treatments of the semantics of context-sensitive claims often treat the possible world in which a claim is uttered (or believed, or thought, etc.) as one of the context parameters. Thus even *de dicto* claims are context-sensitive, since their truth-values vary from world to world. Since our stories do not involve agents moving from one possible world to another, this kind of context-sensitivity will not lead to an agent's being required to be certain of a belief *de dicto* at one time during a story and then less-than-certain of it at a later time. Thus for our purposes here, I will continue to describe beliefs *de dicto* as context-insensitive. But if an agent could move from world to world, stories about such agents would have models in which one could not conditionalize on sentences representing particular *de dicto* claims.

the set of histories of M^+ , because the credence functions in M will be defined over a different set of sentences than the credence functions in M^+ .

Now suppose M is a submodel of M^+ , the credence functions in M are written $P_{t,h}(\cdot)$, and the credence functions in M^+ are written $P_{t,h}^+(\cdot)$. If we take an arithmetic statement concerning the histories in M , we can generate an **analogue** of that arithmetic statement in M^+ by replacing “ $P_{t,h}$ ” everywhere it appears with “ $P_{t,h}^+$ ” and quantifying the statement over the histories in M^+ instead of the histories in M .²² The result is an arithmetic statement concerning the histories in M^+ , which I hope will be clear is in an important sense analogous to the original arithmetic statement concerning the histories in M . Certainly if the arithmetic statement in M represents a requirement of ideal rationality, its analogue in M^+ represents the same requirement of ideal rationality.

The challenge we are concerned with here takes the following form: We have a story and a model M for which we have obtained some verdicts. We believe that L , the modeling language of M , contains sentences representing all the claims relevant to the question posed in the story. Thus we take the verdicts of M to represent requirements of ideal rationality answering that question. Nevertheless, a challenger asserts of some claim, or set of claims, that it is relevant to answering the question and is not represented by sentences in L . The appropriate response to this challenge is to take L and add sentences representing those claims, creating a new modeling language L^+ . We then build a model M^+ representing the same story as M , with the same time-sequence as M , but with L^+ as its modeling language. M^+ is a supermodel of M . Finally, we check whether the verdicts of M have analogues in M^+ that are also verdicts of M^+ .

The most favorable possible outcome of such a check is that the analogue of every verdict of M is a verdict of M^+ . In that case, we will call M^+ a **conservative supermodel** of M . If M^+ turns out to be a conservative supermodel of M , we have responded to the challenge and may continue to rely on the verdicts obtained from M (at least until another such challenge arises).²³ We may also be confident that the question asked by the story can be fully analyzed with a model whose modeling language is L ; the extra sentences included in L^+ are unnecessary for the analysis.

Our sixth and final systematic constraint makes it easier to determine whether such a check will yield the most favorable outcome. It describes some precise conditions in which expanding the modeling language of a model yields a conservative supermodel:

²²Note that the “analogue” relation works in both directions: an arithmetic statement in a supermodel can have an analogue in a submodel. Moreover, if we have an arithmetic statement in M and take its analogue in M^+ , then in turn take the analogue of *that* M^+ statement in M , we get the original arithmetic statement back out again.

²³Of course, there are outcomes possible besides M^+ turning out to be a conservative supermodel of M . If M^+ has verdicts that directly *contradict* the analogues of verdicts of M , this is good reason to conclude that M has failed to represent relevant information and yields verdicts that are unreliable. However, it may also turn out that the analogues of verdicts of M neither are verdicts of M^+ nor are contradicted by verdicts of M^+ . (The analogues will be *false statements* about M^+ , in that they are universally quantified statements that are not true of every history in M^+ , but there may still not exist a true verdict of M^+ that contradicts them.) In such cases, the complexity of M^+ prevents our constraints from restricting its histories enough to yield analogous verdicts. While such a situation may in some cases *suggest* that M is too simple and yields verdicts not representing requirements of ideal rationality, it is not *conclusive evidence* that the verdicts of M are unreliable.

**Systematic Constraint (6),
Conservative Modeling Principle (CMP):**

Given a story, a model M representing that story with modeling language L , and a supermodel M^+ of M with modeling language L^+ and credence functions $P_{t,h}^+(\cdot)$,

if for any time t_k in the time sequence of M^+ and any sentence $Y \in L^+$, there exists an $X \in L$ such that $P_k^+(X \equiv Y) = 1$,
then the analogue in M^+ of any verdict of M is also a verdict of M^+ .

To be entirely explicit about the quantifier order, we can write the conditional in (CMP) as:

$$[\forall(t_k)\forall(Y \in L^+)\exists(X \in L)(P_k^+(X \equiv Y) = 1)] \supset \\ M^+ \text{ is a conservative supermodel of } M.$$

(CMP) plays two important roles in our modeling technique. First, it helps us understand what happens when we augment a modeling language in a particular way. Suppose we have a story and a model M of that story from which we have already obtained some verdicts. Then suppose we create a new model M^+ of the same story by adding sentences to the modeling language of M . (CMP) guarantees that if for every sentence added, at each time under consideration there is already a sentence in the old language that ideal rationality requires the agent to be certain has the same truth-value as the sentence in the new language, then the analogues of the original model's verdicts will still hold true in the new model. Adding new sentences that at every moment in the time sequence have known truth-value equivalents in the old language leaves intact verdicts about ideally rational degrees of belief.²⁴

In this first role, (CMP) can help us predict what will happen when we add new sentences to a modeling language. However, (CMP) is also a systematic constraint of our modeling technique. It therefore helps determine which histories are part of a given model. For that reason, (CMP) has a second role: we can use it to analyze models already in hand and derive verdicts from them. As a further systematic constraint beyond the initial five, (CMP) allows us to recognize requirements of ideal rationality that cannot be obtained from the first five systematic constraints alone.

In the next section, we will see an application of (CMP) in its second role. We will also discuss why (CMP) is an intuitive constraint to place on models whose verdicts are meant to represent requirements of ideal rationality. At this point, though, I'd like to give an example of (CMP) in its first role.

Recall our model D of The Die. D contains two atomic sentences, both about the outcome of the die roll. However, someone might object that over the course of the story Marilyn learns more than that the die came up odd. She also, like any agent

²⁴The antecedent of the conditional in (CMP) actually requires that *all* sentences in the modeling language of the supermodel have certain truth-value equivalents in the modeling language of the submodel. However, we need check only the sentences in the supermodel that are not in the submodel. The others clearly have truth-value equivalents in the language of the submodel, namely themselves. Moreover, among the sentences new to the language of the supermodel, we need check only the atomic sentences. If all the atomic sentences in the language of the supermodel have certain truth-value equivalents in the language of the submodel, the sentences constructible from them via truth-functions must have certain truth-value equivalents as well.

in a story that occurs over a stretch of time, notices that some time has passed. By t_2 , which occurs after the announcement that the die is odd has been made, Marilyn will be certain of the claim “The announcement about the die has already been made,” a claim whose falsity she was certain of at t_1 . Perhaps, someone will object, this information about the passing of time is relevant to Marilyn’s other degrees of belief, and should be represented in our model.

To test this suggestion we create a new model, D^+ :

Model: D^+

Story: The Die

L^+ : Built on these atomic sentences:

TH The die came up 3.

OD The die came up odd.

BM The announcement about the die has already been made.

t -seq: Contains these times:

t_1 After Marilyn is told about the die but before she hears the announcement.

t_2 After Marilyn hears the announcement.

- ES: (1) $0 < P_1^+(TH) < 1$
 (2) $0 < P_2^+(TH) < 1$
 (3) $0 < P_1^+(OD) < 1$
 (4) $P_2^+(OD) = 1$
 (5) $P_1^+(TH \supset OD) = 1$
 (6) $P_2^+(TH \supset OD) = 1$
 (7) $P_1^+(BM) = 0$
 (8) $P_2^+(BM) = 1$

$I_{1,2}^+$: OD , BM , sentences entailed by their conjunction

Our verdicts from model D were generated by (LC). But now we have a problem: once we add BM into the modeling language, it contains a sentence representing a claim that ideal rationality requires Marilyn to be certain at t_1 is true and certain at t_2 is false: $\sim BM$. (This occurs because BM represents a context-sensitive claim.) So the antecedent of the conditional in (LC) is not met, and we cannot use (LC) to draw verdicts from D^+ .

However, we can use (CMP) to evaluate D^+ . D^+ is a supermodel of D , with the added sentence BM . We noted earlier that every modeling language contains at least one tautology and one contradiction;²⁵ let’s refer to one of the tautologies in L as T and one of the contradictions as F . From the last two extrasystematic constraints on D^+ and our synchronic systematic constraints, we can derive that

$$(8) \quad P_1^+(BM \equiv F) = 1$$

and

$$(9) \quad P_2^+(BM \equiv T) = 1$$

Thus there is a sentence in L that ideal rationality requires Marilyn to be certain at t_1 has the same truth-value as BM , and there is a sentence in L that ideal rationality requires Marilyn to be certain at t_2 has the same truth-value as BM . At both times, there are sentences in the old modeling language that are known truth-value equivalents to the new sentence in the expanded language.

²⁵See Section 1.1 above.

We have met the conditions for (CMP), and therefore may conclude that D^+ is a conservative supermodel of D : the analogues of D 's verdicts are also verdicts of D^+ . Since we know $P_2(TH) = P_1(TH | OD)$ is a verdict of D , we can use (CMP) to derive

$$(10) \quad P_2^+(TH) = P_1^+(TH | OD)$$

From there we can go on to derive in D^+ our other requirements of ideal rationality for The Die. It turns out that taking the information Marilyn learns about the passage of time into consideration in our model of The Die doesn't affect our verdicts about the relationships between ideally rational degrees of belief concerning the outcome of the die roll. This matches up well with our intuition that information about the mere passage of time isn't relevant to Marilyn's beliefs about the die roll.

1.7. The Rationale for (CMP). In Sections 1.4 and 1.5, the story *Sleeping In* presented us with a problem. In that story, Olga wakes up, notices that her (perhaps unreliable) clock reads 9am, falls asleep and wakes up again, notices that the clock reads 2pm, then wonders whether it is morning or afternoon. In the model that seemed sensible to apply, **SI**, all the atomic sentences were known by Olga to be context-sensitive, and in many cases ideal rationality required her to move from certainty in their truth to certainty in their falsity. Conditionalization in its unrestricted form therefore yielded absurd verdicts when applied to this model, while Limited Conditionalization failed to yield any verdicts at all.

(CMP) had a relatively easy time with model D^+ of The Die. The claims asked about in that story were context-insensitive, so all (CMP) had to do was rule out the context-sensitive sentences introduced by D^+ as irrelevant to those context-insensitive claims. Past that point, (LC) could be applied to deliver the verdict desired. In *Sleeping In*, however, the question in the story is about degrees of belief in context-sensitive claims, and context-sensitive claims learned by the agent over the course of the story also intuitively seem relevant. Nevertheless, (CMP) can help us answer the question in *Sleeping In*, and the way in which it does so will also help explain why it represents a requirement of ideal rationality.

To obtain the verdicts we want about *Sleeping In*, we'll need to work with models other than **SI**. We'll start with a more complex model, **SI'**:

Model: **SI'**

Story: *Sleeping In*

L': Built on these atomic sentences:

F2 The first time I awaken today, the clock reads 2pm.

FA The first time I awaken today, it is afternoon.

S2 The second time I awaken today, the clock reads 2pm.

SA The second time I awaken today, it is afternoon.

N2 The clock now reads 2pm.

NA It is now afternoon.

NS Now is the second time I awaken today.

t-seq: Contains these times:

t_1 After Olga first awakens and sees the clock reading 9am.

t_2 After Olga awakens for the second time and sees the clock reading 2pm.

ES: (1) $0 < P'_1(S2) < 1$

- (2) $P'_2(S2) = 1$
- (3) $P'_1(N2) = 0$
- (4) $P'_2(N2) = 1$
- (5) $P'_1(SA \equiv NA) < 1$
- (6) $P'_2(SA \equiv NA) = 1$
- (7) $P'_1(NS) = 0$
- (8) $P'_2(NS) = 1$
- (9) $P'_1(FA \equiv NA) = 1$
- (10) $P'_1(F2 \equiv N2) = 1$

$I'_{1,2}$: $S2, N2, SA \equiv NA, NS$, sentences entailed by their conjunction

Note that the first four atomic sentences listed (those that don't start with an "N") are to be read as tenseless "eternal" sentences. (They are *de dicto* and context-insensitive.) As for the extrasystematic constraints: We assume that the first time Olga wakes up, it would not be ideally rational for her to be certain that the second time she wakes up the clock will read 2pm; this accounts for Constraint (1). Constraint (5) expresses the fact that when Olga first wakes up, it would not be ideally rational for her to be certain that the claim "The second I awaken today, it is [will be] afternoon" has the same truth-value as the claim "It is now afternoon." After all, at that point she knows neither whether it is afternoon nor whether it will be afternoon the next time she awakens. While she might puzzle out some kind of correlation between those claims, it would be bizarre for her to be *certain* that they have the same truth-value.

Expressed in terms of SI' , the question in Sleeping In asks us to relate $P'_2(NA)$ to some $P'_1(\cdot)$ credences. However, given the extrasystematic constraints on SI' , we can express that question in a different way. Since $P'_2(SA \equiv NA) = 1$, our synchronic systematic constraints tell us that

$$(11) \quad P'_2(NA) = P'_2(SA)$$

Thus if we can relate $P'_2(SA)$ to some $P'_1(\cdot)$ credences, our task will be complete.

Unfortunately, that cannot be accomplished directly using SI' . Language L' contains a number of sentences representing claims Olga knows are context-sensitive between t_1 and t_2 (all the sentences with an "N"). Ideal rationality requires her to move from certainty in the truth of some of these sentences ($\sim N2, \sim NS$) at t_1 to certainty in their falsity at t_2 . So the antecedent of the conditional in (LC) is unsatisfied and we cannot use (LC) to relate t_2 credences to t_1 credences. Instead we must work indirectly, through a submodel of SI' I'll call SI^- :

Model: SI^-

Story: Sleeping In

L^- : Built on these atomic sentences:

$F2$ The first time I awaken today, the clock reads 2pm.

FA The first time I awaken today, it is afternoon.

$S2$ The second time I awaken today, the clock reads 2pm.

SA The second time I awaken today, it is afternoon.

t -seq: Contains these times:

t_1 After Olga first awakens and sees the clock reading 9am.

t_2 After Olga awakens for the second time and sees the clock reading 2pm.

ES: (1) $0 < P_1^-(S2) < 1$

$$(2) P_2^-(S2) = 1$$

$\mathbf{I}_{1,2}^-$: $S2$, sentences entailed by $S2$

\mathbf{SI}^- is a submodel of \mathbf{SI}' , where all the context-sensitive sentences have been removed from the modeling language. The question is whether \mathbf{SI}' is a conservative supermodel of \mathbf{SI}^- . To apply (CMP), we need to check that for every moment in the time sequence and every sentence in \mathbf{L}' not in \mathbf{L}^- , there is a sentence in \mathbf{L}^- that ideal rationality requires Olga to be certain at that moment has the same truth-value as the sentence in \mathbf{L}' . The chart below displays the truth-value equivalents at each moment in the time sequence for the three sentences in \mathbf{L}' not in \mathbf{L}^- . “T” stands for any tautology in \mathbf{L}^- and “F” stands for any contradiction.

$$\begin{array}{ll} P_1(N2 \equiv F2) = 1 & P_2(N2 \equiv S2) = 1 \\ P_1(NA \equiv FA) = 1 & P_2(NA \equiv SA) = 1 \\ P_1(NS \equiv F) = 1 & P_2(NS \equiv T) = 1 \end{array}$$

With all these certainties in place, the antecedent of the conditional in (CMP) is satisfied, and we can conclude that \mathbf{SI}' is a conservative supermodel of \mathbf{SI}^- . Analogues of verdicts in \mathbf{SI}^- are verdicts of \mathbf{SI}' . In particular, since there are no context-sensitive sentences in \mathbf{L}^- and therefore no sentences that ideal rationality requires Olga to be certain of at t_1 and less-than-certain of at t_2 , we can apply (LC) to obtain:

$$(12) \quad P_2^-(SA) = P_1^-(SA | S2)$$

The analogue of this verdict of \mathbf{SI}^- for \mathbf{SI}' is

$$(13) \quad P_2'(SA) = P_1'(SA | S2)$$

Combining this with the fact that $P_2'(NA) = P_2'(SA)$ from Equation (11) above, we have

$$(14) \quad P_2'(NA) = P_1'(SA | S2)$$

This equation proposes an answer to the question in Sleeping In; it relates $P_2'(NA)$ to $P_1'(\cdot)$ credences. And I think this answer succeeds in representing a requirement of ideal rationality. Suppose that upon first awakening, Olga were to ask herself, “Let’s suppose that when I next awaken the clock reads 2pm. How confident am I right now that, given the truth of that supposition, it will be afternoon when I next awaken?” When Olga then falls asleep, reawakens, and sees that her clock actually *does* read 2pm, intuitively I think it is rational for her to assign a degree of belief to “It is now afternoon” equal to whatever degree of belief was her answer to that question. And this is precisely the requirement of ideal rationality represented by Equation (14).²⁶

²⁶I have actually simplified the question in quotes above a bit to make the intuitive idea behind Equation (14) more clear. Technically, Equation (14) above constrains only those histories that appear in model \mathbf{SA}' and therefore meet all the constraints representative of ideal rationality. Thus Olga’s question should probably be something like “Given that supposition, how confident is it ideally rational for me to be right now that it will be afternoon when I next awaken?” Even with that modification to the question, I think the intuitive argument behind Equation (14) still goes through, though its path is a bit more tortured.

However, the modified version of the question brings up a limitation of our modeling technique. It is possible for agents to assign degrees of belief to claims like “The ideally rational degree of belief for me to assign to claim X right now is 1/2.” Our modeling technique presumes that for every claim that is represented in a model of a story, we can determine prior to applying the modeling technique whether it is ideally rational for the agent to be certain of that claim.

Given that model SI^- yields verdicts that represent requirements of ideal rationality, our next question is why it does so, and why in general (CMP) should be accepted as generating requirements of ideal rationality. We will start by looking specifically and in fairly informal terms at Sleeping In, then gradually generalize and formalize from there. We will draw on relevance intuitions concerning the *semantic* content of claims to help motivate a purely *syntactic* (CMP) relating models.

Once Olga awakens the second time and is certain it is the second time she has awoken, ideal rationality requires her to be certain that the second time she awakens it is afternoon if and only if it is now afternoon. This allows us to shift our attention from her degree of belief that it is now afternoon to her degree of belief that the second time she awakens it is afternoon, since our synchronic systematic constraints tell us that ideal rationality requires her to have equal degrees of belief in each. Moreover, since Olga is certain now is the second time she awakens, her certainty that the clock now reads 2pm makes it a requirement of ideal rationality that she be certain that the second time she awakens, the clock reads 2pm.

Our modeling technique requires us to model a story in two stages. First, we determine which claims ideal rationality requires the agent to be certain of. Second, we use that information as background in applying the modeling technique. In Sleeping In, it is important that by t_2 Olga has learned that now is the second time she awakens and that the clock now reads 2pm. It is hard to imagine, for example, how in the context of this particular story she could deduce without that information that the second time she awakens the clock reads 2pm. But once we have taken that information into account and used it to establish that at t_2 ideal rationality requires Olga to be certain that the second time she awakens the clock reads 2pm, we move on to the second stage of the modeling technique.

In that latter stage, where requirements of certainty are already established and we look instead to determine ideally rational degrees of belief, the claim that the second time Olga awakens the clock reads 2pm carries all the information relevant to whether it is afternoon the second time she awakens. The fact that the clock reads 2pm the second time she awakens is evidence enough for whatever change is required in her degree of belief that it's afternoon on her second awakening. From the point of view of establishing that degree of belief, the information about which

It seems to me that ideal rationality requires an agent to be certain of claims that express the verdicts of our modeling technique. And yet the entire point of our modeling technique is to yield verdicts like the meta-claim just quoted, so if we can determine whether it is ideally rational for an agent to be certain of such meta-claims prior to the application of our modeling technique (as the presumption supposes), what is the point of that technique?

The answer is that I have not designed our modeling technique with an eye to, nor have I tested that technique with respect to, modeling languages that include meta-claims like the one quoted in the previous paragraph. My hope is that in the story in question, one could establish whether it is ideally rational for the agent to assign a degree of belief of $1/2$ to X using a model whose language did not include any meta-claims. One could then build a model whose language included such claims, and apply the results of the simpler model to the more complex model as extrasystematic constraints. Thus while one would not be determining whether it is ideally rational for an agent to be certain of every sentence in the complex modeling language prior to applying the modeling technique *at all*, one would be making those determinations prior to applying the modeling technique *to the more complex model*. However, there may be complicated stories in which degrees of belief in meta-claims are closely correlated with those in beliefs “on the ground” (such as some of the more complicated stories involving appeals to the Reflection Principle). I cannot guarantee that our modeling technique will be able to sort out the requirements of ideal rationality for such stories.

moment it is right now is redundant and therefore unnecessary. In fact, whatever degree of belief is required of Olga that it's afternoon the second time she awakens, that degree of belief is required of her based on the evidence that the clock reads 2pm when she re-awakens regardless of what time it is now or what awakening she's on when she happens to consider that evidence.

For this reason, once we've established what ideal rationality requires Olga to believe for certain we can *set aside* any information about what moment it is now and examine her degree of belief that it's afternoon the second time she awakens exclusively in terms of her degrees of belief about first and second awakenings. We can take a model of the story (SI') that represents some claims about what's going on now and some claims about Olga's various awakenings, and remove from its modeling language all the sentences about what's going on now. The model that remains (SI^-) still contains all the sentences needed to establish correct verdicts concerning Olga's degree of belief that it's afternoon the second time she awakens. And since these verdicts represent requirements of ideal rationality, it is appropriate for their analogues to constrain histories in the more complex model SI' . Which is precisely what (CMP) achieves in this case.

This situation can be described more generally, bringing us closer to a syntactic representation of the principle at work. The reason we have to apply (CMP) to generate verdicts for model SI' is that a number of sentences in its modeling language represent claims that are context-sensitive in this story, making a direct application of (LC) impossible. Context-sensitive claims typically derive their context-sensitivity from context-sensitive expressions contained within them; for example, the context-sensitive claims in *Sleeping In* represented by sentences in SI' all contain the context-sensitive expression "now."²⁷

However, something in *Sleeping In* is working in our favor. At each moment in the time-sequence we are interested in, there is an expression not containing "now" that ideal rationality requires the agent to be certain uniquely picks out the denotation of "now" at that time. For example, at t_2 ideal rationality requires Olga to be certain that "the second time I awaken" uniquely picks out the same time as "now." Since such uniquely denoting expressions are available to Olga at each time in the time-sequence, at each such time there is a claim not containing "now" that ideal rationality requires Olga to be certain has the same truth-value as each claim containing "now" — namely a claim that substitutes the available uniquely denoting expression for "now." For purposes of determining degrees of

²⁷Following the discussion of MacFarlane's terminology in the note in Section 1.5 above, I am using "context-sensitive expression" to pick out those expressions whose *denotation* changes from context to context. An expression whose *content* changed from context to context would be an "indexical expression."

I should also note that this talk of claims' "containing" expressions violates my earlier promise of neutrality concerning whether claims (possible objects of doxastic states) are themselves linguistic expressions. The argument at this point in the paper is an informal argument intended to impart an intuitive sense of what's behind (CMP). Maintaining complete neutrality in the explanation would torture the prose and make that sense less clear, so I have abandoned the attempt. Nevertheless, I believe my explanation is just as valid when put in terms of what's contained in *linguistic expressions representing claims*. And the modeling technique it helps explain ultimately makes no reference to claims' containing expressions in any of its formal definitions or principles. Thus the modeling technique itself remains entirely neutral.

Finally, when I discuss a claim's "containing" an expression, please read in wherever necessary the caveat that the expression is contained in a non-intensional context.

belief in claims not containing “now,” the substituted claims contain all the relevant information contained in the claims containing “now.” Once we have established what certainties are required of the agent, we can leave the claims containing “now” out of arguments about ideally rational degrees of belief in claims not containing “now.”

Let’s characterize this situation even more generally and more formally: Suppose the modeling language \mathbf{L}^+ of a model \mathbf{M}^+ defined over time-sequence (t_1, t_2) includes an atomic sentence X representing a claim that contains an expression “ x ”. Further suppose that at t_1 , ideal rationality requires the agent in the story being modeled to be certain that an expression “ y ” uniquely picks out the denotation of “ x ” at t_1 . Then there will be a claim Y like X but containing “ y ” in place of “ x ” throughout. Let’s suppose that \mathbf{L}^+ contains a sentence representing Y , as well as a sentence representing the claim Z formed by replacing “ x ” throughout X with an expression “ z ” that ideal rationality requires the agent to be certain at t_2 uniquely picks out the denotation of “ x ” at t_2 . Moreover, let’s assume that for every other atomic sentence in \mathbf{L}^+ representing a claim containing “ x ”, \mathbf{L}^+ includes sentences representing the claims obtained by substituting “ y ” and then “ z ” for “ x ”. The central idea animating (CMP) is that in such a case, we can obtain verdicts concerning degrees of belief in claims not containing “ x ” by working with a modeling language that contains all the sentences of \mathbf{L}^+ except those representing “ x ”-claims.

But that characterization doesn’t operate at the level of syntax we need. Our modeling technique works at the level of whole sentences (read as unanalyzable formal symbols with no internal structure) and their truth-values. We reach a systematic constraint at this level of syntax by noticing the following: given that ideal rationality requires the agent to be certain at t_1 that “ y ” uniquely picks out the denotation of “ x ” in the situation above, it also requires the agent to be certain at t_1 that Y has the same truth-value as X . Similarly, in the situation above ideal rationality requires the agent to be certain at t_2 that Z has the same truth-value as X . And it is conditions like these with which (CMP) operates. Since at each moment in the time sequence ideal rationality requires the agent to be certain that another claim has the same truth-value as X , (CMP) allows us to obtain verdicts not concerning X by working with a modeling language that includes these other claims and excludes X .²⁸

One strong indication that this line of reasoning is correct is that for modeling languages involving only context-insensitive claims, (CMP) can be proven as a theorem from our other five systematic constraints. A detailed proof appears in the Appendix to this paper, but its basic idea is simple. Suppose X is a sentence in \mathbf{L}^+ but not in \mathbf{L} . If the antecedent of (CMP) is met, there exists a $Y \in \mathbf{L}$ such that ideal rationality requires the agent to be certain that $Y \equiv X$ at t_1 (or whatever is the earliest time in the model’s time sequence). But if all the sentences in \mathbf{L}^+

²⁸Besides obtaining the required level of syntax, (CMP) also further generalizes the ideas in the previous paragraph about claims’ containing various expressions. Clearly the great majority of context-sensitive claims derive their context-sensitivity from context-sensitive expressions within them. But if there are context-sensitive claims that cannot be understood as being context-sensitive because of context-sensitive expressions they contain, (CMP) is designed to be flexible enough to apply to them as well. I should admit that I have not thoroughly tested (CMP) on claims of this kind, in part because I am unsure whether I have been able to find any. So I will set them (and the possibility of their existence) aside for the rest of this paper.

represent context-insensitive claims, $Y \equiv X$ is context-insensitive, so if the agent is required to be certain of it at t_1 she continues to be so required at all subsequent times. At every moment in the time sequence, ideal rationality requires the agent to be certain that X has the same truth-value as Y . From the point of view of the model, X is redundant; everywhere X appears in a verdict of M^+ , Y could just as easily appear.

This line of thought indicates a suggestive way of thinking about what's going on in (CMP). Consider a situation in which we begin with a model M and then extend its language by adding a sentence X to generate a model M^+ . First, take the case where all the sentences in L^+ are context-insensitive. As we've just seen, there exists an $Y \in L$ such that at every time in the time-sequence ideal rationality requires the agent to be certain that $Y \equiv X$. From the point of view of the modeling technique, all we have done in adding X to our language is introduce a *synonym* for Y . If we are investigating a degree of belief involving the claim represented by X — if, say, we want a verdict on a conditional credence in a conjunction that features X as a conjunct — it is important to have X represented in our language. But for degrees of belief not involving the claim represented by X , the presence of X in the modeling language is unnecessary. Since X is just a synonym for Y , the only new piece of *information* representable by L^+ not representable by L is the fact that there's this synonym for Y , X . And that piece of data, which we might characterize from the model's point of view as purely *linguistic* data about synonymy, isn't relevant to degrees of belief not involving the claim represented by X .²⁹

Similar remarks apply to the case in which L^+ contains context-sensitive sentences, but there is still an $Y \in L$ that ideal rationality requires the agent to be certain has the same truth-value as X at every moment in the time sequence. The interesting case is the one in which X is context-sensitive and there is no $Y \in L$ with the same truth-value as X at *every* moment in the time-sequence. In this case, it remains suggestive to think of X as having a synonym in L at each moment in the time-sequence; it's just that *which* sentence X is synonymous with periodically changes. Again, model M^+ will be able to represent some information not representable by M , namely information about how X switches its synonyms in L^+ from time to time. But again, from the point of view of the model we might characterize this information as purely linguistic, and so irrelevant to degrees of belief not involving the claim represented by X . Thus we can obtain verdicts about those degrees of belief using M , a model whose language does not include X .

I don't want to put too much emphasis on this line of thought. After all, from a point of view that reflects a deeper understanding of the *content* of the claims represented by these sentences, the claim represented by X could be much more than a mere synonym for other claims with the same truth-value. For example,

²⁹In response to questions from early readers of this manuscript, I'd like to emphasize that the point made in the last sentence of this paragraph is not simply an assertion on my part — it is a corollary of the proof that appears in the Appendix. In the case of a model whose language contains only context-insensitive sentences, it deductively *follows from* systematic constraints (1) through (5) that under the conditions in the antecedent of (CMP) X is irrelevant to establishing credences in sentences in which X does not appear. What appears in the next paragraph is the unproven extension of this idea. My hope is that once the reader understands how (CMP) works in the context-insensitive case, this will lend some plausibility to its extension to context-sensitive cases.

in our SI' model of Sleeping In above, at t_2 ideal rationality requires Olga to be certain that “Now is the second time I awaken today” has the same truth-value as any tautology in the modeling language of SI^- . Various views in the philosophy of language will have different things to say about the content of “Now is the second time I awaken today” when it is uttered at t_2 , but most of them will admit that on some level there is an important distinction between that claim and a claim of the form $P \vee \sim P$. From that point of view, “Now is the second I awaken today” is much more than a mere *synonym* for a tautology. From the point of view of the modeling technique, however, the sentences that represent claims are just unanalyzable symbols whose only properties are their truth-values, the relations their truth-values bear to the truth-values of other sentences, and the credences assigned to them by various credence functions. From that point of view, a new sentence like X brings with it nothing other than its synonymies and the power to represent information about those synonymies. Thus despite the deeper considerations about content floating around, the effects on the verdicts of M of adding X to the modeling language are nil.

In the end, the best test of a modeling technique is whether it yields verdicts that match our intuitive judgments about the requirements of ideal rationality in stories where such requirements are obvious and uncontroversial. In this section and the last, we have shown how including (CMP) as a constraint of our modeling technique allows that technique to yield intuitively correct verdicts for The Die and Sleeping In. I have also in this section tried to provide a couple of intuitive suggestions about how (CMP) works and why we should accept it as representing a requirement of ideal rationality. The suggestions here are intended not as proofs, but simply as approaches that will help readers understand (CMP) and perhaps come to support it. Ultimately, (CMP) will have to be judged by the acceptability of the verdicts it yields and by the fruitfulness of the modeling technique in which it plays a part. As far as fruitfulness goes, we have an important problem to address, which it is the business of the next section to introduce.

1.8. The Remaining Problem. As was mentioned in the previous section and is proved in the Appendix, in models whose modeling languages include no context-sensitive claims (CMP) simply reaffirms verdicts that could be derived without it from the other systematic constraints. So anyone who finds our first five systematic constraints acceptable should find (CMP) acceptable in those cases as well. As for models with modeling languages containing context-sensitive claims, (CMP) should be heartily supported in those cases by people who find themselves intuitively drawn to the relevance-limiting theses proposed in the introduction — theses claiming that self-locating beliefs have limited relevance to beliefs *de dicto*. For suppose we have a model whose language contains sentences representing self-locating beliefs. Those self-locating beliefs will contain context-sensitive expressions. Suppose further that at each moment in the time-sequence, there are expressions that ideal rationality requires the agent to be certain uniquely pick out the denotations of those context-sensitive expressions, and that themselves are context-insensitive (in this story and every other). Then for every self-locating belief represented in the model, there will be a belief *de dicto* that the agent is required to be certain has the same truth-value as that self-locating belief. (CMP) tells us that we may then establish verdicts about ideally rational degrees of belief in *de dicto* claims using a model whose language contains no sentences representing self-locating beliefs. In this case,

once the background certainties are established, ideally rational degrees of belief in beliefs *de dicto* can be determined without reference to self-locating claims. This seems to validate a position that self-locating beliefs are irrelevant to beliefs *de dicto*.

The relevance-limiting theses hold that self-locating beliefs have limited relevance to beliefs *de dicto* in *every* case. The argument in the previous paragraph shows that (CMP) affirms this judgment for cases in which for each context-sensitive expression the agent has at each moment a context-insensitive expression that uniquely picks out its denotation. Thus the backer of a relevance-limiting thesis should endorse (CMP) for at least the range of cases in which the agent has uniquely denoting context-insensitive expressions available for each of her context-sensitive expressions — after all, (CMP) yields precisely the relevance-limiting verdicts desired in such cases. But what about cases in which the agent does not have such uniquely denoting expressions available?

Let me start by pointing out that such cases are extremely rare in real life. When I use a context-sensitive expression like “today,” I almost always have a context-insensitive expression like “January 17, 2006” available. Even if I don’t know, say, what time it is, I can construct an expression with the same denotation as “now” by referring to experiences I am having at this moment (for example “the time on January 17, 2006 when I was sitting in front of my computer typing the sentence. . .”). And even if I have had such experiences multiple times, I can refer to this one as the second such time, the third such time, etc. Thus for an agent to lack a uniquely denoting context-insensitive expression for a context-sensitive expression, she must have multiple experiences that are subjectively identical in a very strong sense. Not only must the experiences be qualitatively identical from the point of view of the agent; the agent must also be unable to determine where in the numerical series of exposures to such experiences this particular experience falls. And this almost never happens in real life. In our day-to-day lives we constantly have multiple context-insensitive expressions available to pick out who we are, where we are, what time it is, etc. The only situations I have been able to think of in which such context-insensitive expressions are unavailable are hypotheticals involving very precise kinds of forgetting or bizarre experiments with Doppelgangers.³⁰

Thus for purposes of determining requirements of ideal rationality in stories drawn from real life, we need not worry much about cases in which an agent lacks a uniquely denoting context-insensitive expression. Nevertheless, such cases are pertinent to our aims in this paper for two reasons.

First, we need to evaluate the range of applicability of our modeling technique. The discussion in the last section suggested a general strategy for modeling stories in which context-sensitive claims play an important role. For each moment in the time sequence and each context-sensitive expression contained in an important claim, we find a context-insensitive expression that ideal rationality requires the agent to be certain uniquely picks out the denotation of the context-sensitive expression. We then create a model whose modeling language represents not just the context-sensitive claims, but also claims in which the context-insensitive expressions have

³⁰As we will see in the latter half of the paper, the situation need only involve the credible *threat* of memory loss or of the creation of Doppelgangers for an agent to lack the needed context-insensitive expressions. Such Doppelganger examples appear in (Arntzenius 2003), (Bostrom manuscript), (Elga 2004), and (Meacham manuscript).

been substituted for the context-sensitive ones. This expanded model will have a submodel whose language represents all the resulting context-insensitive claims but none of the context-sensitive ones. (CMP) guarantees that the expanded model is a conservative supermodel of this submodel, so we can work with the submodel to generate verdicts whose analogues will hold true of the expanded model. And since the submodel contains no context-sensitive claims, (LC) will be a fertile source of such verdicts.³¹

But this strategy does not tell us how to model stories in which context-sensitive claims are relevant but the agent does not have available uniquely denoting context-insensitive expressions for each of her context-sensitive expressions. If (CMP) cannot be applied to such stories, there will be no way to apply (LC) and thus no way for our modeling technique to generate useful verdicts. We need to examine stories in which the agent lacks uniquely denoting context-insensitive expressions so as to determine whether our modeling technique can answer the questions in such stories at all.

The second reason we need to examine such stories is the evaluation of the relevance-limiting theses proposed in the introduction. The discussion in this section and the last may suggest that our evaluation of those theses in the introduction was somewhat unfair. We have been emphasizing that self-locating beliefs almost never travel alone. Take the case discussed in the introduction, in which I am in a bunker under heavy fire and notice that my watch reads 12:05. What precisely I learn at that moment depends on how one defines “learn,” but clearly if I become certain that it is now 12:05 ideal rationality also requires me to be certain that I live long enough to see 12:05. The latter is a belief *de dicto* (or can be made into one by substituting my name for “I”), and the defender of a relevance-limiting thesis may argue that this belief *de dicto* is in fact the one responsible for the changes to our other beliefs *de dicto* cited as counter-examples to the Strong Relevance-Limiting Thesis. In other words, the defender of a relevance-limiting thesis might argue that when I learn “It is now 12:05,” I also learn “I live long enough to see 12:05,” and that the latter belief, the one *de dicto*, is the one that is relevant to my other beliefs *de dicto*. Thus the bunker example does not demonstrate that self-locating beliefs can be relevant to beliefs *de dicto*.

Following this line of thought, someone who thinks that self-locating beliefs are irrelevant to beliefs *de dicto* might want to reformulate the Strong Relevance-Limiting Thesis as:

³¹This is essentially the strategy we applied in modeling Sleeping In in the last section, moving from SI, to SI' (the expanded model), to SI⁻ (its submodel). I did, however, take some shortcuts in implementing the strategy there for the sake of efficient exposition. First, while the modeling language of SI contains a sentence representing “The clock now reads 9am,” this sentence turns out to be unnecessary in answering the question posed by Sleeping In, so I left it out of SI'. (Including that sentence makes the model and its analysis a bit more complex, but in the end the same verdicts result.) Second, the context-insensitive expression used for “now” at t_2 is “the second time I awaken today.” One of the sentences in SI' represents “Now is the second time I awaken today,” so strictly following the strategy just suggested would require the modeling language of SI' to contain a sentence representing the claim “The second time I awaken today is the second time I awaken today.” Since this claim is a tautology, it can be equally well-represented by any tautology in the modeling language of SI', so I chose not to add an extra sentence representing this claim.

Revised Strong Relevance-Limiting Thesis: It is never rational for an agent who learns *only* self-locating beliefs to respond by altering her degree of belief in a belief *de dicto*.

The idea would be that the bunker example is not a counter-example to the Revised Strong Relevance-Limiting Thesis because in that example the agent learns both self-locating and *de dicto* beliefs. A similar revision could be made to the Weak Relevance-Limiting Thesis:

Revised Weak Relevance-Limiting Thesis: It is never rational for an agent who learns *only* self-locating beliefs to respond by altering her relative degrees of belief in two beliefs *de dicto* compatible with those self-locating beliefs.

Defending these revised theses requires adopting particular positions about when an agent should be understood to “learn” something. To avoid having to debate such positions, we should evaluate the revised relevance-limiting theses by analyzing cases in which, regardless of how inclusive a definition of “learn” one adopts, it is clear that the agent has not learned *any* beliefs *de dicto*. In our formal terms, we want to analyze stories in which between two times ideal rationality requires the agent to become certain of some self-locating claims but *no* claims *de dicto*.

Now suppose that we have such a story, and ideal rationality requires the agent to become certain of some particular self-locating claim. If the agent has a uniquely denoting context-insensitive expression available for every context-sensitive expression in that self-locating claim, there is another claim that substitutes the context-insensitive expressions for the context-sensitive expressions in the self-locating claim. Ideal rationality requires the agent to become certain of that other claim between the two times, and that other claim is *de dicto*. So if we want a story in which it is ideally rational for an agent to become certain of some self-locating claims but no claims *de dicto* between two times, that story will have to leave the agent without uniquely denoting context-insensitive expressions for some of the context-sensitive expressions in those self-locating claims. And so we will have to focus on such stories to test the revised relevance-limiting theses.

Thus despite the infrequency of such cases in real life, there are two reasons we want to examine stories in which an agent lacks some uniquely denoting context-sensitive expressions. First, we want to see if our modeling technique can be applied to generate verdicts for such stories. Second, the revised relevance-limiting theses can only be tested in such stories. The latter half of this paper will be devoted to the analysis of one story, the Sleeping Beauty Problem, in which an agent lacks any uniquely denoting context-insensitive expression for an important context-sensitive expression. There, I will develop two different strategies for using our modeling technique to obtain verdicts for such a story. I will also demonstrate that the relevance-limiting theses, even in their revised forms, are false.

2. THE SLEEPING BEAUTY PROBLEM

2.1. The Problem.

The Sleeping Beauty Problem: Beauty has volunteered for an on-campus experiment in epistemology. She arrives at the lab on Sunday, and the details of the experiment are explained to her in full. She will be put to sleep Sunday; the experimenters will then flip

a fair coin. If the coin comes up heads, they will awaken her on Monday, leave her awake for a bit, then tell her it's Monday, leave her awake a bit longer, and finally put her back to sleep. If the coin comes up tails, they will engage in the same Monday process, then *erase any memory she has of her Monday awakening*, awaken her on Tuesday, leave her awake for a bit, tell her it's Tuesday, leave her awake a bit longer, then put her back to sleep.

Beauty is told and believes with certainty all the information in the preceding paragraph, then she is put to sleep. Some time later she finds herself awake, unsure whether it is Monday or Tuesday. What does ideal rationality require at that moment of Beauty's degree of belief that the coin came up heads?

The two classic answers to the Sleeping Beauty Problem are $1/2$ (defended by Lewis (2001)) and $1/3$ (defended by Adam Elga (2000)). In this section I will first present conclusions that Lewis and Elga both agree on, then discuss the point where their views diverge and the principles they rely upon to obtain their conflicting results.

We will eventually have two models of the Sleeping Beauty Problem. The first, **S1**, concerns Beauty's degrees of belief only while she is awake on Monday:

Model: **S1**

Story: Sleeping Beauty

L1: Built on these atomic sentences:

M Today is Monday.

H The coin comes up heads.

t -seq: Contains these times:

t_1 Monday morning, after Beauty awakens but before she knows whether it is Monday or Tuesday.

t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.

ES: (1) $0 < P_1(M) < 1$

(2) $P_2(M) = 1$

(3) $0 < P_1(H) < 1$

(4) $0 < P_2(H) < 1$

(5) $P_1(H \supset M) = 1$

(6) $P_2(H \supset M) = 1$

I_{1,2}: M , sentences entailed by M

Note that H is to be taken as an eternal sentence, and that its negation indicates that the coin comes up tails. Similarly, given what Beauty knows at t_1 , in the context of this model $\sim M$ indicates that today is Tuesday. Extrasystematic constraint (5) comes from the structure of the experiment: if the coin comes up heads, Beauty is awakened only on Monday.

While M represents a claim that is context-sensitive in some contexts, it is not context-sensitive between the two times represented in our model, t_1 and t_2 . While Beauty learns something about what day it is between t_1 and t_2 , she does not believe that the truth-value of "Today is Monday" has changed between those two times. At t_2 Beauty is certain that it is now the same day that it was at t_1 . Ideal rationality requires Beauty to be certain that "Today is Monday" has had the same

truth-value the entire day; earlier in the day she just didn't know what that truth-value was. So there are no sentences in the modeling language of **S1** that ideal rationality requires Beauty to be certain of at t_1 but less-than-certain of at t_2 . The antecedent of the conditional in (LC) is satisfied, and we can apply (LC) to obtain:

$$(15) \quad P_2(H) = P_1(H | M)$$

Given our synchronic systematic constraints, it is a consequence of Equation (15), extrasystematic constraint (1), and extrasystematic constraint (5) that

$$(16) \quad P_1(H) < P_2(H)$$

We can see why this should be the case. By extrasystematic constraint (1), ideal rationality prohibits Beauty at t_1 both from being certain that today is Monday and from being certain that today is Tuesday. So her t_1 degree of belief that the coin came up heads is required to be a weighted average of her degree of belief that the coin came up heads on the supposition that it's Monday and her degree of belief that the coin came up heads on the supposition that it's Tuesday. However, if it's Tuesday, the coin didn't come up heads (extrasystematic constraint (5)), so ideal rationality requires Beauty's degree of belief that the coin came up heads conditional on the supposition that today is Tuesday to be zero. Thus Beauty's t_1 degree of belief that the coin came up heads is required to be strictly less than her degree of belief that the coin came up heads on the supposition that today is Monday. By Equation (15), ideal rationality requires Beauty's t_2 degree of belief that the coin came up heads to equal her t_1 degree of belief that the coin came up heads on the supposition that today is Monday. So ideal rationality requires Beauty's t_1 degree of belief that the coin came up heads to be strictly less than her t_2 degree of belief that the coin came up heads.

We can also derive some more subtle relationships from Equation (15) and our synchronic systematic constraints. Constraints (1) through (4) imply an equation called Bayes' Theorem, which can be applied to give us

$$(17) \quad P_1(H | M) = \frac{P_1(M | H) \cdot P_1(H)}{P_1(M | H) \cdot P_1(H) + P_1(M | \sim H) \cdot P_1(\sim H)}$$

With our synchronic systematic constraints, extrasystematic constraint (5), and Equation (15), this becomes

$$(18) \quad P_2(H) = \frac{P_1(H)}{P_1(H) + P_1(M | \sim H) \cdot (1 - P_1(H))}$$

Equation (18) represents a relationship between three degrees of belief at play in the Sleeping Beauty Problem. $P_1(H)$ is the ideally rational degree of belief for Beauty to have Monday morning (before she knows what day it is) that the coin came up heads. $P_2(H)$ is the ideally rational degree of belief for Beauty to have Monday night (after she knows that it's Monday) that the coin came up heads. $P_1(M | \sim H)$ is the ideally rational degree of belief for Beauty to have Monday morning that it is currently Monday, on the supposition that the coin came up tails. If the coin came up heads, the structure of the experiment dictates that it must be Monday. But if the coin came up tails, as far as Beauty knows Monday morning today could be either Monday or Tuesday. $P_1(M | \sim H)$ represents Beauty's degree of belief that it's currently Monday on that supposition.

Both Lewis and Elga would agree with all the results obtained so far. Their arguments are a bit different from the ones I've presented here: they use a slightly

different modeling language, and they employ standard Conditionalization as opposed to (LC). But their modeling language translates easily to ours, and since there are no sentences involved that are context-sensitive between the two times in question, using (LC) instead of Conditionalization makes no difference to these results.

However, these results are insufficient to yield the verdicts we are after. Equation (18) relates three values in the problem; knowing any two of them is sufficient to calculate the third. But so far we haven't seen any constraints give a precise value for even one of the three. Lewis and Elga deal with this shortfall by applying principles for ideally rational degrees of belief that are not among our systematic constraints.

First, both Lewis and Elga apply a "highly restricted principle of indifference" (Elga 2000, 144) to argue that Monday morning, ideal rationality requires Beauty to assign equal degrees of belief to its being Monday or Tuesday on the supposition that the coin came up tails. In other words, some sort of indifference principle guarantees a $P_1(M | \sim H)$ value of $1/2$. Lewis's and Elga's arguments then diverge. Each employs the Principal Principle to set one of the other values in Equation (18), but each employs it to set a different value (leading ultimately to their conflicting results).

The Principal Principle was first proposed by Lewis in (Lewis 1980). For our purposes, the Principle says that an agent who is certain that a particular event has a particular objective chance of occurring and who has no inadmissible evidence is required by ideal rationality to have a degree of belief in that event's occurrence equal to that objective chance. We've seen an example already: in *The Die*, when Marilyn is certain that a fair die has been thrown but doesn't yet know anything about its outcome, the Principal Principle suggests that ideal rationality requires her to have equal degrees of belief in each of the possible outcomes of the roll.

Applying the Principal Principle requires differentiating between admissible and inadmissible evidence. Admissible evidence is either wholly irrelevant to the event, or if relevant is relevant in that it plays a role in determining the objective chances. Inadmissible evidence includes information that the event actually occurred (or didn't), information that implies that the event occurred (or didn't), and information positively correlated with the event's occurrence (or non-occurrence). So when Marilyn knows that the die is fair, but doesn't know anything about the outcome of the roll, she has no inadmissible evidence and is required to set her degrees of belief equal to the chances. Once Marilyn knows the die roll came out odd, she has inadmissible evidence and so can deviate her degree of belief in an outcome of 3 away from its objective chance of $1/6$.

Everyone involved in the *Sleeping Beauty* discussion agrees that an application of the Principal Principle requires Beauty to have a Sunday night degree of belief in heads of $1/2$. At that point, she knows the coin is fair, but the coin flip hasn't even occurred yet, so she couldn't possibly have any inadmissible evidence about its outcome. The disagreement between Elga and Lewis is ultimately a disagreement about at which points after Sunday night Beauty is in possession of inadmissible evidence, and so is permitted by the requirements of ideal rationality to have a degree of belief in heads not equal to $1/2$.

Elga argues that it would make no difference to the experiment if the coin were flipped Monday night after Beauty is put to sleep instead of Sunday night. If the

coin were flipped Monday night, the same kind of argument as was given in the previous paragraph would show that on Monday night Beauty had no inadmissible evidence (since the coin hasn't been flipped yet). Thus Elga believes the Principal Principle requires Beauty to set $P_2(H) = 1/2$. Applying his principle of indifference and Equation (18), this generates a $P_1(H)$ value of $1/3$.

But this result implies that Beauty has inadmissible evidence when she awakens Monday morning; otherwise it would conflict with the Principal Principle for her to assign a degree of belief to heads other than $1/2$. Lewis, examining the information Beauty gains between Sunday night and Monday morning, cannot find anything that counts as inadmissible. Thus Lewis concludes by the Principal Principle that Beauty is required to set $P_1(H) = 1/2$. By the indifference principle and Equation (18), this gives him a $P_2(H)$ value of $2/3$. Thus Lewis concludes that Beauty has inadmissible evidence Monday night.³²

The basic problem with these analyses is that we lack a precise enough definition of what constitutes inadmissible evidence. As she progresses from Sunday night, to Monday morning, to Monday night, Beauty gains various pieces of self-locating evidence. Once we've progressed beyond Sunday night, our specification of "inadmissible evidence" is insufficient to determine when in the process she has inadmissible evidence and when she doesn't. To my knowledge, no one has succeeded in fleshing out the criteria for inadmissible evidence to the extent that the Principal Principle could, for example, be formalized as one of our modeling technique's systematic constraints.³³

But it turns out we don't need to formalize the Principal Principle to settle the dispute between Elga and Lewis. The constraints we already have in place are sufficient to do so.

2.2. Sleeping Beauty Solution. There now exists a fairly vast literature attempting to adjudicate the dispute between Elga and Lewis. Most of these attempts involve comparing Sleeping Beauty to another story that feels analogous, drawing out a particular solution's counter-intuitive implications for another case, or applying some unsupported position about relevance directly to the problem. The trouble with these one-off arguments is that they yield no general understanding and no systematic approach that will help us reach verdicts about other stories.³⁴

³²Was Lewis's Sleeping Beauty position based in part on the Revised Weak Relevance-Limiting Thesis? At (2001, 174) Lewis argues for a $P_1(H)$ value of $1/2$ on the basis of a "Premiss" that "Only new relevant evidence, centred or uncentred, produces a change in credence; and the evidence [that today is Monday or Tuesday] is not relevant to HEADS versus TAILS." No further explanation of this relevance judgment is offered, perhaps because Lewis thought it was obvious. I certainly think that those who read Lewis (2001) and find this point obvious often do so on the basis of an intuition like the Revised Weak Relevance-Limiting Thesis.

³³Ned Hall (2004) improves on the notion of admissibility offered in (Lewis 1980), going so far as to offer an alternative formulation of the Principal Principle that does away with the notion of admissibility entirely. However, Hall's analysis will not help us with the Sleeping Beauty Problem. Applying his criterion for admissibility would require answers to the very questions of evidential relevance we are trying to use the Principal Principle to settle.

³⁴(Halpern 2005) is a welcome exception to this complaint. Halpern presents a systematic, general approach to modeling stories in which an agent loses track of the time. He then applies this approach to obtain a solution to the Sleeping Beauty Problem. However, Halpern's approach has two flaws. First, it yields a position about the relation between Beauty's Monday morning degrees of belief and her Monday night degrees of belief that I take to be unsupportable. See Section 2.3 below for discussion of this point. Second, it yields a different ideally rational Monday

In the first half of this paper we developed a systematic, general modeling framework by thinking about stories with questions whose answers I take to be obvious. We will now solve the Sleeping Beauty Problem by simply applying that framework once more.

In our analysis of the Sleeping Beauty Problem to this point, we have applied only systematic constraints (1) through (5). While Lewis and Elga use Conditionalization instead of (LC), we’ve noted that the distinction between the two is immaterial to our analysis thus far. Thus all the systematic apparatus employed so far is common to the disputants in the case.

The disagreement between Elga and Lewis can be settled, however, by applying (CMP) — a principle neither of them employs. To apply (CMP), we need to focus on the transition in the problem from Sunday night to Monday night. Recall that one of the tenets of our system (first mentioned in Section 1.3 above) is that we can model an agent’s degrees of belief at two times without our model’s representing the agent’s degrees of belief between those two times. Thus in the Sleeping Beauty Problem we can construct a model of Beauty’s Sunday night credences and Monday night credences without any reference to her credences Monday morning. We’ll call this model **S0**:

Model: **S0**

Story: Sleeping Beauty

L: Built on these atomic sentences:

M Today is Monday.

H The coin comes up heads.

t -seq: Contains these times:

t_0 Sunday night, after Beauty has heard the experiment described but before she is put to sleep.

t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.

ES: (1) $P_0(M) = 0$

(2) $P_2(M) = 1$

(3) $0 < P_0(H) < 1$

(4) $0 < P_2(H) < 1$

I_{0,2}: M , sentences entailed by M

Between t_0 and t_2 , Beauty learns M . M represents a context-sensitive claim, and ideal rationality requires Beauty to be certain of $\sim M$ at t_0 and certain of its falsity at t_2 . So (LC) cannot be applied to generate verdicts for **S0**. Yet M is context-sensitive because it contains the expression “today.” On Sunday night, Beauty has a uniquely denoting context-insensitive expression for “today” (namely “Sunday”), and on Monday night Beauty has a uniquely denoting context-insensitive expression for “today” (namely “Monday”). So we introduce a submodel of **S0** called **S0⁻**:

Model: **S0⁻**

Story: Sleeping Beauty

L⁻: Built on this atomic sentence:

H The coin comes up heads.

t -seq: Contains these times:

morning degree of belief in heads in the Sleeping Beauty Problem than it does in the Technicolor Beauty story I introduce in Section 2.5 below. See Section 2.6 for arguments that the ideally rational Monday morning degree of belief in heads must be the same in both stories.

t_0 Sunday night, after Beauty has heard the experiment described but before she is put to sleep.

t_2 Monday night, after Beauty has been told it is Monday but before she is put back to sleep.

ES: (1) $0 < P_0^-(H) < 1$

(2) $0 < P_2^-(H) < 1$

$\mathbf{I}_{0,2}^-: \emptyset$

Since \mathbf{L}^- contains no sentences representing context-sensitive claims, and therefore no sentences ideal rationality requires Beauty to be certain of at t_0 and less-than-certain of at t_2 , (LC) can be applied to generate

$$(19) \quad P_2^-(H) = P_0^-(H \mid \emptyset) = P_0^-(H)$$

We can then apply our synchronic systematic constraints to the extrasystematic constraints on $\mathbf{S0}$ to determine that $P_0(M \equiv \mathbf{F}) = 1$ and $P_2(M \equiv \mathbf{T}) = 1$. Since \mathbf{L}^- contains a tautology and a contradiction, the conditions in the antecedent of (CMP) have been met. So we conclude that $\mathbf{S0}$ is a conservative supermodel of $\mathbf{S0}^-$. Thus we obtain the analogue of Equation (19) in $\mathbf{S0}$:

$$(20) \quad P_2(H) = P_0(H)$$

The argument concludes by combining the information from both our models of the Sleeping Beauty Problem. Between Sunday night and Monday night, Beauty learns “Today is Monday.” Our analysis of $\mathbf{S0}$ via (CMP) demonstrates that this information is irrelevant to whether the coin came up heads. Ideal rationality requires Beauty to assign the same degree of belief to heads on Monday night as she assigned on Sunday night. Our earlier analysis of $\mathbf{S1}$ already demonstrated that Beauty is required to assign a higher degree of belief to heads Monday night than she does Monday morning. Putting all this together, ideal rationality requires Beauty to assign a lower degree of belief to heads when she awakens Monday morning than she did when she went to sleep Sunday night. And this conclusion is sufficient to refute Lewis’s answer to the problem.

Notice that this conclusion follows from the analyses of $\mathbf{S0}$ and $\mathbf{S1}$. Those analyses proceeded strictly on the basis of our systematic constraints, without reference to the Principal Principle or a principle of indifference. Thus a technique that correctly models the acquisition of context-sensitive beliefs (by including all six of our systematic constraints) is sufficient to refute Lewis’s position by itself, without supplementation by either of those other principles. That technique concludes that Beauty’s Monday morning degree of belief in heads is required to be less than her Sunday night degree of belief in heads.

If we wanted to apply those principles, we could obtain some more specific results about what Beauty’s Monday morning degree of belief in heads is required to be. For example, we could use the one application of the Principal Principle that everyone agrees on — that Beauty’s Sunday night degree of belief in heads is required to be $1/2$ — to add an extrasystematic constraint to $\mathbf{S0}$ that $P_0(H) = 1/2$. Percolating this constraint through the models, we would wind up with a conclusion that ideal rationality requires Beauty’s Monday morning degree of belief in heads to be less than $1/2$. Further, if we accepted the indifference principle application Lewis and Elga do, we could add an extrasystematic constraint to $\mathbf{S1}$ that $P_1(M \mid \sim H) = 1/2$. Plugging all this information into Equation (18), we would obtain Elga’s result that $P_1(H) = 1/3$. But all this goes beyond the conclusions of our fundamental analysis,

based solely on the formal systematic constraints of our modeling technique, which indicates that for Beauty to assign a Monday morning degree of belief in heads equal to her Sunday night degree of belief would violate the requirements of ideal rationality.

2.3. Objections to This Analysis. I now want to consider two objections to this analysis of the Sleeping Beauty Problem. The first is an attack on the conclusions we drew from model **S1**, especially the conclusion that ideal rationality requires Beauty’s Monday night degree of belief in heads to be greater than her Monday morning degree of belief in heads. Bostrom (manuscript) and Halpern (2005) each hold that Beauty’s degree of belief in heads is required to remain $1/2$ all day Monday, both before and after she learns what day it is. To maintain this position, they argue that there is something wrong with the logic that supports our application of (LC) in **S1**, by which we concluded that ideal rationality requires Beauty’s Monday night degree of belief in heads to be her Monday morning degree of beliefs in heads conditional on its being Monday. In particular, they point out that between t_1 and t_2 Beauty becomes certain of something else besides the claim that it’s Monday; she also becomes certain that she has been *told* it’s Monday. Beauty is required to conditionalize both on this information and on the claim that today is Monday. Model **S1** fails to represent this claim and thereby yields unreliable verdicts about the relation between Beauty’s t_1 and t_2 degrees of belief.

Let’s examine the claim “I have been told that today is Monday” more closely. As we have set up the Sleeping Beauty Problem, Beauty already knows on Sunday night that each day the experimenters awaken her they will eventually tell her what day it is.³⁵ So to capture the information Beauty gains when she is told on Monday what day it is we need to represent two claims in our modeling language: “Today is Monday” and “I have been told today what day it is.” Given that she already knew Monday morning that she was going to be told today what day it is, logical combinations of these two claims are sufficient to capture pieces of information Beauty learns between t_1 and t_2 such as “I have been told that today is Monday.”

The claim “Today is Monday” is already represented in the language of model **S1**. To test Bostrom and Halpern’s suggestion that a relevant piece of information is left unrepresented in that model, we need to create a supermodel of **S1** whose language includes a sentence representing “I have been told today what day it is.” But this extension of **S1**’s language follows a pattern that is eerily familiar. In fact, it is precisely analogous to adding a sentence representing the claim “The announcement about the die has already been made” to our model **D** of The Die. The analogy is so close that I won’t even bother to construct the supermodel **S1**⁺ whose language includes the additional claim. Exactly in the way we did with **D** and **D**⁺, we can show that **S1** and **S1**⁺ meet the antecedent of the conditional in (CMP), and therefore that **S1**⁺ is a conservative supermodel of **S1**. Thus all our verdicts about the relationship between Beauty’s Monday morning degree of belief

³⁵One could set up the story without this stipulation; Beauty could be surprised on Monday that the experimenters decide to reveal what day it is. But to keep the relations between degrees of belief from being distorted, conditions would have to be added so that Beauty is certain when the experimenters reveal this information that their decision to do so has not been influenced by the outcome of the coin flip or the day of the week. With those conditions added, the results would ultimately be the same as what I describe in the main text. Building Beauty’s foreknowledge of the revelation into the description of the story simply streamlines the arguments and the analysis.

in heads and her Monday night degree of belief in heads remain intact when we add a representation of Bostrom and Halpern’s neglected claim to our modeling language. That claim does nothing more than mark the passage of time over the course of Monday. While the fact that it is Monday is relevant to the outcome of the coin flip (since it could have been Tuesday if the coin came out tails), what time it is on Monday is completely irrelevant to Beauty’s degree of belief in heads.³⁶

The first objection is thus easily dismissible using our existing modeling technique. Handling the second requires a bit more care. In Section 1.5 above, we noted that there are two broad categories of exceptions to Conditionalization. One category involves context-sensitive claims. Our alterations to Conditionalization that resulted in (LC) dealt with that exception. The second category of exceptions involves cognitive mishaps by the agent such as forgetting. We now need to treat this second category of exception.

When an agent forgets some information between two times, or loses information due to some other cognitive mishap, she can go from being certain in a claim to being less-than-certain in that claim, leaving her degrees of belief incapable of being modeled by Conditionalization. Our move from Conditionalization to (LC) was motivated by a need to model context-sensitive claims that suffered such a degree-of-belief decrease, but context-insensitive claims can be forgotten just as easily as context-sensitive claims. So it is unclear whether our modeling technique is capable of representing requirements of ideal rationality on agents who undergo cognitive mishaps.

This problem is particularly pressing because Frank Arntzenius has shown that conditionalizing can fail not just when an agent has suffered a cognitive mishap, but also when the agent hasn’t suffered a cognitive mishap but is required by ideal rationality to entertain the possibility that she may have. And this is precisely the position Beauty finds herself in when she awakens Monday morning. Required by ideal rationality to be uncertain whether it is Monday or Tuesday, Beauty knows that if it is Tuesday she has undergone memory erasure since Sunday night. Thus ideal rationality requires her to be uncertain on Monday morning whether she has suffered a cognitive mishap. If our modeling technique is unable to successfully

³⁶If left unchecked, the type of argument Bostrom and Halpern make here could be used to derail almost any application of Conditionalization (or Limited Conditionalization). For presumably in any case in which an agent learns some claim, we might think that ideal rationality also requires her to learn that she has learned that claim. Bostrom implicitly acknowledges this threat when he writes,

In ordinary cases, such changes in indexical information are irrelevant to the hypotheses being considered and can hence be safely ignored. The standard elliptic representation of Bayesian conditionalization can then be used without danger. In certain special cases, however, such delicate changes in indexical information can be relevant, and it is then crucial to recognize and make explicit the hidden intermediary step. Sleeping Beauty, on the model proposed here, turns out to be just such a special case. (Bostrom manuscript, 12)

Bostrom needs Sleeping Beauty to be a special case because of arguments of a very different kind he has presented elsewhere in the paper. But he offers no guidance on systematically separating the “special cases” from cases like The Die, in which we all agree that conditionalization can proceed without taking into account information of the form “I have learned” Our modeling framework provides the needed systematic approach, and as was demonstrated in Section 1.6 handles the extra information in The Die exactly right. Yet when it is applied to the inference in the Sleeping Beauty Problem with which Bostrom and Halpern find fault, that framework refutes their argument by ruling the extra information irrelevant there as well.

model agents who have either suffered a cognitive mishap or are required by ideal rationality to believe that they may have, that technique’s verdicts about Beauty’s Monday morning degrees of belief will be unreliable.

Arntzenius argues that conditionalizing can fail when an agent believes she *may* have suffered a cognitive mishap by analyzing a story he calls “Two Roads to Shangri La.” (Arntzenius 2003) Stripping it down to its structural essentials and leaving out some of the details Arntzenius adds for color, the story runs like this:

Shangri La: You have reached a fork in the road to Shangri La. The guardians of the tower will flip a fair coin to determine your path. If it comes up heads, you will travel the Path by the Mountains; if it comes up tails, you will travel the Path by the Sea. Once you reach Shangri La, if you have travelled the Path by the Sea the guardians will alter your memory so you remember having travelled the Path by the Mountains. If you travel the Path by the Mountains they will leave your memory intact. Either way, you will remember having travelled the Path by the Mountains.

The guardians explain this entire arrangement to you, you believe their words with certainty, then they flip the coin. Suppose the coin comes up heads and you travel the Path by the Mountains. What does ideal rationality require of the relationship between your degree of belief in heads while you are on the path and your degree of belief in heads once you reach Shangri La?

Stories involving cognitive mishaps can be categorized into three groups. The first group is the most straightforward kind. At its most basic level, our modeling technique is designed to model stories in which an agent gains information over time. The claims the agent learns she becomes certain of; the question in the story is about how ideal rationality requires her degrees of belief in other claims — claims she is not certain of — to be affected by the information she gains. The most straightforward cognitive mishap stories involve the opposite process: an agent loses some information she previously possessed. In our models, this is represented as a claim’s having an unconditional credence of 1 at one moment in the time sequence, then an unconditional credence less than 1 at a later time.

The Shangri La story provides two excellent examples of this kind of credence change, one caused by actual cognitive mishap and another caused by cognitive mishap the agent merely believes may have occurred. Let t_1 be a time when you are walking down the Path by the Mountains, and t_2 be after you have reached Shangri La. If the coin had come up tails instead of heads, we would have a case of actual cognitive mishap. At t_1 you would be walking down the Path by the Sea, and ideal rationality would require you to be certain that the coin came up tails. Once you reached Shangri La, the guardians would alter your memory and you would be uncertain whether you had walked the Path by the Sea or the Path by the Mountains. Thus ideal rationality would require you to be uncertain at t_2 whether the coin came up tails.

Now consider the actual case, in which the coin lands heads. At t_1 ideal rationality requires you to be certain that the coin came up heads. When you reach Shangri La, you suffer no actual cognitive mishap; the guardians do not tamper with your memory at all. However, you know that if you had travelled the Path by the Sea, they would have altered your memories so they would be like the memories

you have now. Ideal rationality requires you to be uncertain whether you actually travelled the Path by the Mountains. So ideal rationality requires you at t_2 to be less-than-certain that the coin came up heads. This is despite the fact that you have suffered no cognitive mishap; the effect is caused by the fact that ideal rationality requires you to be less-than-certain that no cognitive mishap has occurred.

Our primary concern about cognitive mishap stories is that applying our modeling technique will yield absurd results, like those we obtained in Equation (6) when we tried to use Conditionalization to model Sleeping In. But with cognitive mishap stories falling into the first group — stories in which a cognitive mishap or the possibility thereof requires the agent to become less-than-certain of a claim she once was certain of — our modeling technique is already protected against such results. The threat is that modeling such stories by conditionalizing will cause trouble because of the decrease in certainty. But (LC) already restricts us from conditionalizing in cases where ideal rationality requires an agent to move from certainty in a claim to less-than-certainty in that claim. So we need not worry that applying our modeling technique to these stories will yield results unrepresentative of ideal rationality.

And in fact, we can do more than just avoid absurd results. Applying some care and a bit of ingenuity, we can actually generate verdicts using our modeling technique that represent requirements of ideal rationality for stories in the first group. We'll take Shangri La as an example. Because of the required move from certainty in heads at t_1 to less-than-certainty at t_2 , any model of Shangri La will be incapable of relating t_1 credences in heads and t_2 credences in heads directly via (LC). But we can relate such credences through an indirect method. The key is to include in our model's time-sequence not just t_1 and t_2 , but also a time t_0 after the guardians have explained the arrangement to you but before they have flipped the coin.

I won't actually give such a model here, but I'll outline the important moves. Once any context-sensitive sentences were removed from the modeling language by an application of (CMP), (LC) could be applied to relate t_0 and t_1 credences. The results would not be very impressive — they would simply echo the extrasystematic constraints requiring you to be less-than-certain of heads at t_0 and certain of heads at t_1 . The more interesting results would come when we related t_0 and t_2 credences. Recalling that our modeling technique allows us to relate credences at two times without relying on information about credences in the interim (see Sections 1.3 and 2.2 above), we can derive verdicts relating t_0 and t_2 credences directly. Once (CMP) had cleared out any context-sensitive sentences again, t_0 credences and t_2 credences could be related using (LC). Ideal rationality does not require you to be less-than-certain of any claims at t_2 that you were required to be certain of at t_0 ; in fact, you have no more or less information relevant to the outcome of the coin flip at t_2 than you had at t_0 . Applying (LC) would yield a verdict that $P_2(H) = P_0(H)$.

Finally, we might include an extrasystematic constraint in our model reflecting an uncontroversial application of the Principal Principle to yield $P_0(H) = 1/2$. Without directly relating t_1 and t_2 credences via (LC), but working indirectly by relating each separately to t_0 credences, our modeling technique suggests that ideal rationality requires you to be certain as you walk the Path by the Mountains that the coin came up heads, then to have a degree of belief of $1/2$ in heads once you

reach Shangri La. This squares perfectly, I think, with what intuition tells us about the story.³⁷

The second group of cognitive mishap stories doesn't involve your losing any information at all. In these stories, the threat of cognitive mishap merely prevents you from gaining some information you would have gained otherwise. The Sleeping Beauty Problem provides an excellent example of this group. When we consider the transition from Monday night to Tuesday morning, we have an example of a first-group cognitive mishap: the memory erasure causes Beauty to lose such pieces of information as "I have already been awakened once during the experiment." But now consider the transition from Sunday night to Monday morning. There are claims that ideal rationality requires Beauty to be certain of Sunday night and less-than-certain of Monday morning. These include claims like "Today is Sunday." But the required decrease in certainty of those claims is due to their context-sensitivity, and is of the kind we have been examining in great detail. It turns out there are no claims Beauty is required to go from certainty in Sunday night to less-than-certainty in Monday morning *due to cognitive mishap or its possibility*.

As we have already noted, when Beauty awakens Monday morning ideal rationality requires her to entertain the possibility that her memory has already been erased. But the effect of this possibility is simply to give her less specific information than she might otherwise have had. If the experiment had involved no prospect of memory erasure (and Beauty had known that), she would have awakened Monday morning and concluded "Today is Monday." With the prospect of memory erasure, when Beauty awakens Monday morning all she can conclude is "Today is Monday or Tuesday." But this is still information gained; the possibility of cognitive mishap has merely reduced its specificity. That possibility does not cause Beauty to *lose* any certain information between Sunday night and Monday morning.

In fact, if we confine our analysis of the Sleeping Beauty Problem to the period running from Sunday night to Monday night, cognitive mishap and its possibility do not cause Beauty to lose certain information at any point. The only claims that become certain and then uncertain during that time do so because of their context-sensitivity, which is ably handled by our modeling technique with (LC) and (CMP) included. And since our analysis in Sections 2.1 and 2.2 above concerned only that time period, we need not worry that the presence of cognitive mishap in the Sleeping Beauty Problem invalidates our analysis of its solution.³⁸

Finally, there is the third group of cognitive mishap stories. Since stories may be completely fabricated by us, any conceptual possibility seems fair game for inclusion in a story (as long as the story conforms to the definition of its kind laid out in Section 1.1). As long as we are allowing cognitive tampering by guardians of secret territories, crazed experimenters, and the like, we must admit the conceptual possibility that someone could tamper with an agent's cognitive state between two times such that the agent did not gain or lose any certain information, but merely had her degree of belief altered in some uncertain claim. We also have confined our attention so far to cognitive mishaps involving memory implantation or loss;

³⁷For a more thorough discussion of cognitive mishap cases, including a completely different way of obtaining intuitive Shangri-La results using our modeling framework, see (Titelbaum manuscript).

³⁸Halpern (2005) also argues that cognitive mishap is a red herring in analyzing the Sleeping Beauty Problem, though for somewhat different reasons.

we have not considered more drastic malfunctions such as insanity or temporary derangement. Thus there is a vast “Other” group of cognitive mishap stories out there collecting those mishap stories that do not fall into our first two groups. I don’t know whether our modeling technique can yield reliable verdicts for stories in this group, or even whether there are requirements of ideal rationality for stories such as these.

2.4. Modeling Strategies. I suggested in Section 1.8 that the Sleeping Beauty Problem would illustrate two strategies for obtaining verdicts about stories in which an agent lacks a uniquely denoting context-insensitive expression for some context-sensitive expression central to the story’s question. This circumstance arises in the Sleeping Beauty Problem between Sunday night and Monday morning. Clearly information about what day it is plays an important role in Beauty’s reasoning between those two times. For Beauty, this information is contained in claims about what day “today” is. Yet when she awakens Monday morning, Beauty does not have a context-insensitive expression that she can be certain uniquely picks out the day denoted by “today”. And so the modeling procedure outlined in Section 1.8 cannot be applied here. That procedure required us to find uniquely denoting context-insensitive expressions for the context-sensitive expressions in important context-sensitive claims, use those context-insensitive expressions to generate claims that ideal rationality required the agent to accept as truth-value equivalents of the context-sensitive claims, then use (CMP) to move to a model whose language lacked those context-sensitive claims. If there is no context-insensitive expression that the agent is required to be certain uniquely picks out the denotation of an important context-sensitive expression, that procedure runs aground at the start. And this is the case with Beauty between Sunday night and Monday morning.³⁹

When there are only two distinct times in a story, and the transition between the two involves the acquisition of important information containing a context-sensitive expression for which the agent lacks a uniquely denoting context-insensitive expression at one of the times, our modeling technique is incapable of correctly modeling the story. At that point, the only modeling strategies available involve altering the story to some extent. And while the order of presentation has obscured this fact somewhat, this is exactly what we have done with the Sleeping Beauty Problem. When Adam Elga first proposed the Problem in (Elga 2000), it contained only two distinct times: Sunday night and Monday morning. As that story stood, it could not be effectively modeled using our technique. Elga then had the idea of adding a Monday night to the story when Beauty would be told what day it was. We have

³⁹In fact, I believe the only contribution the threat of memory erasure makes to the Sleeping Beauty Problem is that it allows a circumstance to arise in which the agent lacks a uniquely denoting context-insensitive expression for an important context-sensitive expression. As I argued in Section 1.8, stories in which this occurs are very difficult to arrange. That this is the only contribution of the cognitive mishap in the Sleeping Beauty Problem is borne out by the fact that one can arrange close analogues to the Problem in which no forgetting occurs at all. In these stories, instead of Beauty’s being awakened twice if the coin lands tails, tails prompts the experimenters to make a perfect copy of Beauty, all the way down to the memories the copy possesses. (See (Bostrom manuscript) and (Meacham manuscript) for examples of such stories.) After this procedure has occurred, Beauty has no trouble with context-insensitive expressions for “today” but lacks a uniquely denoting context-insensitive expression for “I,” which turns out to be context-sensitive in this case.

followed him in adding this time into the story, and doing so makes our analysis possible.

In fact, the analysis of the Sleeping Beauty Problem (with Monday night included) that results bears a close similarity to our analysis of the Shangri La story in the previous section. In Shangri La, the possibility of cognitive mishap prevented us from directly relating your degrees of belief while you were on the path to your degrees of belief once you'd arrived. So we related those degrees of belief indirectly by relating each separately to your degrees of belief before the coin was flipped. In the Sleeping Beauty Problem, we cannot relate Sunday night credences to Monday morning credences directly because of Beauty's lack of a uniquely denoting context-insensitive expression. So we relate Beauty's Sunday night credences to her Monday night credences with model S0, then relate her Monday morning credences to her Monday night credences with model S1. Working indirectly, we come to the conclusion that ideal rationality requires Beauty's Monday morning credence in heads to be strictly less than her Sunday night credence in heads.

The addition of Monday night to the story is designed to fill a very precise role. On the one hand, Beauty has a uniquely denoting context-insensitive expression for "today" on both Sunday night and Monday night, so the fact that she learns claims between the two times that contain "today" does not impede our modeling technique. On the other hand, Beauty does not learn any relevant claims containing "today" between Monday morning and Monday night, so the fact that she lacks a uniquely denoting context-insensitive expression for "today" on Monday morning also fails to impede our modeling technique. This addition of Monday night to the Sleeping Beauty Problem illustrates our first strategy for obtaining verdicts about stories in which the agent lacks a uniquely denoting context-insensitive expression: add a time to the story (along the pattern of the Monday night addition to the Sleeping Beauty Problem) such that the degrees of belief of interest can be indirectly related via degrees of belief at the added time without any impediments to our modeling technique.

Both our strategies for modeling troublesome stories will involve adding some element to the story that wasn't already there. When we change a story in this fashion, there is always the threat that we will disturb the very phenomenon we are interested in modeling. We must be very careful that the added element does not alter the ideally rational degrees of belief the question in the story concerns. Because our modeling technique is incapable of modeling the story before the addition is made, I know of no formal way to guarantee that no such alteration has occurred. We must rely on informal, intuitive arguments made outside the modeling framework to establish that the addition has not prevented our obtaining verdicts true to the original.

Thus I suppose it is open to someone to argue that the addition of Monday night to the Sleeping Beauty Problem in its original version (containing only two times) alters the relationship required by ideal rationality between Beauty's Sunday night degree of belief in heads and her Monday morning degree of belief in heads. Since the alleged alteration is to degrees of belief Beauty assigns before the time added into the story (Monday night), the argument would have to be that merely knowing in advance that she would in the future be told what day it was would alter Beauty's Monday morning degree of belief in heads. It strikes me as very difficult to make such an argument out.

The second strategy for modeling difficult stories also involves adding something to the story. But this time, instead of adding another time to the story to indirectly relate the two times already present, we add another feature to the story to give the agent a uniquely denoting context-insensitive expression where she lacked one in the original. An example of this kind of modification to the Sleeping Beauty Problem appears in the next section. Again, the key point is to ensure that the added element remains independent of the degrees of belief asked about in the original problem.

2.5. Technicolor Beauty. To apply the second modeling strategy to the Sleeping Beauty Problem, we add one small element, giving us the following story:

Technicolor Beauty: Beauty is brought in on Sunday and the experiment is described to her just as in the Sleeping Beauty Problem. But one of the experimenters is Beauty's friend, and before she is put to sleep Sunday night he agrees to a request. While the other experimenters flip their fateful coin Sunday night, Beauty's friend will go into another room and roll a fair die. (The outcome of the die roll will be independent of the outcome of the coin flip.) If the die roll comes out odd, Beauty's friend will place a piece of red paper in the room with Beauty where she will see it when she awakens Monday morning, then replace it on Tuesday with a blue paper she will see if she awakens Tuesday morning. If the die roll comes out even, the process will be the same, but Beauty will see the blue paper on Monday and the red paper if she awakens Tuesday morning.

Certain that her friend can be relied upon, Beauty falls asleep Sunday night. Some time later she finds herself awake, unsure whether it is Monday or Tuesday, but with a red piece of paper before her. What does ideal rationality require at that moment of Beauty's degree of belief that the coin came up heads?

I have chosen to have Beauty see a red piece of paper on the awakening we're asked about just to give us a definite case on which to focus our discussion. But this choice is made without loss of generality, and our analysis below would yield identical results if the paper seen were blue or if no paper color were specified at all. The analysis begins with a model of Technicolor Beauty, model TB:

Model: TB

Story: Technicolor Beauty

L: Built on these atomic sentences:

H The coin comes up heads.

MR Monday is the red paper day.

AR I am awake to see the red paper.

TM Today is Monday.

TR Today is the red paper day.

t-seq: Contains these times:

*t*₀ Sunday night, after Beauty has heard the experiment described and made her arrangements with her friend but before she is put to sleep.

*t*₁ Monday morning, after Beauty awakens and sees the red paper but before she knows whether it is Monday or Tuesday.

- ES: (1) $0 < P_0(H) < 1$
 (2) $0 < P_1(H) < 1$
 (3) $0 < P_0(MR) < 1$
 (4) $0 < P_1(MR) < 1$
 (5) $0 < P_0(AR) < 1$
 (6) $P_1(AR) = 1$
 (7) $P_0(TM) = 0$
 (8) $0 < P_1(TM) < 1$
 (9) $P_0(TR) = 0$
 (10) $P_1(TR) = 1$
 (11) $0 < P_0(TM \equiv MR) < 1$
 (12) $P_1(TM \equiv MR) = 1$
 (13) $0 < P_0(MR | H) < 1$
 (14) $P_0(AR | H) = P_0(MR | H)$
 (15) $P_0(AR | \sim H) = 1$

I_{0,1}: $AR, TR, TM \equiv MR$, sentences entailed by their conjunction

Note that the first three sentences listed in **L** (the ones without a “T”) are meant to be tenseless. AR , for example, does not represent the claim that “I am awake, seeing a red paper right now;” it represents a claim along the lines of “Whichever day is the red paper day, I am/will be/was awake on that day.” Extrasystematic constraint (11) results from the fact that while ideal rationality requires Beauty to be certain at t_0 that “Today is Monday” is false, it also requires her to be uncertain whether “Monday is [will be] the red paper day” is true. Constraint (12) results from the fact that by t_1 Beauty knows that today is the red paper day, so Monday is the red paper day just in case today is Monday. Constraint (13) represents the fact that since the coin flip and the die roll occur independently, ideal rationality requires Beauty not to be certain whether the die roll comes out odd on the supposition that the coin flip comes out heads. Constraints (14) and (15) come from the structure of the story: if the coin comes up heads, Beauty will awaken on Monday only, so she will be awake on the red paper day just in case Monday is the red paper day. On the other hand, if the coin comes up tails, Beauty will be awake both days, and so is guaranteed to be awake for the red paper day.

Looking closely at the list of extrastematic constraints, we can see that ideal rationality requires Beauty to be certain of the claim represented by $\sim TM$ at t_0 but less-than-certain of that claim at t_1 . Thus (LC) cannot be applied to generate verdicts of **TB**. This is no surprise, since **L** contains representations of a number of context-sensitive claims. Those claims owe their context-sensitivity to the shifting denotation of the context-sensitive expression “today.” In the original Sleeping Beauty Problem, this prevented us from reaching verdicts relating t_0 and t_1 credences without adding t_2 into the model. Now, however, the presence of the red paper ensures that Beauty has uniquely denoting context-insensitive expressions for “today” at both t_0 and t_1 . The t_0 expression is “Sunday,” while the t_1 expression is “the red paper day.” We can thus move to a model containing no context-sensitive claims, model **TB⁻**:

Model: **TB⁻**

Story: Technicolor Beauty

L⁻: Built on these atomic sentences:

H The coin comes up heads.

MR Monday is the red paper day.

AR I am awake to see the red paper.

t-seq: Contains these times:

t_0 Sunday night, after Beauty has heard the experiment described and made her arrangements with her friend but before she is put to sleep.

t_1 Monday morning, after Beauty awakens and sees the red paper but before she knows whether it is Monday or Tuesday.

- ES: (1) $0 < P_0^-(H) < 1$
 (2) $0 < P_1^-(H) < 1$
 (3) $0 < P_0^-(MR) < 1$
 (4) $0 < P_1^-(MR) < 1$
 (5) $0 < P_0^-(AR) < 1$
 (6) $P_1^-(AR) = 1$
 (7) $0 < P_0^-(MR|H) < 1$
 (8) $P_0^-(AR|H) = P_0^-(MR|H)$
 (9) $P_0^-(AR|\sim H) = 1$

$\mathbf{I}_{0,1}^-$: *AR*, sentences entailed by *AR*

Since there are no claims in language \mathbf{L}^- that ideal rationality requires Beauty to be certain of at t_0 and less-than-certain of at t_1 , we can apply (LC) to generate:

$$(21) \quad P_1^-(H) = P_0^-(H|AR)$$

Next, in a move that should surprise no one by now, we look to apply (CMP). To do so, we have to find truth-value equivalents at both t_0 and t_1 for the two atomic sentences in \mathbf{L} that are not in \mathbf{L}^- , namely *TM* and *TR*. Those equivalents are:

$$\begin{array}{ll} P_0(TM \equiv \mathbf{F}) = 1 & P_1(TM \equiv MR) = 1 \\ P_0(TR \equiv \mathbf{F}) = 1 & P_1(TR \equiv \mathbf{T}) = 1 \end{array}$$

Having met the conditions in the antecedent of (CMP), we conclude that **TB** is a conservative supermodel of \mathbf{TB}^- . The analogue of the verdict of \mathbf{TB}^- expressed in Equation (21) is therefore a verdict of **TB**. And so we have

$$(22) \quad P_1(H) = P_0(H|AR)$$

Applying Bayes' Theorem to this equation yields

$$(23) \quad P_1(H) = \frac{P_0(AR|H) \cdot P_0(H)}{P_0(AR|H) \cdot P_0(H) + P_0(AR|\sim H) \cdot P_0(\sim H)}$$

We now apply extrasystematic constraints (14) and (15) of model **TB** to obtain

$$(24) \quad P_1(H) = \frac{P_0(MR|H) \cdot P_0(H)}{P_0(MR|H) \cdot P_0(H) + 1 - P_0(H)}$$

And finally, using a bit of elementary algebra and the fact that both $P_0(MR|H)$ and $P_0(H)$ are between 0 and 1, we conclude that

$$(25) \quad P_1(H) < P_0(H)$$

In moving from the Sleeping Beauty Problem to the Technicolor Beauty story, we added in the colored pieces of paper so as to give Beauty a uniquely denoting context-insensitive expression for “today” when she awakens Monday morning. But by basing the choice of papers on a random event independent of the coin flip, we kept this addition from giving Beauty any extra information about the outcome

of the flip when she awakens on Monday.⁴⁰ The addition of the papers makes it possible to analyze the story without adding an additional time into the models, and the verdict we obtain matches precisely what we determined in our earlier analysis: that ideal rationality requires Beauty's Monday morning degree of belief in heads to be lower than her Sunday night degree of belief.⁴¹ And once more, we have obtained this verdict strictly from our modeling technique, without invoking the Principal Principle or a principle of indifference.

We can, however, get more precise verdicts if we choose to apply the Principal Principle. Again, I trust it will be uncontroversial to apply that principle to generate an extrasystematic constraint on TB that $P_0(H) = 1/2$. Also, since the die is fair and its roll is a chance process completely independent of the coin flip, it seems uncontroversial to apply the Principal Principle and generate an extrasystematic constraint that $P_0(MR|H) = 1/2$. Plugging these values into Equation (24) above yields

$$(26) \quad P_1(H) = \frac{1}{3}$$

We now have a precise degree of belief that ideal rationality requires Beauty to assign to heads when she awakens Monday morning.

2.6. Objections to Technicolor Beauty. Our analysis of Technicolor Beauty may seem suspicious for a couple of reasons. First, consider model TB^- . Language \mathbf{L}^- contains no sentences representing claims about today, only sentences representing claims about the coin flip and about the red piece of paper. We have already stated that the red paper's appearance on a particular day is independent of the coin flip's outcome. Yet according to (CMP) language \mathbf{L}^- contains all the information relevant to Beauty's changing degree of belief in heads. How can this be?

First, recall that (CMP) does not suggest in this case that information about today plays *no* role in Beauty's Monday morning degrees of belief. It is hard to imagine how Beauty could come to be certain Monday morning that she is awake to see the red paper without being certain that today is the red paper day. The point is only that once Beauty's certainty in various context-insensitive claims is established, the context-sensitive claims can be set aside in determining her ideally rational degree of belief in heads.

And this is not so preposterous. According to model TB^- , the key piece of information Beauty learns between Sunday night and Monday morning is that she is awake to see the red piece of paper. Way back in Section 1.4, we noted that Conditionalization allows us to determine a required degree of belief at a later time from a required degree of belief at an earlier time, specifically the earlier degree of belief conditional on the supposition of the information learned between the two

⁴⁰Kenny Easwaran deserves credit for suggesting colored bits of paper to me as an expedient for giving Beauty a uniquely denoting context-sensitive expression for "today." Though I hadn't read their work at the time, Brian Kierland and Bradley Monton had already pointed out in (Kierland and Monton 2005) that the Sleeping Beauty Problem does not require Beauty's awakenings to be subjectively indistinguishable. In fact, (Kierland and Monton 2005, 391) offers a color-coding idea involving pajamas quite similar to the pieces of paper idea presented here.

⁴¹If we *were* to extend model TB to include t_2 (Monday night) in its time-sequence, the verdicts we obtained in Section 2.2 above concerning Beauty's Monday night degrees of belief would all be obtainable again.

times. Model TB^- allows us to relate Beauty's Sunday night and Monday morning degrees of belief in heads by conditionalizing, so we ought to be able to understand Beauty's required Monday morning degree of belief in heads in terms of an ideally rational conditional degree of belief on Sunday night. In particular, since TB^- suggests that the only information relevant to heads Beauty learns between Sunday night and Monday morning is AR , we ought to be able to obtain Beauty's required Monday morning degree of belief in heads by considering her ideally rational Sunday night degree of belief in heads conditional on AR .

Imagine that on Sunday night, after making her arrangements with her friends, Beauty supposes that she will be awake to see the red paper.⁴² Beauty's seeing the red paper is twice as likely to happen if the coin flip comes up tails than if it comes up heads: if the flip comes up tails, Beauty is guaranteed to see the red paper, whereas if the flip comes up heads she will see the red paper only if the die roll comes out odd. So Beauty is required by ideal rationality to have a higher degree of belief in tails than heads on the supposition that she is awake on the red paper day. If we apply the Principal Principle and require her degrees of belief to match the objective chances, Beauty is required to set a degree of belief in heads on the supposition that she is awake to see the red paper at exactly $1/3$. Conditionalizing has done its work: Beauty's ideally rational Monday morning degree of belief in heads is exactly equal to her ideally rational Sunday night degree of belief in heads on the supposition of all the relevant information she learns in-between.

Yet it may seem intuitively like there is something wrong here. We stated in the previous section that our choice of the red paper for Beauty to see upon awakening was arbitrary and did not introduce a loss of generality. Thus the same results should arise if Beauty sees the blue paper instead. And if we re-run the analysis with blue instead of red, we do indeed get the same results. For instance, Beauty's ideally rational Sunday night degree of belief in heads on the supposition that she will be awake to see the blue paper is also $1/3$. But now there seems to be a problem: Beauty knows on Sunday night that she will see the red paper or she will see the blue paper. Her ideally rational degree of belief on the supposition that she will see the red paper is $1/3$, and her ideally rational degree of belief on the supposition that she will see the blue paper is $1/3$. Since she knows one of those is bound to happen, shouldn't her unconditional Sunday night degree of belief in heads already be $1/3$?

This intuitive objection applies an idea that can be nicely summed up with a theorem derivable from our synchronic systematic constraints:

Disjunctive Conditionalization Theorem: If A and B are mutually exclusive, $P_t(X | A) = x$, and $P_t(X | B) = x$, then $P_t(X | A \vee B) = x$.

Notice the condition of mutual exclusivity in this theorem. The statement isn't a theorem without this condition, and we can see why with a simple example:

⁴²Darren Bradley reminds me that Beauty's *supposing* she will be awake to see the red paper is very different from someone's *telling* Beauty on Sunday night that she will be awake for that paper. If Beauty gets some sort of *evidence* that she will see the red paper, she has to ask how she got that evidence and whether the process that produced that evidence is independent of all the other factors at play in the story. Various answers to those questions will affect her Sunday-night degrees of belief in various ways. And so we will imagine Beauty's just idly supposing that she will be awake to see the red paper day, not her learning somehow on Sunday night that that will be the case.

Suppose a fair die is rolled. It's reasonable to assign $P_t(1 | 1 \vee 2 \vee 3) = 1/3$, and to assign $P_t(1 | 1 \vee 3 \vee 5) = 1/3$. But that doesn't make it reasonable to assign $P_t(1 | (1 \vee 2 \vee 3) \vee (1 \vee 3 \vee 5)) = 1/3$.⁴³

The trouble with the objection is that Beauty's being awake to see the red paper and Beauty's being awake to see the blue paper are not mutually exclusive. If the coin comes up tails, both occur. And so Disjunctive Conditionalization does not apply. It is perfectly consistent with our systematic constraints for Beauty's ideally rational Sunday night degrees of belief in heads conditional on "awake for red" and conditional on "awake for blue" to both be $1/3$, while her ideally rational unconditional Sunday night degree of belief in heads remains $1/2$.

The second reason Technicolor Beauty might look suspicious is that it seems to provide us with stronger results than we could get from the original Sleeping Beauty Problem. In our original analysis of the problem, even with t_2 included in the time sequence of our model the most precise verdict we could obtain without applying a principal of indifference was $P_1(H) < 1/2$.⁴⁴ Yet with Technicolor Beauty we get a precise value for Beauty's required Monday morning degree of belief in heads without applying an indifference principle at all. This may suggest that by adding the pieces of paper into the story we have somehow altered the relationships between ideally rational degrees of belief, perhaps to the extent that requirements from Technicolor Beauty should not be assumed to be requirements of the original Sleeping Beauty Problem at all.

As for the strength of the verdicts obtainable from Technicolor Beauty without any indifference principles, my hunch is that the Technicolor Beauty analysis makes up for the lack of an indifference principle by applying the Principal Principle twice. To get a verdict with a precise value of $1/3$ in Technicolor Beauty, we had to apply the Principal Principle not only to require that $P_0(H) = 1/2$ (an application we made in our analysis of the original problem), but also to require that $P_0(MR | H) = 1/2$. Perhaps the story has remained unchanged at the fundamental level, but a double dose of the Principal Principle has done the work previously achieved by applying the Principal Principle and an indifference principle one time apiece.⁴⁵

But the general point of the objection is well taken. Adding in the colored pieces of paper does change Beauty's story, and we cannot just assume that we have been successful in keeping that change from altering the answer to the question the story asks. As I noted in Section 2.4 above, there is no formal way of guaranteeing that an addition to a previously unmodelable story will not alter the answer to its question. Nevertheless, I will now informally argue that adding in the pieces of paper does

⁴³The Disjunctive Conditionalization Theorem bears a family resemblance to Savage's "Sure-Thing Principle." (Savage 1972, 21) There are differences: Disjunctive Conditionalization concerns credences, while Sure-Thing concerns preferences; Disjunctive Conditionalization is a theorem derived from other principles of our modeling framework, while Savage discusses Sure-Thing when he is introducing his system's postulates. Nevertheless, there's an important parallel in that Savage always discusses the Sure-Thing Principle as applying to two events B and $\sim B$, which must of course be mutually exclusive.

⁴⁴See Section 2.2 above.

⁴⁵Interestingly, the Technicolor Beauty analysis with two doses of the Principal Principle validates the results of applying of Elga's "highly restricted principle of indifference" to this case. Via an equation much like Equation (18), we can demonstrate in Technicolor Beauty that on Monday morning ideal rationality requires Beauty to assign $P_1(TM | \sim H) = 1/2$.

not alter the Monday morning degree of belief in heads required of Beauty by ideal rationality.

When we move from the Sleeping Beauty Problem to Technicolor Beauty, Beauty gets additional information at two points in the story. First, on Sunday night she knows not only about the setup of the experiment, but also about the arrangements she has made with her friend. If we examine her ideally rational Sunday night degrees of belief concerned exclusively with the outcome of the coin flip and the passing days of the week, it strikes me as implausible to argue that these have been changed in any way by the extra information. Beauty knows that her friend is going to put some pieces of paper in the room with her later on, and one might try to argue that those pieces of paper will indirectly give her some evidence about the outcome of the flip. But on Sunday night she hasn't actually seen the papers yet, and it's hard to imagine that just knowing she's *going* to see them changes her ideally rational degrees of belief concerning the flip at all.

So let's move to the second point at which Beauty gains extra information in Technicolor Beauty: Monday morning, when she awakens and sees the colored piece of paper. Someone might challenge our analysis by arguing that seeing the paper indirectly gives Beauty some information about the outcome of the flip that she didn't have in the original Sleeping Beauty Problem. The argument would be that, due to this extra information, Beauty's ideally rational Monday morning degree of belief in heads in Technicolor Beauty might be different from her ideally rational Monday morning degree of belief in heads in the original problem.

I think this hypothesis can be tested, in the following way: Imagine that after Beauty awakens Monday morning, there is a very brief period of time when she is awake but hasn't yet seen her colored paper for the day. Call this time $t_{0.5}$. If we were correct above in claiming that Technicolor Beauty doesn't give Beauty any extra information relevant to the outcome of the coin flip on Sunday night, then at $t_{0.5}$ Beauty's information relevant to the outcome of the coin flip is no different from the relevant information she has on Monday morning in the original problem. (She acquires the alleged extra damaging information between $t_{0.5}$ and t_1 .) So Beauty's $t_{0.5}$ ideally rational degree of belief in heads in Technicolor Beauty should equal her ideally rational degree of belief in heads on Monday morning of the original problem.

We cannot set up a fruitful model relating Beauty's $t_{0.5}$ degrees of belief to her t_0 degrees of belief, for exactly the reason we couldn't relate Beauty's Sunday night degrees of belief to her Monday morning degrees of belief in the original problem: at $t_{0.5}$ Beauty does not yet have a uniquely denoting context-insensitive expression for "today." However, we can relate $t_{0.5}$ degrees of belief to t_1 degrees of belief and work backwards. I won't actually set up the model here, but the analysis would go as follows: First, let's suppose that Monday is the red paper day. We would set up a model with a time-sequence of $t_{0.5}$ and t_1 , then use (CMP) to eliminate from its language sentences representing context-sensitive claims about the passage of time between $t_{0.5}$ and t_1 . In the resulting model, the atomic sentences in Beauty's learned information set would be TR and AR . There would be no sentences in the model representing claims that ideal rationality requires Beauty to be certain of at $t_{0.5}$ and then less-than-certain of at t_1 , so we could apply (LC) to generate $P_1(H) = P_{0.5}(H \mid TR \& AR)$. At $t_{0.5}$ ideal rationality requires Beauty to be certain

that TR implies AR , so by our synchronic constraints this can be simplified to $P_1(H) = P_{0.5}(H | TR)$.

Our initial analysis of Technicolor Beauty (with the Principal Principle) tells us that once Beauty awakens and sees a red piece of paper, ideal rationality requires her to have a degree of belief in heads of $1/3$. So if we are supposing that Monday is the red paper day, we can conclude that $P_{0.5}(H | TR) = 1/3$. By an exactly parallel analysis supposing that Monday is the blue paper day, we can conclude that $P_{0.5}(H | \sim TR) = 1/3$. And now we are in a position to use our Disjunctive Conditionalization Theorem. TR and $\sim TR$ are clearly mutually exclusive, so we conclude that $P_{0.5}(H) = 1/3$. Seeing the colored paper does not provide Beauty with extra information that alters her ideally rational degree of belief in heads; ideal rationality requires that exact degree of belief Monday morning before she even sees the paper. And this makes sense. We set up the story about the colored pieces of paper so that the color of paper Beauty sees on a particular experiment day is independent both of what day it is and of the outcome of the coin flip. Seeing a particular colored paper on one of the days of the experiment gives Beauty no information about which day it is and therefore no information about the outcome of the coin flip. The only thing the colored paper does is make it easier for *us* to create a model and discover how noting that she's awake on a day of the experiment affects Beauty's ideally rational degree of belief in heads.⁴⁶

Conclusion. The best test of a modeling technique is to see whether it can provide results for a variety of cases, and whether those results match our intuitions in cases whose solutions are obvious. As we have built up our modeling technique over the course of this paper, I have tried to supply intuitive considerations in favor of each of its individual pieces as we added them to the system. But more importantly, we have now tested the technique against a wide variety of stories. Our first five systematic constraints yielded a technique sufficient to model stories in which the important claims are context-insensitive. The addition of (CMP) as a sixth constraint allowed the technique to model stories involving context-sensitive claims where at each moment the agent has a uniquely denoting context-insensitive expression for each context-sensitive expression. Finally, we saw two strategies for modeling the remaining stories. The first strategy involved adding a time to the

⁴⁶Technicolor Beauty stipulates that the chance process Beauty's friend uses makes it equally likely that Monday will be the red or the blue paper day. What if the story were generalized, so that the chance process gave Monday probability r of being the red paper day, where r need not equal $1/2$? Analysis with our framework reveals that the ideally rational degree of belief in heads once Beauty awakens and sees a red paper varies with the value of r (though as long as r is non-extreme it remains below $1/2$). But r values other than $1/2$ alter the relevant information available to Beauty when she awakens, so that the resulting requirements of ideal rationality do not mirror those in the original Sleeping Beauty Problem. If one color of paper is more likely on Monday than the other, awakening and seeing a particular colored paper gives Beauty information about what day it is, and therefore about the outcome of the coin flip. Notice that our analysis of Technicolor Beauty in Section 2.5 above started by assuming *without loss of generality* that Beauty awakens to see a red paper. If $r \neq 1/2$, Beauty's awakening to see a red paper is either more or less likely than her awakening to see a blue paper, and so that generality disappears. Further, if $r \neq 1/2$, it will no longer be the case that $P_{0.5}(H | TR) = P_{0.5}(H | \sim TR)$, and so we will not be able to apply the Disjunctive Conditionalization Theorem to establish $P_{0.5}(H)$. The argument above that introducing the colored papers does not alter the Monday-morning degree of belief in heads required by ideal rationality goes through only if $r = 1/2$; this is because the colored papers *do* in fact alter that ideally-rational degree of belief when $r \neq 1/2$.

time sequence of the story; the second left the time sequence unaltered but added an element to the story that gave the agent the needed context-insensitive expressions. The element added to the story in each strategy allowed our modeling technique to yield verdicts but remained independent of the phenomenon we originally set out to model.

Among our other results, we concluded that David Lewis’s solution to the Sleeping Beauty Problem was incorrect: when Beauty awakens Monday morning, ideal rationality requires her degree of belief in heads to be lower than it was on Sunday night. This result is also sufficient to disprove the relevance-limiting theses we considered at various points in the paper, both in their original and in their revised forms. Whatever criteria one uses to specify what an agent “learns” in a story, it is clear that in the Sleeping Beauty Problem Beauty does not learn any beliefs *de dicto* between Sunday night and Monday morning. The information Beauty possesses Monday morning is logically compatible with all the same beliefs *de dicto* as the information she possessed Sunday night. Yet ideal rationality requires her to change her degree of belief in a belief *de dicto* (that the coin comes up heads) between Sunday night and Monday morning. Learned self-locating beliefs can, all by themselves, require an agent to change her relative degrees of belief in beliefs *de dicto*, without at the same time being incompatible with any such beliefs. This is sufficient to disprove all the relevance-limiting theses we have considered.

The defender of a relevance-limiting thesis would presumably try to challenge some of the arguments we have used to reach these results. But at least when it comes to the Sleeping Beauty Problem, that would be very difficult for him to do. In each of our strategies for solving that problem, the key move was to argue at particular points that some self-locating information was *irrelevant* to a belief *de dicto*. On the strategy that added Monday night to the story, we argued first that Beauty’s learning that a day had passed between Sunday night and Monday night was irrelevant to her degree of belief in heads; then we argued that Beauty’s learning that some time had passed was irrelevant to her conditional degrees of belief between Monday morning and Monday night. On the Technicolor Beauty strategy, we argued that with the colored pieces of paper involved *all* the self-locating information Beauty learns between Sunday night and Monday morning (after she sees the paper) is irrelevant to her degree of belief in heads. Someone who thought that self-locating beliefs could not be relevant to beliefs *de dicto* would have to support these analyses, and yet it is these very analyses that yield our Sleeping Beauty results showing that the relevance-limiting theses are false. This demonstrates that the relevance-limiting theses are in a very real sense self-contradictory.

Some sort of error theory might be appropriate here — why did a relevance-limiting thesis seem intuitively attractive to begin with? First, as we noted in Section 1.8, cases like the Sleeping Beauty Problem capable of disproving a relevance-limiting thesis outright are extremely rare (if not wholly absent) in real life. Our intuitions can be excused for not being finely-tuned to a set of situations they are unlikely ever to encounter. Second, our intuitions on this point may be responsive to simpler ways of understanding learning situations. The simplest way to interpret learning is to understand an agent as either believing a claim or not believing it (which is different from *disbelieving* it, something more akin to believing its negation), and to understand all inference as showing that two beliefs are mutually exclusive, that one deductively entails the other, or neither of the above. On this

coarse-grained understanding of learning, it is impossible to construct an example that conclusively shows a naked self-locating belief inferentially altering a rational agent's attitude towards a belief *de dicto*. For the only way a self-locating belief can have this effect is to either imply some belief *de dicto* or be mutually exclusive with it, and in either case one could argue that the self-locating belief is not the only thing learned: the belief *de dicto* or its negation was learned by the agent as well.

Modeling agents as having fine-grained degrees of belief responsive to probabilistic reasoning allows subtle phenomena to emerge that are hidden by a more coarse-grained understanding. The relevance relations that result are complex; with enough ingenuity we can construct stories in which almost any kind of information is relevant to any other. To properly understand these relations we should not start with the assumption that there are strict barriers in place — high walls preventing inference from one type of claim to another. Instead, we should develop a modeling technique that is general enough to model a wide variety of cases, sufficiently formal that there is no debate what it says about any given case, and responsive to what we think are our most settled and obvious intuitions. Having achieved this, we can let the models teach *us* about the relevance relations.

References

- Arntzenius, Frank. 2003. Some Problems for Conditionalization and Reflection. *Journal of Philosophy* 100: 356-370.
- Bostrom, Nick. Manuscript. Sleeping Beauty and Self-Location: A Hybrid Model. Forthcoming in *Synthese*. Currently accessible at <http://www.anthropic-principle.com/preprints/beauty/synthesis.pdf>.
- Elga, Adam. 2000. Self-locating Belief and the Sleeping Beauty problem. *Analysis* 60: 143-7.
- Elga, Adam. 2004. Defeating Dr. Evil with Self-Locating Belief. *Philosophy and Phenomenological Research* 69: 383-396.
- Fitelson, Branden. Manuscript. Logical Omniscience and the Bayesian Stance. Forthcoming in *Stance and Rationality*, edited by O. Bueno and D. Rowbottom. Oxford University Press.
- Hall, Ned. 2004. Two Mistakes about Credence and Chance. *Australasian Journal of Philosophy* 82: 93-111.
- Hájek, Alan. 2003. Interpretations of Probability. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2003.
- Halpern, Joseph Y. 2005. Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems. *Oxford Studies in Epistemology, Volume 1*, edited by T. S. Gendler and J. Hawthorne, 111-142. Oxford University Press.

Kierland, Brian and Monton, Bradley. 2005. Minimizing Inaccuracy for Self-Locating Beliefs. *Philosophy and Phenomenological Research* 70: 384-395.

Lance, Mark Norris. 1995. Subjective Probability and Acceptance. *Philosophical Studies* 77: 147-179.

Lewis, David. 1979. Attitudes De Dicto and De Se. *Philosophical Review* 88: 513-543.

Lewis, David. 1980. A Subjectivist's Guide to Objective Chance. *Studies in Inductive Logic and Probability, Volume 2*, edited by Richard C. Jeffrey, 263-294. Berkeley, Ca.: University of California Press.

Lewis, David. 2001. Sleeping Beauty: reply to Elga. *Analysis* 61: 171-6.

MacFarlane, John. 2005. Making Sense of Relative Truth. *Proceedings of the Aristotelian Society* 105: 321-39.

Meacham, Christopher. Manuscript. Sleeping Beauty and the Dynamics of De Se Beliefs. Currently accessible at <http://philsci-archive.pitt.edu/archive/00002526>.

Savage, Leonard J. 1972. *The Foundations of Statistics*. New York: Dover Publications, Inc.

Talbott, William. 2001. Bayesian Epistemology. *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2001.

Titelbaum, Michael. Manuscript. Unlearning What You Have Learned.

Appendix: (CMP) in the Context-Insensitive Case. This Appendix proves that when all the sentences in a model's language represent context-insensitive claims, and the agent in the story suffers no cognitive mishaps and is required by ideal rationality to remain certain throughout that she has suffered no cognitive mishaps, (CMP) can be derived as a theorem from systematic constraints (1) through (5).

For the purposes of this argument, we are going to suppose that the only systematic constraints on our models are systematic constraints (1) through (5) described in the text. We are then going to show that under particular circumstances (CMP) follows from those constraints as a theorem.

Suppose we have a model M^+ (systematically constrained only by constraints (1) through (5)) whose modeling language L^+ contains only context-insensitive sentences. That is, suppose that for any sentence in L^+ and any two moments in the time-sequence of M^+ , ideal rationality requires the agent to be certain at the later time that the truth-value of that sentence has not changed between the two times. Suppose further that the agent in the story represented by M^+ suffers no cognitive mishaps during the story, and is required by ideal rationality to remain

certain throughout the story that she has suffered no cognitive mishaps during the story.

Now suppose the antecedent of (CMP) is met. That is, suppose we have a submodel M of M^+ such that for any time t_i in the time sequence and any sentence $Y \in \mathbf{L}^+$, there exists an $X \in \mathbf{L}$ such that $P_i^+(X \equiv Y) = 1$. We will show that in this case the analogue in M^+ of any verdict of M is a verdict of M^+ .

Lemma 1: Under these circumstances, an arithmetic statement in M is an extrasystematic constraint on M just in case its analogue is an extrasystematic constraint on M^+ . An extrasystematic constraint is obtained by arguing that ideal rationality requires an agent's degree of belief in a particular claim to have a particular property, or by arguing that ideal rationality requires an agent's degrees of belief in various claims to bear a certain relation. The extrasystematic constraint itself is an arithmetic statement in the language of our model representing that requirement of ideal rationality. Because M is a submodel of M^+ , sentences that appear in the languages of both models represent the same claims in both models. Therefore if an arithmetic statement in M is an extrasystematic constraint, it represents a requirement of ideal rationality whose representation in M^+ is the analogue of that arithmetic statement in M . For the same reason, the analogue of any extrasystematic constraint on M^+ is an extrasystematic constraint on M .

Lemma 2: Under these circumstances, for any claim represented by a sentence in \mathbf{L}^+ that ideal rationality requires the agent to be certain of at a particular time in the time sequence of M^+ , the agent is also required to be certain of that claim at any later time in the time sequence. Suppose we select two times in the time sequence of M^+ and a sentence in \mathbf{L}^+ . Ideal rationality requires the agent to be certain at the later time that the claim represented by that sentence has the same truth-value at the later time as it did at the earlier time. Moreover, the agent experiences no cognitive mishaps between the two times and is required to be certain at the later time that she has suffered no cognitive mishaps between the two times. Therefore if ideal rationality requires the agent to be certain of the claim represented by the sentence at the earlier time, it also requires the agent to be certain of that claim at the later time.

Lemma 3: Under these circumstances, for every $Y \in \mathbf{L}^+$ there exists an $X \in \mathbf{L}$ such that at every time t_i in the time sequence of M^+ , $P_i^+(X \equiv Y) = 1$ and $P_i^+(X) = P_i^+(Y)$. Since the time sequence is finite, there will be an earliest time in the sequence. Call this time t_1 . By hypothesis, for any $Y \in \mathbf{L}^+$ there exists an $X \in \mathbf{L}$ such that $P_1^+(X \equiv Y) = 1$. For any arbitrary time t_i in the time sequence later than t_1 , Lemma 2 gives us $P_i^+(X \equiv Y) = 1$. (And for $i = 1$, this result is trivial.) By our synchronic systematic constraints, we therefore have $P_i^+(X) = P_i^+(Y)$ for any i .

Main Proof: We want to show the consequent of (CMP): that any verdict of M has an analogue that is a verdict of M^+ . Any verdict of M can be derived by the following process: First, specify a set of premises, each of which is either an extrasystematic constraint or follows directly from a systematic constraint. (For example, for any $X \in \mathbf{L}$ and t_i in the time-sequence, $P_i(X \vee \sim X) = 1$ follows

directly from systematic constraint (2).) Next, apply algebraic techniques to derive the desired verdict from the premises.

Since analogous verdicts can be derived algebraically from analogous premise sets, it will suffice to show that if each premise is a verdict of \mathbf{M} , its analogue will be a verdict of \mathbf{M}^+ . The extrasystematic constraints are easy. By Lemma 1, any premise that is an extrasystematic constraint of \mathbf{M} will have an analogue that is an extrasystematic constraint (and therefore a verdict) of \mathbf{M}^+ .

On to premises that follow directly from systematic constraints. We'll start with systematic constraints (1) through (4). Suppose a verdict of \mathbf{M} follows from one of these systematic constraints, and the set $\{X_1, X_2, \dots, X_m\}$ contains just those sentences of \mathbf{L} that appear in the verdict. The fact that that verdict follows from that systematic constraint is independent of what other sentences appear in the modeling language, of what extrasystematic constraints apply to the model, and of what sentences appear in the learned information sets. All the sentences in $\{X_1, X_2, \dots, X_m\}$ are contained in \mathbf{L}^+ . So the only differences between \mathbf{M} and \mathbf{M}^+ are in the other sentences that appear in the modeling language, what extrasystematic constraints apply to the model, and what sentences appear in the learned information sets. Thus the analogue of the verdict in question will also follow directly from that systematic constraint in \mathbf{M}^+ .

The tricky systematic constraint is systematic constraint (5), (LC). The verdicts issued by (LC) depend on the learned information sets in the model, and these change when modeling languages change. The argument of the previous paragraph will therefore not apply to (LC), so we take the following approach:

Take a verdict of \mathbf{M} that follows directly from (LC). That verdict will take the form

$$(27) \quad P_k(Z) = P_j(Z \mid \mathbf{I}_{j,k})$$

Where Z is a sentence in \mathbf{L} and t_j and t_k are two times in the time-sequence common to \mathbf{M} and \mathbf{M}^+ . (Recall that when the sign for a set of sentences appears within a credence expression, it is abbreviating a sentence logically equivalent to the conjunction of all the sentences in the set.)

Consider a sentence $Y_b \in \mathbf{L}^+$, and assume for reductio that $P_j^+(Y_b) = 1$ and $P_k^+(Y_b) < 1$. By Lemma 3, there exists an $X_a \in \mathbf{L}$ such that $P_j^+(X_a) = 1$ and $P_k^+(X_a) < 1$. Both of these will be extrasystematic constraints on \mathbf{M}^+ , so by Lemma 1 their analogues will be extrasystematic constraints on \mathbf{M} . Therefore we have $P_j(X_a) = 1$ and $P_k(X_a) < 1$. But then the antecedent of the conditional in (LC) is not met and we cannot derive *any* verdicts from (LC) in \mathbf{M} , much less Equation (27). And so we have a contradiction.

The result of this argument is that there does not exist a $Y_b \in \mathbf{L}^+$ such that $P_j^+(Y_b) = 1$ and $P_k^+(Y_b) < 1$. Since Z is also a sentence in \mathbf{L}^+ , we can apply (LC) to obtain

$$(28) \quad P_k^+(Z) = P_j^+(Z \mid \mathbf{I}_{j,k}^+)$$

Now consider $\mathbf{I}_{j,k}^+$. We can rewrite $\mathbf{I}_{j,k}^+$ as the set $\{X_1, X_2, \dots, Y_1, Y_2, \dots\}$, where the X_a are members of \mathbf{L}^+ that are also in \mathbf{L} , while the Y_b are members \mathbf{L}^+ that are not in \mathbf{L} .

By Lemma 3, for every $Y_b \in \mathbf{I}_{j,k}^+$, there exists an $X_a \in \mathbf{L}$ such that $P_j^+(Y_b \equiv X_a) = 1$, $P_j^+(Y_b) = P_j^+(X_a)$, and $P_k^+(Y_b) = P_k^+(X_a)$. Since Y_b is a member of $\mathbf{I}_{j,k}^+$,

we have $P_j^+(Y_b) < 1$ and $P_k^+(Y_b) = 1$. So $P_j^+(X_a) < 1$ and $P_k^+(X_a) = 1$. But then by the definition of a learned information set, $X_a \in \mathbf{I}_{j,k}^+$.

The upshot of the previous paragraph is that for every $Y_b \in \mathbf{I}_{j,k}^+$, there exists an $X_a \in \mathbf{I}_{j,k}^+$ such that $P_j^+(Y_b \equiv X_a) = 1$. By our synchronic systematic constraints,

$$(29) \quad P_j^+(\mathbf{I}_{j,k}^+ \equiv \{X_1, X_2, \dots\}) = 1$$

Combining Equation (28) and Equation (29) and applying our synchronic systematic constraints once more, we obtain the verdict

$$(30) \quad P_k^+(Z) = P_j^+(Z \mid \{X_1, X_2, \dots\})$$

An X_a belongs to $\mathbf{I}_{j,k}^+$ just in case there are extrasystematic constraints on \mathbf{M}^+ stating that $P_j^+(X_a) < 1$ and $P_k^+(X_a) = 1$. By Lemma 1, this will occur just in case $P_j(X_a) < 1$ and $P_k(X_a) = 1$. Thus $X_a \in \mathbf{I}_{j,k}^+$ just in case $X_a \in \mathbf{I}_{j,k}$. In other words, the set $\{X_1, X_2, \dots\}$ is just the set $\mathbf{I}_{j,k}$. And so from Equation (30) we have

$$(31) \quad P_k^+(Z) = P_j^+(Z \mid \mathbf{I}_{j,k})$$

This is the analogue in \mathbf{M}^+ of Equation (27), our original verdict that followed directly from (LC) in \mathbf{M} . And since Equation (27) was an arbitrary verdict following directly from (LC), we have shown that any premise following directly from (LC) has an analogue that is a verdict of \mathbf{M}^+ .

This completes our argument. Any verdict of \mathbf{M} is derivable via algebraic steps from a set of premises that are verdicts of \mathbf{M} . The analogue of each premise will be a verdict of \mathbf{M}^+ , so by similar algebraic steps we can derive the analogue in \mathbf{M}^+ of our original verdict. Thus the analogue in \mathbf{M}^+ of any verdict of \mathbf{M} is a verdict of \mathbf{M}^+ . We obtained this result by supposing that the only systematic constraints on our models were systematic constraints (1) through (5). Yet we showed on that basis that if the antecedent of the conditional in (CMP) is met, the consequent follows. Thus for stories lacking cognitive mishap and for models all of whose sentences represent context-insensitive claims, (CMP) follows from systematic constraints (1) through (5) alone.

Michael Titelbaum
University of California, Berkeley
titelb@berkeley.edu