Temporally Continuous Probability Kinematics

by

Kevin Blackwell

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2020

Doctoral Committee

Professor James Joyce, Chair
Professor Gordon Belot
Professor Sarah Moss
Professor Leopoldo Pando Zayas

Kevin Blackwell

kevblack@umich.edu

ORCID iD: 0000-0001-7605-7220

DEDICATION

In loving memory of my Grandpa Cope, who shared his lifelong passion for learning with me when I was young and impressionable.

ACKNOWLEDGEMENTS

This dissertation has greatly benefited from helpful questions, comments, and suggestions from many sources. I am not sure that I can list everyone whose help has improved this dissertation in some way, but I attempt here to identify some of the most impactful contributions of others.

Thank you to Kenny Easwaran: my hearing you mention the possibility of neighborhood norms of rationality at FEW 2015 was the genesis of this entire project.

I have shared drafts of several parts of this dissertation with friends and colleagues at the University of Michigan – including during Candidacy Seminar, a meeting of the Graduate Student Working Group, my brown bag talk, and the Epistemology Work in Progress group. Thank you to anyone who participated in any of these events and gave me feedback.

At one point during my research on this project, Alex Meehan and I became concerned that we might have been working on very similar ideas – I am thankful both that that turned out not to be the case, and for your feedback on my project!

Thanks to Catrin Campbell-Moore for catching an absolutely awful mistake in the proof of what is here presented as Chapter II; thank you, also, for helping me try to figure out how to fix it!

Special thanks are due to the members of my committee. Thank you to Gordon Belot for both pressing me on and helping me to better present the mathematical formalism of Continuous Probability Kinematics. Thank you to Sarah Moss for many helpful questions and pushback on the epistemological assumptions and beliefs underpinning this project. Thank you to Leo Pando Zayas for agreeing to serve on the committee on such short notice. And, finally, special thanks to my chair, advisor, and mentor Jim Joyce for helpful comments on and conversation about an absurdly large number of drafts of this project at every stage of development. Your advice and feedback have been immeasurable sources of help.

On a personal note: thanks to all of the family and friends whose love and support have been sources of encouragement and motivation, through times both dark and bright.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

ABSTRACT

The heart of my dissertation project is the proposal of a new updating rule for responding to learning experiences consisting of continuous streams of evidence. I suggest characterizing this kind of learning experience as a continuous stream of stipulated credal derivatives, and show that Continuous Probability Kinematics is the uniquely coherent response to such a stream which satisfies a continuous analogue of Rigidity – the core property of both Bayesian and Jeffrey conditionalization.

In the first chapter, I define *neighborhood norms* of rationality with reference to Kenny Easwaran's definition of *neighborhood properties*. I summarize and comment on some of the key arguments in the dispute between time-slice epistemologists, who argue that there are no fundamentally diachronic norms of rationality, and the proponents of diachronic norms. I am sympathetic to two of the key motivations often given in support of the synchronist position: mentalist internalism and the idea that metaphysical disputes about the identity of persons in bizarre puzzle cases should not play a central role in epistemologists' assessments of the rationality of agents. However, I argue that time-slice epistemology cannot adequately address the rationality of temporally-extended processes like reasoning and learning. Neighborhood norms present a viable third way between these two positions, capturing much of the spirit of the previously-discussed synchronist motivations while still providing *just enough* temporal structure to meaningfully guide and evaluate temporally-extended rational processes. Continuous Probability Kinematics is an example of one such neighborhood norm.

In the second chapter, I develop my updating rule CPK and establish many of its core properties. Of special note here are the deep connections to Jeffrey's Probability Kinematics, as well as some key differences. The net result of any CPK updating process will always be representable as a Jeffrey shift on the refined partition generated by the propositions that the agent is receiving direct evidence concerning. However, one crucial difference is that CPK

provides an intuitive account of how to *combine* the effects of learning experiences that are each about fundamentally different underlying partitions. In CPK's formalism, an agent can receive simultaneous evidence streams about an arbitrary (finite) number of propositions, which can themselves be evidentially related in any way. At any given instant, the result of the combination is a simple sum of the effects that learning about the individual propositions would have separately.

CPK is concerned with a novel kind of learning experience and involves a novel characterization of evidence. The third and final chapter of this dissertation is concerned with explaining what this characterization of evidence means and with arguing that it can be the basis for genuine learning. I begin by characterizing learning experiences in terms of the Value of Information, and prove a Value of Information theorem for CPK learning experiences under the assumption of a Martingale constraint on the agent's prior distribution over the signals that they might receive. I examine Timothy Williamson's arguments that evidence must be propositional and express my skepticism. I then explore two different routes to model agents who update by CPK *as if* they are learning some propositional content and updating the rest of their credences by Bayesian conditionalization on this content. The second of these two routes provides a very interesting lens to reexamine the evidential commitments that underwrite updating by CPK, which I analyze.

CHAPTER 1

Neighborhood Norms of Rationality: A Third Way Between the Synchronic and the Diachronic

1. Time-Slice Epistemology vs. Diachronic Norms: an Overview of the Criticism

In this section, I begin by introducing the time-slice position, and presenting some of the major synchronist arguments against diachronic norms and for synchronicity. I briefly evaluate how persuasive I find these arguments in attacking the viability of diachronic norms.

The Time-Slice Position

Many common norms of epistemology are *diachronic*: according to them, what an agent should believe at one time depends on facts about the agent at other times. For example, Bayesian conditionalization is usually understood as a diachronic norm. Suppose that at t_0 an agent has a prior credence function c_0, and then at t_1 the agent learns that some proposition *E* is true. The credence function that Bayesian updating requires the agent to adopt at t_1 is a function of the agent's prior, c_0.

Roughly speaking then, the time-slice (or synchronist) position is simply the denial that agents are beholden to diachronic norms. Here is Moss (2015)'s statement of time-slice epistemology:

> "… at a first pass, we define this theory as the combination of two claims. The first claim: what is rationally permissible or obligatory for you at some time is entirely determined by what mental states you are in at that time. This supervenience claim governs facts about the rationality of your actions, as well as the rationality of your full beliefs and your degreed belief states. The second

> claim: the fundamental facts about rationality are exhausted by these temporally local facts. There may be some fact about whether you are a rational person, for instance. But this is a derivative fact, one that just depends on whether your actions and opinions are rational for you at those times." (172)

Similarly, Hedden (2015) puts the position like this: "how you rationally ought to be at a time depends only on your mental states at that time, not on how you (or time-slices psychologically continuous with you) were in the past or will be in the future" (7).

Although these statements of the view seem simple enough, there is a subtlety related to what counts as a mental state. Williamson (2000, chapter 1) famously and controversially argues that knowledge is not only a genuine mental state, but a more paradigmatic kind of mental state than belief. So, consider a Williamsonian view where what your doxastic state at any moment in time should be depends on your current evidence, and your current evidence consists of all and only the propositions you know. Now consider the following two possible worlds. World 1 contains Agent 1 that, at time t, possesses a veridical memory of eating a bagel for breakfast that morning. In World 2, there's some time-slice of the universe, occurring at time t', that's physically identical to the state of World 1 at t; however, the supposed memory of Agent 2 (who is, at t', physically identical to Agent 1 at t) in World 2 is nothing of the kind – the agent didn't even have breakfast that morning. (This could be for all kinds of reasons: the agent didn't exist prior to time t'; the agent's memory has been overwritten by some device; etc.) Very plausibly, the agent in World 1 knows that she had a bagel for breakfast that morning. (In fact, let's just stipulate that Agent 1 does know this. If there is any possible precisification of the case as outlined so far where the reader thinks it's determinate that Agent 1 knows she had a bagel for breakfast, feel free to fill in the details.) The agent in World 2 *definitively* doesn't know that she ate a bagel that morning, because she didn't. A Williamsonian view places different demands on the two agents: Agent 1 must adopt the doxastic state consistent with some total body of evidence including the fact that she ate a bagel, while Agent 2 must not.[1] Is such a view really consistent with the time-slice picture?

[1] I've been deliberately ignoring the following wrinkle: if content externalism is correct, even if Agent 2's memory of eating a bagel were veridical, the obvious propositions that Agent 2 would be in a position to know would not be the same as the propositions that Agent 1 knows, because their concepts would have different referents. I'm

Hedden (2015) seems to think it is: “I have no quarrel with these [Williamsonian] epistemologists, and indeed I am sympathetic to their views. … Mentalist Internalists should say that whether your perception is reliable, and whether your memory is veridical, should affect what you ought to believe now only in virtue of affecting your present mental states. This is compatible with the claim that in fact you and the BIV, or you and Swampman,[2] are not in the same mental states after all” (26). Kelly (2016) argues that classifying Williamsonian views as synchronic is suspect:

> “On the view in question, the fact that you are now in a position to justifiably believe this proposition is ultimately grounded in a set of facts that includes purely historical facts, for example, facts that a certain past learning event actually occurred. It is at the very least unclear that such a view should be classified as a current time slice account as opposed to an historical theory. Contrast the view just described with a different account of memory-based justification. According to this alternative account, you are justified in believing that *p* on the basis of memory when (1) you have a *current apparent memory as of p*, a state that provides prima facie justification for believing *p*, and (2) you currently lack any reason to distrust this apparent memory. Such a view is *clearly* a current time slice view, in a way that the epistemological view described in the preceding paragraph is not.” (47-8)

There are two separate criticisms of classifying Williamsonian views as synchronic that this contrast makes apparent. As we will see in the next subsection, one of the most natural ways of motivating the synchronist view is by appeal to internalist intuitions. However, agents in the skeptical scenarios discussed in the previous paragraphs are in a state that is *internally*

not going to try to make this precise, but I think it’s obvious that there’s an important epistemic difference beyond this mere difference in content: roughly, the proposition about eating a bagel for breakfast that Agent 2 believes is structurally analogous (in some ways) to the proposition that agent 1 believes about eating a bagel for breakfast. Agent 1 has, and agent 2 lacks, a certain kind of knowledge about her own history; Agent 2 could have had this *kind* of knowledge, even if the propositions in question would be different.

[2] Both the brain in a vat and Swampman are skeptical scenarios somewhat like the one I’ve sketched. However, you and a BIV that have the same (apparent) experiences would not be in time-slices of the universe that are physically identical. A Swampman (“created when a lightning bolt causes a bunch of molecules to spontaneously arrange themselves into a human form” (25), with the same apparent memories as you) case *could* be an instance of my scenario, but it’s underspecified.

indistinguishable from the case where their beliefs constitute knowledge. I join Hedden in finding this claim very plausible: “what you rationally ought to do or believe should depend on what information you have available, rather than simply on how the world in fact is” (11). On the understanding of availability of information that I find most intuitive, agents in the skeptical and good cases, respectively, have access to the same (or at least, structurally analogous – see fn. 1) information. And the supposed difference in mental states that the Williamsonian view claims to obtain is fundamentally grounded in an aspect of “how the world is” that the agent does not have access to.

Leaving internalist motivations aside, the stronger criticism Kelly is making is this: if whether an agent knows *E* at *t* irreducibly depends on facts about the agent’s history prior to *t*, then having whether the agent’s belief that *E* counts as knowledge determine whether or not *E* is evidence for the agent amounts to making the agent’s doxastic norms irreducibly depend on aspects of the agent’s history that are not really encoded in the agent’s present state. Indeed, the decisive facts are not even encoded in the present physical state of the entire universe. Here’s an obviously mistaken way of arguing that (ordinarily understood) Bayesian conditionalization is a synchronic norm. Let P_r be the family of propositions of the form: that, at t_0, the agent’s conditional credence in *H* on *E*, $c_0(H|E) = r, r \in \mathbb{R}$. Here’s a synchronic version of Bayesian conditionalization: $(\forall\, r \in \mathbb{R})$ if, at the present moment the agent has just learned *E,* then the agent is required to have credence r in *H* iff P_r is true at the present moment. Now, the P_r are not even arguably part of an agent’s present mental state, so the time-slice positions we have been discussing (Moss’s and Hedden’s) both correctly reject this supposedly synchronic norm. But if the property of knowing *E* at *t* is grounded partly in non-mental facts about the agent at earlier times, it’s not clear to me that it should be any less troubling for the synchronist.

Synchronist Arguments Against Diachronic Norms

In this subsection, I present some of the major arguments that proponents of the time-slice position deploy against the validity of diachronic norms of rationality.

Considerations from Personal Identity

Hedden (2015, Chapters 2-3) presents several puzzle cases for diachronic norms that involve ambiguity about whether one time slice is the same person as another time slice. We will look at some of these cases in more detail in a moment, but the general thrust of these arguments is to attack diachronic norms, like Conditionalization, that are *intrapersonal*. As it's ordinarily understood, Conditionalization treats time slices of a single agent differently than it treats time slices of different agents. When you learn some piece of evidence *E*, the prior conditional credences $c_0(x|E)$ that Conditionalization instructs you to adopt as your current credences, $c_1(x) = c_0(x|E)$, are usually understood as *your* prior conditional credences. Your prior, c_0, is treated as relevant to what your current doxastic state should be in a way that c'_0, the credence of some other agent at t_0 is not. And if there are multiple time slices that have competing claims to being you at t_0, then applying Conditionalization requires you (at t_1) to adjudicate their claims. As we will see, this can sometimes be quite tricky. I now consider two of these puzzle cases that Hedden discusses.

The Combined Spectrum

Parfit (1984, 236) asks us to consider a spectrum of medical procedures that he might undergo. On one extreme, no operation occurs – Parfit exits the operating room unchanged. On the other, Parfit's entire body is replaced with an exact cellular replica of the body of Greta Garbo as she was at 30 years old. In all of the intermediate cases, some number of Parfit's cells are replaced with Garbo cells, and some are left in place. It's supposed to be obvious that, on the first extreme, the person that exits the operating room is identical to Parfit; and on the second extreme, the person that results is certainly not identical to Parfit. In many of the intermediate cases, it is supposed to be quite tricky to identify whether the result of the operation is Parfit or not.

Double Teletransportation

> "One person (call her 'Pre') enters the teletransporter. Her body is scanned. Then, at the instant her body is vaporized, the information about her molecular state is beamed to two locations, Los Angeles and San Francisco. In each city, a molecule-for-molecule duplicate of Pre is created. Call the one in Los Angeles 'Lefty' and the one in San Francisco 'Righty'. Lefty and Righty are each qualitatively just like Pre is before her body is vaporized." (Hedden 2015, 16)

Hedden elaborates on Double Transportation, crafting a more explicit challenge to Conditionalization. Now, suppose that before entering the teletransporter, Pre has the following beliefs. When Lefty and Righty awake to take their first views of the world, they will each see the décor of the mad scientist responsible for the teleportation procedure. For each of several prominent medical schools, Pre's credence that the scientist is an alumnus of that school conditional on the scientist's home featuring the colors of said school is very high. Lefty sees crimson; should she have high credence that the scientist graduated from Harvard? (32)

Hedden's answer: "If Conditionalization is right, then it depends. If Lefty and Pre are the same person, then Conditionalization says that Lefty indeed ought to have high credence that the mad scientist is a Harvardian. But if not, Conditionalization is silent, for it is as if Lefty just suddenly came into existence" (2015, 32). Hedden goes on to claim that, according to Subjective Bayesianism, if Lefty is not the same person as Pre, Lefty is within her epistemic rights to choose some prior – that need not be related to Pre's final credence function before death in any important way – and then update accordingly. Although Hedden focuses on Conditionalization, the argument should generalize to most diachronic norms that are intrapersonal in the sense of taking the agent's prior doxastic states to play a special role in prescribing or justifying the agent's current doxastic state.

Identity: A Quick Reply on Behalf of the Conditionalizer

My goal in this paper is not to fully defend diachronic norms as ordinarily understood. Later, I am going to introduce a new kind of norm, neighborhood norms, that have some of the advantages of both the diachronist and time-slice positions. However, I think the arguments from personal identity are not entirely fair to the proponent of diachronic norms.

Conditionalization is, fundamentally, a norm about learning experiences. Although there is some controversy about whether there are kinds of learning that Conditionalization (including Jeffrey updating) is not well-equipped to handle, it is fairly *uncontroversial* that Conditionalization is really not intended to model what an agent should do when they take themselves to be undergoing epistemic misfortune: losing evidence, forgetting, having beliefs changed in ways they don't endorse, etc. Parfit's Combined Spectrum case is not just an example of Cronenberg-worthy body horror but is also epistemically horrific. In the intermediate cases, it seems unlikely that the resulting chimera will even have anything resembling ordinary human thought – assuming it even lives. But if it does, it would be frankly astonishing if the resulting credal state wasn't wildly incoherent; it is the result, after all, of mashing together the minds of two different agents. Although much less disturbing than the typical Combined Spectrum case (at worst, it only involves a near-instant death), *the act of being teleported* in Double Teletransportation is similar in not being a learning experience of any kind. (Both Lefty and Righty undergo learning experiences *after waking up*, but ordinary conditionalization can handle that without any reference to Pre at all.) In fact, it's stipulated that Lefty and Righty are qualitatively identical to Pre. Although this is, perhaps, insufficient grounds to establish that Lefty and Righty have the same mental states as Pre[3], I am going to assert that *learning* without internal physical change is impossible. This claim should be unobjectionable to anyone who is not a fairly strong dualist.

So, I claim that the proponent of Conditionalization should respond that her norm is silent about the teleporting part of Double Teletransportation, *whether Pre is identical to Lefty or not*; Conditionalization should also rest mute on the entire sequence of the Combined Spectrum: the agent should not regard any of the operations as learning experiences. Once Lefty awakes with the credence function that she in fact has after the teleportation, she undergoes an ordinary learning experience of seeing crimson; Conditionalization then applies

[3] E.g., If Williamsonian views about what can count as a mental state are admissible, two agents with the same physical state may have different mental states.

straightforwardly, using her actual prior. (The same is obviously also true of Righty.) And the diachronist need not be embarrassed that her diachronic norms only apply in certain situations; she never committed to the claim that there are only *universal* diachronic norms. The (very strong) position that the synchronist is arguing for is that there are no fundamental diachronic norms at all.[4] Although Hedden may be correct that Double Teletransportation is a puzzle for a certain variety of Subjective Bayesianism, the problem is not really that Conditionalization is diachronic. The real mistake is assuming that (modulo choice of prior, which is permissive), conditionalization must identify what credence function it's rational for the agent to have at every moment in time, in response to every possible kind of situation. But there's just no reason to expect conditionalization to be applicable to arbitrary mental changes; I think that understanding of Subjective Bayesianism was doomed from the beginning, and it's not clear that any Bayesian has ever held such a position.

Arguments from Mentalist Internalism

The second major criticism of diachronic norms is that they fail to make what it would be rational for an agent to believe at a time supervene on an agent's current mental states. Hedden (2015) calls this supervenience mentalist internalism. A similar point is also made in Moss (2015): "The problematic cases for diachronic norms are exactly those cases where your past opinions do not have their usual effects on your current mental states. … Instead of restricting diachronic norms to cases where your past credences have their usual effects on your current mental states, we should admit that your current mental states are what determine whether your current credences are rational" (176). Moss is also explicit that her view is intended to be neutral on the stance of epistemic externalism vs internalism (179). Both Moss and Hedden give Arntzenius's Two Paths to Shangri La and ordinary forgetting as examples intended to demonstrate that Conditionalization violates this supervenience. Both Moss and Hedden are also clear that this is not merely a problem for Conditionalization, but an instance of a broader problem: any diachronic norm will sometimes make demands of an agent

[4] The word "fundamental" was important – Hedden, for instance, does admit that there may be diachronic norms that are derivative of the fundamental synchronic norms. See Section 2.

that seem at odds with what is intuitively rational from (in Hedden's terminology) a mentalist internalist perspective.

Two Paths to Shangri La

> "There are two paths to Shangri La, the Path by the Mountains, and the Path by the Sea. A fair coin will be tossed by the guardians to determine which path you will take: if heads you go by the Mountains, if tails you go by the Sea. If you go by the Mountains, nothing strange will happen: while traveling you will see the glorious Mountains, and even after you enter Shangri La, you will forever retain your memories of that Magnificent Journey. If you go by the Sea, you will revel in the Beauty of the Misty Ocean. But, just as you enter Shangri La, your memory of this Beauteous Journey will be erased and be replaced by [an apparent] memory of the Journey by the Mountains." (Arntzenius 2003 356)

Suppose you go by the Mountains. Intuitively, you should be very close to certain that you're taking the Path by the Mountains while you're actually on it: you see the Mountains, and you have no reason to think that anything has gone wrong with your perception at this point. Once you enter the city, your memory of traveling by the Mountains remains veridical – but you are now in a position where you have no reason to believe this is true. You know that you would have qualitatively indistinguishable "memories" as of the same journey if you had taken the Path by the Sea. Your memory has lost its evidential import because it now fails to discriminate at all between the two cases. It seems that the best you can do is base your current credence in the path you took on the fact that the coin was fair: there was chance $1/_2$ you would take each path, and you now have no other evidence (from your perspective) about which path you took. "Note the internalist intuition here: that what you ought to believe depends on what your evidence is, and your evidence supervenes on your present mental states, which are the same no matter which route you took" (Hedden 2015 36).

Conditionalization is not well-equipped to deal with this case. My diagnosis is that this is because Rigidity, the core property of conditionalization, is a bad fit for the kind of learning experience that this case sets up. Rigidly learning about a proposition *E* means changing your credences in such a way that your credences conditional on *E*, $c(\cdot\,|E)$, remain constant. There is an important sense in which this amounts to maintaining your beliefs about what kind of evidence *E* is for various propositions.[5] This is a case where the most natural understanding of what the agent has learned is that the evidential import of his evidence *has changed*: prior to entering the city, his memory is very strong evidence about which of the two Paths he took – his credence in having gone by the Mountains, conditional on his memory of seeing the mountains should be close to 1; after entering, his evidence does not discriminate between the two Paths at all – his credence on having gone by the Mountains, conditional on a memory with precisely the same propositional content, should now be $1/_2$. Rigidity is desirable precisely when an agent is learning about some evidence in a way that *preserves* her beliefs about the import of said evidence, which is not an especially natural way of understanding what has happened here.

Hedden diagnoses the problem slightly differently: "But upon entering Shangri-La, you do not gain any new evidence that bears on whether you traveled by the Mountains, and hence Conditionalization does not kick in. So, according to Conditionalization, you ought to just retain your credence 1 that you traveled by the Mountains. The problem is that you do not learn anything new that is evidentially relevant to the question of which route you took" (2015 36). However, Hedden's own proposed solution to the problem is actually inconsistent with this claim that passing through the gate is not evidence relevant to which path you took. Hedden argues that his norm of Synchronic Conditionalization explains the intuitively correct response to entering Shangri La better than diachronic conditionalization. Here's the norm:

> **Synchronic Conditionalization** Let *P* be the uniquely rational prior probability function. If at time *t* you have total evidence *E*, your credence at *t* in each proposition *H* should equal $P(H|E)$. (Hedden 2015 138)

[5] For more on the sense in which Rigidity involves maintaining your evidential commitments, see Chapter II of this dissertation.

And here’s the explanation: “The thought is that which route you took was determined by the result of a coin toss. And your current evidence that you seem to remember traveling by the Mountains does not discriminate between your having traveled by the Mountains and your having traveled by the Sea. So your credence that you traveled by the Mountains ought to equal $^1/_2$” (Hedden 2015 141). Now, I very strongly agree with Hedden that this is the intuitive story about why your credences should change when you enter Shangri La. But I think it’s very unclear as an explanation of how Synchronic Conditionalization is supposed to resolve the problem.

Hedden can reasonably claim that the total evidence the agent has before and after entering the gate of the city are different in virtue of facts like where or when the memory was had, and so we can’t easily represent the agent undergoing the kind of cumulative learning required by ordinary diachronic conditionalization. But for Hedden’s account to work, the uniquely rational credence function does still have to treat the proposition “I have, at some point in the past, crossed through the gate” as evidence against having taken the path by the Mountains in cases where that’s the only thing that agent learns; the fact that the agent also plausibly has other changes in their total evidence only obscures that fact.

To see this, consider the case of a visitor to Shangri-La who has lost track of where they are and what time it is – it may help to imagine that they have lost the use of their sight. Consider two versions of this agent: one who remembers that they have passed through the gate (call this proposition G) with certainty, and another who has credence 0.5 in G. Their total body of evidence is in all other respects the same; in particular, *both agents have qualitatively identical memories as of the same journey by the Mountains.* Now, what should each of these agents conclude about the proposition M: “I took the path by the mountains”? Very clearly, for the same reasons we have been discussing, the first agent should have (nearly) credence 0.5 in M. If we think that the uniquely rational prior should satisfy Reflection, then the second agent is required to have credence 0.75 in M – they think there’s probability 0.5 that they’re in a situation where the required credence is 1 and probability 0.5 that the required credence is 0.5, so the required credence is $0.5 \cdot 1 + 0.5 \cdot 0.5 = 0.75$. But even without committing to Reflection, it seems clear that that agent is rationally required to have a credence greater than

0.5 in the second case. But this is to say that the uniquely rational credence function treats learning G as evidence against M for the second agent. Let E be the second agent's total evidence. We have $P(M|E \wedge G) < P(M|E)$.

Ordinary Forgetting

> "Suppose you are now certain that you had cereal for breakfast. At some point in the future, you will no longer remember having had cereal today, but since you will not have learned anything new that bears on what you had for breakfast today, Conditionalization says that you ought to retain your certainty that you had cereal. But this is crazy! Surely once you no longer remember having eaten cereal, you ought to drop your confidence that you had cereal" (Hedden 2015 42).

Hedden also mentions what I take to be the correct response on behalf of the Conditionalizer: Conditionalization tells you what to do in certain cases of gaining evidence; it makes no recommendation about what to do when your evidence remains the same or when you lose evidence, because that is simply not the task for which it was built. However, Hedden argues that this response is a cop-out: the truly fundamental norms of rationality *would* address all such cases (2015 43). Moss (2015) makes a point in the same vein: "From the point of view of theory building, the repeated restriction of diachronic norms is unsatisfying. … Time-slice epistemology is a natural response to this pattern of observations" (175-6). For my part, I find it unclear why we would expect global, universally applicable rational norms. There are many different kinds of learning situations, and still more kinds of belief change that are epistemically undesirable (including at least some kinds of forgetting); these cases don't seem especially easy to unify. Counting a failure to apply to all such kinds of situations as a black mark against a norm that performs admirably in a well-defined area of applicability is strange to me; this is especially true given that the cases Bayesian updating handles are some of the most paradigmatic kinds of learning experiences. And I am highly skeptical that there is any unified synchronic norm which actually handles all of the cases well. However, I completely agree with Hedden and Moss that, were such a unified synchronic norm to exist, that would provide excellent reason for us to think that it might be more fundamental than a patchwork quilt sewn together from pieces of various diachronic norms. I also agree that a version of Conditionalization which demanded that agents maintain beliefs on the basis of evidence they no longer have seems implausible. However, I think it's unfair to try to saddle the proponents of Conditionalization with this implausible view on the grounds that it would be a more generalized version of their view; to generalize to outside of a domain of applicability is a mistake. Make everything as simple as possible – but not more so!

So, I'm unconvinced by the objection from ordinary forgetting, and I'm skeptical that synchronic versions of conditionalization provide substantial improvement over diachronic conditionalization in cases like Shangri-La. But regardless of my stance on the specific examples, I am very sympathetic to the underlying mentalist internalist motivation: diachronic norms, by making reference to an agent's actual prior mental states, certainly have the possibility of conflicting with norms that are instead based on the agent's, e.g., beliefs about their prior mental states.[6] And when it comes to action guidance, I agree that it seems preferable to base norms in states that agents have better access to; as a limiting case, if an agent could not even in principle access a state at all, it's irrelevant to what the agent should do.

Similarly, although I'm not convinced by the specific examples of problems with personal identity that I've discussed, I completely agree with time-slice epistemologists that our epistemic evaluations should not depend on the identity facts in these weird, hard puzzle cases. However, I think treating each time slice as an agent complete-in-themselves, independent of previous time slices, is a massive overcorrection; I think the time-slice view has especially weird consequences when we concern ourselves not with action-guidance, but with epistemic evaluation. In the next section, I look at one major diachronist criticism of the time-slice view, and the corresponding synchronist responses.

[6] Part of what is at issue here is *how diachronic* actual proponents of conditionalization take it to be. As I noted in the case of ordinary forgetting, Hedden and Moss create what they take to be a more general version of conditionalization than the restricted version that I take it most proponents of conditionalization actually advocate for.

2. Diachronist Criticism of the Time-Slice Position

The criticism of the time-slice position that I will consider in this section centers around the idea that there are *rational processes*, e.g., reasoning, that are evaluable in ways that the time slice picture cannot accommodate. The properties of a process, the diachronist argues, are irreducibly diachronic – they cannot be captured by aggregating the properties that an agent possesses at various times. Podgorski (2016) presents an analogy with Zeno's paradox of the arrow: the property of moving does not obtain at any instant. At each moment in time, the arrow occupies a single position. Although the positions occupied at each moment are different, at no moment is the arrow moving. And so, Zeno concludes, if the arrow is motionless at each moment, the arrow never moves. Similarly, the synchronist claims that rationality for an agent at each moment in time is a property that supervenes on (mental) properties of the agent at said moment. And since what it is rational for an agent to believe at each moment is determined purely synchronically, they synchronist concludes that rationality supervenes purely on the synchronic. Diachronic norms, if they exist at all, must be reducible to fundamental synchronic norms. Podgorski argues that this is the same kind of error as Zeno's: rationality-at-an-instant is not the full story of rationality. There are some rationally evaluable properties that, like motion, are properties that a subject possess in virtue of behavior over an interval and cannot be reduced to the properties that the subject has at any instant. Purely synchronic norms simply lack the resources to address these properties: you can satisfy synchronic norms at each instant without exhibiting the required pattern over the interval. Thus, Podgorksi argues, if there are rational requirements that are fundamentally about *processes* like reasoning, the time-slice view will not be adequate (Podgorski 2016 862-3).

Rationality of Belief Formation

Suppose that Podgorksi has a friend, Minnie, who delights in breaking promises. In most situations, receiving a promise from Minnie is excellent evidence that she will not do what was promised. However, Minnie is very superstitious and seriously attempts to keep all promises she makes on the 13th of each month. On the 13th of some month, Minnie promises Podgorksi that she will attend his birthday party; he knows that it's the 13th when he hears the promise. Call the time at which Podgorski receives the promise t_0. Now, we are assuming that processing this evidence takes time: the soonest that he will be able to form an opinion about whether Minnie is coming to his party or not is some time t_1 after t_0. Unfortunately, at t_1, Podgorksi will suddenly forget that the day's date is the 13th, and have no idea what the date is. At present, he has no inkling this will happen. What belief about Minnie's presence at his birthday party should he form at t_1? (Podgorski 2016 867)

The synchronist, Podgorski claims (and as we will see, Hedden agrees), will say that the belief Podgorksi should have at t_1 is the one supported by his evidence at that very instant: he should believe that she's very unlikely to come to his party. Without a particular belief that today is the 13th, Podgorski should think it probably isn't the 13th; after all, the average number of days in a month is 30.42, so a random day is quite unlikely to be the 13th.[7] And on any day other than the 13th, Minnie's promise is very strong evidence that she won't be coming. The problem is that for Podgorski's process of belief formation to result in a belief at t_1 that Minnie (very probably) won't be coming to his party, the process would have to be insensitive to the evidence that he has *during the actual process*. Throughout the entire time that Podgorksi is reasoning about whether Minnie will come to his party, he knows that it's the 13th; he forgets what day it is only at the moment he forms the belief. And so the conclusion justified by his evidence during the *process of forming the belief* is that Minnie will very likely come to his party. A process that takes that evidence as input and yields as output that Minnie won't come to the party is clearly a defective one; we should not want to reason this way.

Hedden (2016) presents two responses to this argument. The first is more tailored to Podgorski's specific case, whereas the second is a very general claim about the relationship between the time-slice position and supposed epistemic norms governing processes. Hedden's first response is to argue that Podgorski's analysis of the case must be incorrect, because it implies that it's sometimes rational to be in an incoherent doxastic state. If Podgorski is correct,

[7] This is admittedly a huge oversimplification. Depending on the month, which Podgorski presumably knows, we can fill in the more precise probability of 1 in 28, 1 in 30, or 1 in 31. Also, it may be much more realistic that Podgorski forgets what day it is, but has high confidence of being in some interval consisting of a couple weeks, or some similar arrangement. The details of this probability don't really matter; the crucial point is just that Podgorski ends up in a situation where his evidence makes it less likely that the date is the 13th than that it isn't.

then at t_1 he should believe that Minnie is coming to his party, that Minnie promised today to come to his party, and that the day's date is probably not the 13th. But conditional on Minnie promising today to come to his party, and today not being the 13th, Podgorksi is supposed to believe that Minnie will almost certainly break the promise and not come to the party. So it seems as if he is committed to believing both that Minnie will and won't attend the party (Hedden 2016 877). Now, Hedden points out that Podgorksi could reasonably claim that this is a case where there are conflicting rational norms: there's a synchronic norm that requires agents not to have incoherent beliefs, and there's a diachronic norm that requires that agents form beliefs using the evidence present during the formation process. In the Minnie case, it's impossible for Podgorski to satisfy both norms – but maybe rational norms just conflict sometimes, and this is not evidence for either norm being incorrect. Hedden asserts, without argument, that "judging a case to be one in which there is a genuine conflict between requirements of rationality is a last resort" and that it is preferable to simply reject the supposed diachronic norm (Hedden 2016 879).

Although I share Hedden's intuition that a putative normative dilemma should usually be interpreted as evidence that your normative framework needs revision,[8] I find this response puzzling. Although it is technically true that it's impossible for an agent to satisfy *both* norms in the Minnie case, the two norms do not share the burden of that impossibility equally. In cases like the Minnie case, it is *impossible to consistently or deliberately satisfy the synchronic norm by itself*; it is comparatively very easy to satisfy the diachronic norm. At t_1, Podgorski forgets what day it is. As both authors present the case, this happens *without any warning*[9] – there is nothing Podgorksi can do to avoid this happening, and no reason for him to make any prior preparations for his other beliefs at that moment. Essentially, we are not thinking of the forgetting as a belief change that is attributable to Podgorski as an agent, but as an arational change imposed from without. There is, in general, no way to safeguard the coherence of your beliefs against this kind of change. If your beliefs are coherent before such an arational change, they will typically be incoherent after it. Even if you knew that *some* arational change were

[8] I will also join Hedden in not trying to provide any argument supporting this intuition in the present discussion.

[9] Although Podgorski also considers a variant of the case where he knows beforehand that this will happen.

coming, there would be no way of pre-emptively causing future coherence without evidence about what beliefs would be most likely to change and how. It's not *literally* impossible to satisfy: you might get very lucky, and the arational change may happen to result in a coherent doxastic state. But not only is this very unlikely, there is no strategy that is expectedly better than doing nothing; the norm is also impossible to satisfy deliberately. In effect, applying a synchronic norm for coherence to this kind of case treats the agent as exhibiting a failure of rationality *in virtue of arational processes.* In this particular case, it treats forgetting as a rational mistake, which is the exact criticism that Hedden levies against diachronic norms, as discussed in the previous section. If this kind of forgetting is, at least sometimes, not under our voluntary control and not foreseeable, then treating this as a failure of rationality is incompatible with the internalist intuitions that I believe Hedden and I share.

Suppose you have two premises, A and B. An argument by *reductio* starting with A and B shows that your premises are jointly inconsistent; it does not provide any evidence about which is false in the actual world. Now, suppose you have evidence that A is very likely false. This is still not evidence that B is true[10] – or even non-contradictory – but it *does* show that rejecting B is very unlikely to help make any set of propositions that includes A true; choosing to reject B while holding that A is true is completely unmotivated. For all you know, B may or may not be problematic, but you should be very confident that A is. The same structure of reasoning seems applicable to the case of normative dilemmas. Trying to solve the dilemma by rejecting the diachronic norm is unmotivated when the synchronic norm is impossible to satisfy in any consistent or deliberate way.

Hedden's second reply to Podgorski revolves around distinguishing between fundamental and derivative norms. He claims that the only fundamental norms of rationality are synchronic (e.g., at each moment, your current beliefs should be proportioned to your present evidence).

> "If we were perfectly rational, we wouldn't need to engage in reasoning in order to satisfy the requirements of rationality and have beliefs which are

[10] Although it might be evidence that B is true if you have some independent reason to think that the disjunction of A and B is likely to be true.

> proportioned to our evidence. We reason precisely because we fall short of perfect rationality. Reasoning is a tool we can use to get ourselves to come closer to satisfying the requirements of rationality. In this way, its value is contingent and instrumental – contingent because it stems from our contingent cognitive limitations, and instrumental because reasoning serves as a means to the end of having beliefs proportioned to one's evidence" (Hedden 2016 882).

To the extent that we should be concerned with norms governing processes at all, he claims, we should see such norms as merely derivative; in particular, patterns of reasoning are epistemically good or bad precisely to the extent that employing them tends to lead to satisfying the fundamental norms well or poorly. In this way, norms of reasoning are of a kind with norms about how often to nap, how much caffeine to ingest, or how long to brainstorm (assuming we're evaluating these actions in terms of their efficacy in producing the mental states required by the fundamental synchronic norms) (Hedden 2016 883).

Hedden thinks that an ideally rational agent would not need to reason. She would instead adopt the synchronically prescribed, uniquely correct, doxastic state instantaneously upon receiving new evidence; but he admits that because of our cognitive limitations it is at least sometimes impossible for real agents, like us, to do this. He claims we have an *excuse* and so are not blameworthy for our failure to live up to these very stringent epistemic norms. And he argues that it is a virtue of his account that it applies a single unified norm to a huge variety of different cognitive beings. On the alternative picture, where we tailor the norms to the limitations of each agent (so that we only demand of each agent that they satisfy the most stringent norms it is possible for them to satisfy), we would end up with disparate disunified norms: different epistemic norms not only for every species, but for most individuals (2016 881).

I think there's something intuitively compelling about the idea that you should incorporate new evidence into your beliefs as soon as possible. Any delay is time spent with beliefs that you think are less expectedly accurate and less practically useful than the beliefs you will have once you've finished responding to the evidence. And I can see claiming that you should accomplish it instantaneously as a useful kind of shorthand that elides all of the messy factors that determine what the minimum standard that *actually* applies to you should be. Hedden presents the picture as the most stringent, limiting case applying to everyone. But then he satisfies "ought implies can" by establishing derivative standards of blameworthiness, which *are* sensitive to the particular cognitive limitations of each individual. I'm unsure there's any real difference between this position and the position which claims that the only real standards binding on each individual are the blameworthiness standards; it sounds to me like a verbal dispute.[11] But I see the differences in the permissible delay in incorporating evidence as stemming entirely from differences in cognitive architecture: how many parallel processes is each system capable of executing at once, how many operations can each processor perform a second, what are the memory/storage constraints, etc. And the reason that I see these facts as relevant is that it makes a difference to which algorithms can be implemented most efficiently on the different pieces of hardware, and to how much real time running different programs will take. But the idea that the ideal case would be to achieve the correct belief state without any reasoning amounts to the claim that the limiting case is to get the correct result without any algorithm at all. This seems obviously false to me.

Let's consider three different systems that are each supposed to be Bayesian agents. All three "agents" initially have the same prior credence function *c*, which is defined on some finite algebra. Call the atoms of the algebra $A_1, \dots, A_n$; as usual, the algebra is closed under negation and disjunction. Each agent is given the same series of inputs: propositions that they are supposed to learn with certainty, adopting credence 1. Each input is always some element of the algebra. Let's call the first agent **Siri**. **Siri** updates using an algorithm that looks something like this:

Siri's Algorithm

1. Read the input, determine which element of the algebra is specified.
2. Initialize a counter variable *i* to 1.

[11] This isn't intended to be a general claim that the distinction between "wrong, but not blameworthy" and "not wrong" is typically a verbal dispute. But when the stricter standard is one that's impossible for any cognitively realistic agent to satisfy, *even in principle*, it leads me to think of that standard as useful primarily as a more abstract way of representing the real norms. This is, ultimately, what I will claim about my own norm of Continuous Probability Kinematics, developed in the next chapter.

3. Calculate $c(A_i \wedge E)/c(E)$ where *E* is the input from step 1. Write the result to $c(A_i)$.[12]
4. Increment *i*.
5. Repeat steps 3 and 4 until $i = n$.

Siri is kind of slow and also has the disadvantage that she is incoherent in the middle of performing the update. One obvious way to improve on **Siri** is **Parr**. Instead of running a loop that iterates over all of the A_i, **Parr** has *n* modules running in parallel, each of which is dedicated to updating **Parr's** credence in one of the A_i. So rather than performing *n* iterations of step 3, **Parr** performs all of the instances of step 3 at the same time. **Parr** also saves computational cycles by not needing to increment or check the counter – **Parr** doesn't need a loop. Of course, to run his algorithm, **Parr** needs the ability to perform more simultaneous operation and needs more memory locations than **Siri** does. But if **Parr** has that available, his algorithm is much faster. Also, because his credences in each of the atoms of the algebra are updated at the same time, **Parr** is not incoherent at any time during the updating process.

Finally, consider a third system: **Stan**. There is no algorithm that describes Stan's mental states – he doesn't perform any analogue of the kinds of steps that **Siri** and **Parr** do. He doesn't calculate any conditional probabilities, and there are no systems in him that look like they're copying values from one memory location to another. Yet, somehow, every time that the same evidence presented to **Siri** and **Parr** is given to Stan, the values of each of his credences instantly jump to the same values that **Siri** and **Parr** will eventually arrive at after their toils. Now, I claim, **Stan** doesn't appear to be some paradigm of rationality, vastly superior in his rationality to **Siri** and **Parr**. **Stan** doesn't even appear to be any kind of agent! There is no causal story about why **Stan**'s credences end up taking the values they do (in stark contrast to both **Siri** and **Parr**).[13] It's not even clear what calling the inputs that Stan receives "evidence" means: evidence plays a certain functional role, and there's nothing in **Stan** that shows any kind of comprehension, or any kind of use of the inputs he receives. He is just a collection of

[12] I'm ignoring the issue about what to do when $c(E) = 0$. This is another way in which **Siri**'s algorithm could be improved!

[13] To be completely explicit: *there isn't even any causal story that explains why he assigns* credence 1 to *E*. Although *E* is put into his input box, there is no process in **Stan** that says to assign credence 1 to whatever proposition is given to input.

disconnected states which somehow have a magical correlation with the inputs that he's being fed. Yet, at each instant, Stan's "beliefs" are perfectly proportioned to his "evidence." We can even stipulate that he satisfies Hedden's Synchronic Conditionalization at each instant. But I claim the idea that **Stan** is even *assessable* with respect to rationality is highly implausible – let alone the claim that he is rationally superior to either **Siri** or **Parr**. If we encountered a system like Stan, that showed a clear pattern of states that looked like conditionalization, but with absolutely no evidence of doing any kind of internal processing, the obvious conclusion would be that ***Stan*** *wasn't an agent, but a result of the calculations of some other system*. Stan could look to us like a collection of snapshots of something that might be an agent, but we would have no reason to think he was one. And if we couldn't find any connection from some other system to Stan, then Stan would be deeply causally mysterious.

Although the above cases may not be fully decisive, I think they strongly suggest that Hedden's view of synchronic rationality as primary, with the rationality of processes being merely derivative, is precisely backwards. As I will argue for a bit more in the subsection *Why I Believe Relation R is What Matters to Rationality* of the next section, playing certain roles in certain kinds of processes is foundational to what belief is. Merely satisfying a bunch of synchronic constraints, without the right kind of causal connections between those states, is not even sufficient to guarantee that the states in question are beliefs, let alone *rational* beliefs. Our paradigms of rationality should be developed from systems that exhibit certain kinds of connections between states *over time*. At this point, attentive readers might be confused. What I have been saying in this subsection might well sound like a full-throated endorsement of the necessity of diachronic norms, which I explicitly said wasn't what I was going to do. In the next section, I will argue that there may be room for a kind of norm which is neither quite synchronic nor diachronic, but somewhat blurs the line between the two. I will argue that this kind of norm can provide the kind of structure necessary for rational processes, through constraints that are, in a sense that I will make sense of soon, *minimally diachronic*.

III. A Third Way: Neighborhood Norms

In "Why Physics Uses Second Derivatives", Kenny Easwaran defines the notion of a "neighbourhood property"[14] as follows:

> "A (two-sided) neighbourhood property at t is a property of an object that is not grounded in the fundamental properties of the object at t, but, for every interval $(t - \Delta, t + \Delta)$, the fundamental properties of the object across that interval are sufficient to ground it" (847).

Easwaran is concerned with explaining how so-called "instantaneous" velocity in classical physics can play the causal role that it does, given that (at first pass) it seems that some of the facts that are supposed to causally depend on instantaneous velocity (positions at nearby future times) seem to be part of the causal ground of the instantaneous velocity. Although this is certainly not the concern of this paper, instantaneous velocity is an excellent example to introduce the concept of neighborhood properties.

For simplicity, let's focus on the case of motion in a straight line: we will choose our coordinates so that the *x*-axis is along this line, with the initial direction of motion being the *positive* direction; the origin is set to be the initial position of the object with time 0 being the start (of our consideration) of the motion, so that $x_0 = 0$. Now, consider some interval $[0, \tau]$ on which we know the *x*-position at each moment in time: that is, we know the position as a function of time $x(t)$ for $0 \leq t \leq \tau$. We can use our knowledge of the position to calculate various *average velocities*. For any subinterval we like, say with temporal endpoints *a* and *b*, we can calculate $\bar{v}_{ab} = \frac{x(b)-x(a)}{b-a}$. This average velocity is very useful: if we knew the average velocity and the length of the interval, but not the displacement of the object, we could calculate the change in position: $\Delta x_{ab} = (b - a)\bar{v}_{ab}$. Similarly, with the average velocity and the change in position, we can calculate how long that interval was. As the preceding discussion hints at, these average velocities are very firmly properties of the *interval*. They are determined by the

[14] I will use the spelling "neighborhood", except in citations of this paper.

conjunction of two facts: the *displacement* (change in position) of the object over the interval and the *duration* of the interval. Setting a particular average velocity on the interval places no constraints on the position of the object at any particular moment in time; for any time $a \leq t \leq b$, *any* x_t *whatsoever* is consistent with any stipulated value for the average velocity $\bar{v}_{ab}$. Even if we set the initial position, say $a = 0$, then the average velocity still only constrains the final position (e.g., now we have $x_b = (b-a)\bar{v}_{ab}$); for any moment in time strictly after *a* and strictly before *b*, $a < t < b$, it's still possible for the object to have *any* position. But nonetheless, it *is* a strict constraint on the endpoints.

Consider a series of intervals that each have duration Δt centered around time *t*. If the average velocities of each of these intervals converge to some value as the intervals approach length zero, we can call this limiting value the instantaneous velocity at *t*, $v(t) = \frac{dx}{dt} = \lim_{\Delta t \to 0} \frac{x\left(t+\frac{\Delta t}{2}\right)-x\left(t-\frac{\Delta t}{2}\right)}{\Delta t}$. The mathematics that talking this way enables us to do is so useful that doing classical physics without appealing to it seems unthinkable.

But notice that this quantity is conceptually kind of strange. First of all, as we've been discussing, average velocity is a kind of measure of how much an object has moved over a certain interval. As is famously abused by Zeno, motion cannot be a property of an instant: at any particular instant, the object has a single position $x(t)$. Motion consists in having a sequence of different positions *at different times*. So, despite the name and notation, this quantity of "instantaneous velocity" cannot be a property of an instant. Unlike with average velocity, specifying the initial and final positions on any finite interval (a, b) centered around time *t* is *neither necessary nor sufficient to determine* $v(t)$*.* It's not necessary, because any stipulated value of $v(t)$ is consistent with *any pair of initial and final positions* $x(a)$ *and* $x(b)$. It's not sufficient, because any value of $v(t)$ is consistent with any stipulated positions $x(a)$ and $x(b)$; in fact, it's consistent with stipulated initial and final positions on an infinite number of intervals centered around *t* – so long as the intervals have some finite minimum duration. What determines the instantaneous velocity is the displacement of the object over *infinitely many intervals* that are *arbitrarily short in duration*. Put slightly differently: the *entire* *positional history* of the object *over any single interval* containing t, no matter how small, is sufficient to fix $v(t)$; this makes $v(t)$ a neighborhood property.[15] It is a property that is underdetermined by the truly instantaneous properties of the object and overdetermined by the properties that the object has on any finite interval.

So far, I have been considering the dispute between synchronic and diachronic norms of rationality. As I have discussed, one of the essential claims of the synchronic view is that norms of rationality must be expressed in terms that refer only to properties that obtain *at an instant.* By contrast, diachronic norms judge whether agents exhibit certain properties over *extended intervals* of finite duration. But we have just made salient that while these two views of the permissible forms of rational norms are *contraries*, they are not *contradictories* – there is a third possible kind of rational norm, the neighborhood norm. Neighborhood norms can thread the needle: they can avoid some of the challenges that synchronists raise to true diachronic norms, while still being able to capture much of the spirit of diachronic norms and avoiding some of what I take to be the problems with the synchronist position. I will discuss a subset of neighborhood norms that seem to naturally complement a certain view about the existence/persistence conditions of rational agents, and argue that this way of understanding the persistence of rational agents defuses the identity puzzle cases presented earlier in the chapter. Finally, I compare and contrast Parfit's views about the identity of persons over time with my own views.

A *neighborhood norm* is a norm that an agent satisfies, at some instant, by having a certain conjunction of neighborhood properties and instantaneous properties; crucially, to count as a neighborhood norm, the norm cannot place constraints or depend on the properties that the agent exhibits at any other definite time,[16] and so

[15] This is good enough for our purposes, but only the starting point for Easwaran. He goes on to define past and future neighborhood properties, and argues that we should treat velocity as a past neighborhood property, in order for it to play the correct causal role. I will not be dealing with any of the subtleties of causation that arise from the differences between two-sided, past, and future neighborhood properties.

[16] It will, of course, often be true that satisfying a particular neighborhood norm *over an interval* will place constraints on the properties that the agent has at various definite times; the point here is that whether the agent satisfies the norm *at any given instant* does not. I'll elaborate on this a bit below.

cannot be a standard diachronic norm.[17] The most obvious class of neighborhood norms (at least to me) are those that an agent satisfies by conforming to certain "instantaneous" rates of change – although note that the value must also be specified in a way that depends only on the neighborhood and instantaneous properties of the agent at that instant.[18] We can continue to use velocity as our toy example: instantaneous speed limits are a nice example of a neighborhood norm. So, consider the requirement that, at time t_1, the agent must not be traveling at greater than 10 m/s. This norm places no definite restrictions on the positions that the agent is allowed to occupy at any moment in time: that is, for any time t, having any position at t is consistent with satisfying this norm at t_1. Any average velocity, on any particular interval is also permissible. But it is not empty; there are many patterns of motion that this norm rules out. For instance, any constant velocity greater than 10 m/s will violate this norm.

It is also very interesting to see what happens when an agent satisfies this norm not merely at an instant, but over an interval. So, for each $t \in [0, \tau]$, the agent does not have an instantaneous velocity at t exceeding 10 m/s. This continuous series of neighborhood norms puts very firm diachronic constraints on the agent's permissible motion. The average velocity of the agent on any subinterval of $[0, \tau]$ (including the entire interval itself) must be at most 10 m/s. The displacement of the agent over the interval can be at most 10τ meters. The displacement over any subinterval can be at most 10 m/s multiplied by the duration of the subinterval. This is the kind of constraint that we might ordinarily think of as diachronic (it creates a very sharp dependence between the initial and final positions of the agent), but it was arrived at by following a neighborhood norm at each moment in the interval.

What would a neighborhood norm of rationality look like? In *(Temporally) Continuous Probability Kinematics*, I develop a specific example of a neighborhood norm, which is also an example of the kind of neighborhood norm that operates by

[17] Should truly synchronic norms count as a special case of neighborhood norms, or should we add a clause to the definition to exclude them? I don't think anything important turns on which way choose to talk, as long as it's clear that neighborhood norms have the potential to make commitments that outstrip the purely synchronic while not being straightforwardly diachronic.

[18] What does it mean to "adopt" an instantaneous rate of change? Hold that question for a couple of paragraphs.

stipulating rates of change in the agent's credences.[19] Credal rates of change are defined in a very similar way to velocities: both average and "instantaneous". On some interval, the average rate of change for the agent's credence in proposition *x* is $\frac{c(x;\, b)-c(x;\, a)}{b-a}$, where $c(x; t)$ is the agent's credence in *x* at time t. Instantaneous rates of credal change are defined as a limit of these average rates, exactly like with velocity. In the kind of learning experience governed by CPK, at each moment the agent is receiving inputs from nature that require the agent's credences in some propositions should be increasing or decreasing at certain rates.[20] From the pair of the agent's current conditional probabilities and these inputs from nature, CPK specifies a particular rate of change (which might be zero) for each of the agent's credences at that moment. This is not a diachronic norm: satisfying CPK at a given instant places *no requirements on what credences an agent should have at any specific time after the moment in question.* But, much like with the toy case of "instantaneous" speed limits, it does rule out various future evolutions of the credence function as impermissible. And, just like the speed limit case, what happens when an agent satisfies CPK over an interval is very interesting. CPK was constructed to satisfy a temporally continuous analogue of *Rigidity* – the property at the heart of both Bayesian and Jeffrey conditionalization. This results in obeying CPK over some interval being equivalent (in final outcome) to having performed a Jeffrey shift on a certain obviously relevant partition.[21] CPK generates a kind of

[19] As I said, this is the kind of neighborhood norm which makes the most sense to me. But making essential reference to temporal rates of change means that this kind of neighborhood norm seems most naturally suited to certain ways of doing formal epistemology: representing agents as having some kind of degreed mental states, like credences. Are there any interesting and plausible neighborhood norms that would be useful to philosophers who prefer to think in traditional terms about concepts like full belief or knowledge, or in other kinds of formal frameworks that don't make similar use of degreed quantities? I think this is a very interesting question, but not one that I will pursue here.

[20] That's the modeling assumption, anyway. See both of the next two chapters, but especially Chapter III, for much more discussion of how we might connect this way of modelling the agent's doxastic changes to other ways of representing the evidence the agent is acquiring.

[21] Suppose the agent is learning from nature about two propositions A and B, which might be arbitrarily related. Updating by CPK preserves the conditional probabilities on the refined partition $\{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\}$ throughout the duration of the learning experience. The final outcome is always equivalent to a Jeffrey shift on this partition. See the next chapter for details.

learning that looks like a diachronic process, in aggregate, from constraints that are much “thinner” than true diachronic constraints at each moment in time.

How can an agent satisfy such a norm? What does it mean to “adopt” an instantaneous rate of change? What decisions would an agent who decides to conform to an instantaneous rate of change have to make? At this point, I think it’s probably best if we drop the abstraction of truly temporally continuous belief change and talk about what an agent that could only update its credences on discrete timesteps should do if they wanted to approximately follow a credal rate neighborhood norm. At some instant, the norm requires that their right credal derivative be equal to some value. The discrete approximation of this is to make their average credal rate over the interval starting at the present moment and concluding after τ_{min} – the minimal timestep after which the agent is capable of adopting a new credence – equal to the stipulated value. That is, if the norm requires $\frac{dc(x;\,t)}{dt} = u$, the approximation is to make $\frac{c(x;\,t+\tau_{min})-c(x;\,t)}{\tau_{min}} = u$. Notice that this equation fixes a unique value of the credence the agent should have after τ_{min} has elapsed: $c(x;\ t+\tau_{min}) = c(x;\ t) + u \cdot \tau_{min}$. So, to approximate the neighborhood norm, for any agent that can only update discretely (which is, again, almost certainly all possible agents) amounts to obeying a (derived) diachronic norm on the smallest possible future-looking interval. Satisfying the neighborhood norm over an interval is approximated by satisfying a bunch of these tiny diachronic norms at every timestep of the interval (the number of timesteps in the interval is the duration of the interval divided by the minimal time τ_{min}).

Now at this point, the reader may be wondering: what was the point of developing this contrast between neighborhood norms and diachronic norms, then? If the way that any real agent will go about trying to comply with a neighborhood norm is to implement a bunch of diachronic norms, how can neighborhood norms really have any advantage over diachronic norms? The point is that, even in their discrete realization, the demands placed by these norms are *minimally* diachronic. To determine what credence you should have at the next possible timestep, you don’t need to consult some ancient credence function that may now be long-forgotten – all you need is your present credence function. And there’s also no room to suggest that this future time slice is a completely separate agent that some diachronic norm is illicitly and arbitrarily yoking together with your present self. What you are required to do at present, and in the time until that next slice, is to begin the mental operations (calculations/reasoning) that will result in you adopting the stipulated credence at the earliest possible time that can happen. The processes are the ties that bind. Thinking in terms of the abstraction of neighborhood norms allows us to avoid having to deal with the details of a given system’s minimal processing time – which, of course, has no fundamental normative significance, but is just an empirical fact about the system. It also allows us to think of multiple discrete systems as approximating the same neighborhood norm; and, e.g., to make judgments like that one is a better approximation than the other. If we were to focus on the derived diachronic norms, which will be different for most systems, much of the signal will get lost in the noise.

Directing the agent’s behavior as a series of neighborhood norms, while leading to strong diachronic patterns, places *absolutely minimal* demands on the agent’s ability to commit to future plans or remember past actions. They need to remember only actions that happened an arbitrarily short amount of time ago; they need to be able to influence only their most immediate next actions. In Section 1, while I argued against the specific problems from internalism that Hedden raised against ordinary conditionalization, I did confess sympathy with the problem in the abstract: diachronic norms, in being tied to what (mental) properties an agent had in the past, may impose requirements that an agent cannot rationalize with their current mental states. A neighborhood version of credal updating minimizes this problem by demanding only the smallest possible intervals of memory at any instant. However, successfully following the norm over some interval will still, in the aggregate, lead to behavior that looks thoroughly diachronic. Of course, even competent reasoners will sometimes make mistakes, and be unable to meet these demands. But I take it that some basic capacity to tie present mental states to future mental states and actions is at the very heart of what is to be an agent; this is the point that I'll begin arguing in the next couple of subsections.

(Credal Rate) Neighborhood Norms and Personal Identity

I have just argued that neighborhood norms should seem attractive to anyone who, like me, is sympathetic to the internalist view that it's better for rational norms to be expressed in terms of states that agents have a very high degree of access to: they make the access requirements as proximate as possible without collapsing into synchrony. However, it remains to see how neighborhood norms fare with the identity puzzle cases that synchronists also see as evidence of the defects of diachronic norms. Here, I will restrict my attention to the class of neighborhood norms that operate by stipulating credal rates of change: credal rate neighborhood norms, for short(er). One very interesting feature of such norms is that they assume that the credence function of the agent is a differentiable function of time at all points at which they apply. And to be a differentiable function of time at a point, the function must be continuous at that point. So, any solution to a series of credal rate neighborhood norms on some interval must consist of a temporally continuous function. Any agent that obeys this kind of neighborhood norm at all instants in some interval will exhibit a credal state that is a continuous function of time (on this interval).

I will argue that this kind of credal continuity implies an important kind of mental continuity, consistent with Parfit's Relation R. I will show that any agent who has a credence function that is a continuous function of time will satisfy *arbitrarily rigorous* standards of continuity – and thus, any agent that obeys credal rate neighborhood norms should count as a mentally continuous agent under even the strictest of standards.

In *Reasons and Persons,* Parfit argues that the logic of personal identity obfuscates what really matters about personal identity. Personal identity has indistinguishability and transitivity requirements that mean that, in cases like Double

Teletransportation,[22] it's implausible that either Righty or Lefty can be identical to Pre. Both seem to have equally good claims to being identical to Pre, but they are distinguishable (Lefty and Righty are in different positions, for instance), which means they can't both be identical to Pre, because transitivity would then mean that they would be identical to each other. Parfit argues that there is no further fact that could explicate why one is identical to Pre, while the other isn't. His solution is to deny that Pre persists as either Lefty or Righty, but he also thinks that this fact isn't very important; he claims that the way in which Pre ceases to exist is at least nearly as good as (and maybe better than) ordinary survival (261-64). Parfit claims that what matters is relation R: "R is psychological connectedness and/or psychological continuity, with the right kind of cause"; and for Parfit, any cause can be the right kind of cause (262). Both Lefty and Righty are highly psychologically continuous (and connected, for that matter) to Pre, and that is much more relevant to almost all concerns about rationality and morality than whether or not Pre survives. For Parfit, "strong psychological connectedness" obtains between two time-slices A and B if there are many direct psychological connections between the two: B has direct memories of many things that A did, shares many of A's intentions and desires, etc. (205-6). "Psychological continuity" consists in "overlapping chains of strong connection" (206). For B to be psychologically continuous with A, it suffices that there is some series of intermediaries, $C_1, \dots, C_n$, so that A is strongly connected to C_1, C_i is strongly connected to C_{i+1}, and C_n is strongly connected to B. I do not fully agree with Parfit about the (moral and self-interested) unimportance of ordinary survival (my strongest disagreement with him is about teleportation cases), but I find it very persuasive that Relation R, and not identity, forms the necessary condition for persistence of a rational agent.[23]

[22] The discussion I cite here is actually of a case he calls My Division, but they are similar enough that I'm confident Parfit's judgment about Double Teletransportation would be much the same.

[23] I'll say a bit more about this in a later subsection, but for now: I think that two separate persons (e.g., where I believe that A has died, and B is new individual person) can nonetheless be evaluable as a single agent. The clearest case is one where B begins existence with precisely A's mental state, goes on to perform the actions A was intending just before death, and so on. For the norms of rationality, it seems clear to me that it makes no difference whether B is identical to A or merely an R-successor.

Why I Believe Relation R is What Matters to Rationality

As I think of it, the defining characteristic of a rational agent is using information to plan for future contingencies. A rational agent is fundamentally a system that has plans about how to respond both to various new pieces of information it might receive, and about how it intends to use the information currently at its disposal to make choices when posed with various kinds of decision problems it might encounter. What we typically call beliefs or credences are states that play certain functional roles in processes of this kind. One constitutive role of belief is that it's used in certain ways by the agent's decision principles: to count as the kind of system that is a proper subject of prudential rationality, it must have decision principles that take as input both desire-like features (goals, objects of pursuit, some sense of value), and predictions about what kinds of outcomes are likely to result from various actions it might take. Any system that does not have the capacity to use a framework with roughly this structure is not the kind of thing that makes decisions in the sense that I'm concerned with; treating it as a target of norms governing rational actions is pointless, at best. Another constitutive role of belief is to be used in certain patterns of reasoning – this includes having some kind of "updating rule": some commitment to a rule that maps from the pairwise input of the agent's current doxastic state and some new piece of information to the output of a revised doxastic state. As I indicated towards the end of Section 2, I am deeply skeptical of updating rules that are not either implementable or at least capable of approximation by algorithm. Seeing the algorithm is what makes me believe that the agent is a rational system, making adjustments to its belief in accordance with its current commitments. The reasoning is what shows *how the current doxastic state is being used*, how the result depends on the inputs.

On my view of what rationality consists in, a certain amount of psychological continuity is a necessary precondition for rationally assessable behavior. If, e.g., the states that would play the functional role of being beliefs are constantly being exogenously changed to random values, the most natural thing to say is that these states *cannot be beliefs*. The system as a whole just isn't the kind of thing that allows for the kind of states that beliefs are. And, again, as I hinted at towards the end of Section 2, this is why I find the time-slice view of rationality so unsatisfactory: when you uncouple all of the time-slices, I cease to believe that the so-called "beliefs" are what they claim to be. Just as forming an intention that you believe your future self will have no reason to follow seems paradoxical, "beliefs" that are not constitutive of commitments about how your future self will respond to evidence and decision problems don't strike me as beliefs at all; as I understand it, those commitments are what believing involves.

But such a system need not comprise a single person: corporations are typically composed of persons and can easily be rational agents on my view, but *Citizens United* notwithstanding, they certainly are not themselves persons. Such a system need not contain *any* persons: many very simple animals are assessable as rational agents on my view – they make predictions, choose, and learn. And, probably more controversially, I think such systems can even transcend death. Consider the case of ordinary teleportation: A steps into a box in Ann Arbor, where she is scanned and a perfect molecular blueprint of the structure of her body is created. The body in the box is vaporized, which kills A. The blueprint, now stored online, is accessed a few seconds later in Tokyo to build an exact molecular copy of A's body, which we'll call B. A had been planning to travel (by teleporter) to Tokyo for business, and B now executes A's plans. Now, unlike Parfit, I do not think that A should regard this as anywhere near as good as ordinary survival; I think A has made a horrible mistake and thrown her life away. I believe, roughly, that my mind either is, or is an effect of, a certain pattern of neurological activity. My mind might be able to survive the total replacement of my brain in gradual stages. I am quite convinced[24] that my mind can survive certain kinds of unconsciousness (e.g., sleep) that exhibit certain neurological patterns, but there are certain other kinds of unconsciousness (e.g., brain death) that it seems likely my mind cannot survive, and still others (e.g., long-term comas) where I don't know whether I

[24] Usually – occasionally the thought keeps me up at night.

would survive or not.[25] I take everything I just said to be fairly controversial, but I think the claim that being vaporized utterly destroys the mind really shouldn't be. In any case, it also seems clear to me that the system jointly consisting of A and B (who are two different persons – A died just before B came into existence) are naturally considered as a single rational agent. A and B are connected in all of the right kinds of ways: B inherits A's beliefs and executes A's plans. For almost all purposes of rational evaluation, it makes the most sense to consider them as a single system. Here's one more example: consider a computer program intended to sort some data set. Suppose that the program can be interrupted, which will cause it to store a partially sorted set, and the program can later be resumed from this partially sorted state without issue. Someone starts running the program on one computer, stops, transfers the file to another computer, and runs a new instance of the program on that second machine. There are clearly two different computers in this case, but for evaluating the task of sorting the data, the relevant system is comprised of both of them. Now we can imagine it is somewhat ambiguous whether or not the second computer is the same as the first one (say, the second computer resulted from various hardware upgrades to the first). Our identity/persistence conditions for computers are absolutely irrelevant to whether we should treat it as a single system for the purposes of evaluating the task: even if we think the computer didn't survive, there was a persistent agent (consisting of a combination of two different computers) which worked as a unit to perform the task. What the computer should do at the later time to complete the process of sorting doesn't depend on whether it's the same computer as earlier – only on the continuity of the file and how the sorting procedure works.

Connectedness in the Time of Credences

How should we translate Parfit's conception of strong psychological connectedness to a framework where we are representing agents as being modelled by

[25] To be completely clear: I don't think you would survive brain death even if the brain could be "rebooted" fairly quickly after it occurred, and I'm unsure someone who wakes up after certain kinds of comas persisted as a single mind throughout that time.

degreed credences? What's the credal analogue to the persistence of a memory, or an intention? I suggest that the obvious answer is to think of each memory as encoding a bunch of credal values. The persistence of a memory should not, ordinarily, be thought of as requiring the precise maintenance of all of these values over an extended period of time. We all know that our memories change in their details as years go by: certain details become fuzzier and fuzzier, until they may eventually be completely forgotten, and what we believe about certain other details may even change – we may end up falsifying some aspects of our memories to various degrees, while still ordinarily counting the result as a version of the same memory. Sometimes, through repeatedly focusing on certain aspects of a memory, we even become more confident that an event transpired a certain way years later than we were within minutes of the experience the belief is based on! For strong psychological connectedness to obtain over some interval, it was already true on Parfit's view that we need not maintain all of our memories, etc – we just need a large number of direct connections, with precisely how many being left intentionally vague. But when we move to frameworks using credences, there now seems to be an obvious second degree of freedom: how similar do the credences that comprise, e.g., some memory need to be for it to count as a direct connection. So, I suggest that we think of psychological connectedness as a two-parameter *family* of conditions:

A is psychologically *n*-connected to B in proportion *m* if, of the propositions that A has credences in, it's true of at least a proportion *m* of them that the $|c_B - c_A| \leq n$.

For short, we say that A is psychologically (m, n)-connected to B. By varying m, we vary how much of A's credal state B has to preserve to count the two as psychologically connected. By varying n, we vary how similar B's credences have to be to count as being similar enough to be a direct connection. The strictest possible condition is $(m = 1, n = 0)$, which would require that B would perfectly inherit all of A's credences. The weakest possible condition is $(m = 0, n = 1)$, which counts literally any two (normalized) credence functions as connected.[26] These two extremes are obviously both useless. A condition of connectedness that counts all credence functions as connected is patently silly. But a credence function that demands perfect inheritance of credences makes any kind of meaningful mental life impossible – there can be no learning, no forgetting, no reasoning or belief acquisition of any kind. However, it seems plausible to me that there may be various contexts in which different members of the family may find useful niches; I don't think there's any unique standard that's most appropriate. Just as Parfit does, we can now use our family of conditions of psychological connectedness to generate a family of criteria for psychological continuity: B is psychologically (m, n)-continuous with A if there is some sequence of intermediaries $C_1, \ldots, C_n$, so that A is (m, n)-connected to C_1, C_i is (m, n)-connected to C_{i+1}, and C_n is (m, n)-connected to B.

For our purposes, it turns out that it doesn't matter which standards in this family are most plausible, so long as we can rule out $(1, 0)$. This is because we can show that any agent which follows a credal rate neighborhood norm on some interval is guaranteed to satisfy arbitrarily strict standards of psychological continuity on that interval. The agent will satisfy standard $(1, n)$, for any $n > 0$. The crucial point is that an agent who satisfies a credal rate neighborhood norm will have credences that are a continuous function of time. First, some notation: let X be the set of all propositions that A has any credence in. $c(x; t)$ is the credence in some proposition $x \in X$ that the system holds at time t.[27] Let $(0, \tau)$ be some interval on which the system is following a series of credal rate neighborhood norms. Because the credence function is continuous

[26] Technically, I suppose, we could weaken the condition even further by choosing larger values for *n*, which could then allow credence functions that contain credences greater than 1 or less than 0 to count as connected.

[27] What is the "system" I'm referring to? I'm assuming that there is some kind of causally unified system that has the credal states we're talking about. This assumption involves the claim that the states of the system at earlier times are causally relevant to the states at later times, and various assumptions to the effect that the system has certain basic capacities necessary for causally connecting the states that we're (somewhat prematurely) calling credences to outputs necessary for decision making and to the kind of inputs that learning might happen through. What we are deliberately *not* assuming is that the system has the right kind of continuity in these internal states for the system to count as a temporally extended agent in the sense discussed earlier in this section; whether the states we are calling "credences" will end up playing the functional role that makes them credences depends partly on whether the system exhibits this kind of psychological continuity or not.

on the interval, $\big(\forall t \in (0,\tau)\big)(\forall n > 0)(\exists \delta > 0)(\forall x)|c(x;t+\delta) - c(x;t)| \leq n$. For every time t in the interval, for an arbitrarily strict *n*, there is some later moment in time $t+\delta$, where all of system's credences at this time differ by at most *n* from the system's credences at t. So, we let the state of the system at $t+\delta$ be the first intermediary, C_1. By construction, A is $(1,n)$-connected to C_1. We now repeat this process: continuity guarantees us that there's an intermediary C_2, such that C_1 is $(1,n)$-connected to C_2, and so on. This chain of $(1,n)$-connectedness will cover the entire interval $(0,\tau)$, and so the system is psychologically $(1,n)$-continuous on $(0,\tau)$. Obeying a series of credal rate neighborhood norms on $(0,\tau)$ is sufficient to guarantee that $c(x;t)$ is a continuous function of time on $(0,\tau)$, and so obeying credal rate neighborhood norms guarantees $(1,n)$-continuity on $(0,\tau)$ – for any strictness of n. If we believe that what matters to the persistence of rational agents is not personal identity, but psychological continuity, then obeying credal rate neighborhood norms is a way in which a system can organize itself as a rational agent.

Psychological Continuity and Identity Puzzles

I've argued that systems that obey credal rate neighborhood norms will exhibit the kind of continuity in their mental states appropriate to being a temporally extended agent. I think it's worth taking a moment to note that this kind of continuity is *not* all there is to being an agent. As footnote 27 began discussing, only systems that have certain kinds of causal properties are even the kinds of systems that could have the kinds of states that might play mental roles. So, in essence, the continuity considerations we've just gone through are assuming that the states we're talking about are present in a system with the right kinds of capacities and causal interrelations to be the kind of system that might have mental states. Continuity is then just a further precondition on those states persisting in ways that allow them to fulfill the necessary functional roles to count as belief-like states. So, seeing that a system meets very strict standards of continuity in certain states is not itself a reason to think that the system is an agent. But, if we already believe that a system has the right kind of causal properties to be an agent, showing that mental continuity is preserved in strange cases like fission and fusion *is* evidence that there are uninterrupted rational processes that survive whatever the strange event is. Because these processes are uninterrupted, it makes sense to think of the causally and mentally unified system that is responsible for these processes as a persisting agent – even if we cannot resolve the logic of identity, or have independent reason to think that organisms or persons involved died. And as we will see, that the systems in question continue to have the relevant capacities and causal structure is not in question in the kinds of puzzle cases that we've been considering so far (except in Combined Spectrum, where in most of the intermediate cases, the most likely outcomes seem to be that nothing which could be regarded as an agent survives).

What should the view of agenthood that we've just developed say about the particular identity puzzle cases mentioned in Section I: the Combined Spectrum and Double Teletransportation? As it turns out, not much. As I already claimed back in Section I, I think the correct response to both of these cases is that they involve either arational belief change (in Combined Spectrum) or no belief change (Double Teletransportation), and so asking about the rationality of the temporal evolution of belief is pointless in both cases. There are a couple primary purposes that talk about what beliefs it would be rational for an agent to have at a time serve: action guidance (principles that agents can consult when they are *deciding* or *reasoning* about what to believe) and evaluation (judging how well an agent that is following some action-guiding principles is succeeding at certain epistemic goals). In the Combined Spectrum case, whatever beliefs the agent ends up having (if any), are not up to it. There is no guidance to perform, no process attributable to the agent to evaluate. Talking about the rationality of the agent's belief change just isn't very meaningful in this context. In Section 1, I claimed that the proponent of ordinary conditionalization should reply that their norm is about learning experiences, and needn't say anything about the Combined Spectrum. I think this is still the right thing to say even with the more powerful machinery we've just developed. Depending on what exactly happens in the Combined Spectrum, the result may or may not be psychologically continuous enough with the victim, Parfit, to count as a single agent.[28] But whether that's true or not, asking about the rationality of the resultant beliefs is kind of silly – they're stipulated to be arrived at by an exogenous, arational process. In Double Teletransportation, it's similarly true that the beliefs that Lefty and Righty wake up with aren't up to them, and so questions about what they should believe are similarly inert. What happens to Pre is much less horrific than what happens to Parfit in (intermediate cases of) the Combined Spectrum, but the case is alike in there being nothing interesting to say about the rationality of the belief change – there's no room for action guidance and no point to evaluation. Now, of course, my view of agential persistence does have something interesting to say about how Lefty and Righty are related to Pre – but it's very similar to what Parfit's view says. Although neither Lefty nor Righty can be identical to Pre (and I have additional reasons for believing Pre's dead that Parfit doesn't share), both the systems consisting of Pre succeeded by Lefty and Pre succeeded by Righty are psychologically continuous agents. Both Lefty and Righty will have no trouble in continuing to use various evidence that was originally learned by Pre, and they both will and should continue to follow through with the plans they remember making before stepping into the teleportation booth. On both Parfit's and my understanding of psychological continuity, this case is trivial: Pre is fully psychologically connected to, not merely continuous with, both Lefty and Righty. To really put my criteria to work, we need a harder puzzle case. So, consider:

The Doxastic Amoeba

The doxastic amoeba is a very smart creature and has a mental life at least as complex as that of an ordinary human. Much unlike a normal human, however, there comes a point in its life where it starts slowly dividing into two copies: I say copies, because at the moment of division, both of the "children" – call them B and C – have precisely the same credence function, which is also arbitrarily similar to the mental states that the "parent" amoeba – call her A – has in moments leading up to the

[28] And again, mostly, we should expect Parfit to die and not be succeeded by anything that resembles an agent.

division.[29] This is, of course, very tricky to accomplish. In the days leading up to her division, A has to create a constantly-updated duplicate of her entire credal set, as well as building within her enough systems to keep two separate organisms alive after the separation occurs. She has to do all of this, while continuing to function as a doxastic agent. Because unlike in the previous example of personal fission that we've discussed, *A* is *continually learning throughout the entire process of her division.* At each instant, she's processing continuous streams of information about her environment – changes in the pH, changes in salinity, monitoring the distribution of various nutrients, etc.[30]; she is continuously revising her beliefs about these facts as the information comes in.[31] She may also be simultaneously revising her beliefs on other topics – perhaps she's wondering whether either of her "children" could be considered identical to her, and her opinion is vacillating as she considers various arguments and counterarguments. In any case, her credences are in a state of constant but gradual flux until the precise moment at which she divides. And even at the moment, both of her children continue continuously learning, without skipping a beat. How should we evaluate the rationality of B's and C's beliefs? Are either of B or C identical to A? Does that matter to how we should evaluate their beliefs?

My answers to the above question might be obvious from what I've said so far in the chapter in conjunction with how I've set the case up, but let's go through it anyway. I think it is fairly clear that neither B nor C are identical to A, for reasons very similar to the reasons already mentioned by Parfit. Do I think this case is as good, or close to as good, as ordinary survival for A? Not really, but I don't have the same kind of visceral disagreement with Parfit as I do in the teleporter cases, and I don't have any compelling arguments. What I am, again, convinced of is that these hard questions about identity

[29] A little more precisely: call the moment of division $t = 0$. The credence functions that the children each have at this moment are $c_B(x; 0) = c_c(x; 0)$. The value of the parent's credence function $c_A(x; t)$ approaches the value $c_B(x; 0) = c_c(x; 0)$ as time gets arbitrarily close to 0, despite the fact that A's credence function doesn't exist at $t = 0$. Viz., $\lim_{t\to 0} c_A(x; t) = c_B(x; 0) = c_c(x; 0)$.

[30] I have no reason to believe real amoebas do any of this.

[31] For much, much more on how it might be possible to continuously update on multiple continuous streams of information, see both Chapters II and III of this dissertation.

and the survival of *persons* don't matter in the slightest to whether there are continuous agents. It is perfectly clear that both B and C are psychologically continuous with A, by as strict a standard as we'd like; the case was constructed so that this would be true! Both B and C obviously have the relevant causal capacities (at least, if we judge that A did, before engorging herself to prepare for fission) to be agents, and so the systems consisting of A succeeded by B and A succeeded by C both seem like continuous, persistent, rational agents. How should we evaluate the rationality of B or C's credal state at the moments just after division? In exactly the same way that we were evaluating A's processes just before division! In particular, if we think that it would be rationally (permissible, obligatory, etc.), for A's learning processes to be governed by a credal rate neighborhood norm like my Continuous Probability Kinematics[32], we should think that it would be rationally (permissible, obligatory, etc.) for B and C to both continue updating according to this same norm. Both of the two persisting agents can (and can permissibly, and should, etc.) even satisfy the norm at the moment of division.

I don't think it's worth running through all of the details, but the view of agential persistence developed in this section can also accommodate puzzle cases involving the fusion of persons/organisms. But, as the reader can probably imagine from the discussion so far: these are cases where the questions of *identity* are quite fraught, but the question of whether there is some system that can reasonably be regarded as a single temporally-extended *rational agent* is not difficult at all. (Indeed, much like in the fission case, there will be two such systems.) And so, by tying the application of our epistemic norms to the persistence of *epistemic agents*, we simply avoid having to settle the metaphysics of personal identity.

[32] See Chapter II.

IV. Conclusion

In this chapter, I've considered two major kinds of arguments that time-slice epistemologists bring against diachronic norms: arguments that diachronic norms violate a commitment to mentalist internalism, and arguments that diachronic norms rely on personal identity in a problematic way. Although I don't think that either kind of argument is as telling against ordinary conditionalization as the synchronists do, I share much of the commitments underpinning these arguments: I also believe that rational norms should supervene on the agent's mental state, and that our judgments about the rationality of an agent's mental state at some instant shouldn't depend on our judgments about whether that time slice is the same person as some other time slice – especially in puzzling cases like the ones we've discussed. On the other hand, I find the purely synchronic view impoverished. I share Podgorski's worry that it simply cannot properly handle the evaluation of rational processes like learning and reasoning. And as I've tried to explain, I strongly disagree with Hedden that these kinds of rational processes should be seen as an activity that only imperfectly rational agents need to engage in; these kinds of rational processes are at the very core of what it is *to be an agent.*

I have argued that neighborhood norms should be attractive to anyone who, like me, sees the synchronic picture as inadequate – and believes our norms must impose some kind of connections between time slices to properly capture what is involved in rational processes like learning – but who also, like me, is sympathetic to the idea that norms that are action-guiding should be accessible from an agent's current mental states. In recapitulation, the idea is that neighborhood norms require an agent to be able to remember her previous mental states, or bind her future actions, on only the smallest possible time scales. Yet nonetheless, neighborhood norms are fully capable of governing long, complex processes of learning. I have also argued that neighborhood norms are readily compatible with a view that ties the rationality of time slices not to obscure questions about personal identity but to the simpler question of what it takes for a system to comprise an agent; what matters in the identity puzzle cases should not be, as Hedden correctly notes, personal identity. And I believe that Parfit's Relation R gives a crucial part of the answer to what it takes for a system to comprise a temporally extended agent: mental continuity and/or connection is the key. I have also argued that the importance of Parfit's R Relation is deeply complementary to my view of rational processes as being foundational to agenthood. Finally, I argued that systems that obey a certain class of neighborhood norms (credal rate neighborhood norms) will exhibit the necessary kind of continuity to persist as temporally extended agents. And it is the persistent agents that are the proper target of these kinds of rational norms – the questions about which persons persist simply don't matter to what the various time slices should believe.

And so, I see neighborhood norms as largely compatible with the concerns that motivate the time-slice epistemologists (though perhaps not in a way that they will find fully satisfactorily), while still having *just enough* temporal structure at each instant to meaningfully guide and evaluate processes like learning. In this sense, neighborhood norms are *minimally diachronic* at each instant. But when followed on some interval, the neighborhood norm can govern a process that appears robustly diachronic.

Works Cited

Arntzenius, Frank. 2003. “Some problems for conditionalization and reflection”. *Journal of Philosophy* 100: 356-370.

Easwaran, Kenny. 2014. “Why Physics Uses Second Derivatives”. *The British Journal for the Philosophy of Science* 65 (4): 845–62.

Hedden, Brian. 2015. *Reasons Without Persons.* Oxford University Press.

Hedden, Brian. 2016. “Mental processes and synchronicity”. *Mind* 125 (499): 873-88.

Kelly, Thomas. 2016. “Historical versus current time slice theories in epistemology.*”* In B.P. McLaughlin and H. Kornblith (eds.), *Goldman and His Critics*. John Wiley & Sons.

Moss, Sarah. 2015. “Time-slice epistemology and action under indeterminacy”. In Tamar Szabó Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology*: 172-94. Oxford University Press.

Parfit, Derek. 1986. *Reasons and Persons*. Oxford University Press USA – OSO. ProQuest Ebook Central.

Podgorski, Abelard. 2016. “A reply to the synchronist”. *Mind* 125 (499): 859-71.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.

CHAPTER II

(Temporally) Continuous Probability Kinematics

1. Introduction

The setting is Victorian London. You are concerned about your dipsomaniacal uncle who has not yet come home, despite the hours having grown small. You muster your resolve to find him, checking several of his usual haunts. After some searching, you happen upon an alleyway filled with fog, only dimly illuminated by the streetlamp. At the far end of the alley is a shadowy figure drunkenly singing. There are two signs by which you hope to determine if this figure is your uncle: your uncle wears a very distinctive red frock coat; and when he sings after drinking, he usually sings a song of his own invention – although many of your uncle's associates know the song, your uncle is drastically more likely to sing it that anyone else. You walk towards the figure, while gazing intently into the fog and with ears closely tuned to the song echoing through the alley. As you approach, your confidence that you see a red coat increases; you also gradually come to believe that the song you hear is your uncle's. You become confident that you have found him at last.

Richard Jeffrey famously realized that agents often acquire evidence that is difficult to represent as propositional and developed an updating rule that allows for treating the impact of evidence as direct manipulation of the agent's credences in propositions. In a similar spirit, I claim that agents often have learning experiences that consist of *gradually changing* confidence in propositions. In this paper, I develop an updating rule for this kind of learning experience that is able to integrate streams of information about propositions that can be correlated in arbitrary ways. In the next section, I begin with an overview of Jeffrey's probability kinematics. In Section 3, I consider how to formulate a temporally continuous version of Jeffrey conditionalization. In Section 4, I raise the problem of how to incorporate multiple streams of information. Section 5 develops my updating rule, which answers the question raised in Section 4, and has the continuous rule from Section 3 as a special case. Finally, in Section 6, I attempt to provide answers to some questions and objections.

2. Jeffrey's Probability Kinematics

There are many cases of seemingly rational change in belief where it is difficult to represent the change as conditionalization on some proposition. Modelling the agent as performing Bayesian conditionalization on the content of her visual and auditory experiences would require that she had conditional priors over the possible content of each sensory modality, which seems wildly implausible.[33] This aspect of our problem has already been solved by Jeffrey's Probability Kinematics. In this section, I will work through the motivation for Probability Kinematics (or Jeffrey conditionalization) and discuss some important properties that it has: **rigidity** (or **sufficiency**) and **non-commutativity**. After explaining the role that rigidity plays in Probability Kinematics, I present what I take to be the most compelling motivation for rigid updating, with an eye towards what we should try to preserve in the continuous case. I also discuss the difference between **soft** and **hard** learning experiences, as distinguished by Joyce (2004).

Observation By Candlelight[34]

An observer is interested in finding out whether the color of a piece of cloth is green, blue, or violet – let *G, B,* and *V* represent the propositions that the cloth is green, blue, and violet, respectively. These three propositions are *mutually exclusive*: at most one can be true. Although the propositions are not, in general, *jointly exhaustive* (the cloth could be some other color, like red), suppose that the agent starts out confident that the cloth is definitely one of those three colors. The set $\{G, B, V\}$ is thus a *partition*: exactly one of the propositions must be

[33] This makes mundane cases of learning possible only with unimaginable foresight. I've had many sensory experiences I was incapable of imagining prior to their occurrence. Having priors conditional on them would require my having detailed representations of those events before they happened, which seems implausible.

[34] Paraphrased from Jeffrey (1983), p. 165.

true (according to the agent). Before looking at the cloth, the agent has some prior credences in each of the propositions, say, $c_0(G) = c_0(B) = 0.3$ and $c_0(V) = 0.4$; he thinks it's most likely that the cloth is violet, but not by much, and blue and green are equally likely. The agent looks at the cloth in dim candlelight and becomes much more confident, but nowhere near certain, that the cloth is green. Because of the poor lighting, it wouldn't be a huge shock to learn that it was actually blue. The cloth being violet is only barely consistent with the visual impression he had, but can't be ruled out. Let's say the agent's final credences are $c(G) = 0.7, c(B) = 0.25$, and $c(V) = 0.05$. This seems like a very plausible kind of learning, but it's not best thought of in terms of becoming certain of some proposition.

Clearly, the agent has not become certain of any of *G, B,* or *V*. But maybe we can think of the agent as becoming certain of the proposition that his visual experience was *such-and-such-a-way*, where *such-and-such-a-way* is elliptical for some incredibly fine-grained description of the character of his experience – call this proposition *E*. Then we regard the agent as having prior conditional credences $c_0(G|E) = 0.7, c_0(B|E) = 0.25$, and $c_0(V|E) = 0.05$, so that he obtains his final credences by Bayesian conditionalization on *E*. As Jeffrey notes, "there need be no such proposition *E* [under consideration]; nor need any such proposition be expressible in the English language. … It seems that the best we can do is to describe, not the quality of the visual experience itself, but rather its effects on the observer …" (Jeffrey 1983, 165). Rather than modelling the agent as having a prior over the infinitely many incredibly fine-grained specifications of varieties of visual experience, Jeffrey proposed a way of accounting for the learning experience directly in terms of the changes in credences.

Probability Kinematics[35]

Suppose an agent begins with prior credence function c_0 and undergoes a learning experience that assigns new credences to the elements of some partition $\{A_1, \ldots, A_n\}$. Then, the agent's resultant credence in an arbitrary proposition, *x*, is given by

$$c(x) = \sum_{i=1}^{n} c_0(x|\, A_i)c(A_i),$$

[35] This presentation of the updating rule more closely follows Jeffrey (1992) than Jeffrey (1983).

where the $c(A_i)$ are the credences stipulated by the learning experience and the $c_0(x|A_i)$ are the agent's conditional prior credences for *x*, conditional on each of the A_i. If $c_0(A_i) \neq 0$, then $c_0(x|A_i) = \frac{c_0(x \wedge A_i)}{c_0(A_i)}$.

Probability kinematics is a strict generalization of Bayesian conditionalization: Bayesian updating on learning some proposition *E* with certainty is just a special case of Jeffrey conditionalization on the partition $\{E, \neg E\}$ – namely, the case where the stipulated credences are $c(E) = 1, c(\neg E) = 0$. This is true because Probability Kinematics is the unique updating rule for the class of learning experiences Jeffrey identified (which, as just mentioned, includes all Bayesian learning experiences) which satisfies the same core property as Bayesian updating: **rigidity**, sometimes also called **sufficiency**.

Rigidity (Sufficiency). $c(x|A_i) = c_0(x|A_i)$ for each $A_i \in \{A_1, \ldots, A_n\}$ and any proposition *x* in the agent's algebra.[36] Updating on the elements of the partition $\{A_1, \ldots, A_n\}$ maintains the credences conditional on all elements of that partition.

It's fairly simple to prove that, for Jeffrey's class of learning experiences, an agent can satisfy rigidity if and only if they update by Probability Kinematics. But that raises the question: why should an agent satisfy rigidity? In the next subsection, I present what I find to be the most compelling motivation for updating rigidly.[37]

The Probative Value of Evidence

In the context of discussing probabilistic confirmation theory, Hájek and Joyce (2008) draw a distinction between the *incremental* value of a piece of evidence and the *probative*

[36] Take the set of all (either finitely many, or perhaps countably infinite) propositions that the agent has beliefs about. The agent's algebra is the closure of this set under negation and disjunction.

[37] I'm not talking about the tradition of diachronic Dutch book arguments for Jeffrey conditionalization – e.g., Skyrms (1987), but this isn't because I don't find this kind of argument compelling. The main reason is that it's difficult to apply this kind of justification to the kind of learning experience that this paper is ultimately interested in.

value of the same.[38] The incremental value of evidence, measured by $c_E(H) - c(H)$, answers the question: ‘if you were to learn that *E* is true, how would your confidence in *H* change’? This is obviously one very important sense of the quality of evidence: among other things, if you actually acquire *E,* the incremental value measures how much your confidence in your hypothesis changes. The probative value of a piece of evidence, measured by quantities $c_E(H) - c_{\neg E}(H)$, is a bit more subtle, though also very useful. One of the kinds of questions it’s useful for answering is ‘how much does your present confidence in *H* depend on your confidence that you will observe *E*’?

The distinction can be illustrated clearly with an example:

> “Suppose that Ellen is a randomly chosen citizen of a town inhabited by 990 Baptists, 2 Catholics, and 8 Buddhists. Let *H* say that Ellen is not a Buddhist. According to all incremental measures, the datum *E* that Ellen is a Baptist provides exactly the same amount of evidence for *H* as does the datum *E** that she is a Catholic. The probative measures disagree, saying Ellen’s being a Baptist provides a great deal of evidence for *H* whereas the datum that she is Catholic provides hardly any.” (Hájek and Joyce 2008, 153)

At present, when all I know about Ellen is that she’s chosen randomly from the population, I am quite confident she’s not a Buddhist: $c(H) = \frac{992}{1000}$. If I learn that she’s either Catholic or Baptist, I will then be certain that she’s not Buddhist (assuming that the religious affiliations are all mutually exclusive), so $c_E(H) = c_{E^*}(H) = 1$. So, *E* and *E** have the same incremental value for *H*. But, $c_E(H) - c_{\neg E}(H) = 1 - \frac{2}{10} = \frac{4}{5}$, while $c_{E^*}(H) - c_{\neg E^*}(H) = 1 - \frac{990}{998} = \frac{4}{499}$. What’s going on here is that much of my confidence that Ellen isn’t a Buddhist is attributable to my belief that she’s probably a Baptist. If I were to learn that she weren’t a Baptist, it would drastically change how likely I think she is to be Buddhist. But because I’m already so confident that she’s Baptist, becoming *certain* she’s a Baptist wouldn’t change my confidence in *H* very much; I can only be so confident that she’s not a Buddhist because I think I almost definitely won’t learn she’s not a Baptist. On the other hand, because being Catholic is already such a small share of

[38] The core ideas are originally due to Joyce (1999) and Christensen (1999).

the ways in which she might not be Buddhist, learning that she's not Catholic makes an almost negligible difference to how likely I think *H* is. The probative value is sensitive to how your present credence in the hypothesis depends on whether the evidence obtains, while the incremental value isn't.

Using Jeffrey conditionalization on a binary partition $\{A, \neg A\}$ has a special connection with the probative value of evidence.[39] First of all, it will clearly hold the probative values of both A and $\neg A$ constant for any arbitrary hypothesis that the agent is interested in. By the definition of rigidity, we have that $c(x|A)$ and $c(x|\neg A)$, for any proposition *x* in the agent's algebra, are held constant by the probability kinematics. Thus the probative value of *A* for *x*, $c(x|A) - c(x|\neg A)$ is also constant. In fact, it turns out that the change in the agent's credence in *x* is equal to the change in the agent's credence in *A* multiplied by the probative value of *A* for *x*: $\Delta c(x) = [c(x|A) - c(x|\neg A)]\Delta c(A)$.[40] This means that, in the case of a binary partition, there is a special justification for using a rigid updating rule: it's the way of changing your credences in $\{A, \neg A\}$, and any other propositions that might depend on them, that respects your current commitments about A and $\neg A$ as evidence. Other methods of updating will, in general, change the probative values of *A* for various propositions. Thus, rigid updating rules are especially appropriate if you find yourself in a situation where you're learning that some proposition is more (or less) likely than you initially thought it was, but you think that learning the proposition (or its negation) with certainty should still have the same degree of evidential support for the hypotheses you're interested in. There may be other kinds of learning experiences where the evidence you're acquiring changes the evidential support relations that obtain between the proposition and various other hypotheses; rigid updating rules won't be applicable in those situations. Rigidity "should not be thought of as a universal and mechanical rule of updating, but as a technique to be applied in the right circumstances, as a tool in what Jeffrey terms the 'art of judgment'" (Bradley 2005, 362).

[39] There are related quantities that are preserved by Jeffrey shifts on a partition with $n > 2$ elements, but it's a bit more complicated. For partitions of arbitrary size, the $c(x|A_i) - c(x|A_j)$ are preserved for all *i* and *j*. You can think of these quantities as jointly encoding the probative values of the elements of the partition for *x*, but I choose to focus on the simpler binary version.

[40] Jeffrey (1983) notes this relationship, though he calls the quantity $c(x|A) - c(x|\neg A)$ the "relevance" of *A* for *x* (170).

(Non)Commutativity

In general, performing two successive Jeffrey updates on distinct partitions ***A***, and then ***B*** will not yield the same result as Jeffrey conditionalizing on ***B***, then ***A***. This is because the shift on $\{A_1, \dots, A_n\}$ will, typically, change the conditional credences on the elements of $\{B_1, \dots, B_m\}$, and vice versa. For our purposes, the key result is this: two shifts on *binary* partitions ***A*** and ***B*** commute only if ***A*** and ***B*** are probabilistically independent – viz., $c(A_i|B_j) = c(A_i)$ and $c(B_j|A_i) = c(B_j)$, for all *i* and *j*.[41]

Shifts, Hard and Soft

When presented as a direct assignment of new credences to the elements of some partition, successive Jeffrey shifts *on the same partition* also don't commute. It's easy to see why: the first shift will assign credences $c(A_i) = a_i$, the second assigns $c(A_i) = a_i^*$. Unless $a_i = a_i^*$ for all *i*, in which case the supposedly two shifts are in fact the same, changing the order the shifts are performed in will differ in at least the credences assigned to the elements of $\{A_1, \dots, A_n\}$. This way of characterizing Jeffrey's probability kinematics also treats it as what Joyce (2004) terms a "hard" shift (or learning experience): "it ignores the prior and resets [the credences assigned to each element of the partition] *de novo*, thereby requiring the posterior to satisfy [the imposed credences] *for any prior*" (448). If a hard Jeffrey learning experience directly sets $c(A_i) = a_i$, then two agents with completely different priors who are both exposed to this same evidence will end up with the same resultant credences for the elements of ***A***. In contrast, Joyce (2004) defines a *soft* learning experience as one where the constraints on the agent's resultant credences are sensitive to the agent's prior.

As Field (1978) shows, the fundamental dynamics of Jeffrey's probability kinematics do not require that we interpret evidence as imposing hard constraints. He suggests we think of nature as providing the agent with an input parameter that fixes the agent's resultant credences as a function of their prior credence (364). Jeffrey shifts that can be described in this

[41] See Diaconis and Zabell (1982), 825-26, for much more detail.

way are soft and commute.[42] Wagner (2002, 2003) pursues a similar strategy: rediscribing the content of learning experiences as specifying the Bayes factor. In the next section, we will see that a differential version of Jeffrey updating is also soft. To be clear: these soft "versions" of Jeffrey updating describe qualitatively different learning experiences but share the core principle of rigidity.

3. A Temporally Continuous Version of Jeffrey Updating

As I presented it in the previous section, Jeffrey conditionalization is a temporally discontinuous process: you start with some prior credence function c_0, you have a learning experience where you assign new credences to the members of some partition $\{A_1, \dots, A_n\}$, and then you have a new credence function *c* which reflects what you learned in the experience. Now let's consider a different example: observation under brightening light.

Just as in Observation by Candlelight, our agent is interested in the color of a piece of cloth and she has various other beliefs about the cloth which depend on its color. Unlike in the previous case, let's assume that all that matters is whether the cloth is blue or not – initially, she thinks both possibilities are equally likely. A lamp that can be dimmed/brightened continuously is aimed at the piece of cloth. The cloth initially receives no illumination, so that the image the agent sees is pitch-black. The illumination is gradually, almost imperceptibly, increased until she can see the cloth's color quite clearly. By the end, she is nearly completely confident that the piece of cloth is blue. Let's say her final credences are something like $c_f(B) = 0.999, c_f(\neg B) = 0.001$. How should we model the agent's credences throughout the duration of this process?

One option is to think of the agent as undergoing continuum-many learning experiences that update the credences the agent should have for the elements of the $\{B, \neg B\}$ partition; we can think of our agent as obeying Jeffrey conditionalization at each moment in time. So, the kind of evidence nature is giving our agent consists of a function, $c_t(B)$, that specifies at each

[42] Field's reparameterization also differs in not having Bayesian conditionalization as a special case – see p. 365 for his discussion of this point.

time $t \in [0, t_f]$ the credence that the agent should assign to the cloth being blue. (Coherence requires that $c_t(\neg B) = 1 - c_t(B)$, so one function suffices to set the credences for the partition.) At each moment in time, the agent's credences in some proposition *x* in her algebra is given by $c_t(x) = c(x|B)c_t(B) + c(x|\neg B)\big(1 - c_t(B)\big)$. There are no temporal indexes on the conditional credences, because (as shown in the previous section), updating via Jeffrey conditionalization will preserve $c(x|B)$ and $c(x|\neg B)$ as constants.

If the functions given by nature are differentiable, we have a second option. Rather than treating the process as myriad Jeffrey shifts, we can think of nature as presenting the agent with temporal *rates of change* in $c(B)$; we can think of our agent as satisfying an equation relating her temporal derivatives of each of her credences to the temporal derivatives being supplied by nature. So, nature is supplying a function $r_B(t)$ that specifies at each time $t \in \big[0, t_f\big]$ the rate of change that $c(B)$ should be undergoing; viz., $r_B(t) = \frac{dc(B;\, t)}{dt}\Big|\, t$. Again, coherence demands $\frac{dc(B)}{dt} = -\frac{dc(\neg B)}{dt}$. The rate of change for the agent's credence in an arbitrary proposition $x \in \boldsymbol{A}$ is then given by $\frac{dc(x;\, t)}{dt} = [c(x|B) - c(x|\neg B)]\frac{dc(B;\, t)}{dt}$.[43] *All expressions in the previous equation are evaluated at time t.*[44]

Another somewhat subtle difference between these two kinds of continuous Jeffrey updating procedures is that the differential "version" of the many-hard-shifts process is automatically a soft learning experience. The many-normal-shifts version is hard for the same reason that the default way of presenting the probability kinematics is: the credences assigned by nature are completely insensitive to the agent's priors on the partition. Because the differential version specifies rates of change at each moment in time, the total effect over the duration of the interval is to require that the agent's credence in $c(B)$ *changes* by a certain amount. Her final credence in *B* at the end of the interval is the sum of her initial credence

[43] I'm using the fairly cumbersome notation $\frac{dc(x;\, t)}{dt}$ for the temporal derivatives of the credences to try to make the structure of what's happening as clear as possible: we're thinking of the agent's credences as a function from the product of the agent's algebra and the temporal interval $[0, t_f]$ to real numbers in the interval $[0,1]$. This means that, say, $c(B)$ is itself a function of time: $c(B; t)$: $\big[0, t_f\big] \rightarrow\ [0,1]$ that specifies which credence the agent has in *B* at each moment. I will often use the simpler notation $\frac{dc(x)}{dt}$, leaving the temporal dependence implied.

$c_0(B)$ with this change – thus, her final credence in *B* depends both on the quality of the evidence and on her prior. Also, although every evolution governed by the differential process could be represented as satisfying the hard process, the converse is not true; one obvious difference is that the hard process is free to require jumps that aren't even continuous – let alone differentiable.

So, now we have an updating rule that can handle continuous streams of information about a single binary partition. The final result of updating in this way is always equivalent to the Jeffrey shift that sets the final credence in *B*, $c_f(B)$. This is because the conditional credences for all propositions $c(x|B)$ and $c(x|\neg B)$ are constant over the duration of the entire interval. So, the agent's final credence, at time t_f, in an arbitrary proposition is given by $c(x;\ t_f) = c_0(x|B)c_f(B) + c_0(x|\neg B)\left(1 - c_f(B)\right)$. But this is exactly the result of the Jeffrey shift on $\{B, \neg B\}$ that sets $c_f(B)$. So far, so boring. Although the differential version and continuum-many Jeffrey shifts version are conceptually very different (they involve radically different kinds of evidence from nature, with one consequence being that the former is always a soft shift while the latter can be hard), moving to the differential version doesn't solve any problems that the many shifts version couldn't. However, in the next section I will begin introducing a closely related updating rule that has the benefit of being able to handle multiple streams of continuous information about *arbitrarily related propositions.*

4. The New Problem: Integrating Multiple Continuous Streams of Evidence

Let's consider a third example: Ori, the blind florist. Ori is a talented florist with years of experience who has lost her sight. Now that she cannot see, she relies on both touch and a variety of tools to determine which flowers she's working with. For a certain arrangement, she wants a flower that is both a particular shape and blue. For the shape, she relies on touch. For the color, she is using a blue detector. The blue detector is a very sophisticated device that uses machine learning to try to identify whether the predominant color of an image it's viewing is blue. It performs this analysis in real time, and outputs its current confidence about the color of what it's looking at as the pitch of a sound. The program has a highest pitch, which corresponds to certainty that the image is blue, and a lowest pitch representing certainty that the image is not blue. The pitch may change quite rapidly, or rather slowly, depending on how quickly the confidence of the program is changing. Ori picks a flower from a box that she believes has flowers that are mostly *either* blue or the shape that she wants; unfortunately, being both blue and that particular shape is quite rare. Let's say that her initial credences on the two partitions in question $\{B, \neg B\}$ and $\{S, \neg S\}$ are given by the table below:

Table II.1: Ori's prior

	B	not B
S	0.15	0.4
not S	0.4	0.05

The pitch of the blue detector gradually increases, telling her to increase her confidence that the flower is blue; and her touch experiences slowly lead her to the conclusion that the flower is the correct shape. This process takes several seconds. How should Ori's credences change over the course of the interval?

Modelling Ori as undergoing many different Jeffrey shifts isn't as promising a strategy as in the previous case. If we include the agent's credences about the sound she's hearing and the tactile experience she's having, it's very difficult to identify a single partition that could provide the basis for the agent's total evidence at each moment in time (such a partition would have to have uncountably many elements), and it seems somewhat implausible that the agent would really have prior conditional probabilities for every pairwise combination of each possible touch sensation and pitch. Of course, Jeffrey's great insight was that we don't need to do this in cases like Observation by Candlelight.

An alternative, then, is to have Ori perform successive alternating Jeffrey shifts on the two partitions $\{B, \neg B\}$ and $\{S, \neg S\}$. This idea runs into a couple of problems. As I discussed in Section 2, Jeffrey shifts do not generally commute. (And the shifts will not commute for this case in particular, as we will see in a moment.) To pursue this strategy, we must arbitrarily choose one of the two partitions for Ori to update on first. And this choice, for which there can be no real motivation, will result in a different final credence function for Ori than if we had chosen the other. Why, then, think that either method is correct? Another, perhaps more interesting, consequence of this strategy is that it will lead to Ori's credences "zig-zagging". For definiteness and simplicity, let's assume that the blue detector and Ori's touch are both telling her to increase the respective credences at a rate of 0.05 per second, and that the entire process lasts for 8 seconds. Suppose we model Ori with the alternating update method, using timesteps of 1 second; we'll update on $\{B, \neg B\}$ first. Here are the results of the first three updates:

Table II.2: The shift from Ori's prior to c(B)=0.6

	B	not B
S	0.164	0.356
not S	0.436	0.0444

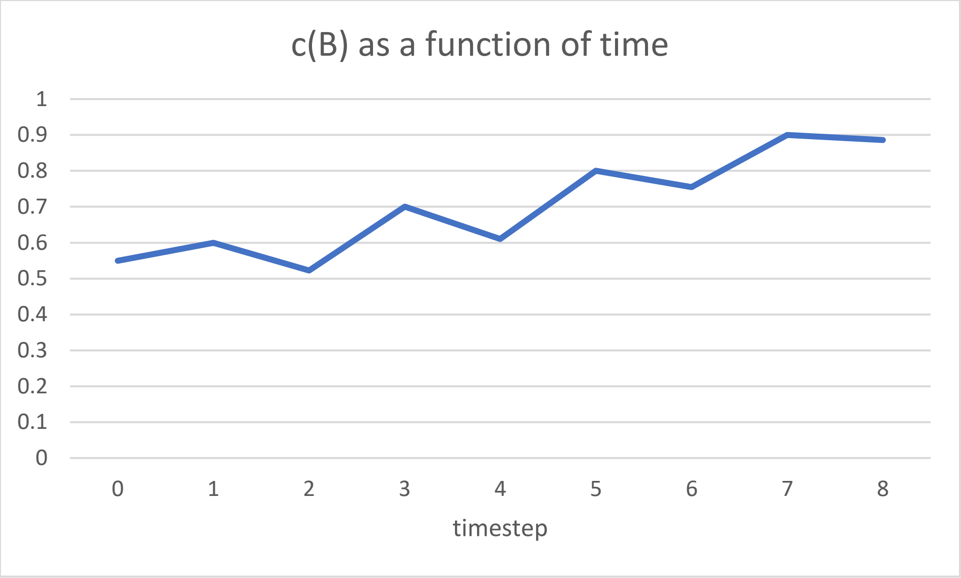
Table II.3: The shift from Table II.2 to c(S)=0.65

	B	not B
S	0.205	0.445
not S	0.318	0.0323

Table II.4: The shift from Table II.3 to c(B)=0.7

	B	not B
S	0.274	0.280
not S	0.426	0.0203

Figure II.1: The "zig-zag"



In Table II.3, we can see that Ori's credence in the flower being blue is approximately 0.523 – she is less confident that the flower is blue than when she started! In Table II.4, Ori's credence that the flower is the right shape is about 0.554 – significantly less that 0.65, which was how confident she was a second before that, in Table II.3. Each update on $\{B, \neg B\}$ causes Ori to lose some of her previous confidence in the flower's shape; each update on $\{S, \neg S\}$ results in losing some confidence that the flower is blue. The result is that the graphs of $c(B)$ and $c(S)$ as functions of time exhibit a "zig-zag" pattern; see Figure II.1 for an example. This process gets something right: Ori is initially convinced that the flower having the desired shape and being blue are anti-correlated, in the sense that $c(S|B) < c(S|\neg B)$ and $c(B|S) < c(B|\neg S)$. This means that Ori is correct to view evidence that the flower is blue as *evidence against* the flower being the desired shape, and vice versa.

But this process also gets something very wrong. At each moment in time, Ori's total evidence for the flower being blue is monotonically increasing. It's true that the evidence from touch is evidence against the flower being blue, but the evidence from the detector is stronger – the net effect is that Ori's confidence that the flower is blue should increase at each moment, but at a slower rate than if she were only listening to the detector and not touching the flower to determine its shape. Similarly, the net effect on Ori's confidence in the flower having the right shape should also be positive, but less than if she were feeling the shape without listening to the detector. Ori's credences should not be zig-zagging. What we want is to somehow combine the two streams of information in a way that simultaneously respects the contributions of both.

Why not merely model Ori as successively updating on the more fine-grained partition $\{S \wedge B, S \wedge \neg B, \neg S \wedge B, \neg S \wedge \neg B\}$?[45] In a way, I think that's exactly what we should do! As we will see, the final result of updating according to the rule I develop in the next section will ultimately be equivalent to a Jeffrey shift on this very partition, and the rule is rigid with respect to this partition throughout. So, it is completely possible to redescribe what I think Ori should do as many Jeffrey shifts on the refined partition. But the key questions are: which shifts? And how are those shifts on the refined partition related to the information that Ori is receiving? Ori is receiving information directly about how her credences in *S* should change and how her credences in *B* should change, at each instant. As most naturally understood on the many-shifts picture, these demands are mutually incompatible – this is one way of understanding why the "zig-zag" occurs. So, another way of understanding the problem that I am proposing a solution to is: can we come up with a principled way of generating required changes on the fine-grained partition $\{S \wedge B, S \wedge \neg B, \neg S \wedge B, \neg S \wedge \neg B\}$ from the inputs from nature that seem to be about changes on $\{S, \neg S\}$ and $\{B, \neg B\}$? And can we do so in a way that makes sense of the result as a direct combination of what we think Ori should do with the signals about $\{S, \neg S\}$ and $\{B, \neg B\}$, if she were to receive them separately? My updating rule is, fundamentally, an attempt at answering these questions in the affirmative. I especially want to avoid the view that what Ori "really" learns about at each moment is the refined partition, where this is understood as a rejection of the claim that she's receiving two separate streams of information. She could, at each moment, choose to focus only on what she heard or felt and ignore the other source. Combing the information *is a choice* (the rational one, I think), so we should be able to represent the evidence as separable.

[45] Thank you to an anonymous reviewer for raising this question. See also the subsection of Section 5, *The Relationship Between CPK and Ordinary Jeffrey Conditionalization.*

5. Differential Rigidity and Continuous Probability Kinematics

Let's return to the second option from Section 3, where we think of nature as giving agents credal rates of change. The goal will now be to answer the question: suppose nature tells an agent to change her credence in some proposition *A* at a certain rate in time through one channel of information (e.g., whether the flower is a certain shape) at the same time as telling her to change *B* at some rate (e.g., whether the flower is blue or not). These propositions may, in general, be related in arbitrary ways. *B* might entail *A*, *A* may be strong evidence for *B*, *A* and *B* might be jointly inconsistent, etc. What rates of change in $c(A)$ and $c(B)$ should the agent adopt at each moment in time? What rate of change should the agent adopt in any proposition *x* that depends in arbitrary ways on $c(A)$ and $c(B)$?

Outside of special cases, we can no longer interpret the rates of change given by nature as the total temporal derivates $\frac{dc(A)}{dt}$ and $\frac{dc(B)}{dt}$, because the logical relationships between *A* and *B* may make it impossible to evolve the agent's credences in a coherent way. Suppose, for example, that *A* and *B* are logically equivalent, and the rates that nature hands the agent at some moment in time are $r_A(t) = 2$ and $s_B(t) = -0.1$. If the agent is initially coherent and tries to conform her credences to $r_A(t) = \frac{dc(A)}{dt}\Big| t$ and $s_B(t) = \frac{dc(B)}{dt}\Big| t$, she will end up with incoherent credences. Since *A* and *B* are logically equivalent, to be coherent the agent must have $c_t(A) = c_t(B)$ for all moments in time. Thus, a coherent agent will also satisfy $\frac{dc(A)}{dt} = \frac{dc(B)}{dt}$ for all moments in time when *A* and *B* are equivalent. However, we can still interpret $r_A(t)$ and $s_B(t)$ as specifying *partial* temporal derivatives: $r_A(t) = \frac{\partial c(A)}{\partial t}\Big| t$ and $s_B(t) = \frac{\partial c(B)}{\partial t}\Big| t$. What this means is that nature is telling the agent that $r_A(t)$ is the instantaneous rate of change in $c(A)$ that the agent would be obligated to adopt if she weren't learning anything else, and similarly for $s_B(t)$ and $c(B)$.[46] However, the total rate of change for the agent's credences in

[46] In general, the partial derivative of $z = f(x, y)$ with respect to *x*, $\frac{\partial z}{\partial x}$, expresses how tiny changes to *x* would "cause" *z* to change, assuming constant *y*. $\frac{\partial z}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x,\ y) - f(x,\ y)}{\Delta x}$. Of course, if *y* is also a function of *x*, then the actual rate at which varying *x* changes *z*, $\frac{dz}{dx}$, will not typically equal $\frac{\partial z}{\partial x}$. You have to combine the change in *z*

$c(A)$ and $c(B)$ will, in general, be some functions of $r_A(t)$ and $s_B(t)$; those functions may also depend on the agent's present credences.

An analogy may be helpful: consider a clock with two hands that are coupled together by a system of gears so that manually moving one hand also causes the other to move. Let's call the angular position of one hand α and let β be the position of the other. If someone directly moves the first hand, then you can think of them setting some rate of angular motion of that hand, call it $r_\alpha = \frac{d\alpha}{dt}$. You can calculate the rate of motion of the other hand, $\frac{d\beta}{dt}$, if you know what the gear ratio between the two hands are – call this ratio $\frac{\partial\beta}{\partial\alpha}$. Similarly, if you moved the other hand, you could directly impose s_β and calculate the motion of the first hand, $\frac{d\alpha}{dt}$, as long as you know $\frac{\partial\alpha}{\partial\beta}$. What happens if you impose r_α and s_β at the same time? Well, now you're directly moving the first hand at the rate r_α, but the first hand is also being moved by the direct input on the other hand. So, the total angular motion of the first hand is $\frac{d\alpha}{dt} = r_\alpha + \frac{\partial\alpha}{\partial\beta} s_\beta$. Similarly, the motion of the other hand is $\frac{d\beta}{dt} = s_\beta + \frac{\partial\beta}{\partial\alpha} r_\alpha$. Put another way: now we're thinking of α as a function, both of time directly, and of β, which is itself a function of time: $\alpha(t; \beta(t))$. So the total temporal derivative of α is given by $\frac{d\alpha}{dt} = \frac{\partial\alpha}{\partial t} + \frac{\partial\alpha}{\partial\beta}\frac{\partial\beta}{\partial t}$: the sum of a direct rate of change in terms of time and the contribution to α's change that direct change in β makes. We can think of r_α and s_β as playing the roles of $\frac{\partial\alpha}{\partial t}$ and $\frac{\partial\beta}{\partial t}$.

How should the agent combine $r_A(t)$ and $s_B(t)$ to generate $\frac{dc(A)}{dt}$ and $\frac{dc(B)}{dt}$? There are some special cases that I hope will be intuitive to the reader:

1. When *A* and *B* are logically equivalent, $\frac{dc(A)}{dt}$ should be given by $r_A(t) + s_B(t)$; and $\frac{dc(B)}{dt}$ *must* equal $\frac{dc(A)}{dt}$ to ensure coherence.

that changing *x* directly induces with the change in *z* that occurs "because" of the changes to *y* that changing *x* induces: $\frac{dz}{dx} = \frac{\partial z}{\partial x} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$. This is known as the multivariate chain rule.

2. When *A* and *B* are logical complements (*A* is equivalent to $\neg B$, and vice versa) $\frac{dc(A)}{dt}$ should be given by $r_A(t) - s_B(t)$. $\frac{dc(B)}{dt}$ *must* equal $-\frac{dc(A)}{dt}$ for the agent to be coherent.

3. Suppose *B* is incremental evidence for *A* in the sense that $c(A|B) > c(A|\neg B)$, then: $\frac{dc(A)}{dt} > r_A(t)$ when $s_B(t)$ is positive; $\frac{dc(A)}{dt} < r_A(t)$ when $s_B(t)$ is negative.

4. When *A* and *B* are independent ($c(A|B) = c(A|\neg B) = c(A)$, and similarly for *B*), $\frac{dc(A)}{dt} = r_A(t)$ and $\frac{dc(B)}{dt} = s_B(t)$.

Here are some brief comments on my judgments about the four cases:

Cases 1 and 2: These are basic compositionality principles. It might sometimes be convenient to model one stream of information as two or more streams of information about the same proposition. Whether we choose to break up a stream of information or not should have no impact on the result of the updating rule. Similarly, whether we combine instructions to increase $c(A)$ and $c(\neg A)$ into a single instruction (to increase or decrease $c(A)$, depending on the relevant magnitudes), or treat the instructions as two separate streams of information, we should get the same result. Without obeying these two constraints, we have no way of consistently combining or breaking up streams of information, even about a single proposition (and its negation). The constraints that $\frac{dc(B)}{dt} = \frac{dc(A)}{dt}$ in Case 1 and $\frac{dc(B)}{dt} = -\frac{dc(A)}{dt}$ come directly from coherence. If *A* and *B* are logically equivalent, then a coherent agent must have $c(A) = c(B)$; if they're complements, the agent needs to satisfy $c(A) = 1 - c(B)$. Differentiating those equalities yield said constraints.

Case 3: If *B* is probative evidence for *A* in the sense that $c(A|B) - c(A|\neg B) > 0$, then increasing an agent's confidence in *B* should increase his confidence in *A*. If we had $\frac{dc(A)}{dt} = r_A(t)$, this would mean that the agent was experiencing the same rate of change in $c(A)$ that he would undergo if he only increased his confidence in *A*. Effectively, this would be completely neglecting the probative value of *B* for *A*. (And $\frac{dc(A)}{dt} < r_A(t)$ would be even worse: it would mean that the agent becomes *less* confident when getting both direct and indirect evidence in favor of *A* than they would be if they had only received the direct evidence.) Similarly, if $s_B(t)$ is negative, this means that the agent is losing confidence in a proposition that provides part of their current evidential support for *A*. If we had $\frac{dc(A)}{dt} = r_A(t)$, the agent would be ignoring the loss of this support.

Case 4: As we've seen in the discussion of the previous cases, deviating from $\frac{dc(A)}{dt} = r_A(t)$ and $\frac{dc(B)}{dt} = s_B(t)$ means that the agent is taking account of the evidential relations between *A* and *B*, so that, e.g., learning about *B* is also an indirect way of learning about *A*. If *A* and *B* are independent, this means that the agent doesn't believe there is any such evidential relationship; changing his confidence in one proposition should have no bearing on his confidence in the other.

These four constraints are highly suggestive. Together, they suggest that the agent should satisfy:

$$\frac{dc(A)}{dt} = r_A(t) + [c(A|B) - c(A|\neg B)]s_B(t) \tag{1}$$

$$\frac{dc(B)}{dt} = [c(B|A) - c(B|\neg A)]r_A(t) + s_B(t) \tag{2}$$

Note that $[c(A|B) - c(A|\neg B)]$ is the probative value of *B* for *A* which, as discussed in Section 2, is one way of measuring the evidential import that *B* has for *A*. Similarly, $[c(B|A) - c(B|\neg A)]$ is the probative value of A for B. If we interpret $r_A(t) = \frac{\partial c(A)}{\partial t}\Big|\, t$ and $s_B(t) = \frac{\partial c(B)}{\partial t}\Big|\, t$, then equations 1 and 2 are:

$$\frac{dc(A)}{dt} = \frac{\partial c(A)}{\partial t} + [c(A|B) - c(A|\neg B)]\,\frac{\partial c(B)}{\partial t} \tag{3}$$

$$\frac{dc(B)}{dt} = [c(B|A) - c(B|\neg A)]\frac{\partial c(A)}{\partial t} + \frac{\partial c(B)}{\partial t} \tag{4}$$

The form of these equations is also quite suggestive: if we're thinking of $c(A)$ as a function of both time directly, and of $c(B)$ (which is itself a function of time), then the multivariate chain rule tells us that

$$\frac{dc(A)}{dt}=\frac{\partial c(A)}{\partial t}+\frac{\partial c(A)}{\partial c(B)}\frac{\partial c(B)}{\partial t} \tag{5}$$

$$\frac{dc(B)}{dt}=\frac{\partial c(B)}{\partial c(A)}\frac{\partial c(A)}{\partial t}+\frac{\partial c(B)}{\partial t} \tag{6}$$

Comparing equations 5 and 4 suggests that $\frac{\partial c(A)}{\partial c(B)}=[c(A|B)-c(A|\neg B)]$ and $\frac{\partial c(B)}{\partial c(A)}=[c(B|A)-c(B|\neg A)]$. Consider a learning experience where nature only directly varies $c(B)$; viz., $r_A(t)=0$. After an arbitrarily small amount of time Δt, there will be a change in $c(B)$ of $\Delta c(B)=s_B\Delta t$. The corresponding change in in $c(A)$ is $c_{t+\Delta t}(A)-c_t(A)$, where $c_{t+\Delta t}(A)$ is whatever credence in *A* the agent has after Δt has elapsed. $\frac{\partial c(A)}{\partial c(B)}=\lim_{\Delta t\to 0}\frac{c_{t+\Delta t}(A)-c_t(A)}{s_B\Delta t}$. There are plenty of possible updating rules for which this quantity will not even be defined. But when it is, $\frac{\partial c(A)}{\partial c(B)}$ measures, at a given instant, how $c(A)$ changes in response to changes in $c(B)$. If $\frac{\partial c(A)}{\partial c(B)}=[c(A|B)-c(A|\neg B)]$, then at each moment in time, the infinitesimal change in the agent's credence in *B* due to nature's stipulation will also be accompanied by a change in $c(A)$ that is proportional to the probative value of *B* for *A*.[47]

In addition to satisfying the four desiderata I outlined above, there's another reason to like combining the rates of change in this way: it's the unique way of doing so that satisfies a constraint that I call **differential rigidity**.

Differential rigidity. Let *A* and *B* be the propositions that nature is providing rates of change for, and let *x* be an arbitrary proposition in the agent's algebra. Then, differential rigidity is the requirement that $\frac{\partial c(x|A)}{\partial c(A)},\frac{\partial c(x|\neg A)}{\partial c(A)},\frac{\partial c(x|B)}{\partial c(B)},\frac{\partial c(x|\neg B)}{\partial c(B)}$ are all zero at every instant. (This is the constraint for the case where nature is providing rates of change for two propositions; the generalization is obvious.)

[47] Continuing the clock analogy, $\frac{\partial c(A)}{\partial c(B)}$ is like the gear ratio. (Although this is a little misleading, because unlike in the gear ratio case, $\frac{\partial c(A)}{\partial c(B)}$ and $\frac{\partial c(B)}{\partial c(A)}$ may not be inverses.)

One consequence of differential rigidity is that, in a learning experience where nature is only providing information about *A*, $c(x|A)$ and $c(x|\neg A)$ will be constant throughout the duration; the differential version of Jeffrey conditionalization introduced in Section 3 falls out as a special case. In more general cases where nature is providing information about both *A* and *B*, $c(x|A)$ and $c(x|\neg A)$ will generally not be constant. However, the rate at which $c(x|A)$ and $c(x|\neg A)$ change will depend only on $\frac{\partial c(B)}{\partial t}$, and not on $\frac{\partial c(A)}{\partial t}$. Intuitively, the way in which the agent updates on the stream of information that's about *A* respects, at each moment the agent's credences conditional on *A*. However, as the agent learns about *B*, it will likely change her credences conditional on *A*. And so, in the most general case, there need be no non-infinitesimal interval on which any of the conditional credences are constant. Similarly, the agent makes uses of the changing confidence in *B* in a way that doesn't change the credences conditional on *B*; but, because the agent is also learning about *A* at the same time, her credences conditional on *B* will also typically be constantly changing.

We can derive Equations 3 and 4 from differential rigidity. We start with Equations 5 and 6, and with the law of total probability:

$$c(A) = c(A|B)c(B) + c(A|\neg B)c(\neg B) \tag{7}$$

Differentiating both sides with respect to $c(B)$, we obtain

$$\frac{\partial c(A)}{\partial c(B)} = [c(A|B) - c(A|\neg B)] + c(B)\frac{\partial c(A|B)}{\partial c(B)} + c(\neg B)\frac{\partial c(A|\neg B)}{\partial c(B)} \tag{8}$$

So, if the agent satisfies differential rigidity, $\frac{\partial c(A)}{\partial c(B)} = [c(A|B) - c(A|\neg B)]$. Similarly, $\frac{\partial c(B)}{\partial c(A)} = [c(B|A) - c(B|\neg A)]$. Recall once again that $c(A|B) - c(A|\neg B)$ is a measure of the probative value of *B* for *A*, while $c(B|A) - c(B|\neg A)$ is a measure of the probative value of *B* for *A.*[48] By satisfying differential rigidity, we see that direct changes to $c(B)$ will also induce changes in $c(A)$ according to the agent's current beliefs about the evidential import of *B* for *A*, and vice versa. Plugging these equalities into Equations 5 and 6 yields Equations 3 and 4.

[48] See pp. 47-49.

Indeed, we are now in a position to calculate the updating rule for an arbitrary proposition in the agent's algebra. We start with the multivariate chain rule:

$$\frac{dc(x)}{dt} = \frac{\partial c(x)}{\partial c(A)}\frac{\partial c(A)}{\partial t} + \frac{\partial c(x)}{\partial c(B)}\frac{\partial c(B)}{\partial t} \tag{9}$$

We use the law of total probability to calculate $\frac{\partial c(x)}{\partial c(A)}$ and $\frac{\partial c(x)}{\partial c(B)}$, assuming differential rigidity:

$$\frac{\partial c(x)}{\partial c(A)} = [c(x|A) - c(x|\neg A)] \tag{10}$$

$$\frac{\partial c(x)}{\partial c(B)} = [c(x|B) - c(x|\neg B)] \tag{11}$$

Substituting Equations 10 and 11 into 9, we finally have:

Continuous probability kinematics

$$\frac{dc(x)}{dt} = [c(x|A) - c(x|\neg A)]\frac{\partial c(A)}{\partial t} + [c(x|B) - c(x|\neg B)]\frac{\partial c(B)}{\partial t}. \tag{12}$$

The generalization to the *n*-proposition case (viz., updating on *n* streams of evidence each concerned with a binary partition A_i) is:

$$\frac{dc(x)}{dt} = \sum\nolimits_{i=1}^{n} [c(x|A_i) - c(x|\neg A_i)]\frac{\partial c(A_i)}{\partial t}. \tag{13}$$

The Relationship Between CPK and Ordinary Jeffrey Conditionalization

As I already mentioned when I introduced differential rigidity, one interesting property of CPK is that it has the single-partition differential version of Jeffrey updating discussed in Section 2 as a special case. But it is also related to Jeffrey conditionalization in a much more general way. Just as that "version" of Jeffrey conditionalization is equivalent (in outcome, but not conception) to continuum-many ordinary Jeffrey shifts, any credence function over some temporal interval that can be modelled as obeying CPK can also be generated by continuum-many ordinary Jeffrey shifts on a more fine-grained partition.

For convenience, let's continue to consider the case where the agent is learning directly about two propositions *A* and *B*. The first trick is to notice that updating by CPK always results in the conditional credences $c(x|A \wedge B), c(x|A \wedge \neg B), c(x|\neg A \wedge B)$, and $c(x|\neg A \wedge \neg B)$ remaining constant for the duration of the learning experience, *for any proposition x*. To show this, we show that the total derivative of each of the above-listed conditional credences with respect to time is identically 0.[49] In other words, updating by CPK is rigid on the partition $\{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\}$.

Now, because updating by CPK preserves coherence,[50] the credence function of an agent obeying CPK at any moment *t* will always satisfy $c_t(x) = c(x|A \wedge B)c_t(A \wedge B) + c(x|\neg A \wedge B)c_t(\neg A \wedge B) + c(x|A \wedge \neg B)c_t(A \wedge \neg B) + c(x|\neg A \wedge \neg B)c_t(\neg A \wedge \neg B)$. (The conditional credences aren't time-indexed because they are constant throughout.) So, that temporal evolution of the agent's credences could also be described as the result of performing a Jeffrey shift at each moment *t* on the partition $\{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\}$.[51] The input from nature is reinterpreted as four functions that specify the values $c_t(A \wedge B)$, $c_t(\neg A \wedge B), c_t(A \wedge \neg B)$, and $c_t(\neg A \wedge \neg B)$ that the agent is required to adopt at each moment in time.

The primary philosophical motivations for thinking about this process in terms of CPK rather than as many ordinary Jeffrey shifts are conceptual. As I alluded to in the end of Section 4, the point is that CPK gives a unified account that makes learning about *A* and *B* simultaneously a simple matter of combining the evidence that the agent would have received if they were learning only about *A* and *B* individually. On the ordinary Jeffrey shift picture, it's much harder to find a story that makes the evidence in the joint case a straightforward combination of the evidence in the separate cases – this is because the ordinary picture thinks of the specified *values* for credences of the elements of the partitions as what the agent learns,

[49] I leave this as an exercise to the reader. Use either the product or quotient rule to differentiate, e.g., $\frac{c(x \wedge A \wedge B)}{c(A \wedge B)}$ with respect to time and use Eq. 12 to find $\frac{dc(x \wedge A \wedge B)}{dt}$ and $\frac{dc(A \wedge B)}{dt}$.

[50] See the Appendix for the proof.

[51] Note, however, the converse is not true: there are Jeffrey shifts that can be specified on $\{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\}$ that are incompatible with CPK.

and the specified values are incompatible; when we move to thinking about the content of the learning experience as specifying *partial derivatives*, it's much easier to see how to coherently combine them into one unified learning experience.

6. Questions, Objections, and Answers[52]

Q: How does CPK apply to credence 0?

A: There are two issues related to how CPK applies to credence 0. The first is that the usual ratio "definition" of the conditional credence $c(x|A)$ as $\frac{c(x \wedge A)}{c(A)}$ is undefined when $c(A) = 0$. The obvious solution to this (and the one that is typically employed in traditional conditionalization) is to simply restrict our attention to when this is the case; CPK only applies when the values of all of the directly-varied credences are between 0 and 1. A second issue: when $c(x) = 0$ (and the directly-varied credences are nonzero), $\frac{dc(x)}{dt} = 0$. The problem is that the agent's future credal evolution seems to be underdetermined; it's consistent with the agent's credence in *x* increasing, decreasing, or remaining constant in arbitrarily small intervals near the moment in question. As I've presented CPK so far, I don't see a way to prove that it is incompatible with $c(x)$ taking on negative values. Again, however, there is an obvious fix: once $c(x) = 0,$ keeping $c(x)$ identically zero for all later times is always a solution to CPK.[53] So, to maintain non-negative credences, I can place an additional constraint on CPK: treat credence 0 and credence 1 as "gutters". Once you have an extremal credence in some proposition, your credence in that proposition is constant for the remainder of the learning experience.

Are these fixes costly? In comparison to ordinary conditionalization (both Bayesian and Jeffrey), no. Both updating rules run into exactly the same issue about the traditional "definition" of conditional probability being undefined when the denominator is zero; both also have the property that, once a proposition achieves an extremal credal value, your credence in it cannot be changed by updating on any partition it's not an element of. It would be nice to

[52] The questions in this section were either raised by anonymous reviewers, or inspired by questions they raised.

[53] See the Appendix.

extend CPK to apply outside of these restrictions, but that would be a bonus. For now, the official formulation of CPK is equation (12)[54] plus the two restrictions outlined in this section.

Q/O: Why think that what happens during some gradual learning process is relevant to rationally evaluating an agent? Why not just be concerned with the final result of the process?

A: I believe that reasoning, as a process, is a *more* valuable target of rational evaluation than result. I would be very suspicious of a system that arrived at the "correct" (according to me) final result via a process that did not have deep structural similarities to CPK – I would think it was almost certainly either extremely lucky or cheating. See the first chapter of this dissertation for much more discussion of this point.

A more shallow, but obvious reason is this: evaluating the entire process allows us to make judgments about cases where the agent is interrupted.

Finally, as a general rule, agents should incorporate new evidence as quickly as possible. To wait is to lose all kinds of expected value (epistemic, pragmatic, moral). If nature is providing evidence gradually, the obvious ideal would be to incorporate each bit as soon as it comes in. (This ideal is almost certainly unobtainable; see the next question for further discussion.) But waiting longer than you have to, until the end of a potentially fairly long process, is an obvious failure of rationality. See both the first and third chapters of this dissertation for additional discussion of this point.

Q/O: What does it mean for nature to stipulate that the agent adopt a certain credal derivative at each moment? And why think that's possible, let alone rationally required? Isn't it much more plausible that agents learn in discrete chunks?

A: Taking the last part first: yes, of course! I'd be willing to bet that all agents we know of have some minimal timescale on which they're capable of processing incoming information (so, their mental time is functionally discrete), and probably also some minimal change in the value of

[54] Equation (13), if there are more than two streams of information.

credences that they're capable of representing. I would be amazed if humans were capable of learning *truly* continuously; further, I doubt that it's even possible for there to be an agent that could. I like the idea that agents should be representable as implementing computer programs, in which case we should expect discrete timesteps and storage limits. Others may, of course, have very different reasons for arriving at the same conclusion that learning is probably best fundamentally understood discontinuously.

The continuity of CPK is an abstraction of the kind that is ubiquitous in classical physics. In classical electromagnetism, for example, charge is fundamentally discrete – yet Maxwell's equations, whether in differential or integral form, assume continuity. The assumption is that the granularity of the fundamental units is so miniscule compared to the scale of application of the continuous laws that the continuous description of the evolution of the properties of the system is a very good approximation of the much more complicated discrete interactions. One great benefit of this kind of abstraction is that it works even when you're ignorant of the precise details on the fundamental scale! To know whether an agent approximates updating by CPK, we don't necessarily need to know the details of what their mental timestep is, or what the minimum difference in tonal pitch that they can detect is; and two different systems that differed in these fundamentals could still be very well-described as approximating CPK, as long as the chunks are small enough. By abstracting in this way, it both keeps the math simpler and demonstrates a kind of unity between different systems that discretely approximate the continuous property. Keeping the above in mind, then, having a certain credal derivative is just to be understood as a limit of making various small credal changes over small intervals of time.

Appendix: Proof that updating by CPK preserves coherence

Suppose the agent has a prior credence function which is coherent – that is, it satisfies the probability axioms.

1. **Non-negativity.** For every proposition *x* in the agent's algebra, $c_0(x) \geq 0$.
2. **Normality.** Let Ω be the union of all elements of the agent's algebra. $c(\Omega) = 1$.

3. **Countable Additivity.** Let $x_1, x_2, \ldots$ be an arbitrary sequence of disjoint propositions. $c_0(\bigvee_i x_i) = \sum_i c_0(x_i)$.

I show that the resultant credence function will continue to satisfy the axioms. This proof is presented in terms of the case where the agent is receiving evidence about two propositions, *A* and *B*, but the generalization to the *n*-proposition case is obvious.

Non-Negativity

By stipulation of the first restriction, the directly-varied credences are always greater than zero. Now we must show updating by CPK never causes the agent's other credences to fall below zero.

For any proposition *x* in the agent's algebra other than the directly-varied credences, when $c(x) = 0$, $c(x|A), c(x|\neg A), c(x|B)$, and $c(x|\neg B)$ are all 0. Thus, updating by CPK yields $\frac{dc(x)}{dt} = [c(x|A) - c(x|\neg A)]\frac{\partial c(A)}{\partial t} + [c(x|B) - c(x|\neg B)]\frac{\partial c(B)}{\partial t} = 0.$

For *reductio,* assume that there is some time t_2 at which $c_2(x)$ is negative. By assumption, $c_0(x) \geq 0$. The agent's credence function $c_t(x)$ is a continuous function of time, so by the intermediate value theorem, there must be some time t_1 (possibly identical to t_0) in the interval $[t_0, t_2)$ at which $c_1(x) = 0$. But from t_1 onward, the agent was required to have $c_t(x) = 0$,[55] and so the agent's credence in *x* must be 0 at t_2. *Reductio.* In the next paragraph, we show that the gutter stipulation is consistent with equation (12), the core principle of CPK.

As shown in the first paragraph of this section, $\left.\frac{dc(x)}{dt}\right|_{t=t_1} = 0$. Let $c'(x;t) = \begin{cases} c(x;t), & t \leq t_1 \\ 0, & t \geq t_1 \end{cases}$. Then, $c'(x;t)$ satisfies equation (12). By construction, $c'(x;t)$ satisfies equation (12) for $t \leq t_1$. It also works for $t > t_1$. LHS: $\frac{dc'(x)}{dt}$ is identically 0 for all $t > t_1$. RHS: $[c'(x|A) - c'(x|\neg A)]\frac{\partial c(A)}{\partial t} + [c'(x|B) - c'(x|\neg B)]\frac{\partial c(B)}{\partial t}$ is identically 0, because $[c'(x|A) - c'(x|\neg A)]$ and $[c'(x|B) - c'(x|\neg B)]$ are identically 0. LHS=RHS for all $t > t_1$.

[55] This is the content of the "gutter" restriction.

Countable Additivity

Consider an arbitrary time $\tau \geq t_0$. For any proposition *x*, $c_\tau(x) = c_0(x) + \int_{t_0}^{\tau} \frac{dc(x)}{dt} dt$. Updating by CPK, we have that $c_\tau(\bigvee_i x_i) = c_0(\bigvee_i x_i) + \int_{t_0}^{\tau} \Big([c_t(\bigvee_i x_i\,|A) - c_t(\bigvee_i x_i\,|\neg A)]\frac{\partial c(A)}{\partial t} + [c_t(\bigvee_i x_i\,|B) - c_t(\bigvee_i x_i\,|\neg B)]\frac{\partial c(B)}{\partial t}\Big) dt$, where $c_t(\bigvee_i x_i\,|A) = \frac{c_t(\bigvee_i x_i \wedge A)}{c_t(A)} = \frac{c_t(\bigvee_i (x_i \wedge A))}{c_t(A)}$.

For the agent to fail to satisfy countable additivity at τ means that $c_\tau(\bigvee_i x_i) \neq \sum_i c_\tau(x_i)$. Similarly, we have that $\sum_i c_\tau(x_i) = \sum_i c_0\,(x_i) + \sum_i \Big(\int_{t_0}^{\tau}[c_t(x_i|A) - c_t(x_i|\neg A)]\frac{\partial c(A)}{\partial t} dt + [c_t(x_i|B) - c_t(x_i|\neg B)]\frac{\partial c(B)}{\partial t} dt\,\Big)$. Now, by assumption, the agent initially satisfies countable additivity, so $c_0(\bigvee_i x_i) = \sum_i c_0\,(x_i)$. Thus, the only way the inequality can hold is if $\int_{t_0}^{\tau} \Big([c_t(\bigvee_i x_i\,|A) - c_t(\bigvee_i x_i\,|\neg A)]\frac{\partial c(A)}{\partial t} + [c_t(\bigvee_i x_i\,|B) - c_t(\bigvee_i x_i\,|\neg B)]\frac{\partial c(B)}{\partial t}\Big) dt \neq \sum_i \Big(\int_{t_0}^{\tau}[c_t(x_i|A) - c_t(x_i|\neg A)]\frac{\partial c(A)}{\partial t} dt + [c_t(x_i|B) - c_t(x_i|\neg B)]\frac{\partial c(B)}{\partial t} dt\,\Big)$. For this inequality to hold, it is necessary – though not sufficient – that there is some time $t_0 < t_1 < \tau$ such that one of the four following inequalities obtains:

- $c_1(\bigvee_i x_i\,|A) \neq \sum_i c_1(x_i|A)$
- $c_1(\bigvee_i x_i\,|\neg A) \neq \sum_i c_1(x_i|\neg A)$
- $c_1(\bigvee_i x_i\,|B) \neq \sum_i c_1(x_i|B)$
- $c_1(\bigvee_i x_i\,|\neg B) \neq \sum_i c_1(x_i|\neg B)$

But to satisfy any of these inequalities is for the agent to already violate countable additivity at t_1. In general, at each instant, changing your credences in line with CPK can contribute to a divergence from countable additivity (at a later time) only if your credences already fail to obey countable additivity at present. Thus, updating by CPK will never cause an agent who initially has coherent credences to violate countable additivity.

Normality

Since, by assumption, the agent initially satisfies normality, to show that the agent continues to satisfy normality while updating by CPK, it suffices to show that $\frac{dc(\Omega)}{dt} = 0$, and thus that $c(\Omega)$ is a temporal constant. Because Ω is the union of all elements of the agent's algebra, we're free to represent it as $\Omega = A \vee \neg A$. Recall that *A* is one the two propositions whose credal rate of change is being set by nature. Because CPK satisfies countable additivity (see the previous section of this Appendix), $\frac{dc(A \vee \neg A)}{dt} = \frac{dc(A)}{dt} + \frac{dc(\neg A)}{dt}$. It's a very straightforward consequence of equation (12) that $\frac{dc(A)}{dt} = -\frac{dc(\neg A)}{dt}$.[56] So,

$$\frac{dc(\Omega)}{dt} = \frac{dc(A)}{dt} - \frac{dc(A)}{dt} = 0.$$

The above reasoning holds at any moment in time at which the agent is updating by CPK, so $c(\Omega)$ is constant for the duration of any CPK learning experience.

[56] It's not much harder to show that $\frac{dc(x)}{dt} = -\frac{dc(\neg x)}{dt}$, for all *x*.

Works Cited

Bradley, Richard. 2005. “Radical Probabilism and Bayesian Conditioning”. *Philosophy of Science* 72 (2): 342-64.

Christensen, David. 1999. “Measuring Confirmation”. *Journal of Philosophy* 96 (9): 437-61.

Diaconis, Persi and Sandy L. Zabell. 1982. “Updating Subjective Probability”. *Journal of the American Statistical Assocation* 77 (380): 822-830.

Field, Hartry. 1978. “A Note on Jeffrey Conditionalization”. *Philosophy of Science* 45: 361– 67.

Good, I.J. 1967. “On the principle of total evidence”. *The British Journal for the Philosophy of Science* 17 (4): 319-321.

Hájek, Alan & Joyce, James M. 2008. “Confirmation”. In S. Psillos & M. Curd (eds.), *The Routledge Companion to the Philosophy of Science*. Routledge.

Jeffrey, Richard. 1983. *The Logic of Decision*, revised 2nd edition. University of Chicago Press.

Jeffrey, Richard. 1992. *Probability and the Art of Judgment*. Cambridge University Press.

Joyce, James M. 2004. “The Development of Subjective Bayesianism”. In Dov M. Gabbay, John Woods & Akihiro Kanamori (eds.), *Handbook of the History of Logic* 10: 415-475. Elsevier.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory.* Cambridge University Press. Chapter 6: 181-223.

Skyrms, Brian. 1987. “Dynamic Coherence and Probability Kinematics”. *Philosophy of Science* 54 (1):1-20.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation.* Harvard University Press.

Wagner, C. 2002. “Probability Kinematics and Commutativity”. *Philosophy of Science* 69 (2), 266–278.

Wagner, C. 2003. “Commuting Probability Revisions: The Uniformity Rule”. *Erkenntnis* 59 (3), 349–364.

CHAPTER III

CPK: Learning and Evidence

1. Introduction

In this chapter, I consider questions about whether CPK is a learning experience, and what we should think about the evidential character of the rates of change supplied by nature. I first argue that CPK is, indeed, a genuine learning experience in the following sense: as long as, prior to receiving the signals from nature, the agent's expected value for each signal is zero, the agent will regard the credence function obtained by updating by CPK on these signals from nature as expectedly better at making decisions than the agent's present credence function. This condition, that the agent's expected value of the signals be zero, is the analogue of the standard Martingale condition – it is a minimal constraint for the agent's current credence function to be justifiable, under the assumption that the agent's present credences should reflect her expectation of her future credences.

After proving that an agent that satisfies this condition will regard CPK as a genuine learning experience, I investigate the question of whether the rates of change stipulated by nature can properly be regarded as evidence. The main worry here is that there are prominent arguments from epistemologists like Timothy Williamson that evidence must be propositional in order to play the roles evidence fills in epistemic life. But it's not obvious that the content of the kind of signals that CPK responds to can be correctly regarded as propositional: if anything, the signals seem much more like *imperatives* than any kind of statement that encodes factive information about the world. My response to this worry is two-fold: first, I will argue that Williamson's conception of evidence is too narrow; he has successfully picked out constraints that apply to a certain *kind* of evidence – in particular, a kind of evidence that we expect in certain kinds of highly intellectualized discourse. But there is a much broader class of evidence consisting of, roughly, whatever plays the functional role of input to a learning experience in the sense that will shortly be explained. And there is no particular problem with understanding the signals that CPK responds to as evidence of this kind.

My second response will be to show that we can obtain superconditioning results for CPK. Thus, if I can't convince the reader that they should be comfortable with evidence that doesn't have propositional content, I will try to convince you that agents that update by CPK can also be modelled *as if* they are updating on propositional content in a richer space of propositions. Finally, the strategy of providing a constructive superconditioning result for CPK provides an interesting lens to examine the evidential commitments that an agent who updates by CPK has (or can be modelled as having). I use this lens to reexamine the question of the factivity of the kind of evidence with which CPK is concerned.

2. Learning Experiences and the Value of Information

One governing principle of genuine learning that many epistemologists accept is that the agent must regard her future self, after the experience is concluded, as being better-informed than her current self. As I've stated this requirement, it's ambiguous: do we require that the *future* agent regard herself after the experience as better-informed than she was before, or do we require that the *present* agent expects her future self to be better-informed than she is presently? Is it possible to have a genuine learning experience that you don't regard as one beforehand, but only after the fact? Viewed through the widely accepted lens of *Immodesty*, it turns out that the first proffered interpretation is trivial. An agent who satisfies *Immodesty* will *always* regard her own credence function as having the highest expected epistemic utility - and this will be true even in cases where the agent, beforehand, will regard her new credences as resulting from a mistake on her part. In cases where Immodesty applies, we cannot use the agent's retrospective comparison of her present and prior credences as any useful indication of whether her credences have improved, because she will *always* think they have.

However, the second understanding of the requirement is usually regarded as more substantive, and there are at least two major ways of formalizing it. One formalization is part of the foundation of a burgeoning research program in formal epistemology: the agent's *expected accuracy* or *expected epistemic utility*, as evaluated from her prior, must increase as a result of the experience. The idea that many epistemic norms can be derived from the assumption that agents are expected-epistemic-utility maximizers has recently been a popular and fruitful starting point for epistemological research. The other formalization is part of an older tradition of measuring the value of knowledge from its pragmatic consequences - the value of being better-informed is reflected in better practical decision-making, attaining greater *expected utility.* Dutch book arguments for diachronic constraints are rooted in this tradition of using the potential for good decision-making (e.g., bookmaking!) as a measure of epistemic success.

A particularly famous example of this older tradition is the Savage-Good Value of Information Theorem, proved in Good (1967). It shows that, in the context of Savage's decision theory, and assuming updating by Bayesian conditionalization, the expected utility of making a decision after a costless experiment where you learn some outcome with certainty is always greater-than-or-equal-to the expected utility of making the decision before the experiment. Skyrms (1990) gives an overview of a generalization of this concept, in the section "Black-Box Learning and Higher-Order Probabilities". Skyrms considers learning experiences of a much more general form:

> "Suppose that the bookie at t_1 has probabilities over some finite space, W, and anticipates an observational experience such that she cannot describe the possible observational results, or even specify a sufficient partition *à la* Jeffrey for the experiment. But she can think about how her probabilities at t_2 may have been modified by the observation, and we will suppose that at t_1 she also has prior probabilities over the possible posteriors that she may have at t_2 - in other words, over the space *W.*" (121-2)

Skyrms summarizes a Dutch book argument which shows that, to avoid Dutch books, an agent in such a situation must satisfy the Martingale relation:

$$p_1(w|p_2 = p^*) = p^*(w), \text{ for each } w \in W.$$

This Martingale principle has been widely discussed in the epistemology literature, although it is more commonly called the Reflection Principle. In an earlier section of Skyrms (1990), "Dynamic Probability and Learning Generalized", Skyrms argues that this condition represents "your belief in the epistemological validity of the impending belief change" (98), and gives a generalization of the Savage-Good theorem which shows that the expected utility of making a decision after a belief change that conforms to the Martingale principle is greater-than-or-equal-to the expected utility of making a decision on the basis of your prior.

In "Learning Experiences and the Value of Knowledge", Simon Huttegger proves what one might call a *reverse* Value of Information theorem. He takes the conclusion of the other Value of Information Theorems as his foundational premise. That is, Huttegger starts by assuming that an experience represents genuine learning *only if* the expected utility of a decision after the experience is higher than the expected utility of making the decision on the basis of your prior. "*Postulate.* If a belief change from p to $\{p_f\}$ constitutes a *genuine learning situation*, then $\sum_f p(p_f) \max_j \sum_i p_f(S_i)u(A_j \& S_i) \; \geq \; \max_j \sum_i p(S_i)u(A_j \& S_i)$"(285), where the A_j are the agent's available actions in the decision problem, the S_i are a partition of possible states of the world, and $u(A_j \& S_i)$ is the utility to the agent of performing act A_j in state S_i. From this assumption, Huttegger derives the Martingale principle. Taking Skyrms (1990) and Huttegger together, we have that the Martingale principle is both *necessary and sufficient* for regarding your future self, after some credal change, as being better-informed than your present self - as measured by ordinary expected utility.

The Martingale Principle and Gradual Learning

What is the relevance of the Martingale principle to CPK? Well, superficially at least, the Martingale principle seems to present a challenge to the idea of gradual learning experiences. The philosophical foundation upon which the CPK formalism has been constructed presupposes the idea that gradual learning experiences are not only possible, but in some important cases provide a more useful model than typical instantaneous/discontinuous learning. To notice the apparent problem, it's helpful first to observe that the Martingale principle entails a requirement that your present credence in some proposition be equal to your expectation of your future credence in said proposition. For now, let's assume that there is a finite set $\{c^i\}$ of credence functions such that you are certain that your credence function after the learning experience, c_2 , will be one of them: $c_2 = c^i$, for some c^i. Assuming you have credences representing how likely each of the c^i are to be your posterior, your expectation of your posterior credence in some proposition x is

$$E[c_2(x)] = \sum_i c_1 \left(c_2 = c^i\right) c^i(x) \tag{1}$$

By the Martingale principle,

$$c^i(x) = c_1\left(x \middle| c_2 = c^i\right) \tag{2}$$

and so,

$$E[c_2(x)] = \sum_i c_1 \left(c_2 = c^i\right) c_1\left(x \middle| c_2 = c^i\right) = c_1(x), \tag{3}$$

by the Law of Total Probability. If we define $\Delta c(x) = c_2(x) - c_1(x)$, the Martingale principle entails that the expected change in your credence, $E[\Delta c(x)]$, in *any* proposition x over *any* finite interval of time (t_1 , t_2) must be zero! This is because

$$E[\Delta c(x)] = E[c_2(x)] - E[c_1(x)] = c_1(x) - c_1(x) = 0. \tag{4}$$

If the Martingale principle is a necessary condition for a *genuine learning* experience*,* this means that there is no genuine learning where an agent expects her credence to gradually increase over some finite interval of time. The intuitive reason for this is that the Martingale principle is an expert-deference principle. If you expect to be better-informed by the end of the interval, and you have a (non-zero) expectation of how your credence will change by the end of the interval, you're rationally required to adopt that change *as soon as possible.* As showcased by the diachronic Dutch book arguments given by Skyrms, delaying just exposes you to losses that could be avoided by adopting the better credences immediately. Or, again in (arguably) more purely epistemic terms: if you regard all of the changes over the interval as epistemic improvements, so that for each $t', t'' \in (t_1 , t_2)$ such that $t'' > t'$, your doxastic state at t'' will be better than your state at t', you should regard your state at , t_2 as best, in which case there's no reason not to adopt it immediately - to the extent to which you can predict what it will be. And a credal change that consisted of specifying the $\Delta c(x)$ for the interval would allow you to predict what your final credence on the interval would be with certainty, assuming you know your updating procedure. It's, of course, a trivial consequence of this requirement that specifying a constant rate of change on some finite interval also cannot be regarded as a genuine learning experience. The Martingale principle requires instantaneous updating; as soon as you know what your final credences would be, those are the credences you should have.

This supposed problem for CPK, though initially compelling, misses a crucial aspect of its formalism. Although I have shown, at some length, that the final resultant credence functions from updating by CPK will be equivalent to the results of a single, instantaneous (soft) Jeffrey shift on the refined partition, *at no stage of the process is the agent who updates by CPK given information like this*. An agent who updates by CPK is given *instantaneous rates of change* for some of her credences, and she learns the rate of change for the present moment *as that moment occurs*. CPK does not consist of being given a credal change to accomplish by the end of some interval but is governed by a differential equation. Formally, it's equivalent to being given an infinite number of infinitesimal shifts that are each accomplished infinitely quickly. Conceptualizing CPK this way, we can show that CPK is consistent with an infinitesimal version of the Martingale principle and that, when this principle holds, each instantaneous CPK credal change will be regarded by the agent as a genuine learning experience.

As in much of this dissertation, I will consider the two-directly-varied proposition case of CPK. The generalization to when the agent is being directly given information about more propositions is readily apparent. At the present moment, the agent has a credence function c defined over a σ -algebra of propositions, including A and B; the agent's algebra also contains a set of hypotheses about the state of the world, $\{ h_i\}$, which is a partition. In the near future, the agent is going to be given a decision problem, with a finite set of of possible actions $\{A_j\}$. The utilities to the agent of the outcomes of her decision are completely determined by which action she performs in which state of the world. The expected utility of an act, given her present credences, is thus $EU(A_j) = \sum_i c(h_i)u(A_j \wedge h_i)$. Additionally, the agent is about to undergo an infinitesimal CPK shift, which we can think of as being comprised of conforming her credences simultaneously to two experimental results. Nature is going to specify two instantaneous rates of change $\frac{\partial c(A)}{\partial t}, \frac{\partial c(B)}{\partial t} \in \mathbb{R}$ and the agent is going to perform an infinitesimal CPK shift, using the stipulated rates of change to calculate her new credences after an infinitesimal time period, dt. This yields final credences $c_f(x) = c(x) + \frac{dc(x)}{dt}dt = c(x) + [c(x|A) - c(x|\neg A)]\frac{\partial c(A)}{\partial t}dt + [c(x|B) - c(x|\neg B)]\frac{\partial c(B)}{\partial t}dt$. Because $\frac{\partial c(A)}{\partial t}$ and $\frac{\partial c(B)}{\partial t}$ each have continuum-many possible values, we cannot require that the agent has credences defined over the possible values results of the two experiments. However, we can and do assume that the agent has a credal *distribution* over the possible results of each experiment: $\rho\left(\frac{\partial c(A)}{\partial t} = r, \frac{\partial c(B)}{\partial t} = s\right)$, which I will henceforth abbreviate as $\rho(r,s)$. We can obtain the agent’s marginal distribution over the various r-values by “integrating out” the s-values: $\rho_A\left(\frac{\partial c(A)}{\partial t} = r\right) = \int_{-\infty}^{\infty} \rho(r,s)\, ds$. Similarly, $\rho_B\left(\frac{\partial c(B)}{\partial t} = s\right) = \int_{-\infty}^{\infty} \rho(r,s)\, dr$. The distribution ρ is normalized, so that $\int_{-\infty}^{\infty} \rho_A(r)dr = \int_{-\infty}^{\infty} \rho_B(s)ds = 1$. We'll notate the final credences resulting from the shift after learning experimental results $\frac{\partial c(A)}{\partial t} = r$ and $\frac{\partial c(B)}{\partial t} = s$ as $c_{rs}(x) = c(x) + [c(x|A) - c(x|\neg A)]r\, dt + [c(x|B) - c(x|\neg B)]s\, dt$. Now that we have this established, we can pose the question: suppose the agent were able to choose whether to make the decision immediately, or to first learn the experimental results and then make the decision. Which would give her higher expected utility?

If we make a pair of assumptions analogous to the Martingale principle in this infinitesimal case, the answer is that we *can* prove that the agent will regard CPK as a genuine learning experience in much the same fashion as the Value of Information Theorems discussed above. The agent's expected utility for first obtaining the experimental results and then making the decision will always be greater-than-or-equal-to making the decision on the basis of her current credences. So, what are the two assumptions? That the expected value of the stipulated instantaneous rates of change are both 0:

$$E[r] = \int_{-\infty}^{\infty} r\rho_A(r)dr = 0 \tag{5}$$

$$E[s] = \int_{-\infty}^{\infty} s\rho_B(s)ds = 0 \tag{6}$$

This pair of assumptions together comprise a weaker assumption than the standard Martingale principle. As shown above, the standard Martingale principle entails (4), which demands that the expected credal change in *any* proposition is 0. (5) and (6) only demand that agent's expectation of the (instantaneous) *directly stipulated* changes be 0. And the rationale for (5) and (6) is very similar to the standard Martingale principle as a kind of expert-deference principle requiring that an agent adopt credences she believes to be better than her own. We are assuming that, before learning the values of $\frac{\partial c(A)}{\partial t}$ and $\frac{\partial c(B)}{\partial t}$ from nature, the agent endorses her present credences - she does not see them as requiring updating. But if, prior to learning the experimental results, she had non-zero expectations for either $\frac{\partial c(A)}{\partial t}$ or $\frac{\partial c(B)}{\partial t}$, this would not generally be the case. She would see her credences which depend on either $c(A)$ or $c(B)$ as needing revision - and she would already have begun the process of moving her credences in the direction of expected adjustment, before even having the learning experience. Although (5) and (6) alone are weaker than the standard Martingale principle, we can show that (5), (6), and CPK together entail the analogous result to (3):

$$E[c_{rs}(x)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c_{rs}(x)\rho(r,s)\;dr\;ds \tag{7}$$

$$E[c_{rs}(x)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (c(x) + [c(x|A) - c(x|\neg A)]r\;dt + [c(x|B) - c(x|\neg B)]s\;dt)\rho(r,s)\;dr\;ds \tag{8}$$

We can rewrite the RHS of (8) as a sum of three integrals:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(x)\rho(r,s)\;dr\;ds + \tag{$I1$}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [c(x|A) - c(x|\neg A)]r\;dt\;\rho(r,s)\;dr\;ds + \tag{$I2$}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [c(x|B) - c(x|\neg B)]s\;dt\;\rho(r,s)\;dr\;ds \tag{$I3$}$$

If the Martingale constraints (5) and (6) hold, (*I*2) and (*I3)* both vanish. I'll work through why (*I*2) must be zero; the reasoning for (*I3)* is exactly parallel.

$$(I2) = [c(x|A) - c(x|\neg A)]dt \int_{r=-\infty}^{r=\infty} r \int_{s=-\infty}^{s=\infty} \rho(r,s)ds\;dr$$

$$(I2) = [c(x|A) - c(x|\neg A)]dt \int_{r=-\infty}^{r=\infty} r\,\rho_A(r)dr$$

Then from (5), we have $(I2) = 0$. Because $(I3) = 0$ for very similar reasons, we are left with

$$E[c_{rs}(x)] = c(x) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(r,s)drds\,.$$

Because $\rho(r,s)$ is normalized, we finally have:

$$E[c_{rs}(x)] = c(x), \tag{9}$$

which is the equivalent of (3). The expected value of an agent's posterior credence in any proposition x is the same as her current credence in x. Just as (3) entails (4), (9) immediately entails:

$$E[\Delta c(x)] = E[c_{rs}(x) - c(x)] = 0. \tag{10}$$

Now, the expected utility of making the decision based only on the agent's current credences, without learning the experimental results is:

$$EU(D) = \max_j \sum_i c(h_i)u(A_j \wedge h_i). \tag{11}$$

The expected utility of making the decision after learning the results of the experiment is:

$$EU(E \wedge D) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \max_j \sum_i c_{rs}(h_i)u(A_j \wedge h_i)\rho(r,s)drds. \tag{12}$$

Using (9), we can rewrite (11) as:

$$EU(D) = \max_j \sum_i \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} c_{rs}(h_i)u(A_j \wedge h_i)\rho(r,s)drds. \tag{13}$$

By Tonelli's theorem,[57] we can rewrite (13) as:

[57] Tonelli's theorem gives a sufficient condition for when nested integrals can be interchanged, resulting in equivalent expressions. Because discrete sums are, of course, just a special case of integration, Tonelli's theorem also gives a sufficient condition for interchanging sums embedded in integrals with integrals embedded in sums. In this application, Tonelli's theorem states that if $f_n(x)$ is non-negative for all n and all x, then $\sum_n \int_X f_n(x)\,dx = \int_X \sum_n f_n(x)\,dx$. Because utility functions are defined only up to positive affine transformation, we are free to represent the agent's utility function so that $u(A_j \wedge h_i)$ is always non-negative. $c_{rs}(h_i)$, $\rho_A(r)$, and $\rho_B(s)$ are all required to be non-negative in virtue of being credences/credal distributions, so $c_{rs}(h_i)u(A_j \wedge h_i)\rho_A(r)\rho_B(s)$ is always non-negative. Thus, we can interchange the outer sum with both integrals without changing the value of the expression.

$$EU(D) = \max_j \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sum_i c_{rs}(h_i) u\big(A_j \wedge h_i\big) \rho(r,s) dr ds \,. \tag{14}$$

Now, we follow Skyrms (1990, p.98) in noting that, if we define $B(c) = \max_j \sum_i c(h_i)\, u(A_j \wedge h_i)$, $EU(D) = B[E(c_{rs})]$, while $EU(E \wedge D) = E[B(c_{rs})]$. Because B is convex, by Jensen's inequality, $EU(E \wedge D) \geq EU(D)$.

And, so we have a Value of Information Theorem for CPK. If the agent satisfies conditions (4) and (5), analogues of the Martingale principle, she will regard each infinitesimal CPK shift as leaving her better-informed in the sense of allowing her to make better practical decisions.

3. CPK and the Propositionality of Evidence

Some epistemologists will be skeptical of the argument that I’ve just given on the following grounds: genuine learning experiences can only occur as a response to evidence, and some may find it difficult to believe that the kinds of rate of change signals from nature which I have been concerned with can count as evidence. The primary source of this skepticism, I believe, is the view that evidence must be propositional in order to fulfill the functional roles that we expect it to play. And from what I have said so far, it may be hard to conceive of the content of the rate of change signals as being propositional. In *Knowledge and Its Limits* 9.5: “Evidence as Propositional”, Timothy Williamson gives three major arguments for thinking that evidence must be propositional:

1. **Inference to the best explanation.** We often use evidence to discriminate between hypotheses on the basis of which hypothesis best explains our evidence. But the relata in the explanation (why) relation, Williamson claims, must be propositions. Thus, evidence must be propositional.
2. **Probabilistic reasoning.** A typical standard for probabilistic evidence is that E is evidence for H iff $P(H|E) > P(H)$. But the probability function P being referred to, whatever it is, is defined over propositions. So E must be propositional.
3. **Restricting possibilities / ruling out hypotheses.** Evidence sometimes reduces what propositions (hypotheses, in particular) we must consider as possible, when our

evidence is inconsistent with the propositions to be rejected. This kind of inconsistency is a feature of propositions, and so evidence must be propositional to play this role.

In this part of the chapter, I want to briefly discuss some worries I have about Williamson's arguments in favor of this point. But before I turn to my thoughts about these arguments, I should point out that Williamson's claim that evidence must be representable as propositional in form has, I think, already been refuted. In *Probabilistic Knowledge,* Sarah Moss compellingly argues that probabilistic contents (understood as sets of probability spaces) can themselves directly play many of the functional roles that epistemologists had previously thought to be reserved to propositional content. Among other things, probabilistic contents can be the contents of belief, can be the objects of assertion, can be evidence, and can be knowledge. If this is correct, and I think it is, then Williamson is wrong that evidence must be propositional, because there will be kinds of evidence on Moss's view that have *thoroughly probabilistic*[58] content, and so cannot be adequately represented by any proposition. However, this refutation of Williamson's claim is in more or less the opposite spirit of the problems that I have with it. Moss's framework establishes that the kinds of objects that can play the functional role of evidence can have a *richer* structure than merely propositional content. Probability spaces are ordered triples, where one of the elements of the triple is an *algebra of propositions* (p. 2, fn 2). As already mentioned in my footnote 57, probability spaces are strictly *more expressive* than merely propositional content – any content expressible as a proposition is representable as a nominally probabilistic content, but there are many thoroughly probabilistic contents that cannot be represented as propositions. My thought is that information with representational structures that do not remotely resemble propositions may nonetheless be evidence; I believe that this includes information with much *less* representational content and structure than the framework of propositions provides.

[58] See *Probabilistic Knowledge,* p. 14 for the introduction of this terminology. The point is that sets of probability spaces include propositional content as degenerate cases: there are sets of probability spaces that assign only extremal credences, and so function to pick out some possible worlds. *Nominally probabilistic* is the term that Moss uses for this kind of content; all other sets of probability spaces are *thoroughly probabilistic.*

One general, overarching worry that I have is that Williamson's arguments rely on a confusion between formal representations of structures of reasoning and the structures being so represented. I take the easiest demonstration of this problem to be the second argument concerning probabilistic reasoning. Reasoning probabilistically, in the sense of treating E as evidence for H just in case $P(H|E) > P(H)$, does not require understanding what probabilities are. It merely requires having some representational states that function like probabilities, and of having systematic procedures for changing these representational states in the appropriate way when provided with evidence. Similarly, although the way we would formally model the probabilities that such a system employs uses variables that represent propositions, the system itself need not represent those propositions. The dispositions to use the probability-like representations to decide on certain actions and to update the representations when given certain kinds of stimuli will lead *us (theorists about the system)* to label the representations as encoding probabilities about certain propositions, but the system need have no idea what these propositions are. And, thus, there is no reason to think that propositions are what such a system is responding to as evidence.

What does count as evidence for such a system? My suggestion is very simple: whatever functions as an input to a genuine learning experience. I will not explore this at any great length here, but I think there are obviously related worries for the other two arguments. An agent that performs reasoning that conforms to the schema of inference to the best explanation need not model itself as doing so, and so need not see its evidence as giving a propositional why-justification; an agent that correctly decides that certain possibilities are no longer worth worrying about on the basis of some evidence need not understand that reduction in possibility-space as a feature of the inconsistency of propositions. Our understanding of propositional content is certainly a highly useful modelling tool in understanding why various kinds of reasoning processes are justifiable, but systems can act according to justified schemas without self-modeling the processes in this particular way.

The reader may wonder if this objection is actually fair to Williamson: isn't his claim merely that evidence "*consists of*" propositions (p. 197), not that any particular system that is capable of processing evidence must *regard* it through the lens of propositionality? I'm genuinely not sure. Throughout this section of the book, Williamson not only claims that evidence must be propositional, but that the propositions must be *grasped* by the agents in question to count as evidence for the agent. It is not entirely clear what "grasping" entails for Williamson, but it at least sometimes seems like it requires something like the ability to express the proposition in some language of thought. E.g., "one grasps the propositions that are one's evidence; one can think them" (p.194). And again, on p. 195, Williamson ties the idea of grasping a proposition to the ability to *use a proposition in an explanation*, where context suggests that this is to be understood as something very much like a verbal/linguistic exercise: "One can use an hypothesis to explain why A only if one grasps the proposition that A." In the previous paragraph, in discussing the sense in which a bloodied knife can be evidence in a court of law, he argues that it is the use the knife is put to in the theories of the case advanced by the prosecution and defense which make the knife a source of evidence, qua source of various evidential propositions: "…the bloodied knife provides evidence because the prosecution and defence **offer competing hypotheses**" (p. 195, my emphasis in bold, original emphasis in italics). In the context of discussing point 3 labelled above, Williamson argues:

> "In particular, our evidence sometimes rules out some hypotheses by being *inconsistent* with them. … But only propositions can be inconsistent in the relevant sense. If evidence *e* is inconsistent with an hypothesis *h* in that sense, it must be possible to *deduce* $\sim h$ from *e;* the premises of a deduction are propositions. **Moreover, the subject who deduces $\sim h$ from *e* must grasp *e.*"** (p. 196, my emphasis in bold)

So, in context, Williamson is arguing that evidence must be propositional because only propositions can play the appropriate role in logical deduction that he believes evidence must play in order to rule out hypotheses. But he is also arguing that only grasped propositions can function as evidence, and one of his pieces of evidence for that claim, this last bolded claim, seems to me to very strongly suggest that he regards the ability to perform a certain kind of deduction, in some language with the structure of propositional logic, as a precondition for being able to use evidence in the functional role of ruling out hypotheses.

Although the passages that I have cited so far seem to support the idea that "grasping" is, for Williamson, a very cognitively rich relation, often seeming to involve the ability to express the concepts involved in the potentially grasped proposition, there is also some evidence against this interpretation. In discussing whether certain very simple creatures, which may lack the concept of appearance can grasp the proposition that something appears in a certain way, he suggests a dispositional account of grasping that seems significantly less demanding than the kind of grasp suggested by the earlier passages:

> "Although one's grasp of the property of appearance may be inarticulate, one must have some inkling of the distinction between appearance and reality. For instance, one should be willing in appropriate circumstances to give up the belief that things were that way while retaining the belief that they appeared to be that way. In the absence of such dispositions, it is implausible to attribute the qualified belief that thing appear to be that way rather than the unqualified belief that they are that way." (p. 199)

And so the net effect is that I am quite unsure what exactly Williamson intends grasping to amount to, and hence fairly dubious about whether he thinks that agents must be capable of representing the content of propositions that constitute their evidence in order to count as possessing that evidence.

Still, whether it's Williamson's view or not, the reader might press: even if my objection is correct, don't these three arguments give us some reason to think that if *external theorists* (like us) look at something that can play an evidential role in some genuine learning process, we should be *able to represent it* as having propositional content? I think this is a very good question (and might even be Williamson's original point!), but the best answer I have is that I am deeply unsure. My uncertainty stems from the fact that I am still worried that the model of evidence underwriting the three arguments is based in highly sophisticated traditions of evidence that involve a great deal of meta-analysis. The discussion revolves around questions like "what kinds of things can answer why-questions?" and "what are the sorts of mental objects that might be able to eliminate hypotheses?" These kinds of questions are, of course, hugely important to the functional role that evidence plays in many human intellectual disciplines and in socially-structured, norm-governed decision making (e.g., law). But I think it's crucial to notice that an important aspect of these disciplines is that debates about the nature and standards of evidence are *part of the ordinary practice of these disciplines*. And so my worry is that when we think about what features evidence must have to fulfill the functional role that it plays in this kind of discipline, we are not merely investigating what representational features are necessary for something to serve as evidence – we are imposing a much stricter requirement. We are actually asking what representational features are necessary for the kinds of mental objects that can be *analyzed as evidence* in these kinds of systems. We are asking what features are necessary to *argue* that some putative evidence *is evidence*, to *explain why* some evidence generates the support that it does, etc.; these are much more demanding tasks than merely playing the functional roles of leading to better-informed, more accurate beliefs and decisions (or whatever basic conception of evidence the reader may wish to substitute). And I think this tendency to think about this kind of highly intellectualized use of evidence, where that includes meta-analysis of the evidence, in certain kinds of social epistemic systems as the model of evidence's basic functional features explains the modelling/modelled confusion that I have attempted to identify in the past several paragraphs.

4. Superconditioning on the (Propositional) Content of Experience

Although I am skeptical of the source of our intuitions that evidence should be *representable as propositional* by theorists external to the agent using said evidence, I do not have any compelling arguments against this claim. The claim that I do strongly disagree with, whether it is actually Williamson's or not, is that possession of evidence requires the ability *by the agent* to represent this propositional content, and to make use of it in exercises like logical deduction, answering why-questions, or verbally identifying the elements of the space on which their probabilities are defined. For all that I have said in this section so far, it may well be that any putative evidence must be representable by external theorists in such a way that it is, e.g., possible to use it answer why-questions, etc. So, now I turn to the question of what the proponent of CPK can say about how to represent credal rate-of-change signals as propositional.

The first place to look is, of course, at what Richard Jeffrey had to say about the propositional contents of the kinds of evidence at play in his learning experiences. In *The Logic of Decision,* Jeffrey explains that the reason to model the agent as directly acquiring some new degrees of belief, is not that it is *in principle impossible* to regard the agent's sensory experiences as containing some kind of propositional content – it's that there's no reason to think that the agent *knows what this content is*, and thus that's it's implausible to regard the agent as updating on priors over it.

> "In all such cases there is some definite quality of his sensuous experience which leads the agent to have various degrees of belief in the various relevant propositions; but there is no reason to suppose that the language he speaks provides the means for him to describe that experience in the relevant respects. … and even if [the experience is describable in his language], there is every reason to suppose that the agent is quite unaware of what that pattern is and is quite incapable of uttering or identifying a correct description of it." (p. 166)

And so Jeffrey thinks that understanding such a learning process through the lens of Jeffrey updating is more useful *as a norm that is intended to consciously guide an agent's process of belief revision:* "To serve its normative function, the theory of decision must be used by the agent, who therefore must be able to formulate and understand the relevant propositions" (167). Asking an agent to perform a Jeffrey shift is, the thought goes, a much more plausible task than asking an agent to conditionalize on some proposition that we should very much doubt the agent as being able to even identify.

Of course, whatever it makes sense to require of the agent, there is a separate question about what kinds of modelling are accessible to external theorists: for any learning experience representable as a Jeffrey shift, can a theorist choose to think of the agent's learning process *as if* it were some kind of Bayesian conditionalization on some kind of propositional content? The answer to this question is famously yes – as long as the theorists are comfortable with modelling the conditionalization as taking place on a greatly expanded algebra that they may have no reason to believe that the agent has priors over. This is Diaconis and Zabell's proof of what Jeffrey called *superconditioning* (Diaconis and Zabell, "Updating Subjective Probability" p.

824; Jeffrey, *Probability and the Art of Judgment* pp. 128-9). (Somewhat) informally, a probability function Q is obtainable by superconditioning from probability function P whenever you can:

1. "Translate" P into a probability function M defined on a larger σ algebra.
2. Update M on some element of the larger algebra by Bayesian conditionalization to obtain probability function N.
3. N on the larger algebra corresponds to Q on the algebra P was defined on.

Diaconis and Zabell proved that Q is obtainable by superconditioning from P iff there exists some $b \geq 1$ such that $\frac{Q(\omega)}{P(\omega)} \leq b$, for all elements ω of the algebra Ω that Q and P are defined on (p. 824). In particular, this shows that *any* Jeffrey shift defined over *finitely many* propositions can be obtainable by superconditioning.

One way of thinking about this result is that it gives us a proof that any learning process that is representable as a Jeffrey shift (over finitely many propositions) is also representable as a learning experience that consists of learning some propositional content – although, from what we have said, it is not at all clear why we should be convinced that the proposition that we use to perform the update on the larger algebra correlates in any way with our intuitions about what the content of the agent's learning experience plausibly might be. Nonetheless, this kind of result at least lends some credibility to the idea that Jeffrey shifts can be thought of as involving propositional evidence. Of course, as the previous quotes from Jeffrey suggest, the case is also bolstered by the fact that we have an intuitive grasp of what the representational content of an agent's learning experience might be in the cases of sensory learning that he is concerned with. These are propositions that represent the agent as having experienced certain visual perceptions, having heard sounds with certain characteristics, etc. So, next I will show that the Diaconis and Zabell superconditioning result can be shown to apply to CPK *on any finite algebra*; afterwards, I will try to give an intuitive sketch of how we might relate the intuitive propositional content of sensory experiences to the rate-of-change signals that CPK updates on.

First, the proof: let P be an agent's coherent credence function prior to some CPK learning experience, and let Q be the result of the learning experience. Because P is coherent, and updating by CPK preserves coherence, Q must also be a probability function. Assume further that P is defined over some finite algebra $\mathcal{A}$. The existence of an upper bound is then trivial: consider the ratio $\frac{Q(\omega)}{P(\omega)}$ for all $\omega \in \mathcal{A}$. Because there are finitely many such ratios, there will be a greatest. Let $b = \max \frac{Q(\omega)}{P(\omega)}$; by definition, $\frac{Q(\omega)}{P(\omega)} \leq b$ for all $\omega \in \mathcal{A}$. We are also guaranteed that $b \geq 1$ by the coherence of P and Q. Consider some tautology $\Omega \in \mathcal{A}$. $\frac{Q(\Omega)}{P(\Omega)} = 1$, and $\frac{Q(\Omega)}{P(\Omega)} \leq b$, so $b \geq 1$. The conditions of Diaconis and Zabell's proof are met, and thus Q can be obtained from P by superconditioning.

What might the propositional content of a CPK learning experience consist in, and how should we think of this content as related to the rate-of-change signals that the agent is updating on? Consider an agent who is interested in whether a particular piece of cloth is blue or not. The agent is getting a continuous stream of sensory representations of the cloth, which we can think of as continuum-many visual representations, one for each moment in some interval of time. We might also idealize the agent as having, at each moment in time, a certain kind of *conditional visual expectation* that somehow characterizes the kind of experience they would expect to have if the cloth is blue. There are many other kinds of visual experience that the agent would regard as *not matching* this expectation: these are experiences that are different enough from what the agent expects to see if the cloth were blue, and similar enough to what the agent might expect see if the cloth were some color other than blue, that the agent will regard these images as *evidence against* the cloth being blue. How strong is each of these pieces of evidence? Perhaps we can use some measure of similarity to the blue *conditional visual expectation*: when the observed images are very good matches for typical blue experiences, the agent is more confident that the cloth is blue. If the agent is just ambivalent between classing the image as a match with the blue expectation vs. a match with what they would expect to see if the cloth weren't blue, perhaps this would correlate to a Jeffrey experience of assigning credence 0.5 to both blue and not blue.

Of course, rather than processing each image fully and comparing it separately against the visual expectation, we can imagine a system that is primarily interested in *easily observable differences between the images*. So, as each new image comes in, some quick processing delivers a judgment as to whether the just received image is *more or less similar* to the blue expectation than the agent's current model of the cloth, and then estimates the *magnitude of this change in similarity*. Thus, rather than directly tracking the similarity of each new image to the blue expectation and performing the appropriate Jeffrey shift, the agent could track whether the similarity of images to blue is increasing or decreasing, and how quickly this change in similarity is occurring. And if the agent chooses to process this visual information as increasing/decreasing similarity to the blue expectation, the natural way to understand how this should affect their credences is as a rate that their credence that the cloth is blue is changing at with respect to time; this way of processing the information naturally lends itself to thinking about the temporal derivative of their confidence that the cloth is blue.

So, at each moment that it might be appropriate to model an agent as receiving a signal that $\frac{dc(B)}{dt} = x$, we could instead choose to think of the agent as instead learning some propositional content derived from the image that she is processing. But, of course, there is really no reason to think that the agent has access to that content in all its overabundant richness of detail – just as Jeffrey argued about the cases of sensory experiences that he discussed. And, although we might be able to formally model the agent as updating as if they were conditionalizing the content of this experience on some algebra, it will not – in general – be an algebra that we have any reason to believe they have credences defined over.

The preceding sketch of how the propositional contents of visual perception might plausibly generate the kinds of credal derivatives with which CPK is concerned is, unapologetically, extremely speculative. My only real goal in providing it is to help the reader imagine that some such process might be possible. However, there is some support for the basic ideas underwriting the suggestion that can be found in the psychology and neurobiology literature. To begin with, there is a fairly long tradition of arguing that the raw signals from sensory organs are to be regarded as tests of predictions that the agent makes prior to observation, with the processed contents that are visually presented to the agent containing extrapolations, interpolations, and predictions. The thought is that this processed content is generated by some kind of *inference* from the raw data that makes use of the agent's prior beliefs. For instance, in "Perceptions as Hypotheses", the psychologist R.L. Gregory writes:[59]

> "Are perceptions like hypotheses of science? … It may be said that hypotheses *structure* our accepted reality. More specifically, it may be said that hypotheses allow limited data to be used with remarkable effect, by allowing interpolations through data-gaps, and extrapolations to be made to new situations for which data are not available. These include the future. … I shall hold that all these statements are true, and that they apply to perception. In addition, both the hypotheses of science and the perceptual processes of the nervous system allow recognition of familiar situations or objects from strictly inadequate clues, as signaled by the transducer-instruments of science and the transducer-senses of organisms. … To suggest that perceptions are like hypotheses is to suppose that the instruments and the procedures of science parallel essential characteristics of the sense organs and their neural channels, regarded as transducers transmitting coded data; and the data-handling procedures of science may be essentially the same as the cognitive procedures carried out by perceptual neural processes of the brain" (pp. 181-2)

The idea that these predictions of what will be observed take a specifically probabilistic form, and that systems are most interested in tracking deviations from these predictions, are important aspects of the paradigm of predictive coding in both psychology and neuroscience. In "Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects", Rao and Ballard argue that there are visual effects in processed images that are well-explained as resulting from hierarchical processing of images involving detection of discrepancies ("errors") between predictions by higher levels of the activity of lower levels and the actual neural

[59] Interestingly, in the same article, Gregory also expresses skepticism about whether the representational content of "perceptual hypotheses" must be propositional: "The example of a graph illustrates that hypotheses – for the accepted curve or function may *be* a predictive hypothesis – can be *non-propositional*. Perhaps hypotheses are generally thought of as sets of propositions, but there seems no reason to restrict hypotheses to propositions as expressed in language. … There seems no reason to hold that 'perceptual hypotheses' require a propositional brain language, underlying spoken and written language, though this might be so" (p. 186).

activity at lower levels of the hierarchy. In the abstract for "Bayesian Surprise Attracts Human Attention," Itti and Baldi write:

> "We propose a formal Bayesian definition of surprise to capture subjective aspects of sensory information. Surprise measures how data affects an observer, in terms of differences between posterior and prior beliefs about the world. Only data observations which substantially affect the observer's beliefs yield surprise, irrespectively of how rare or informative in Shannon's sense these observations are. … Bayesian surprise is a strong attractor of human attention, with 72% of all gaze shifts directed towards locations more surprising than average, a figure rising to 84% when focusing the analysis onto regions simultaneously selected by all observers. The proposed theory of surprise is applicable across different spatio-temporal scales, modalities, and levels of abstraction."

I am not claiming that the details of any of the accounts that I am citing go any significant way towards confirming that my just-so story about how the contents of perception might generate signals encoding credal derivatives is psychologically realistic; I merely hope the reader is convinced that such a connection is somewhat plausible.

Further, we have the result that the final product of updating by CPK is always some Jeffrey shift on the refined partition – that is, if A and B are the propositions that the agent receives direct signals about, then the entire outcome of the learning experience will always be representable as some Jeffrey shift on the partition $\{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\}$. But, of course, that means we can also represent the outcome of the *entire experience* as the result of superconditioning. What might the content of such a proposition be? Well, again, we have no guarantee from the formalism that the proposition that plays the role in superconditioning will be related in any obvious way to our intuitive understanding of the propositional content of the learning experience. But if our goal is just to sketch what kind of content we might find it *plausible* to play such a role, there is certainly an obvious candidate: the conjunction of the content of all of the sensory impressions that the agent received over the interval. Again, this will be a conjunction with continuum-many terms, and so there is absolutely no reason to expect any finite agent to even be able to represent this content – let alone do anything useful with it! But if our goal is just to satisfy epistemologists that demand that evidence must be in-principle representable as some kind of proposition, this intractability should be of no special concern.

5. Discretized-Signal CPK: A Superconditioning Result

One very natural thought is that we should be able to represent the result of a CPK shift as obtainable by superconditioning on the partition whose elements are the propositions specifying the pairs of signals from nature (r, s) that specify rates of change $\left(\frac{\partial c(A)}{\partial t}, \frac{\partial c(B)}{\partial t}\right)$. The content of these propositions are $\frac{\partial c(A)}{\partial t} = r$ and $\frac{\partial c(B)}{\partial t} = s$. Unfortunately, I'm not quite sure how to make this idea work, because we've been assuming that $r, s \in \mathbb{R}$ and that the agent has a continuous probability *distribution* $\rho(r, s)$ over the values these signals might take. The agent's credence in any specific ordered pair of precise values $c(r, s)$ should thus presumably be zero, and so it is impossible to represent the agent as updating on the elements of this partition via standard Bayesian conditionalization. However, if we constrain the possible values of (r, s) to a *finite grid* so that we allow some finite number of signals (r_i, s_j), we can approximate any single infinitesimal CPK shift by conditionalization on the elements of this discretized grid, to any desired level of accuracy.[60] For finite, cognitively limited agents such as ourselves, we will only be capable of representing finitely many different signals anyway, and so the choice of step size between signals could be chosen to match the cognitive capabilities of the system in question. There may similarly be cognitive limitations on the maximum and minimum values of signals representable by the system (perhaps generated by facts about the cognitive architecture, facts about the physical implementation of the system, or both). Also, as we will see shortly, if we assume that the system is approximating the CPK update on some small finite timescale, rather than in infinitesimal time, this will also naturally place constraints on the maximum and minimum value of the signals that the system will regard as consistent

[60] Throughout the entirety of this section, I will be discussing a CPK shift with respect to two propositions, but the generalization to CPK shifts with respect to a larger number of propositions is obvious.

with maintaining coherence – and thus of needing to be entertained as signals that might be part of a genuine learning experience.

First, we assume that the algebra $\mathcal{A}$ that the agent's prior credence function, c, is defined over has some number of finite atoms. We interpret the atoms as truth assignments to some finite number n of propositions P_i, $i \in \{1, \dots, n\}$. The atoms are propositions of the form $X_1 \wedge \dots \wedge X_n$, where each of the X_i is what is often termed a *literal*: each X_i consists of either P_i or its negation, $\neg P_i$. There are 2^n such atoms, corresponding to the 2^n possible truth assignments to the P_i. The algebra is closed under negation and disjunction. As in earlier in the chapter, we assume that the agent has some prior distribution $\rho(r, s)$ over the values that might be specified in an infinitesimal CPK shift with respect to two propositions in this algebra, A and B. Now, consider the *discretization* of $\rho(r, s)$ onto a grid with $l \times m$ elements, where l represents the number of allowed signals for *r* and and m is the number of allowed signals for *s*, $l, m \in \mathbb{N}$. Let $c(r_i)$ be some suitable discretization of $\rho_A(r)$, for $i \in \{1, \dots, l\}$, and $c(s_j)$ be some suitable discretization of $\rho_B(s)$, for $j \in \{1, \dots, m\}$. We require that $c(r_i)$ and $c(s_j)$ satisfy discrete analogues of the Martingale constraints (5) and (6):

$$E_{c(r_i)}(r) = \sum_{i=1}^{l} r_i c(r_i) = 0 \tag{15}$$

$$E_{c(s_j)}(s) = \sum_{j=1}^{m} s_j c(s_j) = 0 \tag{16}$$

The pseudo-credences $c(r_i)$ and $c(s_j)$ must satisfy these constraints for the same reason that (5) and (6) must hold: if the agent had non-zero expectations for the values of the signals that she is about to receive in the learning experience, an agent with those credences would regard the agent's current credence function[61] as already in need of revision before the learning experience occurs. We also require that $c(r_i)$ and $c(s_j)$ are normalized: $\sum_{i=1}^{l} c(r_i) = \sum_{j=1}^{m} c(r_i) = 1$, which guarantees that the entire discretized distribution $c(r_i, s_j)$ is normalized. In order to ensure that we can update on any allowed combination of signals, we

[61] The agent's *real* credence function *c* defined on the algebra $\mathcal{A}$.

also demand that $c(r_i, s_j)$ is nowhere zero. See the table below for an example of $c(r_i, s_j)$ for the *very* course discretization $r_i, s_j \in \{-1,0,1\}$.

	$r = -1$	$r = 0$	$r = 1$
$s = -1$	xy	zy	xy
$s = 0$	xw	zw	xw
$s = 1$	xy	zy	xy

Table III.5: A toy, extremely coarse, discretized prior over CPK signal inputs. This example assumes r and s are independent.

$2x + z = 2y + w = 1$. The very strong symmetry of this table arises from the assumption that $c(r_i)$ and $c(s_j)$ are independent, the Martingale constraints – (15) and (16), and the very small number of elements of the grid. This symmetry will not necessarily be a feature of $c(r_i, s_j)$ that don't treat *r* and *s* as independent, or that are defined over more elements.

Now, we construct a refinement, $\mathcal{A}^\#$, of the algebra $\mathcal{A}$ and a new credence function $c^\#$ defined on $\mathcal{A}^\#$ with the following properties:

1. $c^\#(x) = c(x)$, for all $x \in \mathcal{A}$.
2. $c^\#(r_i, s_j) = c(r_i, s_j)$.
3. $c^\#(x|r_i, s_j) = c_f(x)$, for $x \in \mathcal{A}$, and where c_f is the credence function that results from an infinitesimal CPK shift with respect to A and B, with $\frac{\partial c(A)}{\partial t} = r_i$ and $\frac{\partial c(B)}{\partial t} = s_j$.

First, the algebra: we subdivide the atoms of $\mathcal{A}$ to make the atoms of $\mathcal{A}^\#$. Let Z be an arbitrary atom of $\mathcal{A}$; then, the atoms of $\mathcal{A}^\#$ include $Z \wedge r = r_i \wedge s = s_j$, for $i \in \{1, \dots, l\}$ and $j \in \{1, \dots, m\}$. Where $\mathcal{A}$ has 2^n atoms, $\mathcal{A}^\#$ has $2^n \times l \times m$. Close $\mathcal{A}^\#$ under negation and disjunction. Then, we identify any proposition $x \in \mathcal{A}$ with $\bigvee_{i,j}(x \wedge r = r_i \wedge s = s_j) \in \mathcal{A}^\#$.

We construct $c^\#$ as follows: for any $x \in \mathcal{A}$,

$$c^\#(x r_i s_j) = c(r_i, s_j) \cdot \Big([c(x|A) - c(x|\neg A)]r_i dt + [c(x|B) - c(x|\neg B)]s_j dt + c(x)\Big) \quad (17)$$

We define:

$$c^{\#}(x) = c^{\#}\left(\bigvee_{i,j} x r_i s_j\right)$$

$$= \sum_{i,j} c(r_i, s_j) \cdot \Big([c(x|A) - c(x|\neg A)]r_i dt + [c(x|B) - c(x|\neg B)]s_j dt + c(x)\Big) \tag{18}$$

Because of the Martingale constraints (15) and (16), both of the dt terms vanish:

$$\sum_{i,j} c(r_i, s_j) \cdot [c(x|A) - c(x|\neg A)]r_i dt = [c(x|A) - c(x|\neg A)]dt \sum_i r_i \sum_j c(r_i, s_j)$$

$\sum_j c(r_i, s_j) = c(r_i)$, so:

$$\sum_{i,j} c(r_i, s_j) \cdot [c(x|A) - c(x|\neg A)]r_i dt = [c(x|A) - c(x|\neg A)]dt \sum_i r_i c(r_i) = 0,$$

from (15).

$$\sum_{i,j} c(r_i, s_j) \cdot [c(x|B) - c(x|\neg B)]s_j dt = 0$$

in much the same way. So, we're left with:

$$c^{\#}(x) = \sum_{i,j} c(r_i, s_j) \cdot c(x) = c(x) \tag{19}$$

because $c(r_i, s_j)$ is normalized, by construction.

Property 1 holds.

We define:

$$c^{\#}(r_i, s_j) = c^{\#}(\Omega r_i s_j), \tag{20}$$

where $\Omega \in \mathcal{A}$ is some tautology. Assuming *c* is coherent,

$$c(\Omega|A) = c(\Omega|\neg A) = c(\Omega|B) = c(\Omega|\neg B) = c(\Omega) = 1. \tag{21}$$

Thus,

$$c^{\#}(r_i, s_j) = c(r_i, s_j) \cdot \big([1-1]r_i dt + [1-1]s_j dt + 1\big) = c(r_i, s_j). \tag{22}$$

Property 2 holds. As long as $c^{\#}(xr_is_j)$ is nowhere zero, we can define:

$$c^{\#}(x|r_i,s_j)=\frac{c^{\#}(xr_is_j)}{c^{\#}(r_i,s_j)} \tag{23}$$

From (17), (22), and (23), we have:

$$c^{\#}(x|r_i,s_j)=[c(x|A)-c(x|\neg A)]r_idt+[c(x|B)-c(x|\neg B)]s_jdt+c(x), \tag{24}$$

which is the $c_f(x)$ that results from an infinitesimal CPK shift with $\frac{\partial c(A)}{\partial t}=r_i$ and $\frac{\partial c(B)}{\partial t}=s_j$.

Property 3 holds.

We can also define $c^{\#}(r_i)$ and $c^{\#}(s_j)$ in the obvious ways, and check that $c^{\#}(r_i)=c(r_i)$ and $c^{\#}(s_j)=c(s_j)$:

$$c^{\#}(r_i)=c^{\#}(\bigvee_j r_is_j)=c^{\#}(\bigvee_j \Omega r_is_j)=\sum_j c^{\#}(\Omega r_is_j) \tag{25}$$

From (22),

$$c^{\#}(r_i)=\sum_j c(r_is_j)=c(r_i), \tag{26}$$

because $c(s_j)$ is normalized. $c^{\#}(s_j)$ is defined in the parallel way, and the reader is invited to check for themselves that $c^{\#}(s_j)=c(s_j)$.

So, we've shown that $c^{\#}$ satisfies the three properties outlined above. These together entail that c_f is obtainable from $c^{\#}$ by ordinary Bayesian conditionalization on the proposition (r_i,s_j). Now, if all we care about is a constructive superconditioning result for CPK – viz., giving a recipe for *some* way to embed c in a larger algebra and obtain c_f by Bayesian conditionalization on *some* element of the larger algebra, this result is *exact*. For any infinitesimal CPK shift, we can retroactively choose a grid that happens to include the exact (r,s) that were specified, cook up any arbitrary $c(r_i,s_j)$ that is normalized and satisfies the Martingale constraints, and construct a $c^{\#}$ such that $c_f(x)=c^{\#}(x|r,s)$.[62] But I think this result is actually more interesting than that.

[62] Even better: we never need a $r_i\times s_j$ grid larger than 3×3. Just let $r_i\in\{-r,0,r\}$ and $s_j\in\{-s,0,s\}$.

6. Superconditioning on $(\boldsymbol{r}, \boldsymbol{s})$: CPK's Evidential Commitments

First of all, the elements of $\mathcal{A}^{\#}$ that we are updating on are not just arbitrary: they are, intuitively, propositions that *actually pick out* different CPK experiences that the agent might undergo. Second, the $c(r_i, s_j)$ are also not arbitrary. Although the toy example 3×3 $c(r_i, s_j)$ matrix that I cooked up in *Table III.1* is obviously farcical as a genuine approximation of the agent's priors over CPK inputs they might receive, much more fine-grained discretizations can claim to be approximations of the agent's priors over possible CPK inputs with a straight face. Neither the choices of how finely to discretize (viz., the step value $r_{i+1} - r_i$) nor the choice of maximum and minimum values for the signals need be arbitrary. The step value might be chosen according to either limits in the discriminatory power of the system (e.g., the system is only capable of distinguishing some physical signal that plays the *r*-role up to some level of precision), informed by some memory/storage limits (e.g., the system only has the resources to store a certain number of values for this task), etc.

More interestingly, the maximum and minimum allowed values for the r_i and the s_j can be given *epistemic* significance, under the assumption that this updating procedure is being performed by a real system that takes finite time to perform calculations. Such a system will obviously not be able to perform infinitesimally fast CPK shifts, but will have to approximate CPK updating in finite chunks of time. In equation (17), etc., we can replace dt with Δt, and think of the shift specified that way as a time step in a linear approximation of some CPK process. The appropriate value of Δt will depend on some facts about the computational speed of the system in question, which will determine the rate at which the system can effectively execute calculations in the CPK approximation. Unlike in the infinitesimal case, where any assignments of r_i and s_j could be epistemically reasonable inputs, in the finite time approximation, there are values of r_i and s_j that amount to instructions to violate coherence: they will result in credences above 1 or below 0. And so in the finite time approximation, we can place upper and lower bounds on the signals that the agent needs to take seriously as follows:

The result of updating $c(A)$ according to a finite time CPK approximation that specifies only a value for *r* is given by $c^{\#}(A|r)$.

$$c^{\#}(A|r) = \frac{c^{\#}(Ar)}{c^{\#}(r)} = \frac{c(r)[r\Delta t + c(A)]}{c(r)} = r\Delta t + c(A) \tag{27}$$

We can get an upper bound on r by considering the inequality $c^{\#}(A|r) \leq 1$, and a lower bound from $c^{\#}(A|r) \geq 0$. These yield:

$$-\frac{c(A)}{\Delta t} \leq r \leq \frac{1 - c(A)}{\Delta t} \tag{28}$$

And thus any signals outside of this range can be rejected by the agent as not epistemically reasonable. What should an agent do if they receive such a signal? Given that this is a linear approximation of a temporally continuous process where arbitrarily large signals could make sense, it might make sense to “assimilate” all signals above the upper bound to the highest r_i that the agent has in their discretized model; similarly, perhaps it makes sense to collapse all signals less than the lower bound to the smallest modelled input. I haven’t thought very thoroughly about the potential costs and benefits of this proposal – it is merely a tentative suggestion. But the details of how to implement this kind of restriction to epistemically reasonable signals in the approximation case is not really my interest here. The point is just that there are non-arbitrary reasons to establish certain cutoffs for the allowed values of the signal inputs. Just as in the choice of step value between the r_i, there may also be reasons for narrower constraints (viz., lower upper bound or higher lower bound) that emerge from various physical or computational limits of the system.

If we have a fine-grained enough $c(r_i, s_j)$ to take it seriously either as (1) a decent approximation to some agent’s continuous prior over (r, s), or (2) as the *actual prior* of some realistically constrained system, we can use this conditionalization result as a lens into how CPK treats r and s as evidence. We can glean this insight by looking at some conditional probabilities according to $c^{\#}$. First, let’s look at $c^{\#}(r_i|A)$:

$$c^{\#}(r_i|A) = \frac{c^{\#}(Ar_i)}{c^{\#}(A)} = \frac{c(r_i)[r_i\Delta t + c(A)]}{c(A)} \tag{29}$$

Now, a bit of useful terminology: by r_+, I mean any positive-valued r_i; by r_-, I mean any negative value; r_0 is 0. We can see several very interesting things. First, notice that

$$c^\#(r_+|A) > c^\#(r_+), \tag{30}$$

$$c^\#(r_-|A) < c^\#(r_-), \tag{31}$$

because $\frac{r_+\Delta t + c(A)}{c(A)} > 1, \frac{r_-\Delta t + c(A)}{c(A)} < 1$, and $c^\#(r_+) = c(r_+)$, from equation (25). The most obvious interpretation of (30) is that $c^\#$ thinks it is more likely to receive a positive r-signal under the assumption that A is true than the current probability it assigns to a positive r-signal. As is well-known, E is incremental evidence for some hypothesis *h* iff $c(E|h) > c(E)$.[63] So, we can see that $c^\#$ regards any positive value of the r-signal as incremental evidence for A – upon receipt of any positive r-signal, $c^\#$ regards its current credence in A as too low.[64] In (29), holding fixed the prior probability of the signal $c(r_+)$, $c^\#(r_+|A)$ increases as r_+ increases; assuming A increases its confidence in larger signals more than it increases its confidence in smaller positive signals. Put another way, if $c^\#$ initially assigns the same prior to two positive signals, assuming A is true leads it to think that the larger signal is *more likely* than the smaller one.

We can see parallel facts about r_-: r_- is incremental evidence *against* A, and assuming that A is true *decreases* $c^\#$*'s* confidence in more negative signals *more* than in less negative signals. Perhaps somewhat more surprisingly, (29) also entails that

$$c^\#(r_0|A) = c^\#(r_0). \tag{32}$$

The reader might find this surprising, because you might have expected some of $c^{\#\prime}s$ increased confidence in various positive r-signals under the assumption that A is true to come "at the expense" of $c^{\#\prime}s$ credence that it will receive a null signal; this shows that all of the extra weight on positive r-signals must be taken from negative r-signals. But there is an intuitive gloss: (32) also means that r_0 cannot be incremental evidence for A, and this is clearly as it should be.

[63] I think this standard of incremental evidence is usually presented the other way around: $c(h|E) > c(h)$. But it is a consequence of Bayes' Theorem that these two formulations are equivalent. The (incremental) *evidence-for* relation is symmetric: A is evidence for B iff B is evidence for A.

[64] It's also easy to see that any positive r-signal is incremental evidence for A from (26): $r_+\Delta t + c(A) > c(A)$.

I won't rehearse it here, but $c^{\#}(s_j|B)$ encodes a parallel evidential relationship between the s_j and B. Another absolutely crucial thing to notice about the $c^{\#}(r_i|A)$ and $c^{\#}(s_j|B)$ is that they *depend on* $c(r_i)$ and $c(s_j)$, respectively. In a way, this is not very surprising – in order to believe that the signal r_i, generated by some process that is sensitive to the truth of A is advising the agent about how they should be changing their current credence in A, it seems reasonable to assume that the process generating this signal must also be in some way sensitive to what the agent's current credence is. It would, after all, be deeply mysterious how a process could advise the agent about how rapidly to increase their credence in A without any access to what the agent's current credence in A is. However, this does mean that only certain kinds of very special sources of information are going to be remotely plausible candidates for generating CPK signals. The most realistic kind of case, I think, is a system internal to the agent that is processing some kind of stream of evidence with reference to predictions about said evidence encoded by the agent's current mental state. The paradigm example of this kind of process is informational processing of sensory evidence, as discussed toward the end of the previous section. Although I think this kind of sensory processing is the most natural fit, there is no reason that CPK could not also be useful in modelling various kinds of purely internal deliberation, evaluation, or re-evaluation. Plausible sources of CPK signals are not confined to systems internal to some agent *in principle*; it's merely that for a process external to the agent to plausibly provide useful CPK signals, it would need a great deal of informational access to the agent's mental states. At present and in the near future, this may be especially plausible in cases involving computer systems that are in constant communication. But the barriers to imagining this kind of continuous monitoring of mental states for biological agents like humans is really a matter of the present condition of technology, not some kind of fundamental distinction between biological systems and computers.

For arbitrary $x \in \mathcal{A}$, we also find the results for $c^{\#}(x|r_i)$ and $c^{\#}(x|s_j)$ that we should expect:

$$c^{\#}(x|r_i) = [c(x|A) - c(x|\neg A)]r_i\Delta t + c(x) \tag{33}$$

$$c^{\#}(x|s_j) = [c(x|B) - c(x|\neg B)]s_j\Delta t + c(x) \tag{34}$$

Whether some signal r_i is incremental evidence for/against an arbitrary proposition $x \in \mathcal{A}$ depends both on whether the signal is positive or negative and on whether A is evidence for x or not. If x and A are independent (according to c and, hence, $c^{\#}$), then the probative value[65] of A for x, $[c(x|A) - c(x|\neg A)]$ is zero, and so *no* r_i-signal of any valence will be evidence for or against x. If A is evidence for x, then positive r_i signals are incremental evidence for x, negative signals are evidence against x, and the zero-valued, r_0, signal has no effect on $c^{\#}$'s confidence in x. Exactly the opposite is true if A is evidence against x: r_+ signals are evidence against x, r_0 has no effect, and the r_- are evidence for x. In all cases, the magnitude of the change is proportional to the product of the probative value and the strength of the signal. As usual, the story about whether s_j is evidence for or against x is exactly the same, except that the main character is the probative value of B for x.

Again, this is not at all surprising – this evidential dependence is exactly what CPK was designed to exhibit. An agent that regards the r_i and s_j as inputs in a CPK-learning experience thinks of the r_i as evidence that their confidence in A should change, *without changing the evidential relations between* A and any of the other propositions in $\mathcal{A}$.[66] In exactly the same way, the agent interprets the s_j as signals that are directly about B, but don't effect the extent to which B is probative evidence for any other propositions.

And so, this representation result gives us an alternative way of characterizing a CPK learning experience. An agent who updates by CPK can also be thought of *as if* they were conditionalizing on the proposition that picks out which signals they receive, given a prior that represents the signals as having certain evidential connections to the propositions in $\mathcal{A}$ that the agent is interested in. We can think of the $c^{\#}(Ar_i)$ and $c^{\#}(As_j)$ as encoding the beliefs that make it reasonable for the agent to treat the r_i as specified rates of change for $c(A)$ and the s_j as specified rates of change for B. In general, the $c^{\#}(xr_is_j)$ encode the agent's commitment to use the changes to $c(A)$ and $c(B)$ directly caused by r_i and s_j in a way that respects the agent's

[65] For a discussion of the significance of probative value, see the second chapter of this dissertation.

[66] Again, this is in the sense of probative evidence. Changing $c(A)$ will, of course, change the extent to which learning A or $\neg A$ can be incremental evidence. The larger $c(A)$ is, the less that becoming certain (assigning $c(A) = 1$) will change any $c(x)$; while for very large $c(A)$, learning $\neg A$ can have a much larger impact on confidence in other propositions that depend on A. Similarly: for small $c(A)$, learning $\neg A$ will have a much smaller impact on other propositions, while learning A could effect drastic changes.

prior beliefs about whether, and to what extent, A and B are evidence for all of the other $x \in \mathcal{A}$. And of course, by construction, this superconditioning result also gives us a way of understanding the content of the learning experience as propositional – we can model the agent as if they were performing Bayesian conditionalization upon learning *which* signals they in fact received, among several signals that they might have.

7. Factivity and Fittingness of CPK Signals

A question that I am sometimes asked by other philosophers[67] who read about CPK is: "evidence is usually understood as factive. But credal rates of change aren't the kind of thing that can be true or false. How should we understand the credal rates of change that are the inputs to CPK as evidence?" In the past, I've responded roughly as follows: it's true, of course, that it doesn't make sense to ask whether or not an instruction to increase $c(A)$ at some rate is true or false. But we can assess whether or not such an instruction is *fitting* in the conditions the agent finds themselves in. In very broad strokes, I just mean that we can ask whether or not the recommendation to increase the agent's credence is *good advice*. And there are various ways that we might reasonably fill in this question of whether the suggested changes are good advice which are intuitively related to different standards for belief/credence that we might be interested in.

First, we can ask whether such a recommendation actually increases the agent's real accuracy in the situation at hand. According to this standard, a CPK process is fitting/good advice just in case the agent's resultant credences are more accurate, given the actual truth values of the propositions the agent has credences over, than they were before the CPK process. Of course, we could also ask the more restricted questions of whether the raw rate-of-change inputs are good advice in this way: are the agent's resultant credences in A and B more accurate than they were before? If this is true, and the agent's resultant credences are less accurate overall, then we might think that the problem wasn't with the specified rates of change – the problem was that the agent had inaccurate priors about how other propositions

[67] E.g., Sarah Moss – thank you!

were related to A and B such that becoming more confident in these two propositions misled them about the world in general. And, of course, this can happen according to any standard updating rule or learning process.

Another question we might be interested in is not actual accuracy, but something like fit to objective chance. So, suppose there is some genuinely non-deterministic process, such that it is empirically impossible to determine what the outcome will be prior to the result. Obtaining a highly accurate credence in the outcome that happens to obtain is arguably not any kind of credit to an agent trying to predict the outcome. I think it's fairly natural to think that we should attribute this success merely to good fortune. I am partial to the idea that the epistemically best prediction in this kind of case would be to accurately match the objective chance of the outcome according to the non-deterministic laws that govern the process. If this is correct, that's another natural notion of whether the signals the agent receives are good advice or not: are the agent's credences after the CPK shift closer or further away from the objective chance function than they were prior? And again, we can ask the more restricted question: would following the *direct advice* about $c(A)$ and $c(B)$ make these credences better or worse aligned with the objective chance?

Now, I still think that all of that is more or less correct – as far as it goes. But the superconditioning result I presented in the previous section reveals another very interesting feature of CPK: an agent that updates their credences according to my proposal is acting *as if* they have certain conditional priors about how the r and s signals they might receive are correlated with the truths of A and B, respectively. As mentioned in the analysis starting under equation (30), among other things, holding fixed the respective priors, they must believe that stronger (more positive) r signals are likelier, assuming A, than weaker (closer to zero) signals;[68] the same correlation must hold for s_j and B. This correlation is something that the agent might be wrong about. So, if we have a system that is responding to some real signals generated by some process (perhaps as the output of some sensor or organ, a computational output of some

[68] A little more carefully: $\frac{c^{\#}(r_+|A)}{c(r_+)}$ increases as the magnitude of r_+ increases.

other system within the agent, etc.), we can ask: is the agent correct to treat this signal as a CPK input? And just as above, there are multiple things we might mean.

First, we might mean something like: are the $c^{\#}(r_i|A)$ and $c^{\#}(s_j|B)$ accurate? Are these good predictions of the actual frequencies with which the relevant process will tend to generate those signals when A or B are true, respectively? And yet again, here there are two distinctions to be made. Even if the $c^{\#}(r_i|A)$ are not accurate predictions of the signal-generating process' propensity to transmit the various signals, the process could still have a frequency profile that *does* recommend itself as a CPK input. That is, there might be some c^* defined on $\mathcal{A}^{\#}$ that estimates the real frequencies fairly well and which, when conditionalized, outputs the results of the corresponding CPK shifts. In this case, it's not clear that the agent is mistaken to treat the signals as CPK inputs; we might want to say that the agent merely has unfortunate priors. If there is no such c^*, then we can say that treating the signals as inputs to a CPK process is a mistake.

Second, we might be interested in questions about whether the system is justified in treating the signals as inputs to a CPK process, irrespective of whether the beliefs we can model them as having are accurate. And there are many ways that we might approach this. We might be interested in evidential justification: if an agent that updates by treating some signals as inputs to a CPK process repeatedly finds the beliefs it forms this way to be very inaccurate, maybe we should think that the agent isn't justified in continuing to treat the signals this way, because it has a very large amount of evidence that the signals aren't operating the way it thinks they are. We might be interested in some kind of reliabilist justification: if some channel almost always outputs signals that do work well for CPK, but sometimes malfunctions and outputs noise, maybe we think the system is justified in updating by CPK on the noise. And so on…

8. Conclusion

In explaining what it means for the inputs to a CPK process to be evidence, I first explain what it means for an agent to *think* that they are evidence / for the agent to *treat* them as evidence. This is the Value of Information theorem with which I open the chapter: the agent who updates by CPK believes that the signals are evidence in the sense that she believes she will be better informed after making the recommended revisions to her current credences than she is at present; she expects decisions made on the basis of her revised credences to have higher expected utility than facing the same decisions with her current credences. Second, I consider a worry that the kinds of signals that serve as the inputs to CPK cannot actually *be* evidence, because that would require them to have propositional content. I examine Timothy Williamson's arguments that evidence must be propositional. My first impulse in responding to this worry is to disagree with the idea that evidence must have propositional content; the most fundamental role of the kind of evidence with which I am concerned is to require credal change. What the agent *regards* as evidence depends very simply on what credences she believes she should adopt in response to various pieces of information she might acquire. When she thinks that, upon receipt of some information, increasing her confidence in some hypothesis will lead to more accurate beliefs and better-informed decisions, that is *constitutive* of believing that the information is (incremental) evidence for the hypothesis. Whether she is *correct* about this evidential relation depends on facts about things like accuracy and objective chance; this, of course, is an external matter. But I think that there are no special structural characteristics that the representational content of this information needs to have. I think Williamson's arguments that evidence must be propositional are imposing a stricter standard of evidence that is rooted in the representational features that something must have not merely to be evidence, but that are features necessary to *discuss and analyze whether and why* certain pieces of information are or are not evidence. It is this kind of analysis that might plausibly require an agent to be able to represent certain kinds of logical connections between propositions, etc. I argue this point because I think it is true, but I proceed under the assumption that I have not convinced the reader to adopt my position.

So, next I assume that evidence must be propositional – in the very weak sense that, in any case of supposed evidence acquisition, an external theorist should be able to come up with some proposition that is the "real" evidence. I first provide a (trivial) proof showing that any CPK shift defined over finitely many propositions can be obtained through superconditioning on the element of some expanded algebra. Then I pursue, in sequence, two different strategies for identifying what kind of propositional content we might think of as the evidence that could be learned in a CPK experience. The first strategy is deeply analogous to some of Richard Jeffrey's comments on how to understand the non-propositional character of his Probability Kinematics: we can assume that CPK is just an alternative way of characterizing how an agent responds to learning some content that could, in principle, be represented by propositions. So, I sketch an example of how we might relate propositional characterizations of what the agent learns in a continuous serious of visual perceptions to the characterization of that experience provided by CPK. There at least three virtues of choosing to represent the experience in terms of directed rates of change instead of learning various propositions that are very similar to the virtues of Jeffrey's approach: (1) it is a useful abstraction, which allows us to characterize the learning experience in terms of its effects without having to characterize the content of input in, e.g., precise psychological terms; (2) it is more plausible that an agent could have access to (approximate, discretized) CPK inputs than to the "underlying" propositions that characterize the sensory content; and (3) this way of characterizing the learning experience requires assuming the agent has priors over the outcomes of the learning experience that are much more manageable than the kinds of priors we would need to model them as having to update on the propositional content of the sensory experiences. As I quote Jeffrey explaining, it is really quite difficult to imagine that an agent could have priors defined over all of the sensory experiences they think they might undergo. There is an additional virtue of my CPK updating approach, which is a significant focus of the second chapter of this dissertation: CPK gives a much more intuitive gloss on combining learning experiences that are directly about different propositions than the standard Jeffrey approach does.

The second strategy is the constructive proof of a superconditioning procedure that involves modelling the agent *as if* they were updating by Bayesian conditionalization on propositions with the content that they received certain values of signals. Once again, if all we want is a *post hoc* superconditioning result of dubious relation to the agent's actual beliefs, the proof in this section provides an *exact* recipe for constructing a suitably expanded algebra to supercondition on after receiving whatever values of signals they happen to receive; and this can be done with recourse to as small as a 3×3 grid for the possible values of the signals.

However, such a small grid will obviously not actually capture the agent's prior beliefs about what signals they might receive. If we choose a much finer grid, we can begin to actually approximate the agent's (presumed) continuous prior distribution over the signals they might receive; or it might model the actual credal distribution that a real agent physically incapable of storing a truly continuous distribution has. In either case, we can think of this version of the superconditioning result as genuinely reflective (by either representation or approximation) of commitments that the agent *really has*, and so looking at the properties of the credence function on this refined algebra sheds genuine light on what an agent who updates by CPK should believe about the source generating the signals to think of them as that kind of evidence. We see that, for it to be reasonable for the agent to think that the signals she is receiving are indicative of credal derivatives that she should adopt, the agent must have certain particular beliefs about how the probability of receiving various signals varies both with the truth of the learned-about proposition and with her current credence in said proposition. And this sheds significant light on what to say about the question of the factivity of CPK signals as evidence.

When epistemologists discuss the factivity of evidence, there are really two related ways in which evidence is usually assumed/argued to be factive: (1) evidence must have the kind of representational content which allows it to be either true or false and, (2) evidence must be true.[69] Although I have suggested two ways in which we might be able to relate CPK signals to the kinds of objects that can be true or false, I regard this mostly as a formal maneuver that is not indicative of the actual conceptual features that make CPK signals viable as evidence. My suggestion is to replace these conditions with the following: (1) for an agent to *treat* some signals as CPK evidence is for the agent to believe, or to behave as if, there is a certain kind of specific correlation between the truth of the proposition learned about, their current credences, and what signals they receive. It is these beliefs, or beliefs as if, that underwrite why it makes sense for the agent to treat the signals as indications of how their confidence in the proposition should change. (2) for an agent to be *correct* in treating these

[69] Thanks to Sarah Moss for emphasizing this point in a meeting of Michigan's Epistemology Work in Progress (E-WIP) group.

signals as evidence in this way is then a question about whether these beliefs, or beliefs as if, are accurate. We can treat this as a binary (their beliefs about the correlation are correct or incorrect), but it is probably usually more useful to treat it is a matter of degree: how close to the true correlation are their conditional probabilities? Additionally, many epistemologists demand that evidence is not merely true, but that the evidence has some kind of epistemically "*creditable* standing" (Williamson, p. 187). (For Williamson, it turns out that evidence must be known. Other epistemologists may desire evidential beliefs to be well-justified, or safe, or etc.) In the CPK framework, much as what is interesting to say about whether the agent is correct in treating the signals as evidence is not really about the content of the signals themselves (whose content is really merely picking certain real numbers), the question of this kind of *creditability* is also not primarily about the agent's belief in the content of the signals. The interesting questions are whether the agent's beliefs about, or beliefs as of, the correlations between the signals and propositions learned about are creditable. Does the agent have evidence that supports what they take the correlation to be? Are these beliefs about the correlations reliable, safe, etc.? (We could even ask if the agent *knows* these correlations.) Through this lens, the agent's conditional probabilities concerning the signals they might receive are the nexus of all of the questions about the evidential status of the signals: a certain pattern of conditional probabilities is constitutive of taking the signals to be the particular kind of evidence with which CPK is concerned, the accuracy of these conditional probabilities determines whether the agent is – or, rather, to what degree they are – correct in treating them as this kind of evidence, and the creditability of these conditional probabilities is the key to evaluating the praiseworthiness of the agent in relying on these signals as that kind of evidence.

Works Cited

Diaconis, Persi and Sandy L. Zabell. 1982. “Updating Subjective Probability”. *Journal of the American Statistical Assocation* 77 (380): 822–830.

Good, I.J. 1967. “On the Principle of Total Evidence”. *The British Journal for the Philosophy of Science* 17 (4): 319–321.

Gregory, R. L. 1980. “Perceptions as Hypotheses”. *Philosophical Transactions of the Royal Society of London, B: Biological Sciences* 290 (1038):181-197.

Huttegger, Simon M. 2013. “Learning Experiences and the Value of Knowledge”. *Philosophical Studies* 171 (2): 279–88.

Itti, Laurent and Pierre Baldi. 2009. “Bayesian Surprise Attracts Human Attention”. *Vision Research* 49 (10): 1295-1306.

Jeffrey, Richard. 1983. *The Logic of Decision*, revised 2nd edition. University of Chicago Press.

Jeffrey, Richard. 1992. *Probability and the Art of Judgment*. Cambridge University Press.

Moss, Sarah. 2018. *Probabilistic Knowledge*. Oxford University Press.

Rao, R., and D. Ballard. 1999. “Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects”. *Nature Neuroscience* 2**:** 79–87.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation.* Harvard University Press.

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press.