UNLEARNING WHAT YOU HAVE LEARNED

MICHAEL TITELBAUM

Draft of January 2, 2007.
Please do not circulate or cite without permission.

When Bayesians set out to model rational constraints on the ways agents' degrees of belief evolve over time, they usually start by stipulating that the agents they are modeling never forget information. But Frank Arntzenius has shown that there can be substantive constraints on the relation between an agent's degrees of belief at two times even if the agent has forgotten some relevant information between those two times. Consider the following example, adapted from (Arntzenius 2003):

> *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, once in Shangri La you will remember having travelled the Path by the Mountains.
>
> The guardians explain this entire arrangement to you, you believe their words with certainty, they flip the coin, and you follow your path. What does ideal rationality require of your degree of belief in heads once you reach Shangri La?

According to David Lewis's Principal Principle (Lewis 1980), when the guardians initially explain this arrangement to you, at what we'll call t_0, your degree of belief that the coin will come up tails should be 1/2, as should be your degree of belief that you will travel the Path by the Sea. Now suppose the coin comes up tails, and you travel the Path by the Sea. While you are on the Path by the Sea, at what we'll call t_1, you are certain both that the coin came up tails and that you are travelling the Path by the Sea. But once you reach Shangri La, at what we'll call t_2, your memory is erased, and you lose the information that you travelled the Path by the Sea. Arntzenius argues convincingly that at this point, your degree of belief that the coin came up tails should revert to what it was before you had any information about the coin flip's outcome. That is, your t_2 degree of belief in tails should equal your t_0 degree of belief in tails.

A traditional Bayesian modeling approach, based upon a Conditionalization updating rule, is unable to capture this relation between your ideally rational t_0 and t_2 degrees of belief. Conditionalization is usually expressed as something like the following:

> **Conditionalization:** A rational agent's degree of belief in x at t_2 is her degree of belief in x at t_1 conditional on all the information she learns between t_1 and t_2.

It has been widely noted that an agent who becomes certain of a particular claim will remain certain of that claim as long as she updates by Conditionalization. (This is, for instance, what generates the "Problem of Old Evidence.") Thus in the Shangri La story, Conditionalization will require you to be certain at t_2 that the coin came up tails because you were certain at t_1 that it did so. But as we have seen, it is rational for your t_2 degree of belief in tails to equal your t_0 degree of belief in tails, which is lower than your t_1 degree of belief. Thus Conditionalization yields an incorrect verdict in a case in which the agent forgets some information.

But the news for Conditionalization is even worse than that. Suppose the coin lands heads, and you travel the Path by the Mountains. At t_1 you are certain that the coin landed heads and that you take the mountain route. At t_2 you retain memories of the Path by the Mountains, but because you are unsure whether those are genuine memories or memories implanted by the guardians, you are no longer certain which path you traveled. Thus your degree of belief at t_2 that the coin came up heads should once agan match your t_0 degree of belief. And again, because you were certain at t_1 that the coin came up heads Conditionalization will yield the wrong verdict, requiring that your t_2 degree of belief in heads equal your t_1 degree of belief. But notice that in this case no actual memory loss has occurred: because you did in fact travel the Path by the Mountains, the guardians have not had to tamper with your memory at all. As Arntzenius points out, one need not forget any information for Conditionalization to yield incorrect verdicts; the mere *threat* of memory loss can be sufficient for Conditionalization to misfire.

The Shangri La story provides examples in which a relation holds between a rational agent's degrees of belief over time and yet in which the presence of a forgetting episode — or even the threat of such an episode — leaves Conditionalization unable to capture that relation. We might read these examples as *counter-examples* to Conditionalization; we might argue that rationality requires a particular relation to hold here, that Conditionalization gets that relation wrong, and therefore that Conditionalization is a mistaken representation of the requirements of rationality. However, I prefer to think of the Bayesian framework based upon Conditionalization as a modeling technique, and like any modeling technique it has a limited domain of applicability. There are a wide variety of cases that are modeled correctly and elegantly by a Conditionalization-based modeling framework, and we should feel free to retain that framework when modeling cases that fall within that domain. However, the Shangri La story exemplifies a class of cases, involving forgetting or the threat thereof, that lie outside the domain of applicability of a Conditionalization-based modeling framework. The goal of this paper is to offer a modeling framework that extends this domain of applicability, allowing us to capture diachronic relations between rational degrees of belief in stories like Shangri La. After describing the framework and some of its applications, we will examine the domain of applicability of this new, expanded modeling technique.

1. The Modeling Framework

1.1. **Elements of the Framework.**

The goal of our modeling framework is to produce evaluative verdicts about the degrees of belief of agents in situations I call "stories." A story describes an agent who starts with a particular set of claims of which she is certain, then becomes certain of particular claims and/or ceases to be certain of particular claims at various times during the story. (I will describe the objects of agents' doxastic states as "claims" so as to remain neutral among views that take those objects to be propositions, sentences, or something else.) Once we have a story we intend to model, our modeling process proceeds in six steps:

Step 1: Choose a time sequence and modeling language. We begin by specifying a time sequence $(t_1, t_2, \ldots, t_n)$. The time sequence is a finite set of moments in the story at which we are going to model the agent's degrees of belief. We then determine a modeling language by specifying a finite set of atomic sentences. The atomic sentences are strings of symbols each of which represents a claim to which the agent might assign some degree of belief at some point during the story. Atomic sentences are joined by truth-functional connectives in the usual iterative way to fill out the modeling language. For example, the sentence $x \& y$ represents the claim that is the conjunction of the claim represented by x and the claim represented by y.

Step 2: Define unconditional and conditional credence functions. An unconditional credence function $P_k(\cdot)$ is a function from sentences in the modeling language to real numbers. It represents the agent's degree of belief at time t_k in the claim represented by the sentence. A conditional credence function $P_k(\cdot \mid \cdot)$ is a partial function from ordered pairs of sentences in the modeling language to reals, representing the agent's degree of belief at time t_k in the claim represented by the first sentence conditional on the supposition of the claim represented by the second sentence.

Step 3: Determine certainties. The central function of our modeling framework is to determine how the agent's changing sets of certainties shape her degrees of belief in claims of which she is uncertain. I am going to presume that we have available in the background a system of deductive logic that allows us to specify for every claim represented in the modeling language and every time in the time sequence whether the agent is required by ideal rationality to be certain of that claim at that time (if she assigns a degree of belief to the claim at that time at all). We do this as follows: at any given time, the agent is required to be certain of any claim the story says she is certain of at that time, any claim she is required to be certain of at all times (such as "I exist"), and any claim deductively implied by members of the previous two categories.[1] The agent is required to be less-than-certain of any claim not belonging to one of these three categories.

Step 4: Apply systematic and extrasystematic constraints. There are two types of constraints on our models. Systematic constraints apply to every model we construct using this framework, regardless of what story the model represents. I view these constraints, taken together, as representing a particular set of consistency requirements of ideal rationality. They will be enumerated below. We

[1] Note that in determining whether the agent is reqiured to be certain of the claim represented by x at time t_k, we may argue that the claim can be deduced from other claims the agent is required to be certain of, even if those other claims are not represented by sentences in the modeling language.

also have extrasystematic constraints. Extrasystematic constraints represent requirements on the agent's doxastic states derived from the details of the story and requirements of ideal rationality not reflected in the systematic constraints. For example, if the agent believes at t_k that a particular coin is fair, an extrasystematic constraint based on the Principal Principle might require her to assign its coming up heads a degree of belief of 1/2. Most importantly, the set of extrasystematic constraints includes representations of the determinations about certainty made in Step 3. For each sentence in the modeling language and each time in the time sequence, there will be either an extrasystematic constraint setting an unconditional credence in that sentence at that time of 1, or an extrasystematic constraint requiring that unconditional credence at that time to be less than 1.

Step 5: Generate histories and model. A history is a set of unconditional and conditional credence functions containing exactly one unconditional and one conditional credence function for each time in the time sequence. Our model of a story is the set of all possible histories whose credence functions meet both our systematic and extrasystematic constraints.

Step 6: Derive verdicts. An algebraic statement is an equality or inequality relating two expressions composed arithmetically from credence values and/or constants. An arithmetic statement contains no variables and no quantifiers. For example, if x and y are sentences in the modeling language, $P_1(x) + P_2(x \mid y) = 1/2$ will be an arithmetic statement. An arithmetic statement that is true in every history of a model is called a verdict of that model. The systematic and extrasystematic constraints on a model give rise immediately to verdicts of that model, and we can derive further verdicts algebraically from those.

Once we have some verdicts in hand, we can use them to evaluate agents' doxastic states. If when we represent an aspect of an agent's doxastic state (either a single degree of belief assignment or the relation between multiple degree of belief assignments) as an algebraic statement in the language of the model, that algebraic statement contradicts one of the model's verdicts, then the agent's doxastic state fails to be ideally rational.[2] Note that the requirements of "ideal rationality" are stronger than the requirements of rationality as we use the concept ordinarily. Ideal rationality forbids agents to assign inconsistent degrees of belief, whereas in ordinary parlance we may deem an inconsistent agent rational if, for example, she has not recognized that her beliefs are inconsistent.[3] Moreover, the verdicts of our models represent necessary but not sufficient conditions for ideal rationality. An

[2] It need not be that the particular aspect of the agent's doxastic state whose representation contradicts a verdict fails to be ideally rational. The fact that one aspect of the agent's doxastic state, when represented in the model, contradicts one of the model's verdicts indicates only that the agent's doxastic state as a whole fails to be ideally rational. If the verdict involves credences at multiple times, the contradiction indicates only that the agent's doxastic state fails to be ideally rational at at least one moment in the time sequence of the model.

[3] Ideal rationality does not, however, require logical omniscience — for example, it does not require the set of claims of which an agent is certain to be closed under deductive implication. Our evaluative test for ideal rationality involves only those claims to which the agent actually assigns degrees of belief; it does not negatively evaluate agents for failing to assign degrees of belief to other claims.

agent may comply with the verdicts of a model without being ideally rational, as the model does not attempt to represent *all* the requirements of ideal rationality.

Before moving on, a quick note on notation: In what follows, an unitalicized uppercase letter names a model in our framework. An italicized lowercase letter is an atomic sentence in a modeling language. An italicized uppercase letter is a set of sentences. And a bold lowercase letter represents a real number.

1.2. **Systematic Constraints.**

The first four systematic constraints of our modeling framework are synchronic constraints, which taken together represent requirements of consistency between degrees of belief an agent assigns at the same time. Given a modeling language L and a time sequence $(t_1, t_2, \ldots, t_n)$, these constraints are:

(1) For any $t_k \in \{t_1, t_2, \ldots, t_n\}$ and any sentence $x \in L$, $0 \leq P_k(x)$.

(2) For any $t_k \in \{t_1, t_2, \ldots, t_n\}$ and any tautological sentence $\mathsf{T} \in L$, $P_k(\mathsf{T}) = 1$.

(3) For any $t_k \in \{t_1, t_2, \ldots, t_n\}$ and any mutually exclusive sentences $x, y \in L$,

$$P_k(x \vee y) = P_k(x) + P_k(y)$$

(4) For any $t_k \in \{t_1, t_2, \ldots, t_n\}$ and any $x, y \in L$, if $0 < P_k(y)$ then

$$P_k(x \mid y) = \frac{P_k(x \,\&\, y)}{P_k(y)}$$

If $P_k(y) = 0$, the ordered pair (x, y) is not in the domain of the function $P_k(\cdot \mid \cdot)$. (That is, $P_k(x \mid y)$ is undefined.)

The first three systematic constraints, the Kolmogorov Axioms, require unconditional credence functions to be probability functions. The fourth constraint relates conditional credence functions to unconditional credence functions. Since our modeling languages are based on finite sets of atomic sentences, and we are interpreting a credence of 0 as certainty that a particular claim is false, the traditional relation between conditional and unconditional credences described by Systematic Constraint (4) is acceptable.

We now need a diachronic constraint, which when combined with our synchronic constraints will represent requirements of consistency between degrees of belief an agent assigns at different times. We have already seen that if we want stories involving forgetting to be in our framework's domain, Conditionalization will not suffice as a diachronic constraint. To formulate a better diachronic constraint, however, it will help to look more closely at why Conditionalization fails.

With the other elements of our modeling framework laid out formally, we can re-express Conditionalization in a much more precise fashion. To do so, it will help to have a bit of terminology and a bit more notation in place. First, we will define a "certainty set." Given a model with modeling language L and a time t_i in its time sequence, the agent's certainty set at t_i is the set $C_i = \{x \in L : P_i(x) = 1\}$. Second, we will define an operation on sets of sentences. Given a set of sentences $S \subseteq L$, $\langle S \rangle$ is a sentence in L selected as follows: If S is nonempty, $\langle S \rangle$ must be truth-functionally equivalent to the conjunction of all the sentences in S; if S is empty, $\langle S \rangle$ must be a (truth-functional) tautology. Because of the way our modeling languages are constructed, for any modeling language L and any set $S \subseteq L$, there is guaranteed to be an $\langle S \rangle \in L$. (In fact, there will often be more than one such sentence in L; given our synchronic constraints it will not matter which one serves as $\langle S \rangle$.) We will sometimes refer to $\langle S \rangle$ as the sentence "equivalent" to S.

With that terminology and notation in place, we can re-express Conditionalization as:

> **Conditionalization:** Given a model defined over modeling language L, a sentence $x \in L$, and two times t_j and t_k in the time sequence with $j < k$, $P_k(x) = P_j(x \,|\, \langle C_k - C_j \rangle)$.

Suppose we try to represent the version of the Shangri La story in which you travel the Path by the Mountains with a model that uses this precisified version of Conditionalization as a systematic constraint. Our modeling language L will contain just one atomic sentence, h, representing the claim that the coin comes up Heads. Our time sequence will consist of t_0, t_1, and t_2 as described above. Among our extrasystematic constraints will be the algebraic statement $P_0(h) = 1/2$ (by the Principal Principle) and the algebraic statement $P_1(h) = 1$ (because at t_1 you are certain you are travelling the Path by the Mountains).

Since at t_0 you are certain neither of the claim represented by h nor of its negation, C_0 consists only of the truth-functional tautologies of L. C_1 contains not only these tautologies but also h (as well as various sentences in L implied by h). At t_2, however, you are once more uncertain whether the coin came up heads, so C_2 is identical to C_0. Thus if we apply Conditionalization with $j = 0$ and $k = 2$, we have

$$P_2(h) = P_0(h \,|\, \langle C_2 - C_0 \rangle) = P_0(h \,|\, \langle \phi \rangle) = P_0(h \,|\, \mathsf{T}) \tag{1}$$

where T is some tautology in L. By our synchronic systematic constraints, this yields

$$P_2(h) = P_0(h) = 1/2 \tag{2}$$

But if we apply Conditionalization with $j = 1$ and $k = 2$ and then our synchronic systematic constraints, we have

$$P_2(h) = P_1(h \,|\, \langle C_2 - C_1 \rangle) = P_1(h \,|\, \langle \phi \rangle) = P_1(h) = 1 \tag{3}$$

If we take Conditionalization as one of our systematic constraints in a model of Shangri La, both Equations (2) and (3) will be verdicts of that model. Clearly any degree of belief you assign to Heads at t_2 will contradict at least one of these verdicts, so according to this model there is no degree of belief you can assign to Heads at t_2 and be ideally rational. We might say that when Conditionalization is applied to stories like Shangri La, it contradicts itself.

Again, this just demonstrates that stories like Shangri La are not in the domain to which Conditionalization-based modeling frameworks apply.[4] But now we can pinpoint more precisely where the trouble lies. The application of Conditionalization that generates Equation (2) is perfectly acceptable; it generates a verdict that squares with our intuitions about your degree of belief in Heads at t_2. We run into trouble when we try to apply Conditionalization to relate your t_1 and t_2 credences.

[4] You might put this point another way. You might say that Shangri La is within the domain of applicability of Conditionalization-based modeling frameworks, but that those frameworks represent a standard of rationality that is violated by any agent who has forgotten information or believes she might have forgotten. The modeling framework we are about to develop, then, would represent a looser rationality standard on which agents can forget information without running afoul of the requirements of ideal rationality.

But this makes perfect sense. Conditionalization was designed to apply to situations in which an agent's store of information either holds fixed or increases. (We might see these as situations in which the set of possible worlds the agent entertains either remains the same or strictly shrinks.) Conditionalization fails when an agent loses information, in that she goes from being certain of a particular claim to being less-than-certain of that claim. This can happen when the agent forgets a claim, or in subtle situations involving the threat of forgetting. The first occurs in Shangri La when you travel the Path by the Sea; the second occurs when you travel the Path by the Mountains. Either way, you go from certainty about the outcome of the coin flip while you are travelling your path to less-than-certainty when you reach Shangri La.

Thus Conditionalization should only be applied to pairs of times for which the agent has not lost any certainties from the earlier to the later. Applying this restriction to Conditionalization yields our diachronic systematic constraint:

> **Systematic Constraint (5), Limited Conditionalization (LC)**: Given a model defined over modeling language L, a sentence $x \in L$, and two times t_j and t_k in the time sequence with $j < k$, if $C_j \subseteq C_k$ then $P_k(x) = P_j(x \mid \langle C_k - C_j \rangle)$.

Basing our model on (LC) instead of Conditionalization immediately resolves the problem we had above. In Shangri La, $C_0 \subseteq C_2$, so we can conditionalize from t_0 to t_2 and Equation (2) remains a verdict of our model. But $C_1 \nsubseteq C_2$, so the antecedent of (LC) is not met when $j = 1$ and $k = 2$, we cannot conditionalize from t_1 to t_2, and Equation (3) is no longer a verdict. The conflict has been resolved, and we have retained the verdict that matches our intuitive judgment of what ideal rationality requires.[5]

2. Applications

2.1. Generalized Conditionalization.

2.1. Generalized Conditionalization. (LC) captures relations between degrees of belief held by an agent at different times when that agent loses no certainties from the earlier time to the later time. Despite the fact that in Shangri La you lose certainties from t_1 to t_2, we were able to use (LC) to derive verdicts about your t_2 degrees of belief by relating them to degrees of belief assigned at t_0. But is there any interesting relation that holds between your degrees of belief at t_2 and your degrees of belief at t_1?

Here we can apply a general result obtainable from (LC) and our synchronic systematic constraints. Suppose we have a story, a model of that story with modeling language L, and two times t_j and t_k in the time sequence. As long as there is no $y \in L$ such that $y \in C_j$ and $\sim y \in C_k$, we can imagine a time t_l later than both t_j and t_k at which the agent is ideally rational and her certainty set $C_l = C_j \cup C_k$. Since $C_j \subseteq C_l$, for any $x \in L$ (LC) yields $P_l(x) = P_j(x \mid \langle C_l - C_j \rangle)$. Similarly, since

[5] In their response to (Arntzenius 2003), M.J. Schervish, T. Seidenfeld, and J.B. Kadane (2004) suggest that it "is already assumed as familiar in problems of stochastic prediction" that one should update one's degrees of belief by conditionalizing only when "the information the agent has at t_2 includes all the information that she or he had at time t_1." They argue that this rule is not violated by the Shangri La story; they see the rule as a material conditional that is trivially satisfied when Shangri La falsifies its antecedent. This paper goes beyond Shcervish, Seidenfeld, and Kadane's by showing that a properly limited conditionalizing rule is not only *satisfied* in forgetting stories — it can also yield *positive verdicts* for such stories.

$C_k \subseteq C_l$, for any $x \in L$ (LC) yields $P_l(x) = P_k(x \mid \langle C_l - C_k \rangle)$. Setting these two equations equal, and applying the fact that $C_l = C_j \cup C_k$, we have the following general rule:

> **Generalized Conditionalization Principle (GC)**:
> *Given* a modeling language L, two times t_j and t_k in the time sequence, and any $x \in L$,
> *if* there does not exist a $y \in L$ such that $y \in C_j$ and $\sim y \in C_k$,
> *then* $P_j(x \mid \langle C_k - C_j \rangle) = P_k(x \mid \langle C_j - C_k \rangle)$.

Note that we argued for (GC) using (LC) and our synchronic systematic constraints. Thus (GC) does not need to be added into our modeling framework as a constraint; it simply highlights a pattern that was already present in the sets of credence functions consistent with our framework. For example, when $j < k$ and $C_j \subseteq C_k$, $C_j - C_k$ is just the empty set, so the verdict yielded by (GC) is

$$P_j(x \mid \langle C_k - C_j \rangle) = P_k(x \mid \mathsf{T}) = P_k(x) \tag{4}$$

This is just the verdict we would have in this case from (LC).

(GC) also does a nice job of making explicit patterns in degrees of belief in cases where agents lose information. Suppose $j < k$ and $C_k \subset C_j$; the agent has lost some certainties from t_j to t_k and gained none. Then $C_k - C_j$ is empty, so (GC) yields

$$P_j(x) = P_k(x \mid \langle C_j - C_k \rangle) \tag{5}$$

This equation tells us that if at the later time, the agent conditionalizes on all the certainties she has lost since the earlier time, her degree of belief in the claim represented by x will be just what it was before those certainties were lost.

We can understand this process as a kind of reverse-temporal conditionalization, and it is a revealing way to think about what happens when an agent loses information. Our modeling framework tracks the changes in an agent's partial beliefs driven by changes in her certainty set. When an agent gains information, her certainty set expands; when she forgets information, it contracts. From the point of view of her certainty set, one process is just the other happening backwards in time, and so it is no surprise that the effects of these processes on her degrees of belief mirror each other precisely.

And this is what is occuring between t_1 and t_2 in Shangri La. The story is arranged so that, whichever path you travel, your certainty set at t_2 is identical to your certainty set at t_0. Thus the relation between your t_2 degrees of belief and your t_1 degrees of belief is identical to the relation between your t_0 degrees of belief and your t_1 degrees of belief. If you travel the Path by the Mountains, for example, your t_1 degree of belief in Heads is equal to your t_0 degree of belief in Heads conditional on the information you gain between t_0 and t_1, represented in our model by h. But your t_1 degree of belief in Heads is also equal to your t_2 degree of belief in Heads conditional on the information you lose between t_1 and t_2, also represented in our model by h. (GC) brings out this relation between your t_1 and t_2 degrees of belief:

$$P_1(h) = P_2(h \mid h) = 1 \tag{6}$$

2.2. An Application of (GC).

One of Conditionalization's attractive features as a principle of ideal rationality is that it tells you precisely what to do when you gain new information; given a full specification of your earlier credence function and of the set of certainties gained, you can determine what your entire credence function should look like after an update by Conditionalization. It might have been hoped that (GC) would yield a similar recipe for responding to a forgetting episode; given a full specification of your earlier credence function and of the set of certainties lost, it would be clear precisely what your new credence function should look like.

Unfortunately, (GC) does not yield such a recipe. Even given a full specification of $P_j(\cdot)$ and of $C_j - C_k$, Equation (5) does not pick out a unique function for $P_k(\cdot)$. (GC) places restrictions on a number of conditional P_k values, and thereby on ratios of various unconditional P_k values. However, it will not specify a unique $P_k(x)$ value for every $x \in L$. And when we move to stories in which an agent both gains and loses certainties between t_j and t_k (so that C_j is neither a subset nor a superset of C_k), the relations between degrees of belief brought out by (GC) are even less strict.

Nevertheless, the relations captured by (GC) do place substantive and revealing constraints on how an agent's degrees of belief can change when she loses information. Consider the following example:

> *The Lottery:* Al, Dave, and Frank are the only participants in a lottery. One of their names will be drawn at random from a hat, and that first name drawn is the winner's. However, to heighten the suspense, the lottery's organizers will after drawing the name of the winner draw another name, and announce that that player is one of the losers. A week later they will announce the name of the winner.
>
> The organizers draw one name, then draw another. They announce that Al is one of the losers. Hearing this, Dave is overjoyed that he is still in the running. In fact, he is so overjoyed that after a few days' time he has forgotten entirely which name was announced as a loser's; he remembers only that his name was not announced.

Suppose that after a few days' time, Dave has forgotten so completely whether Al's or Frank's name was announced that he comes to view it as equally likely that either of them is still in the running with him. The following table describes the ideally rational development of Dave's degrees of belief:

Claim	P_0	P_1	P_2
Al wins.	1/3	0	1/4
Dave wins.	1/3	1/2	1/2
Frank wins.	1/3	1/2	1/4
"Al" is the name announced.	1/3	1	1/2
"Dave" is the name announced.	1/3	0	0
"Frank" is the name announced.	1/3	0	1/2

Before the name of a loser has been announced, at t_0, Dave considers it equally likely that any of the three players will win, and equally likely that any of the three will be the one whose named is announced as a loser. After Al's name is announced, at t_1, Dave is certain that Al is not the winner. But by t_2, Dave can only remember that his name wasn't announced, and he thinks it equally likely that Al or Frank was the announced loser. (GC) requires Dave at t_2 to retain his confidence of 1/2 that he is the winner, while dividing his remaining credence equally between victories by Al and by Frank.

Now imagine that as time wears on after t_2 Dave completely loses track of whose name was announced by the organizers, and comes to view it as equally likely that any of the three competitor's names was announced. But suppose that at the same time he retains the confidence he had at t_2 that he is twice as likely to win as either Al or Frank. In other words, suppose that Dave's degrees of belief develop by t_3 to these:

Claim	P_0	P_1	P_2	P_3
Al wins.	1/3	0	1/4	1/4
Dave wins.	1/3	1/2	1/2	1/2
Frank wins.	1/3	1/2	1/4	1/4
"Al" is the name announced.	1/3	1	1/2	1/3
"Dave" is the name announced.	1/3	0	0	1/3
"Frank" is the name announced.	1/3	0	1/2	1/3

The t_3 degrees of belief listed above can all be fit into a credence function that meets our four synchronic systematic constraints. Nonetheless, the development from t_2 to t_3 violates (GC). So according to a modeling framework that takes (LC) as a diachronic constraint, the degree of belief development shown in the table above violates the requirements of ideal rationality. And I believe this squares with our intuitions about the case — Dave's confidence in his chances at t_3 is in tension with his beliefs about the evidence.

Notice that the updating processes deemed ideally rational by our (LC)-based framework do not commit it to a "foundations" approach to belief revision of the sort criticized in (Harman 1986).[6] Harman objects to updating policies that require an agent to surrender a belief when she can no longer remember the evidence that justified that belief in the first place. Our (LC)-based framework yields no negative verdict concerning the development of Dave's degrees of belief in The Lottery from t_0 to t_1 and then to t_2. Yet at t_2 Dave is more confident in his chances than he was at t_0 despite the fact that he can no longer remember the piece of evidence (the announcement of Al's name) that justified that increase in confidence.

(LC) does, however, require a particular kind of consistency among Dave's beliefs over time. While Dave need not remember the piece of evidence that rendered victory more likely between t_0 and t_2, he does have to be more confident at t_2 that the organizers' announcement was evidence of victory than he was at t_0. This is what goes wrong between t_2 and t_3: at t_3, Dave is still more confident of victory than he was at t_0 without assigning a corresponding higher confidence that the organizers' announcement supported his chances. (LC) does not require you to retain the evidence that justified your altering a degree of belief, but it does require you to assign degrees of belief concerning the evidence in a fashion consistent with your having made the alteration for good reason.

2.3. **Generalizing Reflection.** An agent's supposition that she has made, or will make, belief changes in a rational fashion is the subject of a well-known rationality constraint articulated by Bas van Fraasen. van Fraasen's Reflection Principle says that an ideally rational agent will, conditional on the supposition that a future version of herself will assign a degree of belief of $\mathbf{r}$ to the claim represented by x, assign a current degree of belief of $\mathbf{r}$ to the claim represented by x.

[6] I am grateful to John MacFarlane for pressing me to address this point.

Under certain conditions, the Reflection Principle can be argued for from (LC). For example, (LC) can be used to argue for Reflection when the following conditions are met (letting t_c be the current time and t_f be the future time about which the agent is making her suppositions):

(1) The agent is certain that at t_f she will be ideally rational.
(2) The agent is certain that $C_c \subseteq C_f$.
(3) The agent is certain that all the sentences in $C_f - C_c$ will be true.
(4) There is a finite set of sentences E such that any pair of sentences in E is mutually exclusive and the agent is certain that exactly one of the sentences in E is equivalent to $C_f - C_c$. (She need not know *which* sentence in E is equivalent to $C_f - C_c$.)
(5) The agent is aware of the degrees of belief she currently assigns conditional on each of the sentences in E.

When these conditions are met, the agent can reason as follows: The agent supposes that an ideally rational future self certain of everything she is currently certain of will assign degree of belief $\mathbf{r}$ to the claim represented by x. The agent can sort through the elements of E and determine a set $S \subseteq E$ such that for each $y \in S$ the agent assigns $P_c(x \mid y) = \mathbf{r}$. Since (LC) relates $P_c(\cdot)$ to $P_f(\cdot)$, supposing that $P_f(x) = \mathbf{r}$ is tantamount to supposing that $C_f - C_c$ is equivalent to one of the elements of S. And this, in turn, is tantamount to supposing the disjunction of all the elements of S. Since the elements of S are mutually exclusive, and since for each element of S the agent currently assigns a credence of $\mathbf{r}$ to X conditional on that element, by a theorem of the probability calculus derivable from our synchronic systematic constraints the agent assigns a credence of $\mathbf{r}$ to x conditional on the disjunction of all the elements of S. Thus given the conditions listed above, if the agent is ideally rational she currently assigns a degree of belief of $\mathbf{r}$ to x conditional on the supposition that her future self will assign an unconditional degree of belief of $\mathbf{r}$ to x.[7]

We have just used (LC) to argue for van Fraasen's Reflection Principle under certain conditions. Just as (LC) applies to updates in which the agent does not lose any certainties, Reflection applies to suppositions about a future self who is certain of everything you are. But what about future selves who have lost some of your current information? Here, we can use (GC) to formulate a generalized version of Reflection. Reflection involves a supposition about a future self's *unconditional* degree of belief in the claim represented by x. The Generalized Reflection Principle involves a supposition about a future self's *conditional* degree of belief in the claim represented by x — namely the degree of belief in x conditional on the set of certainties your future self has lost since the current time. The Generalized Reflection Principle says that an ideally rational agent will, conditional on the supposition that a future version of herself certain of everything she is currently certain of except for F will assign $P_f(x \mid \langle F \rangle) = \mathbf{r}$, assign a current degree of belief of $\mathbf{r}$ to the claim represented by x.

This Generalized Reflection Principle can be argued for from (GC) when the following conditions are met:

(1) The agent is certain that at t_f she will be ideally rational.

[7] In formulating this argument and generating the list of conditions above, I am indebted to (Weisberg manuscript), which is in turn an analysis of the argument van Fraasseen offers in (van Fraassen 1995) that the Reflection Principle can be deduced from Conditionalization.

(2) For some set of sentences F, the agent is certain that $C_c - C_f = F$.
(3) The agent is certain that all the sentences in $C_f - C_c$ will be true.
(4) There is a finite set of sentences E such that any pair of sentences in E is mutually exclusive and the agent is certain that exactly one of the sentences in E is equivalent to $C_f - C_c$. (She need not know *which* sentence in E is equivalent to $C_f - C_c$.)
(5) The agent is aware of the degrees of belief she currently assigns conditional on each of the sentences in E.

When these conditions are met, the agent can reason as follows: The agent supposes that an ideally rational future self certain of everything she is currently certain of except F will assign $P_f(x \,|\, \langle F \rangle) = \mathbf{r}$. The agent can sort through the elements of E and determine a set $S \subseteq E$ such that for each $y \in S$ the agent assigns $P_c(x \,|\, y) = \mathbf{r}$. By (GC), $P_c(x \,|\, \langle C_f - C_c \rangle) = P_f(x \,|\, \langle F \rangle)$. Thus supposing that $P_f(x \,|\, \langle F \rangle) = \mathbf{r}$ is tantamount to supposing that $C_f - C_c$ is equivalent to one of the elements of S. And this, in turn, is tantamount to supposing the disjunction of all the elements of S. Since the elements of S are mutually exclusive, and since for each element of S the agent currently assigns a credence of $\mathbf{r}$ to x conditional on that element, by a theorem of the probability calculus derivable from our synchronic systematic constraints the agent assigns a credence of $\mathbf{r}$ to x conditional on the disjunction of all the elements of S. Thus given the conditions listed above, if the agent is ideally rational she currently assigns a degree of belief of $\mathbf{r}$ to x conditional on the supposition that her future self will assign $\mathbf{r}$ to x conditional on F.

The conditions listed above are sufficient to derive our Generalized Reflection Principle from (GC); I do not know if they are necessary. The point is simply that (GC) can yield a general principle for how an agent should respond to suppositions about her future degrees of belief even when she is certain her future self will have forgotten some information she currently possesses.

3. Domain of Applicability

3.1. **Mandated Credences.** In Section 1.2 above we saw that stories involving forgetting and even some stories involving the threat of forgetting lie outside the domain of applicability of modeling frameworks that use Conditionalization as their diachronic systematic constraint. In these stories, Conditionalization over-generates verdicts — its models make demands on agents that do not represent requirements of ideal rationality. Our next question is whether there are stories for which (LC) over-generates verdicts, stories that lie outside the domain of applicability of our (LC)-based modeling framework.

To answer this question, it helps to distinguish two types of situations. The difference between them can be illustrated by the following story:

> *Chocolate*: We play the following game: I flip a fair coin. If the coin comes up heads, I decide whether to give you a piece of chocolate. If the coin comes up tails, I give you no chocolate. After the rules of the game are explained but before the coin is flipped, what is your degree of belief that you will receive some chocolate?

The important question about this story is whether there exists a specific degree of belief that ideal rationality requires you to assign to the claim that you will receive some chocolate. Most of us would agree that ideal rationality forbids your assigning a degree of belief of zero, and given the Principal Principle it also forbids your assigning 1/2 or greater. But even taking into account everything you know about me, about human nature, about chocolate, and about how humans feel about chocolate, is there a precise degree of belief (such as 1/4) that ideal rationality demands you assign to the prospect of chocolate? Or are there multiple acceptable degrees of belief, such that if you assigned any one of them your degres of belief would be consistent with ideal rationality?

There are a few other options for what ideal rationality might require of your degree of belief that you will receive chocolate in this story. For example, we might think that instead of assigning a single degree of belief to the claim that you will receive chocolate it is possible for you to assign a *range* of degrees of belief to that claim. We might then conclude that any degree-of-belief range whose minimum is positive and whose maximum is less than 1/2 is rationally acceptable in this situation. Or we might conclude that the only ideally rational response is to adopt the specific degree-of-belief range represented by the interval $(0, 1/2)$. While I think these ranged options are worth exploring, I am going to set them aside, as they are not relevant to establishing the domain of applicability of our (LC)-based modeling framework. I should note, however, that the questions I raise below about the strength of rational requirements and the domain of applicability of our modeling framework are just as pressing if you think agents should assign degree of belief ranges as they are for precise degree of belief cases.

The Chocolate story is just one case in which we might wonder whether the information contained in the certainty set an agent entertains at a particular time rationally requires her to assign a specific degree of belief to a specific claim, independent of any degrees of belief she may have assigned at other times. To formalize this question, it helps to have the notion of "synchronically deriving" a verdict. A verdict can by synchronically derived in a model M just in case it can be algebraically derived exclusively from that model's extrasystematic constraints and its synchronic systematic constraints. In other words, while our systematic constraints (1) through (4) can be used in a synchronic derivation, (LC) cannot.[8]

With the notion of synchronic derivation in hand, we can offer the following definition:

> A set $S \subset L$ *mandates a credence* for x in M just in case:
> - for some t_i in the time sequence of M, $C_i \subseteq S$; *and*
> - there exists a real number $\mathbf{r}$ such that for every t_i in the time sequence of M for which $C_i \subseteq S$, the verdict $P_i(x \mid \langle S \rangle) = \mathbf{r}$ can be synchronically derived in M.

The important part of this definition is the second bulleted condition. The equation in that condition uses a conditional credence function so that the set S will mandate not just the ideally rational degrees of belief of agents whose entire certainty set is S, but also the ideally rational degrees of belief of agents for whom what they take for certain and what they are conditionally supposing sums to the set S. The first bulleted condition is in the definition to keep the universally quantified material

[8] If we wanted to define "synchronic derivation" strictly in terms of the model itself and not in terms of how algebra can be used to derive things about it, we could use the following definition: Given the model M, construct a model M^s with the same modeling language, time sequence, and extrasystematic constraints. However, let the only systematic constraints on M^s be the synchronic systematic constraints (1) through (4). Any algebraic statement that is true of all the histories in M^s will be a verdict that is synchronically derivable in M.

conditional in the second condition from being satisfied trivially by an S that always makes that conditional's antecedent false, as would happen for example if S were the empty set.[9]

A set of sentences' mandating a credence is defined relative to a particular model of a story. When we build a model, we bring to bear principles of ideal rationality beyond those represented in our systematic constraints; these additional principles help determine the model's extrasystematic constraints. What principles of ideal rationality we employ in setting the extrasystematic constraints may affect which sets mandate credences for which sentences in the model. If we build different models of the same story with the same time sequence and modeling language but with different principles of ideal rationality reflected in their extrasystematic constraints, a given set of sentences may mandate a credence for a particular sentence in one model but the other.

A view in epistemology that takes the requirements of ideal rationality to be very strong might hold that for any story, once we understand all the principles of ideal rationality that apply and represent them in our extrasystematic constraints, in the model that results *any* consistent subset of L will mandate a credence for *any* sentence in the modeling language. For example, in the Chocolate story any model that represents the Principal Principle in its extrasystematic constraints will have your certainty set mandate a credence of 1/2 for the sentence representing the claim that the coin comes up heads. But a view that takes the requirements of ideal rationality to be very strong will also hold that whatever you may know about people, their chocolate preferences, etc., there is a precise degree of belief that ideal rationality requires you to assign that you will be receiving some chocolate. If in a particular model the certainty set representing your knowledge doesn't mandate a credence for the sentence representing the claim that you will receive chocolate, that's just because that model doesn't represent *all* the principles of ideal rationality applying to this case.

A view that takes the requirements of ideal rationality to be somewhat weaker, however, might hold that however many requirements of ideal rationality we may discover and codify into principles, we will never come to the conclusion that there is a precise degree of belief that you will receive chocolate mandated in the Chocolate story, because the requirements of ideal rationality just aren't strong enough to yield that level of precision for that case. On this view, two different agents placed in the Chocolate story might have identical certainty sets and still assign different degrees of belief that they were going to receive some chocolate, without either of them violating the requirements of ideal rationality. In other words, there will be no model of the Chocolate story representing requirements of ideal rationality on which your certainty set mandates a credence for the sentence representing the claim that you will receive chcolate. This isn't because we don't know enough about what ideal rationality requires, but instead because the requirements of ideal rationality simply aren't that strong.

3.2. **Mandated Credences and (LC).**

The notion of mandated credences helps delineate the domain of applicability of our (LC)-based modeling framework. Suppose we are given a model M, its modeling language L, and two times in its time sequence t_j and t_k with $j < k$. If for some $x \in L$, C_k mandates a credence for x in

[9] I am grateful to Daniel Warren for pointing out the possibility that the second condition might be satisfied trivially.

M, any verdict (LC) generates relating t_j credences to $P_k(x)$ will be synchronically derivable in M.

Proof: Suppose that C_k mandates a credence for X in M. Then by the definition of mandated credences there exists an $\mathbf{r}$ such that for every t_i in M's time sequence for which $C_i \subseteq C_k$, $P_i(x \mid \langle C_k \rangle) = \mathbf{r}$ is synchronically derivable in M. Thus $P_k(x \mid \langle C_k \rangle) = \mathbf{r}$ is synchronically derivable in M. And since $P_k(\langle C_k \rangle) = 1$, we can use our synchronic systematic constraints to show that $P_k(X) = \mathbf{r}$ is synchronically derivable in M. Now consider some t_j earlier than t_k. If $C_j \nsubseteq C_k$, (LC) does not generate a verdict relating t_j credences to $P_k(x)$. If $C_j \subseteq C_k$, the verdict $P_j(x \mid \langle C_k \rangle) = \mathbf{r}$ can be synchronically derived in M (because C_k mandates a credence for x in M). Since $P_j(\langle C_j \rangle) = 1$, a bit more work with our synchronic systematic constraints shows that $P_j(x \mid \langle C_k - C_j \rangle) = \mathbf{r}$ is synchronically derivable in M. And with a final algebraic step, we can synchronically derive $P_k(x) = P_j(X \mid \langle C_k - C_j \rangle)$. This is the verdict (LC) generates relating t_j credences to $P_k(x)$.

We can think of what is going on here in the following way: Suppose there are principles of ideal rationality we can represent in the extrasystematic constraints of a model so that in that model C_k mandates a credence for the sentence representing a particular claim. (Call that sentence x.) We might think of the information in C_k as giving rise to a sort of probabilistic theory of the world, a theory which specifies a particular unconditional degree of belief for the claim represented by x. (Call that degree of belief $\mathbf{r}$.) At t_k the agent's certainty set is C_k, so the probabilistic theory of the world reflecting her certainties rationally requires her to assign an unconditional credence of $\mathbf{r}$ to x. Now consider an earlier time t_j such that $C_j \subseteq C_k$. If C_j is a proper subset of C_k, the agent's probabilistic theory of the world at t_j does not require her to unconditionally assign $\mathbf{r}$ to x. But what if she conditionally supposes the set $C_k - C_j$? Combining what she actually takes for certain at t_j with what she is supposing to be the case, she is working conditionally with the probabilistic theory of the world generated by C_k. So ideal rationality requires her degree of belief in x conditional on $C_k - C_j$ to be $\mathbf{r}$. Thus we have the rational requirement $P_k(x) = P_j(x \mid \langle C_k - C_j \rangle)$.[10]

Now suppose we have a story for which we are confident that our synchronic constraints represent requirements of ideal rationality. If we know that there exist principles of ideal rationality such that in a model of this story representing those principles in its extrasystematic constraints, the agent's certainty set at some t_k would mandate a credence for some x, we can be confident that verdicts derived from (LC) relating $P_k(x)$ to the agent's earlier credences represent genuine requirements of ideal rationality. In other words, as far as relations between $P_k(x)$ and degrees of belief assigned prior t_k go, the story falls within the domain of applicability of our (LC)-based modeling framework. This is because the relevant verdicts of (LC) can be synchronically derived, and we are confident that the constraints used in synchronic derivations (the synchronic systematic constraints and our extrasystematic constraints) represent requirements of ideal rationality.

I anticipate that the most suspect systematic constraint of the modeling framework I have proposed will be (LC). You might also be concerned about the sychronic systematic constraints — for instance, you might worry that in stories involving infinite domains, systematic constraint (4) will generate verdicts that do not represent

[10] I am grateful to Mike Caie for suggesting this line of argument to me.

requirements of ideal rationality.[11] But we are not going to be working with stories involving infinite domains here, so for the rest of this paper I will set such worries aside and assume that for any story, synchronically derived verdicts represent requirements of ideal rationality. Given that assumption, any story for which there exist principles of ideal rationality such that the certainty set C_k mandates a credence for the claim represented by some x will lie in the domain of applicability of a modeling framework that includes our synchronic constraints and the verdicts of (LC) relating credences at times earlier than t_k to $P_k(x)$. If for a given story there exist principles of ideal rationality such that in a model representing them every consistent set of sentences mandates a credence for every sentence in the modeling language, we can simply say that that story lies in the domain of applicability of our (LC)-based modeling framework. And on a view of rational requirements' strength which holds that every story is like that, every story will lie in the domain of applicability of our modeling framework.

Notice that in these situations we can be confident of verdicts derived from (LC) because (LC) isn't really adding anything to the framework. The verdicts of (LC) we are discussing could be derived synchronically, without invoking (LC). Put another way: even if (LC) weren't a constraint on the framework, our models would still yield these verdicts. In these situations, ideal rationality requires the agent to assign particular degrees of belief (conditional and unconditional) at particular times based solely on the information the agent takes for certain at that time. These requirements are independent of any degrees of belief the agent may have assigned at other times; no relations between past and future degrees of belief influence these degree-of-belief requirements. Nevertheless, structural aspects of our synchronic constraints cause these required degrees of belief to fall into simple mathematical patterns over time, and (LC) captures these mathematical patterns. We might say that in these situations (LC) does not *generate* the relations represented by the diachronic verdicts it yields; (LC) merely draws our attention to patterns already in the model that might be particularly useful to notice.

3.3. **(GC), Mandated Credences, and Interpersonal Relations.** We can also use the concept of mandated credences to identify a set of situations for which (GC) is guaranteed to yield verdicts representing requirements of ideal rationality. If in a given model the set $C_j \cup C_k$ mandates a credence for x, then the verdict $P_j(x \mid \langle C_k - C_j \rangle) = P_k(x \mid \langle C_j - C_k \rangle)$ can be synchronically derived. The proof is simple. By the definition of mandated credence, there exists an $\mathbf{r}$ such that $P_j(x \mid \langle C_j \cup C_k \rangle) = \mathbf{r}$ and $P_k(x \mid \langle C_j \cup C_k \rangle) = \mathbf{r}$ can be synchronically derived, so $P_j(x \mid \langle C_j \cup C_k \rangle) = P_k(x \mid \langle C_j \cup C_k \rangle)$ can be synchronically derived. Since $P_j(\langle C_j \rangle) = 1$ and $P_k(\langle C_k \rangle) = 1$, our synchronic constraints yield $P_j(x \mid \langle C_k - C_j \rangle) = P_k(x \mid \langle C_j - C_k \rangle)$.

For example, in the version of the Shangri La story in which you travel the Path by the Mountains, the certainty set C_1 you possess while on the path includes the fact that the coin came up heads, and therefore mandates a credence of 1 for the sentence h. Since the certainty set you possess once you reach Shangri La, C_2 is a subset of C_1, the set $C_1 \cup C_2$ mandates a credence for h. Thus we can be confident that the (GC) verdict relating your t_1 and t_2 degrees of belief in h (Equation (6) in Section 2.1 above) represents a requirement of ideal rationality. Similarly, if the

[11] This concern is thoroughly discussed in (Hájek 2003).

Principal Principle is represented in the extrasystematic constraints on our model, C_2 will mandate a credence of 1/2 for h. Since C_0 (your certainty set just after the guardians explain their plan) is a subset of C_2, we can be confident that the (LC) verdict relating your t_0 and t_2 degrees of belief in h (Equation (2) in Section 1.2 above) represents a requirement of ideal rationality.

As was the case with (LC), when the relevant credences are mandated for the application of (GC), (GC) merely highlights mathematical patterns that were established already (so to speak) by our synchronic constraints. As a result, the verdicts yielded by (GC) gain two attractive features in situations in which $C_j \cup C_k$ mandates a credence for the sentence in question.

First, in these situations we can prove that the verdicts yielded by (GC) represent requirements of ideal rationality without adding any times into the time sequence. Our argument from (LC) to (GC) in Section 2.1 above required imagining a time t_l after t_j and t_k at which the agent's certainty set is $C_j \cup C_k$, then relating her t_j and t_k degrees of belief to t_l degrees of belief. This might have inspired some concern, as there is no guarantee that adding an extra time into a story will leave the agent's t_j and t_k degrees of belief unchanged. (The very fact that such a time will or could exist might alter the agent's degrees of belief in particular claims.) The proof offered in the first paragraph of this section, however, works directly with the agent's t_j and t_k credences conditional on $C_j \cup C_k$, without assuming there is a time at which the agent assigns that set a degree of belief of 1. Thus in situations in which $C_j \cup C_k$ mandates a credence for x, (GC)'s verdicts can be established without controversial additional stipulations.

Second, in situations with the relevant mandated credences (GC) can be used to model interpersonal credence relations. When the relevant mandated credences are in place, the mathematical patterns represented by (GC) are generated in a very direct way: the information content of the set $C_j \cup C_k$ works directly on conditional P_j values through our synchronic systematic constraints, then that information works directly on conditional P_k values in the same way. No prior relation between P_j and P_k values is assumed in generating the patterns captured by (GC). Thus those patterns will hold even if P_j and P_k represent the degrees of belief of different people.

To model such interpersonal credence relations, we build our models just as before but let each t_i in the "time sequence" represent an agent-time pair. So t_3 might represent Agent A at 1:35pm, in which case P_3 values will represent Agent A's degrees of belief at 1:35. In some models, there will be a subset of t-values that represent different agents at the same time; this allows us to model relations between different agents' simultaneous degrees of belief. Some models will also have a subset of t-values all of which represent the same agent; these t-values allow us to model a single agent's degrees of belief developing over time. Or we might just have a variety of agents' degrees of belief being modeled at a variety of times.

Even with this new interpretation of the "time sequence," our synchronic constraints remain the same. They now represent requirements of consistency between the degrees of belief assigned by just one of our agents at a particular time. The definitions of "certainty set," "synchronic derivation," and "mandated credence" do not change either, so the proof that began this section remains intact. If $C_j \cup C_k$ mandates a credence for x, ideal rationality requires that $P_j(x \mid \langle C_k - C_j \rangle) = P_k(x \mid \langle C_j - C_k \rangle)$.

Thus in particular mandated-credence situations, the familiar (GC) equation can be generalized to express a rationally-required relation between degrees of belief assigned by two different people, either at the same time or at different times. In the special case where $C_j \subseteq C_k$, we can derive an "interpersonal (LC)" equation $P_j(x \mid \langle C_k - C_j \rangle) = P_k(x)$. In another special case, in which t_j and t_k both represent the same agent, we have the familiar (GC) equation we have been working with in our single-agent models, from which we can in turn derive the single-agent (LC).

And once we have interpersonal versions of (LC) and (GC), we can obtain interpersonal versions of the Reflection Principles we saw earlier. Under particular conditions, an ideally rational agent at t_j will, conditional on the supposition that an agent at t_k certain of everything she is currently certain of will assign a degree of belief to $\mathbf{r}$ to the claim represented by x, assign a degree of belief of $\mathbf{r}$ to the claim represented by x. (When I say "an agent at t_i will assign a degree of belief…", I mean that the agent represented by t_i will assign that degree of belief at the time represented by t_i.) The relevant conditions will be analogues of the first five numbered conditions specified in Section 2.3, plus a condition that the agent at t_j is certain that there exist principles of ideal rationality such that in a model representing those principles in its extrasystematic constraints, the set C_k mandates a credence for x.

This interpersonal version of the Reflection Principle is what Adam Elga (manuscript) has called an "expert principle." The agent at t_j views the agent at t_k as an "expert" — someone who is perfectly rational and whose certainty set contains hers — and the ideally rational response (under particular conditions) to suppositions about the degrees of belief of an expert is to defer to the expert.

Elga has also described a rational principle for responding to information about the degrees of belief held by a "guru" — someone who is ideally rational, has certainties you don't, but also lacks some certainties you have. Elga's "guru principle" is the Generalized Reflection Principle of Section 2.3 applied to the interpersonal case. Suppose the agent at t_j is certain that an agent at t_k is ideally rational and is certain of everything she is except for some F. Under particular conditions, an ideally rational agent at t_j will, conditional on the supposition that the agent at t_k will assign $P_k(x \mid \langle F \rangle) = \mathbf{r}$, assign a degree of belief of $\mathbf{r}$ to the claim represented by x. The relevant conditions are the analogues of the second five numbered conditions listed in Section 2.3, plus a condition that the agent at t_j is certain that there exist principles of ideal rationality such that in a model representing those principles in its extrasystematic constraints, the set $C_j \cup C_k$ mandates a credence for x.

Thus in situations in which the appropriate credences are mandated, (GC) reveals both relations between ideally rational intrapersonal degrees of belief and relations between ideally rational interpersonal degrees of belief. If the information in a particular certainty set is strong enough to demand a particular rational degree of belief for a claim, it doesn't matter who entertains that certainty set or when he or she entertains it — rationality will require certain particular mathematical patterns to arise.

3.4. Beyond Mandated Credences.

We have seen that in situations in which particular credences are mandated by particular certainty sets, verdicts generated by (LC) and (GC) are guaranteed to represent requirements of ideal rationality. Thus we can designate some situations as definitely within the domain of applicability of our (LC)-based modeling framework. But what about situations in which the relevant credences are not mandated? Are any such situations within the domain of applicability of our modeling framework?

We have been working under the assumption that our synchronic systematic constraints represent requirements of ideal rationality. Under that assumption, synchronically derived verdicts represent requirements of ideal rationality, so such verdicts will be reliable even in situations in which mandated credences are absent. For example, if we can synchronically derive a verdict of the form $0.7 < P(x) < 0.9$, that verdict will represent a requirement of ideally rationality even if no credence is mandated in our model for x. If a model yields only synchronically derived verdicts (for example, if the model contains only one time in its time sequence), we can be confident that those verdicts represent requirements of ideal rationality.

But clearly it is the diachronic verdicts, those derived from (LC) and (GC), that we are most worried about in situations lacking the relevant mandated credences. Take our Chocolate story, for example. Let's refer to the time just after the game has been explained to you as t_0, and let's represent the claim that you will receive a piece of chocolate as c. Now suppose that no matter how many principles of ideal rationality we discover, there will never be a model representing requirements of ideal rationality in which C_0 mandates a credence for c. That is, suppose that there is no specific degree of belief that ideal rationality requires you to assign to the prospect of receiving chocolate.

Now consider a time t_1, shortly after t_0, at which you learn that the coin came up heads (the outcome consistent with the possibility of chocolate). A model that incorporates the Principal Principle in its extrasystematic constraints will restrict $P_0(c)$ to the interval $(0, 1/2)$, but will leave the entire $(0, 1)$ interval available for $P_1(c)$. However, if (LC) is among the model's systematic constraints, it will yield a verdict that $P_1(c) = 2 \cdot P_0(c)$. The question is whether this verdict represents a requirement of ideal rationality.

It seems intuitively that learning the coin came up heads should require you to double your degree of belief that you will receive chocolate, whatever that degree of belief might previously have been. But why is this? We are working under the assumption that no principles of ideal rationality mandate a credence for c at either t_0 or t_1; at each time the most ideal rationality can require of you is that your degree of belief fall within the prescribed range. Under this assumption, ideal rationality considers it perfectly consistent with what your certainty set at t_0 for you to assign a degree of belief of 0.4 at that time to the prospect of receiving chocoloate. Similarly, it is rationally consistent with your certainty set at t_1 for you to assign a degree of belief of 0.3 at that time to the prospect of receiving chocolate. So why should ideal rationality forbid you from assigning $P_0(c) = 0.4$ *and* $P_1(c) = 0.3$?

We might draw an analogy here to an interpersonal case. Suppose the Chocolate game is described to both Jen and Ken. Let's suppose that initially, Jen and Ken have identical information about the world, but then it is secretly revealed to Ken that the coin came up heads. Under the assumption that no credences are mandated for c in this situation, I think it's clear that neither Jen nor Ken is violating a requirement of ideal rationality if Jen assigns a credence of 0.4 to c and Ken assigns a credence of 0.3. So why should the situation be any different when these two credences are held by the same person at different moments in time? If ideal rationality requires diachronic consistency among degrees of belief even when neither of those degrees of belief is strictly mandated by the agent's certainties at a particular moment — that is, if ideal rationality requires future degrees of belief to respect current degree of belief assignments even when those current assignments are admittedly "judgment calls" that go beyond what ideal rationality requires — where might such a requirement come from?

One answer might be that the act of assigning specific degrees of belief at a particular time is a doxastic action involving a set of commitments on the agent's part. Among those are commitments to set future degrees of belief in particular ways should you receive various sets of information. For example, an ideally rational agent in the Chocolate story might come to assign $P_0(c) = 0.3$ by assigning $P_0(h) = 0.5$ and $P_0(c \,|\, h) = 0.6$. Assigning that particular conditional degree of belief might involve a commitment to assign a degree of belief of 0.6 to the prospect of chocolate should the agent become certain that the coin has landed heads. In general, we might read the verdicts yielded by (LC) as expressing the diachronic commitments an agent makes when she assigns various degrees of belief. And if such diachronic commitments can give rise to requirements of ideal rationality even when the degrees of belief in question are not mandated by the relevant certainty set, we will be able to rely on (LC) to generate verdicts representing requirements of ideal rationality.

3.5. **Commitments and Forgetting.** If degree of belief assignments do involve commitments yielding requirements of ideal rationality, that will expand the domain of applicability of our modeling framework well beyond the bounds of credence-mandating situations. Certainly our (LC)-based framework will apply to stories in which the agent never forgets any information, even if those stories do not involve the relevant mandated credences. But how do doxastic commitments interact with forgetting episodes?

I think that ultimately, that question should be answered by a substantive theory of doxastic commitments, and this is hardly the place to develop such a theory. So the rest of this section will be fairly speculative. I will suggest some questions such a theory might attempt to answer, offer intuitive guesses about what the right answers might be, and then outline how our (LC)-based modeling framework could yield verdicts in line with those answers.

Let's continue working under the assumption that no principles of ideal rationality mandate a credence in c for you in the Chocolate story. Suppose I describe the Chocolate game to you today at t_0, and you assign a degree of belief of 0.3 to the claim that you will receive some chocolate. At t_1 I tell you that the coin has come up heads, and you adjust your degree of belief that you will receive chocolate to 0.6. Now suppose that my final decision about whether to give you the chocolate isn't to be made until tomorrow, so you turn in for the night. When you wake up tomorrow morning, at t_2, you have forgotten how the coin flip came out. What does ideal rationality require of your degree of belief at t_2 that you will receive some chocolate?[12]

Intuitively, it strikes me that the answer here should be 0.3. That intuition is backed up by a strong parallel to the Shangri La case in which you travel the Path by the Sea. In that case, you start at t_0 with an initial degree of belief of 0.5

[12] To make matters simple, let's imagine that you knew at t_0 that I would reveal the outcome of the coin flip to you at t_1 no matter how the coin flip came out. That way your knowledge at t_2 that you were told the flip's outcome at t_1 does not provide you with any evidence in favor of a particular outcome.

that the coin comes up tails. You receive evidence that the coin has come up tails, which boosts your t_1 degree of belief up to 1, but then you forget that evidence. So at t_2 when you reach Shangri La, ideal rationality requires your degree of belief in tails to revert to what it was at t_0. If we'd like to make the analogy stronger, we could imagine that at t_0 the guardians tell you not that they will flip a fair coin to determine your path, but that they will flip a biased coin whose degree and direction of bias they refuse to reveal. Your t_0 degree of belief in tails would then not be mandated by your t_0 certainty set, but whatever judgment you made about the likelihood of tails, ideal rationality would require you to revert to that judgment at t_2. Similarly in the Chocolate story, whatever degree of belief in the prospect of chocolate you settle on at t_0, ideal rationality requires you to revert to that degree belief at t_2 when you have forgotten the results of the flip.

This suggests that doxastic commitments made at an earlier time can rationally bind an agent at a later time even when that agent has both gained and lost information between the two times. Clearly not all doxastic commitments hold across a forgetting episode: In the Chocolate story, when you believed at t_1 that the coin had come up heads, that belief presumably carried with it a commitment to go on believing in heads unless you received evidence otherwise. Forgetting the coin outcome between t_1 and t_2 leaves you in violation of that commitment, and yet we don't judge you irrational at t_2 for having forgotten.[13] So it seems that if you lose information from an earlier time to a later time, you can break doxastic commitments involved in your beliefs at that earlier time without violating the requirements of ideal rationality.

On the other hand, if at a later time you retain all the certainties you had at an earlier time, ideal rationality seems to require you to honor those earlier doxastic commitments, even if you have gained and then lost some certainties in the interim. This is what establishes the relationship between your ideally rational t_0 and t_2 degrees of belief. And this, of course, is what (LC) requires: if your certainty set at an earlier time is a subset of your certainty set at a later time, your unconditional degrees of belief at the later time are required to reflect your conditional degrees of belief at the earlier time. Roughly speaking, we might say that (LC) requires you to honor your doxastic commitments from all times you fully remember. Thus the requirements of ideal rationality that seem intuitive for our extended Chocolate story match the verdicts that would be generated for this story by our (LC)-based modeling framework. The domain of applicability of that framework seems to include some stories that involve forgetting but do not mandate the credences relevant for (LC).

Yet matters can become more complex. Suppose that when you wake up tomorrow morning, at t_2, you have forgotten not only the coin outcome revealed to you at t_1 but also what precise degree of belief you assigned to the prospect of chocolate

[13] I'm putting this point as if there was a doxastic commitment in place at t_1 to go on believing in heads and that doxastic commitment was violated at t_2, but the violation of the commitment did not result in a failure to meet the requirements of ideal rationality. We might equally well put the point another way, saying that the doxastic commitment at t_1 contains an *exception* for forgetting episodes, such that you meet the requirements of ideal rationality at t_2 because you haven't violated any doxastic commitments. Again, I think that choosing the correct interpretation of this case will depend on a substantive theory of doxastic commitments, but as long as the varying interpretations ultimately agree in their verdicts about which agents meet the requirements of ideal rationality, the choice of an interpretation won't affect the modeling results I'm after here.

at t_0. If your original $P_0(c) = 0.3$ assignment was arbitrary from the point of view of ideal rationality, and by t_2 you have forgotten that you made that particular assignment, would you violate a requirement of ideal rationality by assigning, say, $P_2(c) = 0.2$?

My inclination is to say no, for two reasons. First, it strikes me as difficult to make out a coherent position about doxastic commitments on which you are not required at t_2 to honor a t_1 degree of belief in chocolate that you don't remember but are required to honor a t_0 degree of belief you don't remember. Second, when working with rationality constraints we usually require that an agent be able to work out at any given moment what ideal rationality requires of her. (This is tied closely to the notion that rationality constraints are standards of *internal* consistency.) If ideal rationality requires you at t_2 to maintain your t_0 degree of belief that you will receive chocolate, but you can't remember what that t_0 degree of belief was, no matter how hard you try you won't be able to do what ideal rationality requires.[14]

Judging that you may assign $P_2(c) = 0.2$ in this case without violating any requirements of ideal rationality does not immediately put the case outside the domain of applicability of our (LC)-based modeling framework. After all, we might think there was something you were certain of at t_0 that you are now not certain of at t_2: the claim that you assign $P_0(c) = 0.3$. If we enrich our modeling language to include sentences representing claims about what degrees of belief you assign at what times, and we assume that whenever you assign a particular degree of belief you are required to be certain at the time that you do so, then C_0 will not be a subset of C_2 and (LC) will not require $P_0(c) = P_2(c)$.

Now suppose that at t_3 you retain all your certainties from t_2 but also suddenly recall that you assigned a degree of belief of 0.3 to the prospect of chocolate at t_0. In setting your $P_3(c)$ value you face an interesting challenge. You retain all the evidence relevant to setting a degree of belief in chocolate that led you to your judgments at both t_0 and t_2, but you are now aware that at those two times that evidence led your reasoning to different degree of belief conclusions. And under our current assumptions, neither of those conclusions violates a requirement of ideal rationality. Is there a particular way ideal rationality requires you to set your degree of belief in chocolate given this information?

We might get some help here from our working hypothesis that (LC) represents requirements of ideal rationality generated by doxastic commitments even in cases that involve forgetting. C_2 is a subset of C_3, so (LC) can relate your degrees of belief at those two times. Between t_2 and t_3, you gain the certainty that you previously assigned $P_0(c) = 0.3$. So by (LC), $P_3(c)$ should equal your degree of belief in c at t_2 conditional on the supposition that you assigned c a degree of belief of 0.3 at t_0. The question now becomes: If at a given time you assign a rational unconditional degree of belief to a claim, what should your degree of belief in that claim be conditional on the supposition that at another time you assigned the claim a *different* rational unconditional degree of belief?

We now have an intrapersonal analogue to the interpersonal question that drives Elga's paper (manuscript). The main concern of that paper is how a rational agent ought to respond to the information that another rational agent with identical relevant evidence assigns a different unconditional degree of belief to a particular claim. Given our discussion in Section 3.3, it should be clear that this is possible

[14] I am grateful to Alan Hájek for suggesting this second reason to me.

only in a situation in which there are no principles of ideal rationality on which that evidence set mandates a particular credence for that claim (otherwise at least one of the agents is not ideally rational). But supposing the relevant credences are not mandated, this interpersonal case is precisely analogous to the challenge we have described for you at t_2; it's just that the rational agent whose differing unconditional degrees of belief you must confront is an earlier version of yourself.

3.6. **Conclusion.** As I suggested at its outset, the last section was highly speculative, and it became more speculative as it went on. First we supposed that in stories in which the relevant mandated credences for (LC) are not in place, there might still be doxastic commitments in play that could generate diachronic requirements of ideal rationality. Then we began to make conjectures about how such commitments and requirements might interact with forgetting episodes. To represent some of the conjectured requirements using (LC), we imagined adding sentences representing second-order claims about one's own degrees of belief to our modeling language. This was already a worrisome maneuver, since we have not examined the consequences of adding sentences representing such second-order claims into our modeling framework. But then we made the further controversial assumption that if an agent assigns credence $\mathbf{r}$ to claim x, it is a violation of ideal rationality for her to assign a degree of belief less than 1 to the claim that she assigns $\mathbf{r}$ to x. Clearly we have ventured rather far afield in our attempt to outline how an (LC)-based modeling framework might yield verdicts in line with our speculations about the rational force of doxastic commitments in various forgetting cases.

Instead of venturing even farther, then, I want to close by providing a sort of map of our (LC)-based framework's domain of applicability. With this map in hand, those who want to defend various substantive positions about the strength of rational requirements and about doxastic commitments will be able to tell how those positions rate our framework as a modeling tool.

First, we have assumed since Section 3.2 that taken together our synchronic systematic constraints represent requirements of ideal rationality. The following discussion is confined to the domain of stories over which that assumption holds true.

Within that domain, we'll start with situations in which credences for the sentences in question are mandated by the relevant certainty sets (C_k for (LC) and $C_j \cup C_k$ for (GC)). These situations will fall within the domain of applicability of our modeling framework, as proven in Sections 3.2 and 3.3 above. This is true even if one holds a strong theory of doxastic commitments on which any agent who forgets anything violates requirements of ideal rationality. In some cases such agents will fail to meet requirements of ideal rationality without contravening any of the verdicts of an (LC)-based model, but this is acceptable, as those verdicts were intended all along to represent necessary but not sufficient conditions for ideal rationality.

On an extremely strong view of the requirements of ideal rationality, every consistent certainty set mandates a credence for every sentence in the modeling language. On a view like this, every story will lie in the domain of applicability of our modeling framework. If the requirements of ideal rationality are taken to be somewhat weaker, there is a question about whether situations lacking the relevant mandated credences fall within the domain of applicability of our framework.

If there are no requirements of ideal rationality based on diachronic doxastic commitments or something like them, situations in which (LC) generates verdicts but the relevant credences are not mandated will lie outside the domain of applicability of our framework, as those verdicts will not reflect requirements of ideal rationality. If there are doxastic commitments of the kind I have described, stories with no forgetting will lie within the domain of applicability of our framework even if the relevant credences are not mandated.

Moreover, if an agent is required to honor doxastic commitments from just those earlier times that she fully remembers, it may be that all stories fall within the domain of applicability of our (LC)-based modeling framework. One serious challenge here will come from stories in which an agent forgets what degrees of belief she assigned at an earlier time and then remembers those assignments later. To model these stories appropriately using (LC), we may need to add sentences representing second-order belief claims to our modeling language and argue for substantive rational requirements on an agent's awareness of her own degrees of belief. Even then, there will be some difficult questions about how an agent should assign her degrees of belief in such stories, analogous to questions about how a rational agent who learns that another rational agent has drawn different degrees of belief from identical evidence should respond to such information.

Michael Titelbaum
University of California, Berkeley
webfiles.berkeley.edu/titelbaum

References

Arntzenius, Frank. 2003. Some Problems for Conditionalization and Reflection. *Journal of Philosophy* 100: 356-370.

Elga, Adam. Manuscript. Reflection and Disagreement.

Hájek, Alan. 2003. What Conditional Probability Could Not Be. *Synthese* 137: 273-323.

Harman, Gilbert. 1986. *Change in View.* Cambridge, Ma: The MIT Press.

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.

Schervish, M.J., Seidenfeld, T., and Kadane, J.B. 2004. Stopping to Reflect. *Journal of Philosophy* 101: 315-322.

van Fraassen, Bas C. 1995. Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies* 77:7-37.

Weisberg, Jonathan. Manuscript. Conditionalization, Reflection, and Self-Knowledge. Forthcoming in *Philosophical Studies.*