Essays on Belief, Decision, and Learning

by

Christopher R. Nicholson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2023

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Scott A. Hershovitz
Professor Peter Railton
Professor Brian Weatherson

Christopher R. Nicholson

chrisnic@umich.edu

ORCID iD: 0000-0002-0100-3428

DEDICATION

This dissertation is dedicated to Ian Fishback, a casualty of many wars.

TABLE OF CONTENTS

ABSTRACT

Chapter One of this dissertation examines the scope of the epistemic imperative to pursue accurate belief, beginning by arguing that the accuracy that comes from believing self-fulfilling prophecies has no epistemic value. It then extends that argument's reasoning to present a theory of the distinction between belief and decision: pure beliefs concern propositions whose truth values are entirely independent of an agent's actions, pure decisions concern propositions whose truth values are entirely dependent on an agent's actions, and there many cases in between, rendering the distinction between belief and decision a spectrum, not a bright line. The chapter examines a variety of cases to illustrate its point, both original ones and well-known ones drawn from the literature on epistemic consequentialism.

Chapter Two presents a theory of the act of suspending judgment about a proposition, exploring the question of how the concept fits within a notion of degrees of belief. It considers the merits of reducing suspension of judgment to the absence of belief, to middling precise credence, and to imprecise credence, concluding that each reduction fails to capture all paradigmatic cases such as agnosticism about God, jurors' suspensions of judgment about defendants' guilt, and Descartes' suspension of judgment at the beginning of his Meditations. The chapter concludes by presenting an affirmative theory: one who suspends judgment about a proposition chooses not to use their belief about it in all the ways people normally use beliefs. By

doing this, they allow the rest of their beliefs to evolve in the way they would have if the agent held no belief about the proposition they have suspended judgment on.

Chapter Three presents a theory of how agents learn new concepts. It begins by describing the problem: the Bayesian theory of learning by conditioning upon evidence explains how people can improve their existing beliefs but is silent on the question of how they can acquire new beliefs about new propositions. The chapter outlines a theory of conceptual learning that uses the same basic ingredients as learning by conditioning and can work hand-in-hand with it. The approach revolves around the use of a catch-all hypothesis that all one's other hypotheses are wrong. Observing confounding evidence one's regular hypotheses cannot account for raises confidence in one's catch-all hypothesis near 1, driving them to modify their original partitions of hypotheses and observations, and their credences in propositions based on those partitions, in the most minimal way sufficient to lower credence in their catch-all hypothesis toward 0. The chapter presents and refutes influential arguments against the use of a catch-all hypothesis. It closes by describing the mechanisms by which new hypotheses and observations change an agent's confidence.

CHAPTER I

Accuracy Through Self-Fulfilling Prophecy

1. Trade-offs and Epistemic Junk Food

Utilitarians answer the question "What should I do?" by identifying the fundamental ethical good as happiness and advocating actions that maximize the expected happiness in the world. Veritists seek to apply a similar strategy to the question "What should I believe?"[1] They translate the basic approach of ethical consequentialism to the epistemic realm and identify the accuracy of one's beliefs as the fundamental epistemic good to be maximized.[2] Over the last decade or so, the philosophical debate about the merits of veritism has often progressed by translating well-known problems for utilitarianism into the epistemic context and assessing how well the epistemic versions of those objections fare. For instance, there are variations on the organ-harvesting objection to utilitarianism and the trolley problem, each of them asking whether it's sometimes good to sacrifice the accuracy of a relatively small number of beliefs in order to achieve greater accuracy in other beliefs.[3] Another recent line of literature explores the epistemic analogue of Derek Parfit's repugnant conclusion, asking whether it's the average accuracy of one's beliefs that should be maximized or their total accuracy.[4]

---

[1] Pettigrew, Richard. *Accuracy and the Laws of Credence*. Oxford University Press, 2018.
[2] Veritists like Pettigrew build off James Joyce's work showing an epistemic imperative to have accurate beliefs, although they make the stronger claim that all epistemic norms serve the pursuit of accuracy. Joyce, James M. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science*, vol. 65, no. 4, 1998, pp. 575–603.
[3] A paper from Daniel J. Singer gives a good list of examples of epistemic tradeoffs. Singer, Daniel J. "How to Be an Epistemic Consequentialist." *The Philosophical Quarterly*, vol. 68, no. 272, 2018, pp. 583–585.
[4] Pettigrew, Richard. "The Population Ethics of Belief: In Search of an Epistemic Theory X." *Noûs*, vol. 52, no. 2, 2016, pp. 336–372. Talbot, Brian. "Repugnant Accuracy." *Noûs*, vol. 53, no. 3, 2017, pp. 540–563. Dunn, Jeffrey. "Group Epistemic Value." *Philosophical Studies*, 2021.

I'll discuss some of these debates about veritism in this paper but my goal is neither to defend veritism nor disprove it. Instead I want to use some of the insights and intuitions that have emerged from the debate about veritism to shed light on other questions about the epistemic value of accuracy. The debate over epistemic consequentialism has yielded important cases that raise strong intuitions; these cases are underutilized when the intuitions they evoke are applied only toward the debate over whether accuracy is all that matters epistemically. A verititist claims that only accuracy is valuable; my main interest is in figuring out why accuracy is valuable. I argue that accuracy achieved through shaping the world to fit one's beliefs isn't worth anything at all and use that claim to develop a theory of the difference between belief and decision. When determining the truth value of a self-fulfilling prophecy, the question is not "What should I believe?" but "What should I do?"

Hilary Greaves' case of the epistemic imps offers an example of how one can make some of their beliefs about the world more accurate by adopting a different belief that causally affects their truth-values.

Epistemic imps:

Emily is taking a walk through the Garden of Epistemic Imps. A child plays on the grass in front of her. In a nearby summerhouse are n further children, each of whom may or may not come out to play in a minute. They are able to read Emily 's mind, and their algorithm for deciding whether to play outdoors is as follows. If she forms degree of belief 0 that there is now a child before her, they will come out to play. If she forms degree of belief 1 that there is a child before her, they will roll a fair die, and come out to play iff the outcome is an even number. More generally, the summerhouse children will play with chance (1-1/2 q(Co)), where q(Co) is the degree of belief Emily adopts in the proposition (Co) that there is now a child before her. Emily's epistemic decision is the choice of credences in the proposition Co that there is now a child before her, and, for each j = 1, …, n, the proposition Cj that the jth summerhouse child will be outdoors in a few minutes' time.[5]

[5] Greaves, Hillary. "Epistemic Decision Theory." *Mind*, vol. 122, no. 488, 2013, pp. 918.

Greaves asks whether it's epistemically rational to take the imps' bribe – if Emily accepts the bribe and ignores her conclusive evidence for the existence of the child in front of her, she can gain an arbitrarily high amount of expected accuracy by becoming perfectly accurate about whether each imp will play. The veritist form of epistemic consequentialism is often thought to require that Emily should take the imp's bribe. Joyce and Weatherson prove in a recent paper that this is false in Greaves' particular formulation of her bribe, where Emily must accept it by adopting a perfectly inaccurate belief in the atomic proposition that there is a child before her. The crux of their argument is that an epistemic consequentialist should minimize the inaccuracy not of the set of atomic propositions about children and imps that Greaves discusses, but of the Boolean *algebra* formed by all the countable conjunctions, disjunctions, and negations of those atomic propositions.[6] It turns out that being perfectly inaccurate about the atomic proposition that there's a child playing in front of her will be very costly for Emily once its effects on the algebra are correctly scored.

However, as Joyce and Weatherson say, this is only an argument against the particular extreme formulation Greaves uses. Not only are there other epistemic bribery cases that lack the problem, there are nearby versions of Greave's imps case that lack it; they provide one example themselves.[7] More generally, one could reverse-engineer a version of Greaves' imps bribe that survives their objection by accepting any plausible scoring rule, determining the inaccuracy score that scoring rule would apply to the evidence-respecting credal function, and then bribing Emily by saying she must disrespect her evidence for C0 only very slightly, in a way calculated to result in less overall inaccuracy given her reward. Although Greaves illustrates her point by

[6] Joyce, James M. and Weatherson, Brian. "Accuracy and the Imps." *Logos & Episteme* vol. 10, no. 3, 2019, pp. 266.
[7] Id. at 270.

requiring Emily to entirely defy conclusive evidence for a proposition, the debate over the epistemic imps would be similar if she'd required Emily to accept the bribe by adopting a credence in C0 arbitrarily slightly different from whatever the evidence recommends. The key intuition is that it strikes many people as wrong to pursue maximizing accuracy at the cost of defying one's evidence. Some philosophers and non-philosophers[8] have the intuition that Emily shouldn't reject conclusive evidence, no matter how much accuracy she'd gain about whether the imps play. Many would likely also believe that she shouldn't accept the bribe even if it only requires her to slightly downplay conclusive evidence, or reject or downplay weak evidence.

The case of the epistemic imps and the responses it evokes have much in common with the famous organ harvesting counterexample to utilitarianism, where one is given an ethical bribe: sacrifice one patient's life to harvest their organs and save multiple lives. At first glance, the moral of the imps' story seems to be that just as it is wrong to use one person's life as a means to the end of saving others, it's wrong to use one belief's accuracy as a means to the end of accurately believing others. Selim Berker claims these kinds of objections to epistemic consequentialism reveal that our concept of epistemic value prioritizes the "separateness of propositions," meaning that, like people, propositions' utility shouldn't be sacrificed for each other.[9] Berker doesn't say much about why this might be so, or what it even means for the separateness of propositions to have normative significance. Unlike people, propositions aren't the kind of entity that can hold rights, so it's mysterious why accurate belief in them shouldn't be sacrificed for greater epistemic utility through other beliefs.

[8] Andow, James. "Do non-philosophers think epistemic consequentialism is counterintuitive?" *Synthese* vol. 194, no. 2631–2643 2017.
[9] Berker, Selim. "Epistemic Teleology and the Separateness of Propositions." *Philosophical Review*, vol. 122, no. 3, 2013, pp. 337–393.

Berker's method of argument involves using examples to illustrate epistemic trade-offs he thinks even veritists would be uncomfortable with. He focuses more on proving that the accuracy of propositions should be kept separate than explaining why this is so: "[N]o one— not even those epistemologists who most explicitly embrace the consequentialist/teleological framework— is willing to countenance all such trade-offs in the epistemic case. So, I will conclude, this entire approach to normative epistemology is misguided: its advocates don't realize what their approach really commits them to, and if they did realize it, they would abandon the approach rather than incorporate the commitment."[10]

Note the burdens of proof Berker establishes early on: if he can provide even one example where epistemic consequentialists themselves would agree their theory endorses an unacceptable tradeoff, epistemic consequentialism is false. Literally, this is true, since epistemic consequentialism holds that consequences should always dictate actions. But Berker's position subtly shifts later on in the paper. After providing a single counterexample to epistemic consequentialism, he concludes that one should *never* accept epistemic trade-offs: "When it comes to the evaluation of individual beliefs, it is never epistemically defensible to sacrifice the furtherance of our epistemic aims with regard to one proposition in order to benefit our epistemic aims with regard to other propositions."[11] A single case of an impermissible trade-off would be very weak evidence there are no permissible ones.

Berker would probably acknowledge that and maintain that there's no principled way to accept some epistemic trade-offs without accepting all of them. Still, he infers a lot from a little. The fact that he later rejects a few principled ways to accept some trade-offs but not others still falls somewhat short of establishing that no principled distinctions are possible, as he

[10] Id. at 340.
[11] Id. at 365.

acknowledges in his conclusion.[12] The method of "one good example of an impermissible trade-off proves trade-offs are never permissible" is suspect. It would have been better to spend more time giving a theory of *why* propositions are separate instead of putting all his eggs in the basket of proving *that* they're separate. But Berker gives no such theory. The importance of the separateness of persons is explained by their rights, but Berker gives no explanation of what plays the role of rights in the separateness of propositions.[13]

So Berker relies heavily on the power of his examples, especially his first. In his later cases, crafted to respond to potential veritist distinctions to deal with the first, I generally don't think veritism recommends the actions he says it does. But his first case is the cleanest, and the one he relies on most, so it's worth exploring in detail. Borrowing from Richard Fumerton and Roderick Firth, he considers an atheistic scientist who must embrace an inaccurate belief in God in order to get a grant from a religious organization, which would allow him to achieve greater accuracy in many beliefs. Veritism holds that this is a good trade-off, yet even a would-be veritist would obviously want to reject it, Berker says.[14]

This is the case where it's clearest veritism recommends the action Berker says it does. Personally, I'd worry that adopting a false belief about such an important proposition as the existence of God would infect many of my other beliefs with its inaccuracy— the belief in God is a prime example of a belief that tends to change one's other beliefs. A lot hinges on whether God exists or not. What will happen to my beliefs about whether biblical events occurred, or my belief about where life begins, or what happens after death? But let me tweak the case so it really

[12] Id. at 380.
[13] In his conclusion, he briefly hints that the answer might be "evidence." Respecting rights explains why people's ethical utility must be kept separate, and respecting evidence explains why propositions' epistemic utility must be kept separate. But Berker does very little to flesh this idea out. Id. at 380.
[14] Id. at 364.

is clear that accepting the epistemic bribe increases accuracy. Suppose I start out with credence .5 God exists and I simply have to increase it to credence .5000001 to get the grant, or suppose I have some rock-solid guarantee the belief in God won't change many of my other beliefs, or suppose I'm allowed to anticipate I'll go back later and correct my error (after all, I'll remember that before applying for the grant I was sure God didn't exist). Any of these stipulations would make me comfortable accepting Berker's epistemic trade-off. In another straightforward modification of his case, I'd eagerly accept the epistemic bribe if instead of a belief in God I simply had to adopt a false belief about what I had for breakfast. On Berker's theory, I should find it obvious that all these trade-offs are unacceptable, yet I find it obvious they're all acceptable.

We choose to sacrifice accuracy in some beliefs to gain greater accuracy in others all the time and are often eager to do so. Consider someone who is deciding whether to read a general science textbook covering many theories, knowing that although no author is infallible the one who wrote it is very knowledgeable. It is a long book with many claims, so the prospective reader knows it's virtually certain they will make a few of their existing beliefs less accurate by reading it but will make many others more accurate.

They might choose to read the book even if they know something about *which* of their beliefs are likely to become less accurate. For instance, suppose they've read reviews alluding to the fact that the book's theories about say, gravity, sound compelling but are disconfirmed by unspecified new evidence. Without knowing what the new evidence is or studying it in detail, the prospective reader knows they are likely to be persuaded by the book's claims about gravity, and they cannot easily guess which or how many of its seemingly reasonable claims are false. Or they could be told, for instance, that a particular proposition about gravity is very compelling to

anyone who hasn't studied physics for many years, and that even those who know experts disagree with the claim end up finding it more plausible than before after reading the book (this might even involve simply going from credence .1 before reading to .2 after). If the rest of the book is expected to be significantly more accurate than alternative textbooks, the prospective reader might justifiably choose to read it.

These ordinary cases show that we accept epistemic trade-offs all the time without any qualms, and we may even reason in a consequentialist way when doing so. Unlike people, we owe propositions nothing. This is an important disanalogy between epistemic bribes and ethical ones: there are people waiting to be happy, but there are no propositions waiting to be known. Sacrificing propositions' accuracy for greater epistemic utility is much more acceptable than sacrificing people's happiness for greater ethical utility. The propositions don't object.

So why does it seem wrong to accept the imps' bribe and sacrifice accurate belief in the child playing in front of us to gain accurate beliefs about whether the imps will each play? I have an alternative theory from Berker's separateness of propositions: what the imps offer isn't valuable at all. In organ-harvesting style cases, we sacrifice something of ethical value to gain more value elsewhere. But although the imps case is superficially similar to those objections to utilitarianism, once we understand how the epistemic context changes things we will see that accepting the imps' offer would require sacrificing something of epistemic value and receiving nothing of value in return.

To begin showing this, I'll present a similar case to Greaves'. The only difference in mine is that while her imps were mischievous, intent on posing puzzles for epistemic consequentialists, my imps will be helpful and do whatever they can to make veritists better off

by their own lights. The question is this: setting aside all questions about bribes and sacrifices, should one want the accuracy the imps offer if they offer it for free?

Benevolent imps:

Emilia is taking a walk through the Garden of Benevolent Imps. In a nearby summerhouse are n children, each of whom may or may not come out to play in a minute. They are able to hear her voice, and their algorithm for deciding whether to play outdoors is as follows. If she yells "I want the accuracy," or just "Come out and play," they will come out to play. If she remains silent, they will roll a fair die, and come out to play iff the outcome is an even number.

Forget bribes; is this generous epistemic offer worth accepting? It doesn't seem like it. Yelling "I want the accuracy" would undoubtedly result in more accuracy, thanks to the benevolent imps— they would deliver exactly the same reward of accuracy that the mischievous imps offer in Greaves' original case, removing the die rolls from the equation. And unlike her case, this accuracy about whether each imp will play comes with no epistemic cost; there is no evidence we'd have to ignore, no belief to be sacrificed. Yet I feel I would understand the world no better if I asked the imps to play than if I remained silent and allowed chance to make their decision.[15] The accuracy the mischievous and benevolent imps offer seems cheap and hollow, like epistemic junk food for veritists. For an alternative version of this example that works only through belief instead of yelling, we could simply add that the imps are able to read minds again, and if you have credence 1 that they'll come out and play, they'll play.

The imps are offering me the opportunity to causally affect their decision about whether to play so that it will no longer be objectively chancy. But objective chance is not the enemy of someone who wants to have their beliefs about the universe accurately reflect the way it is— if the universe is objectively chancy in some ways, we should simply want our beliefs to accurately describe its chanciness. I already have perfectly accurate beliefs about all the factors that will

[15] I discuss the concept of "understanding" more at the beginning of Section 3.

determine whether the imps play or not. I know that it depends first on whether I say yes to their offer, and if I say no, on individual die rolls. Epistemically, I have all I want. The only reason I don't know whether they'll play for sure or roll dice is that I haven't decided whether to accept their offer yet.

I wouldn't decide whether to accept the mischievous or benevolent imps' offers based on how accurate it would make my beliefs; I would make the decision based only on whether I would like for them to play. Whether to accept the imps' offers are matters for decision, not belief. When I think about whether to accept the imps' offer, I'm not trying to determine what the world is like. I already know everything I want about whether they'll play; the unknown thing I'm trying to ascertain is whether I *want* them to play. Figuring out what I want them to do is just making a decision. Pursuing accuracy is important when forming beliefs, but decisions answer to different criteria, like whether they have the effects we intend and whether those effects further our values.

My analysis is similar when thinking of the epistemic bribe Joyce and Weatherson consider at the end of their article: as with the imps, the accuracy their bribe offers is not worth anything at all. They write,

> For example, imagine that Ankita has, right now, credence 0.9 in D0, and 0.5 in D1. These are good credences to have, since she knows those are the chances of D0 and D1. She's then offered an epistemic bribe. If she changes her credence in D0 to 0.91, the chance of D1 will become 1, and she can have credence 1 in D1. Taking this bribe will increase her accuracy.[16]

Taking the bribe *would* increase Ankita's accuracy, but from my perspective, her current understanding of the world is perfectly good; she has perfectly described its chanciness. Things couldn't be better. In fact, from my point of view, things wouldn't even be better if the chances of D0 and D1 were 1 and her credences in them were 1. The goal of belief is simply to describe

[16] Joyce and Weatherson, p. 275.

however the world is, not to be pleased or disappointed by its chanciness. Belief aims to describe the world, not judge it.

Ankita's briber offers her the opportunity to gain more accuracy by slightly disrespecting her evidence in D0, which she knows would cause the chance of D1 to become 1, allowing her to adopt credence 1 in D1, gaining more accuracy than she loses on D0. But Ankita already knows that the chance of D1 will become 1 if she accepts the bribe; the only reason she doesn't have the perfect accuracy the bribe offers is that she doesn't yet know whether she'll take the bribe. Ankita's beliefs are currently in good order, and the only part of the world she doesn't grasp yet is her own decision. She just has to decide whether she wants to make D1 true or not. The bribe offers epistemic junk food, accuracy that tells her nothing about the world outside of her opinions—in fact, she'll know slightly less about that world if she takes the bribe, since her beliefs will go from perfectly matching the objective chances of the two propositions to being slightly off on D0.

Whether Emily should accept the imps' offer or Ankita should accept her bribe are matters for decision, not belief. This is my starting point in my thinking about the epistemic value of accuracy. In Section 2, I elaborate on my argument that cases like the imps are just questions about what to decide, not what to believe. I discuss what it takes to be a decision. In Section 3 I expand on the distinction between belief and decision. I argue that the two are not fundamentally different in kind, but exist on a spectrum, and that the more the truth value of a proposition depends on one's own actions, the more the question of what to think of it is a decision, while the less the truth value depends on one's own actions, the more the question of what to think of it is a belief. In Section 4 I consider the question of what kinds of mental acts count as pure beliefs. In Section 5 I consider mixed cases, such as the question of whether to

invest in a company, showing that such questions are part belief, part decision. I then apply the distinction to argue that the efficient market hypothesis is false because it treats investments purely as predictions of value when they're also decisions about it.

2. Pure Decisions

In the last section I claimed that the accuracy the imps offer is epistemically worthless because it's not accuracy that describes what the world is like—it's accuracy that you get by forcing the world to be a certain way. In this section I justify and expand on that claim and illustrate it with a variety of cases. Cases like the mischievous and benevolent imps, where one achieves accuracy by shaping the world, are matters for pure decision; the truth-value of the proposition regarding whether each imp will play depends entirely on one's own choice. Cases where the truth-value of a proposition depends not at all on one's own actions are matters for pure belief about what the world is like. Beliefs about past events, for instance, would be matters for pure belief, since one's actions can't causally affect the past. I discuss pure beliefs in more detail in Section 4. Determinations about the truth of propositions whose truth-values are entirely dependent on one's own future actions would be paradigmatic examples of decisions, not beliefs. Since you're aware that the benevolent imps' choice to play depends entirely on whether you want them to, yelling "I want the accuracy" and assigning a credence of 1 to the proposition that each imp will play is primarily a successful decision to have them play, and merely entails and enables a successful prediction that they will.

One of the classic examples of a self-fulfilling prediction comes from William James, a case Greaves considers too. Here's the original version:

Leap:

And often enough our faith beforehand in an uncertified result is the only thing that makes the result come true. Suppose, for instance, that you are climbing a mountain, and have worked yourself into a position from which the only escape is by a terrible leap. Have faith that you can successfully make it, and your feet are nerved to its accomplishment. But you mistrust yourself, and think of all the sweet things you have heard the scientists say of maybes, and you will hesitate so long that, at last, all unstrung and trembling, and launching yourself in a moment of despair, you roll in the abyss. In such a case (and it belongs to an enormous class), the part of wisdom as well as of courage is to believe what is in the line of your needs, for only by such belief is the need fulfilled. Refuse to believe, and you shall indeed be right, for you shall irretrievably perish. But believe, and again you shall be right, for you shall save yourself. You make one or the other of two possible universes true by your trust or mistrust,—both universes having been only maybes, in this particular, before you contributed your act.[17]

James argues that the pragmatic goal of living justifies predicting one will successfully make the leap. The question for my purposes is whether the mental act of assigning credence 1 to the proposition "I'll leap the gorge" is best described primarily as a prediction, or a decision.

Those who'd think of it primarily as a prediction remind me of the engine in the children's story "The Little Engine that Could": as the engine tries to make it up a high hill, it keeps chanting "I think I can, I think I can," using the prediction that it can make it up the hill to turn that prediction true. "I think I can, I think I can" has the psychological feeling of a prediction, not a decision. My own position that we should think of the act of assigning credence 1 to "I'll leap the gorge" as a decision is more reminiscent of Yoda's admonition to Luke in *The Empire Strikes Back*: "Do or do not; there is no try." If Yoda were advising someone about to leap the gorge and he heard them muttering "I think I can, I think I can," he'd object that they're mistakenly thinking of the leap as a matter for prediction: they simply need to decide to leap across the gorge. Even if he spotted someone motivating themselves with "I know I will, I know I will," Yoda would say "Just decide you'll do it." Framing it as a prediction makes it seem as if

[17] James, William. "Is Life Worth Living?" from *The Will to Believe and Other Essays in Popular Philosophy*. New York, London, and Bombay: Longmans Green, 1896, 1899, pp. 32-62. https://ethicsofsuicide.lib.utah.edu/selections/william-james/

the outcome is up to the world, while framing it as a decision makes it clear the outcome is up to you. Yoda would insist we think of self-fulling beliefs as matters for decision, not prediction, no matter how confident the prediction.

That actually strikes me as consistent with the ultimate point James is driving at with his leap example. He concludes it by saying "You make one or the other of two possible universes true by your trust or mistrust." The act of believing you'll leap the gorge is merely instrumental to a decision to leap it and make that universe true. The leap example is part of James' larger argument against committing suicide. He ends his speech with "Believe that life is worth living, and your belief will help create the fact." Just before that, he says "Please remember that optimism and pessimism are definitions of the world, and that our own reactions on the world… necessarily help to determine the definition. They may even be the decisive elements in determining the definition." It seems to me that James is arguing that by believing life is worth living, we can decide that it is. The decision to live, or to leap, is the main mental act, and our beliefs in those things are just the mechanisms by which we enact the decisions. That's why I regard the question of what credence to assign "I'll leap the gorge" as a matter for decision, not prediction—it should be determined by whether you want to leap the gorge, not by the pursuit of accuracy. Once you decide you want to leap the gorge, you can then use an accurate prediction you'll leap as a tool to implement your decision. But the accuracy only has instrumental value; it's not an end in itself.

Jennifer Carr illustrates the pitfalls of regarding these self-fulling beliefs as matters for belief instead of decision. She treats them as matters for belief and that sets her on a course that ultimately causes her to give up the theory that we should try to have accurate beliefs.

She begins by considering an example structurally nearly identical to James' leap case, except that the act one wants to perform is a handstand instead of a leap: she stipulates that one's objective chance of performing a handstand is identical to their credence they'll perform one.[18] Carr thinks consequentialist norms about maximizing accuracy require credence 0 or 1 in performing a handstand and observes that respecting one's evidence would allow any credence.[19] She distinguishes between what she calls consequentialist epistemic decision theory, which considers how adopting a credence changes one's beliefs and their truth-values, with non-consequentialist epistemic decision theory, which doesn't. She claims that arguments like Joyce's defense of probabilism depend only on non-consequentialist norms, and that we can only hold onto rules like probabilism and conditionalization if we give up on the theory that rational belief aims at truth.[20] She ends up preferring probabilism and conditionalization to the idea that beliefs aim at truth, then concludes that we must come up with some theory about why beliefs aim not for truth, but for the truth-inspired mathematical object preserved by probabilism and conditionalization.[21] In the end, Carr seems to think we're free to choose any credence about whether we'll do the handstand because beliefs don't aim at truth.

This is a long and complicated road that leads to a strange place. Instead of giving up on the idea that beliefs aim at truth, I would much rather give up on the idea that epistemic consequentialism requires we maximize the accuracy of self-fulfilling beliefs. I wanted to give up on that anyway. Because I depart from Carr very early in her thought-process it would serve little purpose to examine the intricacies of the later details. On my theory, anytime a fact about

[18] Carr, Jennifer. "Epistemic Utility Theory and the Aim of Belief." *Philosophy and Phenomenological Research*, vol. 95 issue 3, 511-534, p. 515.
[19] Id. at 522-23.
[20] Because I depart from Carr earlier in her argument, I won't evaluate these later claims in any detail, though I think some of them are debatable. Her argument against conditionalization, for instance, seems to actually prove that one should have particular priors regarding self-fulfilling beliefs, a fact she acknowledges. Id. at 519.
[21] Id. at 532.

the world depends on our credences, that's a matter for decision—we have the power to decide how the world is, and I think we should use it to move the world closer to our desires instead of crusading after accuracy. After applying my view, there's no difference left between what Carr calls consequentialist and non-consequentialist epistemic decision theory. Very little about the world is under one's control as it is; there's no need to sacrifice any of that precious influence in pursuit of accuracy. We're free to choose any credence about whether we'll do the handstand because *decisions* don't aim at truth.

I don't regard it as an epistemic loss when I have less accuracy about whether I'll leap the gorge or make a handstand merely because I haven't yet decided whether I want to; I don't regard it as an epistemic gain once I make my decision to leap or handstand. I already knew the truth of all the relevant conditional statements like "If I have credence 1 I'll make a handstand I'll make one" and "If I have credence 0 I'll make a handstand I won't make one." I only lack accurate belief about which consequent is true because I haven't yet decided which antecedent to make true. My only epistemic interest is in having accurate beliefs about the conditionals. The truths of the antecedents and consequents are undetermined by the outside world, and it's up to me to decide what they are.

These are matters for pure decision because the truth values of the propositions "I leap the gorge" and "I make a handstand" depend entirely on what I want them to be. Decisions do tend, by their nature, to result in increased accuracy about one's own future actions, but because this accuracy tells us nothing about what the world independent of our decision is like, it's epistemically worthless—when I refer to cheap and empty accuracy I mean accuracy that's gained entirely through forming a decision about what to do. If we mistakenly view decisions as predictions, subject to epistemic imperatives to increase accuracy, every decision will suddenly

look like a self-fulfilling prophecy. Instead of viewing decisions as highly effective kinds of predictions we should view them as a different mental act with different goals. I elaborate on the distinction in the next section, building off common "direction of fit" considerations.

I'm not alone in noting that all decisions can look like self-fulfilling predictions. As David Velleman puts it, "...the traditional distinction between predicting and deciding breaks down in the case of self-fulfilling predictions."[22] Velleman argues that the existence of self-fulfilling predictions means that people have a kind of epistemic freedom even if the universe behaves deterministically. He illustrates his point by analyzing an example from Elizabeth Anscombe where a doctor tells a patient in front of a nurse "Nurse will take you to the operating theater," and the patient interprets it as a straightforward statement of fact, while the nurse interprets it as a directive and makes it true. Velleman explains what he means by the claim that self-fulfilling predictions allow us epistemic freedom with the following:

> The point of the story, for my purposes, is that although the doctor is correct in asserting that the nurse will take the patient to the operating theater, he would have been equally correct in asserting that the nurse would take the patient to the lab, or to any other destination, within reason. Insofar as the nurse stands ready to do whatever the doctor says, the doctor can truly assert any one of several incompatible things; and to that extent, he is epistemically free.[23]

Of course, all of this is compatible with determinism being true and the doctor lacking what we ordinarily think of as freedom, but the doctor is free in the important sense that, from his perspective, the evidence allows him multiple options about where to think the nurse will take the patient, and he can choose to make any one of those options true by believing it and then uttering it. As Velleman puts it, "What makes him feel free, in short, is his freedom from the evidence."[24]

---

[22] Velleman, David. The Possibility of Practical Reason. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2009, pp. 35
[23] Velleman at 36.
[24] Ibid. at 37.

We encounter that same freedom from evidence in James' leap case, Carr's handstand, and in my modification of Greaves' imps case, although in the original imps case one is not free from the evidence, since there's a child in front of them. Just as the evidence allows the doctor to believe the nurse will take to the lab or the operating theater, the evidence equally allows you to believe or not believe you'll leap across the gorge, or have any intermediate credence in it. As Velleman says, self-fulling beliefs like these can also be described as decisions. He goes even further later in the paper and says "all decisions are self-fulfilling predictions of one sort or another."[25] All decisions involve adopting a high credence that one will perform some action, and adopting that credence makes it more likely to become true. I think the order the events typically go in is that we decide what action we want to do, and on the basis of that settled desire we decide to adopt a high credence we'll do it, then the high credence helps make itself true.

Velleman was writing in the context of debates about determinism, not epistemic consequentialism. In the context of veritism, where it's accepted that one ought to maximize the accuracy of their beliefs, it can be misleading to describe all decisions as self-fulfilling predictions, because then a veritist might conclude that one ought to maximize accuracy by making whatever decisions are best at being self-fulfilling predictions. The pursuit of accuracy would suddenly consume not just belief, but all decision-making, since decisions would just be a type of prediction. I elaborate on the perils of this approach in the next section; in short, treating all decisions as predictions and making whatever decisions maximize accuracy would amount to sacrificing all freedom, epistemic or otherwise. We can see this phenomenon rear its head when the pursuit of accuracy suddenly demands that we either have credence 0 or 1 that we'll make a handstand, but nothing in between, even though the evidence allows any credence. Since I think

[25] Ibid. at 52.

being a rational believer makes no such demands, I think adopting a credence about the handstand is purely a matter for decision, not prediction, though the decision is enacted with a prediction.

3. The Belief/Decision Spectrum

My argument that gaining accuracy through believing self-fulfilling prophecies is worthless because it amounts to making a decision is really just an application of a longstanding principle famously articulated by philosophers like J.L. Austin and Elizabeth Anscombe: the aim of belief is to make the mind fit the world, while the aim of desire is to make the world fit the mind. An agent's decisions enact the desires encoded in its utility function; the aim of decision is to make the world fit better with the mind's desires. So although "direction of fit" discussions usually contrast belief with *desire*, it's equally appropriate to say that *decision* aims to make the world fit the mind, while prediction aims to make the mind fit the world. Generally, the mind ought to fit the world well before one attempts to make parts of the world fit the mind. Decisions ought to be informed by accurate beliefs.

My main claim in this section is that the distinction between belief and decision, while a good one, is much blurrier than people tend to think. Ultimately, I do agree with Velleman that all decisions can be viewed as self-fulfilling predictions; the trick is to flesh this out in a way that doesn't require the pursuit of accuracy to dominate all decision-making. While decisions can technically be described as self-fulfilling predictions, there's a good reason we don't usually describe them that way. Paradigmatic cases of decision and belief have different goals, as the standard direction-of-fit distinction makes clear. Since the line between the two is blurry, in this section I'll try to demark where the imperative to pursue accuracy should begin and end.

My starting point is simply a critique of any veritists who find the imps' offer compelling: in their love for accuracy they have ignored fundamental "direction of fit" concerns and mistakenly valued accuracy that comes from forcing the world to fit their desires. There is no epistemic imperative to tie down parts of the world so that we know exactly where they are. Likewise, there is no epistemic imperative to tie down our future selves so that we know exactly where we and the things that depend on us are. Understanding this basic point more fully will help me develop my more fundamental point that the distinction between belief and decision, while valid, is a spectrum, not a bright line. I'll begin with a couple examples.

Tea:

Suppose I'm thinking about whether to drop by your office for a chat at 2:30 PM, and I know you're planning to visit a local café in the afternoon and will order coffee or tea. I also know that you have a rule against drinking coffee after 3 PM, since it has more caffeine than tea. I'm trying to decide whether my visit would interfere with your plans or not. I predict that you're likely to get tea, since that's what you seem to do more often than not, so my chat probably won't disturb your plans. On the basis of my prediction that you're probably going to get tea, I drop by your office for a chat, the conversation runs long, and you get to the café after 3 PM. You order tea.

Here's the question: did I predict you'd get tea, or decide it? Or was it a mix of the two? My prediction you'll drink tea seems to function somewhat like a decision you'll drink tea if I plug it into my decision-making and act in ways that make you more likely to get tea. My prediction of your actions functions partly as a self-fulfilling prophecy.

I call a prediction at least somewhat self-fulfilling if adopting that prediction makes it more likely to become true. There are different possible causal mechanisms for self-fulfillingness. In cases like Leap and Handstand, the exact mechanism is left unstated, but there seems to be an implicit, direct, *constitutive* way that adopting a particular prediction shapes the world in a way guaranteed to make that prediction true—we are to assume that the determining factor in our chance of making a handstand is the prediction we adopt about making it. In Imps,

the case is artificially constrained so that the prediction the imps will play still perfectly correlates with the probability they'll play, but the causal mechanism is more complex, more indirect—the prediction they'll play only has causal power because the imps choose to entirely attach their decision to it.

In the more realistic case Tea, I have added back in all the complexity of the world, so my prediction that you'll drink tea only plays a contributing causal role in leading you to drink tea. The mechanism by which this prediction fulfills itself is that I plug my prediction about what you'll drink into my decision-making on the separate question of whether I should drop by your office. When I act on my prediction that you'll drink tea, I make you more likely to. The fact that we use our predictions in our decision-making, combined with the fact that the decision-making can causally affect the events we're predicting, often makes those predictions at least partly self-fulfilling. When we make predictions while knowing that they're self-fulfilling, they seem to have an element of decision to them.

Predictions about our own future actions are especially likely to be powerful self-fulfilling prophecies, and this can lead to strange behavior if we subject those predictions to an imperative to maximize accuracy. Consider another ordinary example.

Pizza:

Suppose I'm considering what to believe about what I'll eat for dinner. Suppose I know that beliefs like these about one's future activities are self-fulfilling prophecies, to some degree, so I know that the higher my credence is that I'll eat pizza the higher the probability that I'll eat pizza is, because I'll do things like invite friends to eat pizza with me, or fail to make the plans necessary to eat anything other than the pizza in my freezer, or something like that. Suppose I'm torn between eating pizza and eating burgers but I realize that if I have a higher credence that I'll eat pizza, because it's at least partly a self-fulfilling prophecy my credences that I'll eat pizza and that I'll eat burgers will each become more accurate. And suppose I have a pizza in the freezer, but no burgers, so a belief I'll eat pizza would have greater expected accuracy. A belief I'll eat pizza is a stronger self-fulfilling prophecy than a belief I'll eat burgers would be. Could this be a good epistemic tiebreaker to break my indecision?

If I pursue the expected accuracy that comes from having high credences in self-fulfilling beliefs about my future, I'll have much more accurate beliefs about what I'm having for dinner or what I'm doing that evening than I otherwise would. This will in turn give me greater accuracy in all the propositions affected by my future actions—by having more accurate beliefs about my own dinner plans, I'll also have more accurate beliefs about my friends' plans, for instance, and the future state of my freezer. Is my epistemic utility function constantly telling me I'd be more rational if I had higher credences in self-fulfilling prophecies?

I doubt it; the cheap accuracy gained from self-fulfilling credences wouldn't get me any closer to my epistemic goal of understanding what the outside world is like. "Understanding" is a common concept that seems to me to be epistemic but is not discussed nearly as often as concepts like knowledge and accuracy, perhaps because it's not as well-defined. I use the word here because it comes nearest to describing the epistemic good that's missing from the cheap accuracy one can create by intentionally believing self-fulfilling prophecies or accepting epistemic bribes that involve causing propositions to be true. If Emily or Ankita accept their bribes, each of them gains not only accuracy, but probably knowledge. Emily gets to know all the imps will play because she's taken an action she knows will cause them to play. Ankita gets to know D1 is true because she's taken an action she knows will cause it to be true. But intuitively, I don't think either of them gains any understanding of the world; in fact, they each seem to lose it.

I believe the theory I flesh out in this section will shed light on at least an aspect of understanding. I make no claim it's a complete account; that's a task for a different paper dedicated to nothing else. As a starting point for conceptual analysis, consider that when we talk about people understanding things, we often seem to be referring to the ways they grasp the

interactions of systems. Someone might know a lot about bikes if they know all the different kinds and models of them, and the history of their development. But we might think someone else understood bikes more if they knew fewer true claims, but knew enough about bikes to make one, or to fix any. To understand bikes isn't about how much one knows about them, but whether one knows a variety of the most important things about them and how those facts relate to each other. Someone who understands bikes knows how they work.

Understanding seems to involve knowing how parts of a system relate to each other, and how different systems relate to each other. As a crude test of this claim, we could ask whether perfect accuracy or knowledge of an atomic proposition would ever be called understanding, or whether it would only seem like understanding once accuracy about additional propositions was added. Knowing or accurately believing "Joe Biden is president" is easy. It seems intelligible to say someone understands that Joe Biden is president, but that would involve something beyond knowing the truth of that proposition—if someone knows that Biden is president but questions his authority when he orders the Army to do things, they don't understand that he's president. Or someone might know that Biden is president, know, in an intellectual sense, that the president can veto bills, yet say there's nothing he can do to prevent a narrowly-passed bill from becoming law. This failure to combine the separate propositions they know to arrive at a conclusion that Biden can veto the bill would reflect a lack of understanding that Biden is president. Understanding seems to require both knowledge of the different parts of a system and the ability to apply that knowledge appropriately to describe its interactions.

My general description of understanding is similar to Allison Hills' account of moral understanding. Hills appeals to the difference between knowledge and understanding to explain the puzzle of why it seems wrong to form a moral belief based on moral testimony, when in most

non-moral contexts, we consider it acceptable to form beliefs on the basis of testimony. Hills' answer is that even if mere moral testimony can transmit knowledge, it can't on its own create understanding, and understanding is particularly important in the moral context. Like me, Hills identifies the systematic nature of understanding as a key difference between it and knowledge: "You cannot really understand why *p* is true if *p* and the reasons why *p* are the only things about the subject of which you are aware: it is not possible to understand why some isolated fact is true. If you have this kind of appreciation of moral reasons, you must have, at least to some extent, a systematic grasp of morality."[26] Such a systematic grasp of moral claims, Hills argues, entails abilities such as being able to explain them in your own words, recognizing situations that call for the claims and applying them, and modifying the moral claims appropriately when the typical reasons for them are modified.[27]

Understanding is flexible and adaptable in a way that mere knowledge is not. Knowledge alone can be brittle; someone might know how to fix a bike through mere mimicry but be stumped when asked to fix a slightly newer model of it. Someone might know the recipe to make a great cake and have followed it many times but have no idea which ingredients they can use as substitutes if the store lacks the exact ones on their list.

This is far from a complete account of understanding, but my goal is not really to define understanding or to argue that my theory of the difference between belief and decision is an account of it. "Understanding" is simply the word in my vocabulary that best describes my theory of what beliefs aim at. Ultimately, my interest is in explaining and justifying my theory, not defending it as an account of understanding. It is a theory of how a person's beliefs attempt to describe the systems the world operates by, and how they then use those beliefs to make

[26] Hills, Alison. "Moral Testimony and Moral Epistemology." *Ethics*, vol. 120, no. 1, 2009, p. 101.
[27] Id. at 102.

decisions acting within the world's systems. Beliefs aim to describe how the world works so that we can make decisions working within it.

The epistemic purpose of beliefs is to figure out the truth of everything that doesn't depend on our voluntary and involuntary actions, because we want all our actions that might affect the world to be maximally informed about what the world independent of us is like. By "voluntary and involuntary actions" I mean to include voluntary and involuntary formations of belief. I want the credence I end up with about whether I'll have pizza or not to be informed by as accurate a representation as possible of what the world independent of my credence is like. When figuring out the truth of things that depend on me, I have a uniquely epistemic interest in figuring out what the world independent of my beliefs, actions and decisions is like first.

Facts about the world depend on our actions and choices to different degrees, and in different ways, which gives useful guidance about how to go about becoming less ignorant about what the outside world is like. In general, we should start by learning the truth of the propositions that least depend on us, and after becoming reasonably confident about them, think about the ones that depend on us slightly more. Ideally, we would consider the propositions whose truth-values are most influenced by our beliefs or actions with as much information as possible about the truth of the propositions that depend less on us. When we think about what to do and imagine various conditionals about what the world would be like if we made various choices, for instance, we want our imagination of our possible effects on the world to be informed by a reasonably accurate representation of what all the things that don't depend on us are like. To accurately imagine how the world would be different if we affected it in various ways, we need to understand the many ways it would remain the same. Upon understanding what would remain exactly the same no matter what we do, we can more accurately imagine which

aspects of the world would remain almost exactly the same and work our way towards understanding which would be completely different. Being able to accurately imagine counterfactual situations resulting from different possible decisions allows us to choose the decisions likeliest to lead to the best outcomes. This imaginative capacity is critical to making effective decisions, and it requires having an accurate picture of what the world outside of our opinions is like.

Even if we adopt a pure epistemic perspective that for the sake of argument places no weight on the value of making good decisions, there's still an epistemic benefit to giving priority to possessing accurate beliefs about the propositions whose truth values depend least on us – after all, if this gives us the ability to more accurately imagine the outcomes in counterfactual scenarios involving our own actions, that increased accuracy about propositions involving hypothetical situations improves the accuracy of our credal functions as a whole. Still, it seems artificial to ignore what these considerations imply about the boundary between beliefs and decisions. Instead of stipulating what the epistemic perspective cares about, as some veritists and their critics like Berker sometimes do, claims about epistemic value naturally fall out of my way of distinguishing beliefs from decisions.

If the truth value of a proposition depends primarily on our choices and we're aware of that, reasoning about whether the proposition is true is the process of making a decision about the world, not forming a belief about it. Having a higher credence that I'll have pizza knowing that it's a self-fulfilling prophecy is fine if I do it as the result of a decision to have pizza but misguided if I do it to describe the world more accurately. Letting the pursuit of accuracy influence the decision would be a mistake. Raising my credence I'll have pizza increases my expected accuracy because by doing so I'm removing some uncertainty about my own future

actions since I'm aware it's a self-fulfilling belief—I'm sacrificing some of my future ability to choose my effect on the world for the sake of being better able to predict that effect, forfeiting what Velleman calls epistemic freedom. It would be like gaining more expected accuracy about the parts of the world that depend on my actions by handcuffing myself. In some rare circumstance where it really mattered that I be right about whether I'm having pizza or not it might make sense to let the greater expected accuracy of a higher credence influence my decision. Sometimes accuracy has especially great practical importance. But if I'm undecided between burgers and pizza, all pursuing the higher expected accuracy of the higher self-fulfilling credence that I'll eat pizza amounts to is gaining confidence about my future by giving up some of my ability to decide it. The question is not whether I want accuracy, but whether I want burgers or pizza.

To see this problem in its purest form, consider this question: is there an epistemic imperative to make firm decisions? When we not only decide to do something, but make our decision firm in various ways, we are essentially trying to increase the degree to which our belief about our future actions will be a self-fulfilling prophecy. After reaching a tentative decision I might try to make it a firm one by announcing it, by publicly calling it a final decision, by resolving to ignore additional evidence or resolving not to look for it, or resolving not to think about the issue anymore. Any of these tactics would increase my chances of following through on my decision and would therefore justify a higher credence in my future action, leading to an increase in the expected accuracy of that credence. Firm decisions are more conducive to accuracy about our future actions than tentative ones are. Is the epistemic perspective constantly whispering in my ear telling me that I ought to make firm decisions instead of provisional ones? It's hard to see why the project of having rational beliefs should commit me to making firm

decisions. There are often very good reasons to make decisions that are subject to change; having rational beliefs shouldn't require making irrational decisions.

An epistemic consequentialist who sacrificed their own freedom for the sake of accuracy would treat their decisions as nothing more than instruments of their beliefs—most people probably want accurate beliefs in order to make informed decisions, but a maximally-committed veritist would treat every decision as nothing more than a tool to get more accurate beliefs about their future actions. There's nothing contradictory about that, but it would be a strange thing to do, even in the purely epistemic context veritists operate in. It's hard to see why yelling "I want the accuracy, come out and play!" to the benevolent imps is even epistemically good, let alone mandatory. In that case the one yelling would be sacrificing others' freedom rather than their own, but the fundamental point is the same. When the only reason I don't know whether the imps will play is that they haven't decided yet, I can naturally gain more accuracy if I request that they give up their power to decide and just come out and play, but it doesn't seem like a good thing to do.

The presence of freedom in ourselves and others can't be any kind of obstacle to being a rational agent with rational beliefs—if anything, the fact that we are discussing how agents should form beliefs and decisions that sometimes involve each other indicates we should define the epistemic context in ways that allow all the agents to have freedom. The fact that some agent, whether myself or someone else, has freedom that interferes with my accuracy can't automatically be a bad thing for me, epistemically. This is more a definitional claim than a normative one. We're talking about what epistemic agents should and shouldn't do and believe, and that requires us to presuppose that they have meaningful agency. It doesn't make sense to

treat an epistemic agent's own freedom as an automatic obstacle to their epistemic projects. They're an epistemic agent; they have to have epistemic freedom.

As an extension of that idea, since epistemic agents are operating in a context that involves believing and deciding things about each other, it makes no sense to treat another agent's freedom as an automatic obstacle to one's own accuracy. That epistemic value scheme wouldn't be universalizable, and therefore wouldn't achieve its goals. If I think my own epistemic freedom is no obstacle to my epistemic goals but your freedom is necessarily a problem, I have a strong incentive to deprive you of your epistemic freedom, and you have a strong incentive to deprive me of mine. I will avoid making firm decisions myself but require them from you so that I can more accurately predict your future actions, and you will do the same to me. I will save my RSVP to your party until the last minute but demand that you immediately RSVP to mine. This race to the bottom would leave us both worse off, epistemically. I have a strong epistemic interest in figuring out how your various possible decisions *would* affect the world, just as I want to know how mine would, but the decision itself is up to you.

The fact that the world is uncertain in some ways because agents possess freedom to change it isn't bad news, epistemically; it merely sets the boundaries of what the epistemic project is. We want accurate beliefs about all the parts of the world that don't depend on unmade decisions. Those beliefs will help us all figure out how to use our freedom.

4. Pure Beliefs

So far I've argued that some puzzles about what to believe are only puzzling because they're actually questions about what to decide, and I've given examples of pure decisions and

laid out a theory of the blurry line between belief and decision. Although the line is blurry, there still are plenty of pure beliefs, and in this section I give paradigmatic examples of them.

The original imps case and my variation of it are examples of supposed beliefs that are actually just pure decisions. Treating matters for decision as matters for prediction can lead to strange behaviors like ignoring the existence of children playing in front of you or resolving to only make firm decisions. But we encounter similar problems from the opposite direction: treating matters for belief as matters for decision can lead to very strange behaviors as well. I first noticed this when I heard my friends talking about a children's book called *The Little Prince*. The example they discussed provides a perfect case of apparent decisions that are actually just pure beliefs.

The story follows a lonely young prince who leaves his tiny planet to find people to talk to. He encounters a series of odd characters who inhabit their own lonely little planets. But especially odd was the first of them, the king, the character my friends were discussing for some reason, even though we were all adults and they'd read the book years ago. What distinguishes the king is that he expects absolute obedience to his commands, but he seems to only order things to do what they were going to do anyway. For instance, the prince asks him to order the sun to set, since the king has indicated the sun is his to command, and the king assents, but says he will order the sun to set only at the proper time, consulting an almanac to find out when that will be. When the prince asks him why he can't order it sooner the king explains his strategy with this: "One must require from each one the duty which each one can perform… Accepted authority rests first of all on reason. If you ordered your people to go and throw themselves into the sea, they would rise up in revolution. I have the right to require obedience because my orders

are reasonable."[28] The prince quickly tires of this way of thinking, and, against the king's wishes, leaves the planet. As he goes, the king hastily commands him to serve as his ambassador.

I think the memory of the character stuck with my friends for so many years because there was something confounding about him. But once I read Greaves' imps case, I remembered the king from *The Little Prince* and he suddenly made more sense to me: while a veritist who would accept the imps' offer treats matters for decision as matters for belief, the king is simply doing the opposite. While the maximally committed veritist treats every decision as a mere tool to make their beliefs more accurate, the king treats every belief as a mere tool to make his decisions more efficacious. He comes up with meticulously researched, perfectly accurate beliefs telling him when the sun will set, all so he can decide to have it set exactly then, knowing with certainty that it will comply with his orders.

The king really does believe there is an imperative to have high credences in self-fulfilling prophecies, but he does it for exactly the opposite reasons a veritist might—the veritist would believe self-fulfilling prophecies to have more accurate beliefs, while the king would do so to make more successful decisions. If the king were presented with my pizza-or-burgers dilemma, knowing that a belief he'd have pizza would do a better job creating its own accuracy would make him decide to have pizza. A veritist might believe they'll have pizza to pursue accuracy about their dinner plans, and the king might decide to have pizza to pursue success in his. But they'd agree that the question of whether they like burgers or pizza better is irrelevant.

Ironically, I think the king ends up being a more rational believer than the maximally committed veritist. The king's strange obsession with issuing perfectly efficacious demands requires him to form very accurate beliefs about all the things that don't depend on him so that

[28] Saint-Exupéry Antoine de. *The Little Prince*. p. 29. https://books-library.net/files/books-library.online-12201041Ti6B3.pdf

he can then order them to be the way they are. When the sun rises and sets has nothing at all to do with the king, so the only way he can successfully command it is to know what it will do anyway, and then desire and demand that it do that. The king first makes his mind fit the world by making his beliefs about the outside world accurate, and then he makes his mind's desires fit the world by desiring that the things he cannot change be the way they are. After first figuring out what all the things that don't depend on him are like, the king uses that accurate belief to inform his opinions on all the things that do depend on him, like whether he'll order the sun to set and what time he'll do so. The king is following exactly the procedure I outlined in the previous section, although he's applying it to serve the strange goal of always making successful decisions. He can only maintain the fiction that the world perfectly obeys his commands because he has such perfectly accurate beliefs about it. Despite his intentions, the king's beliefs strike me as much more rational than his decisions.

The king gives us one example of a class of propositions whose truth values are purely matters for prediction: natural events that we can have no causal effect on. On my view, then, laws of nature and the natural events they ordain would be matters for pure belief, and one should pursue the highest accuracy possible about them. Forming accurate beliefs about the natural processes completely outside of our control allows us a better idea of which things we do have some influence over, and we can then use accurate scientific knowledge to inform our opinions about the propositions whose truth-values we have more control over.

The most general way to phrase the category of propositions that are matters for pure belief is this: whatever propositions whose truth an agent has no causal influence on. The laws of nature are one paradigmatic class of cases we have no effect on; that's why we call them laws and try so hard to know them. A second subcategory of propositions we have no causal effect on

is propositions about past events. Whatever has happened in the past is not causally affected by my beliefs and decisions in the present. Take the proposition "The Roman Empire defeated Carthage in the Second Punic War." Whatever happened already happened, so what I think about it won't affect it. Setting aside the possibility of time travel, historical claims are clearly matters for pure belief. I discussed a third category of pure beliefs in the last section: hypothetical claims about what would happen *if* we did certain things.

Interestingly, we should generally figure out the laws of nature and the facts about past events before we figure out the truth of counterfactual claims involving our actions. We don't causally influence any of these three categories. However, if I form beliefs about the hypothetical results of my decisions before accurately describing past events and the laws of nature, my beliefs about counterfactuals are likely to be pretty inaccurate. If I form my beliefs about the other two categories first, my beliefs about the counterfactuals are likely to be more accurate. This third category of pure beliefs really ought to be the third, then.

The strongest objection to my account of beliefs and decisions is that it will tend to deliver results that strike some as counterintuitive in cases where the truth-value of a proposition doesn't depend on your actions causally but does depend on them evidentially. In such cases, to figure out the truth of a proposition that doesn't causally depend on us at all we sometimes have to first figure out the truth of propositions that do causally depend on us.

The Newcomb Problem provides a prime example of this challenge to my theory.[29] The problem goes like this: you are looking at two boxes, one clear one with $1000 in it and the other opaque, its contents unknown, and deciding whether to take just the latter box or the former as well. You know that a nearly perfect predictor placed a lot of money in the opaque box if they

[29] Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." *Essays in Honor of Carl G. Hempel*, 1969, pp. 114–146.

predicted you would take only that box, and put nothing in it if they predicted you'd take both. The problem is usually presented as a paradox about decision: whatever is in the opaque box is already there, so it seems strictly better to take both, but deciding to take both gives you great evidence that you'll end up with less money while deciding to take only the opaque box gives you great evidence that you'll be rich. Your decision doesn't change what the predictor has put in the box, but it gives you evidence about what it put there.

Most philosophers are two-boxers. I can't actually causally affect what the predictor put in the opaque box at all—what matters is not whether I decide to take both boxes, but what the predictor thought I'd decide. The one-boxer acts like their decision in the present can causally influence the past. But the past is already written and whatever the predictor put in the box is already there, so all that's left is a simple decision: take $1000 in the clear box plus whatever's in the opaque one or leave $1000 on the table. The one-boxer has been misled by treating a matter for decision as a matter for belief, basing their decision on predicting what the predictor predicted they'd do. A two-boxer decides to take both boxes, knowing their decision won't change what's in the opaque one because that past event doesn't depend on them, and then they regretfully but logically predict that the opaque box probably has nothing in it.

Even if two-boxing is the right answer to the problem about what to *decide,* the puzzle raises a completely separate problem about what an epistemic consequentialist should *believe* about what's in the opaque box. Newcomb's Problem might appear to give an exception to the epistemic procedure I recommend. I've argued that it makes sense to figure out what all the things that don't causally depend on us at all are like before figuring out the truth of the things that do causally depend partly on our actions. But in this puzzle case, the reliable predictor changes the equation so that we actually have to make a decision before we can form an accurate

belief about what something that doesn't causally depend on us is like. If I really want to know what's in the opaque box, it seems I have to decide whether to take the transparent one first. Causally, the truth-value of "The predictor put a lot of money in the opaque box" doesn't depend on anything I do, but evidentially, my knowledge of the predictor's reliability means that my decision to take one box or two gives me the crucial evidence I need to know what's in the box. So in this case, my epistemic project of having accurate belief about what's in the opaque box requires me to first make a decision about whether to take one box or two. In the original version of the problem, taking zero boxes isn't an option. But if we did include that as a possible choice, a veritist would likely reject it, because they need to take at least one box to learn what's in the opaque one.

I think it's clear that we generally ought to figure out the truth of things that depend less on our opinions before figuring out the truth of things that have more causal dependence on them. There is a separate, more debatable question about whether we should sometimes first figure out the truth of propositions that causally depend on us ("I take the transparent box") in order to then deduce the truth of propositions that don't depend on us causally but do depend on us evidentially ("The predictor put a lot of money in the opaque box"). One could conceivably modify my theory to carve out an exception for these cases of evidential dependency. I'm of two minds about this. I think it's a hard question, whether epistemic rationality demands that we decide to take at least one box.

My initial reaction is that my procedure shouldn't change and I don't need to know what's in the box. I would be satisfied knowing everything I could know about what the predictor put in the opaque box without knowing my own decision about whether to take the transparent one. The whole point of two-boxing is that we don't need to know what's in the

opaque box in order to know we should take both; we already know everything we need to know to make the right decision. Knowing this fact about the outside world—what the predictor put in the opaque box—has no effect at all on my decision about how to interact with the world in this isolated case, although I'll consider complications shortly.

But first, consider an alternative version of Newcomb's Problem that's identical in all ways, except all the money is monopoly money. Now the pure epistemic problem is revealed, since there's no practical reason to one-box or two-box, or take any boxes at all. A die-hard veritist would say "One-box or two-box, I don't care which. But definitely take at least one box so you get to know what's in the opaque one." That accuracy would strike them as low-hanging fruit, while I'm inclined to regard it as epistemic junk food. So, the monopoly-money version of Newcomb's Problem serves as a good diagnostic tool to find strong veritist intuitions, just as the trolley problem is a good diagnostic to reveal utilitarian ones. A die-hard veritist will regard it as obvious one should decide to take at least one box: it's free accuracy.

The backdrop of the epistemic bribe Joyce and Weatherson consider suggests another kind of diagnostic. As I said, from my perspective, if Ankita's credences of .9 in D0 and .5 in D1 perfectly match the objective chances of those propositions, things couldn't be better. But a veritist who values accuracy as an end in itself might think her better off if both her credences and the objective chances of D0 and D1 were 1: accurate belief is like epistemic pleasure, and she gets to feel more epistemic pleasure. Some veritists really do prefer objective chances to be 1, and those veritists would be more likely to prefer taking at least one box to none in the monopoly-money version of Newcomb's Problem.[30]

[30] I asked Richard Pettigrew about his intuitions on these two diagnostics, and he confirmed that he'd prefer objective chances to be 1 and would take at least one box.

I truly have no urge to take any boxes of monopoly money, though I can understand others wanting to and don't think they're necessarily irrational. I already know the truth of the conditional statements "If I take one box there is $1000 in the opaque box" and "If I take two boxes there's nothing in the opaque box." The only reason I don't know which consequent is true is that I haven't decided which antecedent to make true. I don't see why having rational beliefs about the world should demand that I decide to take one box of monopoly money or two. I'd ignore the choice altogether.

This is my initial reaction to this isolated case. However, there are modified versions that would be more challenging for me. Suppose, for instance, that I know that if the predictor didn't put a thousand dollars of monopoly money in the opaque box, it built a statue of itself. There is a proposition about the world that has nothing at all to do with me: "A statue of the predictor exists." If I decide to take one box, I immediately get to know not only that there's a thousand dollars of monopoly money in it, but that I should have credence close to 0 that a statue of the predictor exists. If I decide to take two boxes, I immediately get to know that there's no monopoly money in the opaque box, and I get to have credence 1 that a statue of the predictor exists. But if I don't decide to take any boxes, reasoning that monopoly money is worthless, then the best I can do is have some intermediate credence that a statue of the predictor exists.

There either is or isn't a statue of the predictor out there, built long ago, and its presence may have changed many other facts about the world. Now I have a decision to grapple with beyond the question of how many boxes to take: should I try to visit the statue or not? I can decide that I should two-box to maximize my money (monopoly or otherwise) without knowing what's in the opaque box but deciding whether I'll visit the statue or not is tied to knowing what's in that box. All I have to do to know these things about what the world is like is decide to

take one or two boxes of monopoly money. Then I can make a good decision about whether to try to visit the statue.

Part of me insists that epistemic rationality simply cannot demand that I decide to take boxes of monopoly money, even if making the decision would grant all these evidential benefits about what the world outside my opinions is like. That part of me continues to insist that I don't need to know what's in the box or whether a statue of the predictor exists; it's good enough that I know the conditional statements about how my actions would give evidence of their existence. I don't need to know whether the statue exists or whether anyone has ever visited it; I know that if I take two boxes the statue must exist. It will be computationally difficult and will tax my brain to keep track of all the conditionals about how the world might be instead of determining the antecedents so I can focus only on the consequents, but that's only a practical consideration.

But maybe that's wrong. I have to decide whether to try to visit the statue or not, I need evidence about whether it exists in order to make that decision, and I have to take at least one box to get that evidence. There's nothing strange about deciding to read a book to get evidence about the world, and maybe deciding to take these boxes of monopoly money is much like reading a book. When it comes to the book, though, it's the actual reading of it that gives me accuracy about the world, not the decision to read it. The decision's not the evidence; the book is. With cases like Newcomb's Problem and these variations, it's my mere decision to take one or two boxes that constitutes the evidence about what's in the opaque box and whether a statue of the predictor exists. Strange, but maybe the same epistemic goal that requires me to gain evidence about the past by reading books sometimes requires me to gain evidence about the past by making decisions about the present. Usually, we gain evidence by finding it, but in these

cases involving learning the truth of propositions that depend on our actions evidentially (though not causally) we gain evidence by creating it.

I go back and forth on whether or not epistemic rationality requires me to take at least one box of monopoly money. In the end, I think both views are plausible and reasonable people may differ about the question. My intuition about whether there's epistemic value to knowing what's in the opaque box depends on whether any decisions I'll face might require that knowledge. If not, I'm content with knowing conditional statements. I imagine most veritists would disagree with the way I draw the line between practical interests and epistemic ones.

5. Mixed Cases of Belief and Decision

Although I've given examples of pure beliefs and pure decisions, demonstrated that there are impure cases that combine the two, and given a theory of how belief and decision could be opposite ends of the spectrum, the best way to really prove the existence of that spectrum is to analyze common cases towards the middle. In this section I consider two paradigmatic examples of mixed beliefs and decisions: a hospitalized patient's credence about whether they'll get better and an investor's beliefs about the value of a stock. These cases each have strong elements of decision and belief to them, so much that it would be inapt to describe either of them primarily as predictions or decisions. Here's the first case.

Recovery:

Suppose a very ill patient knows that their credence that they'll get better is to some degree a self-fulfilling prophecy, and that a high credence in their own recovery is necessary but not sufficient to recover. In particular, they have reason to think that a lower credence would cause its own accuracy even more than a higher credence would; optimism is helpful, but pessimism is fatal. Is this an epistemic reason for the patient to have lower credence in their recovery?

I could just as easily have set up the example so that a high credence in recovery would cause its accuracy even more than a low credence would cause its own, but I think the reverse is likely to be more prevalent. Either way, this is an all-too-common situation, and there doesn't seem to be any imperative of rationality to adopt whichever credence would contribute most to its own accuracy. A patient in this situation is being perfectly rational if they choose to have as high a credence in their recovery as they can while still respecting their evidence. There would, of course, be something epistemically irrational about convincing themselves their disease is less serious than it really is, even if that would lead to greater accuracy in their belief they'll recover—that would just be trading the accuracy of one belief for another, and ignoring relevant evidence in the process, which would likely be pragmatically justified. Epistemically, the patient's credence in their recovery must be appropriately limited by their credence in the seriousness of their illness, and that crucial fact about the world doesn't depend much on their opinions. But if the patient has let the evidence dictate their credences about the nature of their illness and there is still some leftover question about exactly what credence they should have that they will recover, they are rationally permitted to believe what they want to believe. They are epistemically free to believe they'll get better, even if it would be more accurate to believe they'd succumb to their illness.

Suppose that, given the gravity of their illness, the patient knows that even the most optimistic patients in their situation only recover 70% of the time, while the most pessimistic die 100% of the time. Accepting those background facts about the world is what I have been calling the process of forming a belief, and they ought to aim for the highest accuracy possible when forming these beliefs about the nature of their illness. But these beliefs about the world leave it

undecided whether the patient should be optimistic or pessimistic; that is still up to them. Having accurate beliefs about the range of possibilities allows the patient to imagine counterfactuals—what will happen to me if I believe in my own recovery as much as the evidence will allow? What will happen to me if I doubt my recovery as much as the evidence will allow? At this stage the question is not whether optimism or pessimism would tend to be more accurate; extreme pessimism would be bound to be perfectly accurate, while optimism would not. But the patient has already done a good job of grasping the range of options that are consistent with their evidence, so pursuit of accuracy is meaningless when deciding their final credence in their recovery. They know all the relevant facts about their illness. The only thing they don't yet know is whether they'll choose to raise their chance of recovery. So if the patient then chooses to have credence .7 that they'll recover, their high credence in the proposition "I'll get better" is the product of their beliefs about the nature of their illness and their decision to try to get better. The facts about the world recommend a particular imprecise credence of 0 to .7 that they'll recover, and their decision to maximize their chance of recovering then sharpens that imprecise credence to a precise one of .7.

Just as a belief about the course of one's illness can be a self-fulfilling prophecy to a degree, so can a belief about the value of one's investments. This claim probably sounds much more debatable, and even foolish and dangerous; thinking that your belief in an investment's value makes it more valuable sounds like a good way to go broke. According to the efficient market hypothesis, a hotly debated theory that tends to be more prevalent among academics than financial professionals, what the professional investors are doing is actually pretty pointless, when they invest in public markets, at least. The efficient market hypothesis can be formulated in various ways. In its strongest form, it claims that the market perfectly reflects all public

information and that all public companies are therefore correctly valued, so that it's theoretically impossible to beat the market through selectively investing in companies, on a risk-adjusted basis—the best one can do is buy the market. Weaker forms state that reliably beating the market is highly unlikely.

Some recent financial events provide evidence that might shake one's faith in the efficient market hypothesis. These disparate events follow a pattern that seems to suggest that individuals or groups of investors can sometimes decide that an asset is more valuable than its market price, buy it, and through various mechanisms actually make it more valuable. Economist John Quiggin, for instance, suggests that the continued rise of bitcoin, which he and many other economists consider worthless, provides conclusive evidence against the efficient market hypothesis.[31] Whether bitcoin has any fundamental value, and what that value is, depends largely on whether it's a store of value or not (or a currency). To a large degree, the more people who have high credence that it's a store of value, the more it becomes one. JP Morgan's CEO Jaime Dimon is a prominent skeptic of cryptocurrency, yet JP Morgan recently started offering its clients access to cryptocurrency funds because those clients wanted to invest in it.[32] In general, the trend of the last decade or so seems to be that financial professionals almost universally regarded cryptocurrencies like bitcoin and ethereum as worthless, yet the public decided they were wrong, and by deciding that cryptocurrency was worth something, made it so. The professionals treated the value of assets like bitcoin as a matter for belief, applying their normal

[31] Quiggin, John. "The Bitcoin Bubble and a Bad Hypothesis." *The National Interest*, The Center for the National Interest, 31 May 2013, https://nationalinterest.org/commentary/the-bitcoin-bubble-bad-hypothesis-8353.
[32] Bhattacharya, Ananya. "Jamie Dimon Thinks Bitcoin Is 'Worthless' but Says Clients Can Do as They Please." *Quartz*, Quartz, 12 Oct. 2021, https://qz.com/2072753/jamie-dimon-says-bitcoin-is-worthless-but-jpmorgan-will-sell-it/.

methods of valuation to arrive at a belief it was worthless, while the public simply overrode that assessment by deciding bitcoin was valuable.

Of course, one key difference between this kind of example and all others in this paper is that it takes a large *group* of people to have high credence in the proposition "Bitcoin is a store of value" (or "Bitcoin is a currency") for that belief to become a self-fulfilling prophecy and become a group decision. An individual's credence that bitcoin is a store of value has almost no ability on its own to become a self-fulling prophecy. But those individual small bits of self-enacting power add up: one person's initial high credence that bitcoin is a store of value causes them to start evangelizing it to all their friends, some of them become converted, and the number of people who believe bitcoin is a store of value continues to grow, eventually at an exponential rate. And some high-profile individuals like Elon Musk clearly have far greater power to have their beliefs about cryptocurrency become self-fulling prophecies. After Musk started tweeting about cryptocurrency frequently, he exhibited significant abilities to move the price of bitcoin and all cryptocurrencies by tweeting positive or negative statements about their value. Musk's tweets about bitcoin acted like self-fulling prophecies, and it becomes appropriate to wonder whether Musk was predicting bitcoin's value or deciding it. Basically, the more money and influence an actor has, the less the efficient market hypothesis seems to apply to them. So bitcoin does provide good evidence against the efficient market hypothesis, but not necessarily because it's a worthless thing that has enduring value; it demonstrates that investors can sometimes choose how valuable an asset will become instead of merely predicting it.

Other recent events challenge the efficient market hypothesis by suggesting that to some degree investors can simply decide that an asset is more valuable than its market price. Tesla, for instance, has been a poster-child for overvaluation for the last few years, according to standard

models that input a variety of factors like assets, debts, and discounted future earnings. But as Tesla shares skyrocketed, the company sold more of them to the public. Suddenly it had billions of dollars show up on its balance sheet, and that treasure trove of cash necessarily had to increase its value according to those same models that had declared it overvalued, increasing its credit rating and production.[33] Suddenly the models considered it less overvalued than before, and famous short-sellers who'd bet against Tesla quietly announced that they had closed their short positions.[34] A similar phenomenon occurred with AMC. Ravaged by pandemic closures, AMC was facing bankruptcy, and short-sellers were betting against it to hasten its demise. Groups of online investors decided to buy AMC shares en masse to save the company. This buying caused the shares to rise, causing brokers to margin call short-sellers and buy tons of shares to close their short positions, causing a further rise in the share price, and eventually AMC raised billions of dollars by offering more of its suddenly-valuable shares to the public.[35] Just as with Tesla, once AMC turned the hype behind its shares into billions of cash on its balance sheet, the analysts using standard valuation models were forced to suddenly declare that its high share price was no longer as unreasonable as before.

At its heart, the efficient market hypothesis treats the question of a company's value as a matter for pure belief, with no room for decision. It holds that all public information relevant to the company's value is fully priced into the shares the moment the information becomes public. Then, since it treats the question of the company's value as a matter for pure belief, a mere

[33] Hull, Dana. "Tesla (TSLA) Elon Musk Raising up to $5 Billion in Third Capital Raise This Year." *Bloomberg.com*, Bloomberg, 8 Dec. 2020, https://www.bloomberg.com/news/articles/2020-12-08/tesla-raising-up-to-5-billion-in-third-capital-raise-this-year.
[34] Lambert, Fred. "People Are Not Betting against Tesla (TSLA) Anymore, Short Interest at All-Time Low." *Electrek*, 4 Oct. 2021, https://electrek.co/2021/10/04/people-are-not-betting-against-tesla-tsla-anymore-short-interest-all-time-low/.
[35] Saldanha, Aaron. "AMC Stock on Rollercoaster, Movie Chain Completes Another Share Sale." *Reuters*, Thomson Reuters, 4 June 2021, https://www.reuters.com/business/amc-shares-jump-another-13-reddit-rally-extends-2021-06-03/.

prediction of value instead of a decision, once it adds the claim that all public information is priced into the shares there is nothing left that could make them gain or lose value, assuming insider trading is appropriately regulated. What the efficient market hypothesis overlooks is that investors have a degree of freedom to decide a company's value; an investment isn't just a prediction that a company will become more valuable, it's a decision to *make* it more valuable.

One investor famous for taking advantage of this fact is Cathie Wood, founder of the popular ARK Invest funds. Wood bets big on young companies pursuing disruptive technology, and the sheer size of her bets makes her own a disproportionate amount of the companies' shares, a double-edged sword—her own demand is sufficient to inflate the shares' prices, but critics worry that if ARK's investors sell, forcing it to sell some of its holdings in its companies, its own selling pressure will cause it's investment's share prices to plummet.[36] Wood's choice of the name "ark" actually reflects her intent to cause the companies she invests in to become more valuable—she views her investments as the creation of a kind of ark that will carry humanity into the future.[37] Wood invests in disruptive companies not just because she predicts they'll become more valuable, but because she has decided to make them so.

Certainly, such decision-making power has strict limits; many investors who try to make a company more valuable fail and lose money, as Wood has lately. That's because there are elements of belief and decision to the question of what a company's value is, and a decision to make it more valuable can only go so far to counteract the belief that public information justifies. The market often wins. But the fact that decision plays a significant role at all means that the

[36] Jakab, Spencer. "Cathie Wood's Ark Wasn't Built for a Flood." *The Wall Street Journal*, Dow Jones & Company, 11 May 2021, https://www.wsj.com/articles/ark-wasnt-built-for-a-flood-11620749548.
[37] So, Dorothy. "Ark Angel: Meet Reddit's Favourite Fund Manager Cathie Wood." *South China Morning Post*, 3 Mar. 2021, https://www.scmp.com/magazines/style/celebrity/article/3123779/cathie-wood-religious-reddit-hit-trump-supporter-meet-ark.

question of the value of a company is a matter of both belief and decision, a common mixed case of the two.

6. **Conclusion**

I began this paper by arguing that there are clearly some cases where one should not pursue accuracy at all costs—cases where one gains more accurate beliefs about the world by making decisions about how to influence it. Such cases are matters for decision, not belief, and answer to standards other than the pursuit of accuracy. This position led me to develop a general theory of the distinction between beliefs and decisions. I argued that, to the degree that a proposition's truth-value depends on one's own actions, determining the truth of that proposition amounts to making a decision, and to the degree that a proposition's truth-value does not depend on one's own actions, determining the truth of that proposition amounts to forming a belief. Because the truth-values of particular propositions often depends partly on one's own actions, and can do so in varying ways and to varying degrees, I concluded that the distinction between decision and belief is often a matter of degree, not kind. I presented cases of pure decisions, pure beliefs, and examples of where one's opinion regarding a proposition's truth-value is both belief and decision. Beliefs aim to make the mind fit the world, and decisions aim to make the world fit the mind, but sometimes a credence about a proposition tries to do both.

BIBLIOGRAPHY

Andow, James. “Do non-philosophers think epistemic consequentialism is counterintuitive?” *Synthese* vol. 194, no. 2631–2643 2017.

Berker, Selim. “Epistemic Teleology and the Separateness of Propositions.” *Philosophical Review*, vol. 122, no. 3, 2013, pp. 337–393.

Carr, Jennifer. “Epistemic Utility Theory and the Aim of Belief.” *Philosophy and Phenomenological Research*, vol. 95 issue 3, 511-534, p. 515.

Dunn, Jeffrey. “Group Epistemic Value.” *Philosophical Studies*, 2021.

Greaves, Hillary. “Epistemic Decision Theory.” *Mind*, vol. 122, no. 488, 2013, pp. 918.

Hills, Alison. “Moral Testimony and Moral Epistemology.” *Ethics*, vol. 120, no. 1, 2009, p. 101.

James, William. “Is Life Worth Living?” from *The Will to Believe and Other Essays in Popular Philosophy*. New York, London, and Bombay: Longmans Green, 1896, 1899, pp. 32-62.

Joyce, James M. “A Nonpragmatic Vindication of Probabilism.” *Philosophy of Science*, vol. 65, no. 4, 1998, pp. 575–603.

Joyce, James M., and Weatherson, Brian. “Accuracy and the Imps.” *Logos & Episteme* vol. 10, no. 3, 2019, pp. 266.

Nozick, Robert. “Newcomb’s Problem and Two Principles of Choice.” *Essays in Honor of Carl G. Hempel*, 1969, pp. 114–146.

Pettigrew, Richard. *Accuracy and the Laws of Credence*. Oxford University Press, 2018.

Pettigrew, Richard. “The Population Ethics of Belief: In Search of an Epistemic Theory X.” *Noûs*, vol. 52, no. 2, 2016, pp. 336–372.

Singer, Daniel J. “How to Be an Epistemic Consequentialist.” *The Philosophical Quarterly*, vol. 68, no. 272, 2018, pp. 583–585.

Talbot, Brian. “Repugnant Accuracy.” *Noûs*, vol. 53, no. 3, 2017, pp. 540–563.

Velleman, David. The Possibility of Practical Reason. Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2009, pp. 35.

CHAPTER II

A Theory of Suspension of Judgment

1. Introduction

Imagine the following conversation, filling in 'X' with any proposition:

Amy: I wonder whether X is true.
Bill: I have no idea.
Claire: I'm 50/50 about it.
Dan: I'm suspending judgment about that.

I'm trying to figure out what people like Dan mean. My sense of things is that all four participants in this conversation are expressing different perspectives about X's truth. It's clear, at least, that Amy and Bill aren't expressing the same thought—one could have no idea whether X is true without wanting to know whether it was, and one could wonder whether X is true while having some idea. Similar reasons explain why Amy and Claire aren't saying the same thing. Bill and Claire aren't expressing the same attitude because one has *some* determinate idea about X when they're 50/50 about it.

But it's less clear whether what Dan is doing is reducible to what any of the others are. On some accounts of suspension of judgment, which reduce it to being in a state of opinionlessness, Dan is doing the same thing Bill is—reporting his absence of belief. Other accounts reduce suspending judgment to adopting a middling credence, which means Dan is saying basically the same thing Claire is. And on Jane Friedman's account of suspension of judgment, where one suspends judgment about whether a proposition is true if and only if they're inquiring about it, I take it Dan is responding to Amy's statement that she wonders whether X is

true by saying something like "Me too."[38] But I can't escape the feeling that Dan isn't just agreeing with any of the three who spoke before him. I think he's expressing a perspective regarding whether X is true that wasn't part of the conversation before. In this paper I try to explain what Dan is saying that none of the others have. What do people mean when they say they're suspending judgment?

It's surprising how poorly defined the notion of suspension of judgment is, given that it's discussed so frequently and its origins are so ancient.[39] I'm interested in questions like when we should suspend judgment and what impact doing so has on the accuracy of our beliefs, but investigating them requires a description of what suspension of judgment is and a theory of when we can do it. My main goal in this paper is just to answer the descriptive questions about suspension of judgment well enough that the concept can be applied to normative ones.[40] Figuring out what we can and can't be doing when we suspend judgment will help me figure out what it is to suspend judgment, which will help me understand how suspending judgment affects accuracy, which will shed light on the questions of when the pursuit of accuracy should lead one to suspend judgment and when it forbids doing so.

There isn't that much work about what suspension of judgment is; for the most part, it seems to be a concept that's meant to be intuitively understood through its contrast with belief

---

[38] Friedman (2017)

[39] Descartes famously suspended judgment on all his beliefs at the beginning of the *Meditations*, and Pyrrhonian skeptics advocated broadly suspending judgment more than two thousand years ago.

[40] Tang (2016) takes the opposite approach and explicitly sidesteps the question of what suspensions of judgment are, and whether they're reducible to credences (364), to go directly to the question of when suspending judgment is justified. He's interested in defending process reliabilists against the charge that evidentialists can better account for justified suspension of judgment than they can.

It may be possible to investigate questions like that without a detailed account of what suspending judgment is. Tang begins with the assumption that suspension is an alternative to belief and disbelief (362). There, at least, we disagree. My theory of what suspending judgment is leads me to conclude that it's both possible and normal to suspend about P while having a belief about it. We may be able to make some progress on normative questions about suspending judgment without having a fleshed-out description of what it is, but I suspect that our answers to the normative questions will be different if we approach them armed with answers to the descriptive ones.

and disbelief.[41] I'll assess what little work there is describing suspension of judgment as I offer my own theory of it. Two features distinguish my approach. First, I'm one of a few people who take up the question of how suspension of judgment fits in a picture of beliefs as degrees of confidence. It's usually portrayed as an alternative to full belief or disbelief, and at first glance one who preferred to think in terms of partial beliefs might have no need for the notion of a doxastic state that functions as some kind of alternative to belief. I'll assess two reductions of suspended judgment to partial belief. The first holds that suspensions of judgment are nothing more than a certain kind of precise credence and the second that they're nothing more than a certain kind of imprecise credence. I argue that both reductions of suspended judgment to partial belief lose something important in the reduction.

One factor they both overlook is that there's a voluntariness to the way people suspend judgment that doesn't seem to be present in the way we believe things. We sometimes ask people to suspend judgment, as we do with jurors, and people sometimes declare that they're deciding to suspend judgment, a decision we sometimes disapprove of. Agnostics sometimes say they're suspending judgment about whether God exists, for instance, and atheists sometimes tell them that they should stop. The actions of jurors and agnostics provide two paradigmatic cases where people suspend judgment, and in both cases, we often discuss the act of suspending judgment as if it's voluntary. Descartes decided to suspend judgment about everything he wasn't completely certain of, which means he thought he had the ability to suspend judgment at will about even matters he was very confident of.[42]

[41] Tang begins his paper with this strategy, saying "…lacking the relevant evidence [about a proposition's truth], you'll neither believe nor disbelieve the relevant proposition. Instead, you'll suspend belief in it—your attitude towards it will be one of agnosticism." (362) It's not clear to me whether he thinks that most people will choose to suspend judgment in these situations or that they'll unavoidably do so.

[42] I analyze what Descartes was doing in greater detail at the end of the paper, in Section 6.

This is the second feature that distinguishes my approach to describing suspension of judgment from others: as far as I can tell, I'm the only one who tries to understand what suspension of judgment is by contrasting its voluntary nature with the involuntary nature of belief. The heart of the idea is that we don't seem to be able to directly choose what we believe, but we do seem to be able to directly choose whether to suspend judgment or not.[43] As nebulous as the notion of suspending judgment is, I take it to be a defining characteristic of it that we can frequently choose to do it. On the other hand, many people, myself included, think that once we've figured out which belief our evidence best supports we can't choose which belief we have; in my view, figuring out which belief the totality of our evidence supports is the process of coming to have that belief.[44] If we can choose to suspend judgment but we can't choose what to believe, understanding which doxastic acts *are* under our control will help us understand what it is to suspend judgment. Instead of arguing for voluntarism for suspension of judgment and involuntarism for belief I'll use those assumptions to create a theory of suspending judgment that has explanatory power.

Whatever suspending judgment is, it's something that we can choose to do, and directly selecting our doxastic state doesn't appear to be something we can choose to do. Taken together, these claims may be puzzling, but I don't think they're contradictory, and resolving the tension between them will reveal what it is to voluntarily suspend judgment. The theory I arrive at after rejecting three reductive accounts is that when we suspend judgment about whether a proposition is true, we aren't choosing to have a particular doxastic attitude about its truth; we're choosing to

[43] Perin (2015) argues that Sextus Empiricus was committed to the view that Pyrrhonian skeptics can't suspend judgment at will, and can only do so when they think the evidence warrants it. He appears to be neutral on the question of whether Sextus Empiricus actually believed people could suspend judgment at will or not. Descartes, at least, certainly believed he could suspend judgment at will, and we ask juries to do it, so I'm comfortable assuming that the modern notion of suspension of judgment holds that we can choose to do it at will.

[44] I'm going to assume doxastic voluntarism is false. For a recent defense of it, see Steup (2012).

treat the doxastic attitude we do have towards that proposition's truth differently from the attitudes we have towards the truths of propositions we haven't suspended judgment about. We have no direct control over what we believe once we've assessed evidence, but we can display a finer degree of control over what we do with the beliefs we're stuck with. The way we choose to use our existing beliefs can indirectly shape the way our sets of beliefs evolve, which means that suspending judgment can be a roundabout way of attempting to guide belief-forming processes that aren't under our direct control. By voluntarily changing aspects of the way we think about propositions we can influence what we involuntarily come to believe about them.

As my focus on the voluntary nature of suspending judgment suggests, my approach to understanding the state of suspended judgment involves first forming a theory of the act of suspending judgment. Discussions of suspension of judgment suffer from the fact that the phrase itself is ambiguous between the act of suspending judgment and the state of suspended judgment. I haven't come across any work that sharply distinguishes between the two, probably because "suspension of judgment" can refer to both.[45] The phrase "I'm suspending judgment" is also ambiguous—it could either be a declaration that one is taking the action of suspending judgment, or a report that one is already in the state of suspended judgment. The phrase "I'm going to suspend judgment" sounds intelligible to me, and it refers only to the act. Although the phrases we use to describe suspension of judgment are often ambiguous between the act and the state, I hope my assumptions that there is such a thing as the act of suspending judgment and that it's distinct from the state of suspended judgment are uncontroversial.

[45] Friedman, who has thought about this issue at greater length than anyone else, displays this blurring between the act and the state of suspension of judgment in the first footnote of her latest article: "Talking about 'the state of suspended judgment/suspension of judgment' can be a bit cumbersome. As such I will often just call the attitude 'suspension' or say that a subject 'suspends about…' or 'is suspended.'" (2017, 322) The title of the article, "Why Suspend Judging?", appears to refer to the act, but she doesn't distinguish between explanations of what the act is and what the state the act aims to put one in is.

Most people who think about these questions offer theories about the state of suspended judgment, contrasting it with the state of belief. By first coming up with a theory about what the act of voluntarily suspending judgment could be I hope to arrive at a more informed theory describing the state it puts one in. The theory that results from my approach is that when we choose to suspend judgment about whether a proposition is true, we're altering the way we treat our belief about it for the purpose of mimicking the way our set of beliefs would evolve if we didn't have that particular belief. I'll argue that the act of suspending judgment about whether P is true is an attempt to sequester your belief about P's truth from the rest of your beliefs in order to eventually arrive at the doxastic state you would've been in if you lacked your actual belief about P's truth.

Much of my thinking about suspending judgment has been informed by the work of Jane Friedman, who has written a series of articles on the subject over the last several years. I share many of her starting points. In her 2013 article "Suspended Judgment" she argues that the state of suspended judgment involves having some doxastic attitude towards a proposition. I agree. It's a sign of how little has been said about suspension of judgment that the point that it involves having some attitude was defended at length so recently. In Section Two I make my own arguments against the most basic reduction of suspended judgment, which attempts to reduce the state of suspended judgment to the mere absence of belief about a proposition.

Having established that suspending judgment involves having an attitude, Friedman's next article, "Rational Agnosticism and Degrees of Belief," considers and rejects a theory that reduces the state of suspended judgment to the state of having a middling precise credence about a proposition's truth. Again, I agree with her, and in Section Three I present some problems with this reduction. At the end of that article Friedman mentions the possibility that the state of

suspended judgment might be reduced to the state of having some kind of imprecise credence, but she never investigates that reduction. In Section Four I take up the problem and explain why despite its many virtues the reduction to imprecise credences overlooks something important.

With these three reductions out of the way, I then turn to positive accounts of suspension of judgment, presenting mine in Section Five. My theory explains the state of suspended judgment by giving an account of what the voluntary act of suspending judgment is and examining the state the act puts one in. I explain how my theory maintains the most important virtues of each of the three reductions while avoiding their shortcomings. Then in Section Six I compare my theory of suspending judgment to Friedman's own, which she presents in her 2017 article "Why Suspend Judging?" Out of intellectual curiosity I avoided reading Friedman's positive theory until I had formed my own to see if we would reach similar destinations from the same starting points. To my surprise, although we agree on much about what suspending judgment isn't, our accounts of what it is ended up being very different. On her account, where one suspends judgment about a proposition's truth if and only if they're inquiring whether it's true, when one person says "I wonder whether X is true" and another responds "I'm suspending judgment about that" the two are expressing similar thoughts; on mine, they're expressing much different ones.

2. Reducing Suspended Judgment to the Absence of Belief

First I'll address two less plausible accounts of suspension of judgment under which the phenomenon is already captured by existing concepts in the credence framework, followed by a third, more plausible reduction. The first theory is that being in a state of suspended judgment with respect to a proposition is just being in a state where one lacks any belief about the

proposition's truth. The second theory is that suspending judgment is just adopting a particular kind of precise credence. The third reductive account is that suspending judgment is just adopting a particular kind of imprecise credence. I'm not compelled by any of the reductions, but the third comes close, and each of them contains useful insights which will inform my theory of the act of suspending judgment and the state it puts one in.

We might be tempted to define suspended judgment simply in opposition to committed judgment, thinking of committed judgment as any belief. So we might be inclined to think that suspended judgment is just the label for the state one is in when they lack committed judgment, meaning that the existing framework of formal epistemology has a perfectly good account of suspended judgment. In traditional epistemology, suspended judgment would be the absence of full belief about a proposition, and in formal epistemology, suspended judgment would just be the absence of any credence about it, including extremal ones.

So far I've said that the theory holds that the state of suspended judgment is the absence of a belief about a proposition's truth. Ralph Wedgwood has a succinct explanation of why it can't be right: "[T]he property of neither believing nor disbelieving p is not a mental state at all—even rocks and numbers have that property."[46] Sextus Empiricus defined suspended judgment as the *considered* absence of belief.[47] Jane Friedman has some objections to the refined view that I agree with, but I won't recount them here.[48] Instead I'll focus on a different line of criticism: even the refined theory reducing the state of suspended judgment to the absence of belief can't give a satisfying explanation of what the act of suspending judgment is.

[46] Wedgwood (2002, 272)
[47] Perin at 109
[48] Friedman (2013, 170)

I suppose the theory would be that someone who intentionally suspends judgment on a proposition, for instance by shifting from belief to suspension, is deliberately moving to a state where they lack belief about a proposition's truth value. Perhaps someone could also choose to suspend judgment without having a belief prior to the suspension. In that case, intentional suspension of judgment would be the conscious decision to remain in a state of opinionlessness instead of leaving it. So intentionally suspending judgment would amount to voluntarily placing one's self into the exact same state of suspended judgment that people are always involuntarily in with respect to a proposition when they aren't even aware of it, and once the act is done the doxastic state is the same no matter how intentionally it was arrived at.

There's something appealing about this picture, something about the connection between suspended judgment and absent belief, but it can't be the right account of intentional suspension of judgment. Here's a simple illustration of why intentionally suspending judgment has to be more than just returning to the state of absent belief one had about a proposition's truth value before becoming aware of the proposition.

A: My evidence says it's a little more likely than not that P, but I'm going to suspend judgment about P.

B: Okay, are you done suspending judgment about P?

A: Yeah. I mean, I know what P means, but I don't have any opinions about it now, just like when I wasn't even aware of P.

B: Okay, now let me remind you about all your evidence about P. It says P is a little more likely than not, so you should stop suspending judgment about P now and believe that P's a little more likely than not instead. This is the same thing you'd normally do when you lack belief about a proposition's truth value and then you get evidence bearing on it.

A: That's a good point. I'll have to do that, or at least consider it, and then suspend judgment again. Don't remind me of my evidence next time.

If choosing to suspend judgment about a proposition is just intentionally moving to a state of lacking opinions about its truth, then just as someone who lacks belief about a proposition should adopt a belief upon becoming aware of evidence bearing on its truth, the one

who has suspended judgment intentionally should also adopt a belief upon becoming aware of evidence about what belief is best. Under this first reductive theory, where intentionally suspending judgment is just moving to a state of absent belief, someone could always choose to intentionally suspend judgment a second time instead of having the belief their evidence best supports, remaking the same choice they made the first time, but that's what they'd have to do, and it doesn't seem to be what anyone who intentionally suspends judgment claims to be doing.

Suspending judgment isn't just doing your best to have no opinion about a proposition, because when someone has no opinion they should adopt one after becoming aware of relevant evidence. Agents who suspend judgment already have relevant evidence in their possession, so if they attempt to return to a state of ignorance, they're obligated to act as an ignorant person would upon suddenly discovering important information, which agents who suspend judgment merely have to remember. One could pursue the further move that suspending judgment on a proposition also involves resolving to pretend as if your evidence that bears on it doesn't exist, but that doesn't seem promising.

My own theory, ironically, ends up resembling this unpromising move in some ways. I end up saying not that suspending involves pretending that one's evidence about the proposition they've suspended on doesn't exist, but that the evidential relationship runs the other direction: it involves pretending that the proposition they've suspended on doesn't have the evidential implications for their set of beliefs that their actual belief about it suggests it does. Although I don't think the state of suspended judgment can be reduced to the state of lacking belief, I do think that the act of intentionally suspending judgment about P is an attempt to place ourselves, over time, in a doxastic state close to the one we would've been in if we lacked our actual belief about P. I expand on this in Section 5.

3. Reducing Suspended Judgment to Precise Credences

The first reduction, from suspended judgments to absent beliefs, responded to both people who think all belief is full and people who think some beliefs are partial. The next two claim that once we conceive of beliefs as degrees of belief, we have no need for a notion of suspended judgment that is an alternative to belief, since a suspended judgment is just a certain middling or indeterminate degree of belief. The second theory reduces intentionally suspending judgment to intentionally adopting a particular kind of precise credence.

On the most naïve version of the reduction to precise credences having a credence of exactly .5 is both necessary and sufficient for suspension of judgment. I'm against most naïve conceptions of things, so a somewhat less naïve theory is that there is some symmetric range around .5, perhaps vague and contextually determined, such that falling within that range is both necessary and sufficient for a precise credence to be a suspension of judgment. The main motivation for the reduction to a middling precise credence is the thought that the concept of suspension of judgment is simply the way the traditional full belief model acknowledges high uncertainty. Since precise credences are able to reflect many degrees of uncertainty, once we move from conceiving belief as full to partial we no longer need to posit some extra doxastic attitude beyond belief that's suited for cases of high uncertainty. High uncertainty is just highly divided belief, which amounts to a middling precise credence.

Before I criticize this second reduction, I want to make an observation about its implications for a bigger picture. There is a strange but real possibility that whether we choose to describe belief as generally full or generally partial makes a huge difference towards what psychological capabilities we think we have. I'm referring to the issue of doxastic voluntarism. If we tend to think of belief as a switch, where you either believe something or you disbelieve it

if you have any belief at all about its truth, then it's natural to think that justified high uncertainty, which makes full belief inappropriate, calls for some epistemic response other than belief. Then it's natural to think that a doxastic attitude other than belief exists, something called suspended judgment, and that we can choose to have that attitude instead of belief because belief would sometimes be unfitting. So, because the structure of belief makes it obviously unfitting in situations of high uncertainty, we deduce that we have an ability to choose not to believe by suspending judgment. If we go this far, perhaps we become friendlier to the idea that we can also choose what to believe, not just whether to believe or suspend. Maybe, at least, if we come to believe in partial beliefs, we continue to hold on to our hypothesis that we can choose to suspend judgment rather than have belief.

But unlike the full belief picture, the degreed conception of belief doesn't invite the presumption that we have an ability to choose to adopt some doxastic attitude other than belief. Full belief is a bad response to high uncertainty, so if that's all the belief there is, we must be able to reject belief in favor of something else. But partial belief is a much more reasonable response to high uncertainty, so adopting the best available belief doesn't seem like an unfitting response to high uncertainty in the formal picture, which means there's less reason to think we must be able to adopt some doxastic attitude like suspension of judgment that's distinct from belief. While thinking of belief as full makes us more inclined towards doxastic voluntarism, imagining it as partial may make an epistemologist fail to understand this mysterious ability to consciously reject belief that full belief suspenders of judgment appear to claim to have. The decision whether to describe belief as full or partial may affect the view we arrive at about our psychological capabilities.[49]

[49] I think something like this may be going on in Selim Berker's paradox example (Berker, 370). Berker asks readers to imagine that they believed three propositions, found out they were incompatible and that one was false, and gave

Here's a more general version of this point about doxastic voluntarism. We have a long-standing notion of voluntary suspension of judgment that dates back to the Stoics. It's possible that regarding belief and suspension of judgment as mutually exclusive doxastic states has caused some people to infer that since we can voluntarily suspend judgment, we must also be able to voluntarily choose our beliefs. Thinking we can choose whether to believe makes us more open to the idea that we can choose what to believe. The ability to choose whether to believe doesn't necessarily entail the ability to choose what to believe, but it's only a small step from the former to the latter.

I suggest that we should start with the observations that we can choose to suspend judgment, but we can't choose our beliefs (including whether to have them or not), and therefore reject the idea that suspending judgment about a proposition is mutually exclusive with having a belief about it—if we can choose to suspend judgment even when we're stuck with a belief, then the two can't involve mutually exclusive doxastic states. I prefer this approach to starting with the assumption the two states are mutually exclusive and then inferring that because we can choose to suspend judgment we can choose whether to believe or not.[50] I opt for modus tollens because I'm more confident that we can suspend judgment and can't choose our beliefs than I am

---

up on figuring out which one was false after a few minutes and went back to believing all three. Berker says that it would be more rational to suspend judgment on all three propositions until you can identify one that's less plausible. He seems to be using notions of voluntary suspension of judgment and voluntary belief here.

The natural way to assess the situation using credences is that the evidence before you supports credence 2/3 in each proposition. This way of speaking doesn't suggest you have any choice about whether to have credence 2/3 after figuring out the evidence supports it. But Berker is using talk of full belief, and the evidence doesn't support full belief in any of the propositions, which I think makes him open to the idea that you can choose not to believe by suspending judgment. Once you think we can choose whether or not to believe, you're not far from thinking we can choose what to believe, and Berker talks as if we can choose to believe all three propositions while knowing they're inconsistent. The decision to use the language of full belief may have led him to doxastic voluntarism.

[50] I also prefer rejecting the mutual exclusivity of suspended judgment and belief to concluding that because the states are mutually exclusive and we can't choose whether to believe, we must not actually be able to choose to suspend. I'm surer that we can choose to suspend and we can't choose whether to believe than I am that belief and suspension are mutually exclusive.

that suspending judgment about something and having a belief about it are mutually exclusive.[51] My theory will attempt to explain how we could do both simultaneously.

Setting that aside for the moment, I have a few concerns about reducing suspensions of judgment to precise credences. The first is about a failure of necessity. The second is about a failure of sufficiency. The third is about how well the reduction can explain what people are doing when they claim to be intentionally suspending judgment.

As I said before, the instruction that jurors suspend judgment looks like a paradigmatic case of the concept in action. It doesn't necessarily involve having a middling precise credence; in fact, having a middling precise credence seems to be at odds with the instruction. It would be strange for lawyers and judges to ask prospective jurors to adopt a precise credence around .5 about whether a defendant is guilty before they hear the evidence, and not just because lawyers and judges don't generally speak in terms of credences. Lawyers and judges aren't asking prospective jurors to do anything that amounts to having a precise credence of .5; the strangeness comes in part from the *precision*. Having a middling precise credence instead of a wide-ranging imprecise one suggests that one has some evidence that justifies narrowing their degree of confidence to .5. The request that jurors suspend judgment about their existing belief regarding whether a defendant is guilty is not a request that they suppose confidence in a precise credence of .5 is justified. The request seems to amount to asking them to be blank slates in some way, but a precise credence is a slate with a statement of precise uncertainty written on it.

[51] Although it's not the predominant view, I'm not alone in thinking suspended judgment and belief can coexist. Atkins (2015, 3041), Salmon (1995, 5), and Friedman (2017, 305) also think that one can suspend judgment about P's truth while having a belief about it. While they think this is possible but irrational, just like it's possible but irrational to have contradictory beliefs, I think suspending about something and having a belief about it is what happens during paradigmatic cases of suspension. When we suspend judgment about P, we decline to use our existing belief about P.

A similar problem arises when applying the reduction to precise credences to agnostics, the other paradigmatic case of suspension of judgment. Agnostics don't seem to be claiming that a precise degree of middling uncertainty about God's existence is justified—precise uncertainty requires some evidence, like the knowledge that a coin is fair. An agnostic who believed in the principle of indifference might take the absolute absence of evidence to support a precise credence of .5, but I have no reason to think that most agnostics both accept the principle of indifference and believe they have no evidence regarding God's existence.

Appealing to the principle of indifference and saying that in the absence of evidence we should suspend judgment by adopting credence .5 does seem like the best way to make the reduction to middling precise credences sound plausible, but it doesn't work at all in explaining what juries are doing when they suspend judgment, so it can't establish the necessity claim. The presumption of a defendant's innocence precludes being indifferent between their guilt and innocence. Starting out with a precise credence around .5 assigns more weight to belief in the defendant's guilt than jurors are supposed to, so lawyers and judges can't be asking jurors to replace their actual belief with credence .5 when they ask them to suspend judgment about whether the defendant is guilty. A similar argument applies to the reduction to imprecise credences. If suspensions of judgment are symmetric wide-ranging imprecise credences centered around .5, instructing jurors to suspend judgment about a defendant's guilt is at odds with the instruction that they should presume the defendant is innocent, because the symmetry of the imprecise credence assigns similar weight to guilt and innocence. After I give my theory of what suspending judgment is I'll explain how it allows the instruction to suspend judgment to be compatible with the instruction to presume innocence.

The claim that a precise credence of or around .5 is *sufficient* for a suspension of judgment seems clearly wrong. If I learn that the objective chance of something is 50%, I adopt credence .5 and feel very comfortable, epistemically speaking; it doesn't feel anything like a suspension of judgment. I feel as committed as I do when I adopt a credence matching any objective chance. This is even more apparent if the reduction to precise credences is the less naïve one that creates a vague, contextually determined symmetric band around .5, location within which is necessary and sufficient for a precise credence to be a suspension of judgment. Learning that a coin is slightly biased towards heads and adopting credence .51 that it will come up heads doesn't feel like a suspension of judgment and doesn't match up with any usage of the term I'm aware of.

Someone could say that thinking there's a 50% chance of something for whatever reason, or having credence .5 about something for whatever reason, always counts as a suspension of judgment. Maybe that's how it feels for them. There's no reason to expect the clusters of concepts and feelings we attach to the phrase "suspension of judgment" to be identical anyway. But this sufficiency claim does seem to have trouble explaining what would be going on if someone started out with credence .5 in a proposition and then claimed to suspend judgment about it or did the reverse. If the reduction is correct, doing either of these doesn't make sense, since there's no change to be made. I think anyone who likes credences and thinks we can intentionally suspend judgment probably will think there's at least some case where we can either move from credence .5 to suspending judgment or move from suspending judgment to having credence .5.

4. Reducing Suspension of Judgment to Imprecise Credences

I agree with Jane Friedman that suspending judgment can't just be reduced to having a precise credence in some range within which everything is a suspension of judgment and outside of which everything is a belief.[52] I also agree with her that suspension of judgment isn't just the absence of belief. Some people act like they can suspend judgment, perhaps at will, and I'd like to know how people who care about accuracy should describe what those people think they're doing and what they hope to gain from it. I also want to understand the distinction between suspended belief and partial belief, and how one becomes the other.

I don't think intentional suspension of judgment can be reduced to adoption of precise credences but I'm less sure about whether it can be reduced to adoption of imprecise ones. At first glance, I see the appeal of saying that intentionally suspending judgment is just a matter of intentionally adopting some wide-ranging imprecise credence—a narrow imprecise credence shouldn't be called suspended judgment. Perhaps another feature of suspensions of judgment is that these wide-ranging imprecise credences are roughly centered on .5.[53]

One virtue of this reduction is that it has a natural corollary that explains non-intentional suspension of judgment: it's just non-intentional adoption of a wide-ranging imprecise credence centered on .5. A second virtue is that the reduction to imprecise credences matches up well with the need to acknowledge degrees of suspension of judgment. It's clear that not all suspensions of judgment are identical; perhaps all are wide-ranging imprecise credences centered on .5 but there is some variation in the ranges and in how evenly centered around .5 they are. A third virtue is that this reduction explains how suspensions of judgment differ from partial beliefs, and a fourth, how the former turn into the latter. The idea would be that there is a vague boundary between

[52] Friedman (2013)
[53] Sturgeon (2010) has an account similar to this one.

suspensions of judgment and partial beliefs, and that committing judgment is a matter of gradually or suddenly narrowing a wide-ranging imprecise credence into either a precise one or an imprecise one too precise to deserve to be called suspended judgment. On this reductive account another way to stop suspending judgment is to take an existing wide-ranging imprecise credence centered on .5 and make its distribution of certainty sufficiently imbalanced around .5 to no longer seem deserving of the label suspended judgment.

Reducing suspension of judgment to adoption of imprecise credences has its appeal. There are many important features of suspension of judgment that it correctly captures, but there may be at least one crucial one that it gives up. Roughly speaking, my suspicion is that there's an important difference between having an imprecise belief and not having a belief, and people who claim to be intentionally suspending judgment seem to be claiming to do the latter in some way. People who suspend judgment at will seem to be asserting in some significant way that they are not playing the belief game. The reduction to imprecise credences attempts to explain how they are in fact still playing a version of the belief game. I worry that this reduction doesn't take their assertion that they're stepping back from the normal project of belief seriously enough and is therefore blind to the question of how they think they're stepping back from it. I suppose the reduction's explanation would be that they're announcing that they're stepping back from the project of having precise belief, and some people might be satisfied with this response, but like I said, I worry that it doesn't take their rejection of the belief game seriously enough. People who intentionally suspend judgment aren't merely taking the backward-looking measure of having the doxastic attitudes their evidence calls for, something which tends to result automatically from figuring out what the evidence calls for anyway—they're taking some forward-looking, optional, voluntary action intended to produce some desirable consequence.

With that said, there's room for disagreement here. There's subjectivity to what counts as suspension of judgment to different people, and there's also subjectivity to what counts as an imprecise credence to different people. Friedman acknowledges the former point after presenting her own account of suspension of judgment as inquiry: "…one might propose that suspended judgment has many faces or can be done in many ways… perhaps one can think of the project here as one of describing one of those many faces—something like 'active suspension of judgment.'"[54]

My project is somewhat more ambitious than Friedman's, because I hope to present an account that captures something that all paradigmatic cases of suspension of judgment have in common, from what juries do to what agnostics do to what Descartes did. But I think it's entirely possible that for some reasonable people, their notion of suspension of judgment maps on well with their notion of a kind of highly imprecise credence—for instance, saying that suspending judgment amounts to adopting an imprecise credence over an interval, which amounts to not ruling out credences in that interval.

I can imagine one fitting the actions of jurors and agnostics into an imprecise credence account like that. It's harder to see how Cartesian suspension could be reduced to the adoption of a highly imprecise credence; as I'll elaborate at the end of the paper, he framed his suspension more as if it were some kind of active rejection of his existing beliefs. Even high imprecision won't capture the kind of rejection Descartes seemed to be attempting, where he supposed that almost all his beliefs were false and tried to amplify his reasons to doubt them. Descartes seemed to be *trying* to rule out all his credences in the interval between zero and certainty. So one whose notion of suspension maps onto a kind of highly imprecise credence might have to end up saying

[54] Friedman 2017, 322.

that Descartes wasn't actually suspending judgment—that is where the many faces of suspension of judgment would present difficulty to their account. In Friedman's terms, I will try to present an account that captures something about *all* the faces of suspension of judgment, but an imprecise credence account may describe some of them. I don't think the reduction to imprecise credences is necessarily wrong; I do think my own account may be a more satisfying unifying explanation of paradigmatic cases of suspension of judgment.

5. A Non-reductive Account

Here are the beginnings of a theory about suspension of judgment and credences and what people who claim to be voluntarily suspending judgment are doing. Suspending judgment is not a replacement for having a belief, full or partial, precise or imprecise. We don't have anything like that degree of control over whether to avoid having whatever belief we think our evidence recommends. But we are doing something when we intentionally suspend judgment, and it is something we ought to have a fine degree of control over. We aren't replacing our beliefs with a different doxastic attitude; we're deemphasizing the importance of whatever beliefs we're suspending judgment about. My theory is that when we intentionally suspend judgment about a proposition is true, we're deciding not to use our belief about its truth. The decision whether to suspend or not wouldn't be about whether to have the belief the evidence suggests or not; that's outside of our direct control. But we can exhibit finer control over what we do with the belief we have, by declining to report it to others, declining to act on it, declining to update our other beliefs with it, declining to use our other beliefs to make it more accurate, or declining to gather further outside evidence to make it more accurate.

So far, this theory is entirely about the act of suspending judgment; I've said nothing about the state of suspended judgment, and after I describe my theory of the act in more detail my hope is that it will shed some light about how we should think of the state that the act places one in. A second gap is that I've been vague about the question of how my theory fits with the idea that suspensions of judgment can come in degrees, and the idea that not all suspensions of judgment are identical to each other. As a starting point, if suspending judgment about a proposition amounts to separating the accuracy of our belief about it from the accuracy of our other beliefs somehow, then perhaps there can be different ways and degrees of separating that accuracy. Suspension of judgment is a notion that's been around a long time, and it likely means somewhat different things to different people in different contexts—I don't want my theory of the concept to find more unity in it than there is. But by examining the mystery of how intentional suspension of judgment can be reconciled with the fact that we can't choose our beliefs I can find a kind of loose unity. We can't choose not to believe, but we can choose how we treat our beliefs.

Before I take up the question of degrees and kinds of suspension of judgment, I should examine what it would look like if we suspended judgment about something as fully as possible and tried not to use our belief about it at all. This is best illustrated through an example.

I've always hated and feared spiders. They have too many legs and too many eyes. When I was very young, I read a story about a woman who had long hair that she kept in a big bun and rarely washed. A spider created a nest there, which she discovered when the eggs hatched. Learning this story made my life worse. For a while afterwards, I would wonder "Is there a spider in my hair? Is it making a nest?" Every once in a while, I still think "Is there a spider in my hair?" Every time, I conclude that there almost certainly is not a spider in my hair. When I

wonder about it against my will, I remind myself of all the evidence I have that there isn't a spider in my hair and I gather new evidence by running my hands through my hair, becoming ever more accurate. As hard as I try, I can't completely forget something I believe that I didn't before I read that story: there might be a spider in my hair. Before hearing that story I'd never even contemplated the possibility.

Now that I have, what I really want isn't to become more accurate about it; I want to do my best to return to the happy naïve state I was in before I became aware of it. I can't really give up my beliefs about whether there's a spider in my hair and what it might be doing there, since we can't choose what we believe, but if I do my best to avoid thinking about those beliefs that's the next best thing. When I learn something new about spiders or hair, I'll do my best not to update the belief I don't want about how there might be a spider in my hair, and I won't use my beliefs about that to make any of my other beliefs better. When I assess how well I've done at believing true things about the world, I won't include my accuracy about those propositions in my assessment. I'm suspending judgment about it, doing my best to act as if I don't have any belief even though I do. I can't choose not to have the belief, but I can choose not to use it.

This theory gets at the kernel of truth in the first reduction I considered, that the state of suspended judgment is just the state we're in when we lack belief about a proposition, and that the act of suspending judgment amounts to moving from a state of belief about something to a state of nonbelief about it. That's exactly what I'd like to do with my beliefs about spiders in my hair, but I can't. I think that the epistemic state I'm in during the moment right after I suspend judgment about spiders in my hair contains exactly the same doxastic attitudes I had before I suspended judgment. What I've changed is not my doxastic attitudes—I've changed my doxastic *commitments*, and adding a commitment to avoid thinking about the belief I've suspended

judgment about, even if thinking about it would make it and my other beliefs more accurate, means that over time I'll end up in a different epistemic state than I would've without suspending judgment.

When we ask someone what they think about something and they respond that they're suspending judgment about it, we sometimes interpret them as saying "I have no belief about that," but I think it frequently means something more like "I don't want to talk about that" or "I don't want to commit myself on that." I encountered a version of this right before the 2022 midterm election when debating my conservative grandparents about whether abortion is ever morally permissible. Eventually, I asked them if abortion was at least permissible if necessary to save the life of a pregnant woman, expecting them to give a little ground there, and instead they both shuddered and said things like "I don't even want to think about that," and, when I pressed them, "Someone else will have to answer those questions." I suspect that in their minds, they did hold the belief that abortion is probably permissible to save the life of a pregnant woman, but they were doing their best to act as if they had no opinion about the scenario. It was their version of my spider-in-my-hair case. I think they were suspending judgment about the question. They would not, for instance, act on their belief by choosing to support pro-life politicians who advocate creating exceptions allowing abortion when necessary to save the life of a pregnant woman. Whatever deep-seated, unwanted belief they had about the question would not make it to their statements or actions.

I don't mean to say that all or even most cases of intentional suspension of judgment are like the spider one or the abortion one. I'll examine some other cases, such as ones where we intend our suspension of judgment to be temporary, as juries do when hearing evidence, and ones where we don't, as agnostics do when suspending judgment about whether God exists. What I

think most or all cases of suspension of judgment have in common is that we're adjusting the way we treat our beliefs in some way to approximate the state we'd be in if we didn't have them.

One possible challenge to my theory is that it doesn't allow for people to suspend judgment about the truth of a proposition they're aware of if they don't have at least some kind of belief about its truth. I'm not sure it's possible to comprehend at least part of a proposition's meaning without having at least some highly or maximally imprecise credence about its truth. I try to imagine propositions whose truth I might have no belief at all about, like "The second king of France liked eggs" and "String theory is correct" and I find that my would-be opinionless states about those propositions don't seem to be identical, because I have beliefs about which propositions they depend on, and beliefs about the truth of those, and those relevant beliefs aren't identical. If I had no beliefs about their truths my doxastic states would be identical. My beliefs about the truths of those propositions are highly indeterminate, but not identically indeterminate.

If you understand anything about the relevant concepts in two separate propositions and you believe the concepts are different from each other, it's hard to truly have no belief at all about the truth of the propositions. Take the proposition "X's can Y," where you don't know what X's or Y's are but you know what 'can' means. You probably think it's more likely to be true than "X's can Y and Z"; if so, you can't be opinionless about either proposition. On the other hand, if you don't comprehend any part of the meaning of a proposition, it's hard to say you're aware of it. I'm comfortable saying that people must have some belief about a proposition they're aware of to suspend judgment about its truth (because suspending judgment is deciding not to use one's belief), since I can't imagine any case where one is aware of a proposition without having at least some kind of imprecise credence about its truth.

One interesting feature of my theory is that it implies that it's possible for one to suspend judgment on a proposition no matter how high their credence in it is, meaning that people could suspend judgment even about propositions that they were certain of. I'm fine with this result, and I think it sheds some light on the relation between the descriptive question of when we can suspend judgment and the normative question of when we should. One way we could decide to treat some of our beliefs differently than the rest is by deciding we won't use them to inform our other beliefs. Here the decision to suspend judgment when one has a high credence could be irrational just because it's hard to ignore the implications of beliefs one is confident of, and it gets more difficult as that confidence rises. If it's impossible for a juror to effectively set aside their prior certainty that a defendant is guilty as they evaluate arguments at a trial, then it's irrational for them to decide to do so.[55]

Even if one could effectively suspend judgment on a proposition they had credence 1 in, there would usually be little or no reason to; it would be like turning down free money, or in this case, free expected accuracy. If suspending judgment about a proposition's truth sometimes amounts to deciding not to include it with one's other beliefs when evaluating one's overall accuracy, or to give it less weight in that evaluation, then my theory predicts that the decision to suspend judgment will be less rational the more extreme one's credence about the proposition is, all else being equal. The correct assessment of someone who claims to suspend judgment in the face of persuasive evidence is that they're being irrational, not that they're dishonest or mistaken about what they're doing. If someone successfully decides not to use a high credence as

[55] I'm not completely sure whether to say that the biased juror suspends judgment ineffectively or that they're attempting to suspend judgment but aren't actually able to. Those who want to say the biased juror isn't really suspending judgment at all face a difficult question: if we gradually adjusted the bias downwards, at what point would the juror go from falsely claiming to suspend judgment to actually doing so? My way of speaking avoids needing an answer to that question, since I can just say that the less biased they are the more effective their suspension is. In the end, though, this may be a matter of taste.

evidence, they're not just failing to get a higher accuracy score, they're turning their back on the thing that accuracy is good for and making all the beliefs and decisions that depend on that credence worse. The more confident one is of the belief they decide not to use, the more the decision costs them and the more irrational it is to make it.

I am turning down free expected accuracy when I decline to use my credence that there's a spider in my hair to inform all my other beliefs, and it is, strictly speaking, epistemically suboptimal, but the value of using the belief is so slight that I'm usually willing to let my desire not to think about spiders override it. I try to live the life I would've lived if I'd never realized there could be spiders in my hair. For the most part it's the exact same life, but with slightly less accuracy and significantly less discomfort. My grandparents' decision to suspend judgment about whether abortion is permissible to save life strikes me as less rational, since more depends on that question. By suspending judgment we can become more ignorant, and although ignorance is an epistemic vice, everyone knows it can be bliss.

This isn't the behavior of an ideal epistemic agent, of course. We'd be playing a strange game if we hoped that everyday concepts used by ordinary beings with multiple interests could be explained primarily in terms of their value to idealized agents with one interest.[56] If we pursue that path too single-mindedly, we'll have no understanding of all the phenomena that arise from the efforts of actual beings to balance their competing interests. There probably are epistemic reasons to suspend judgment, and I'll examine them towards the end of this section and in the next, but I think we often suspend judgment because we choose ignorance for pragmatic reasons.

[56] Kennedy (2013), for instance, argues that there are ethical reasons why physicians should "exercise a compassionate suspension of judgment when a diagnosis cannot be immediately made," instead of thinking patients' concerns are unjustified. Kennedy also argues that there are evidential reasons to do so, but presumably the ethical reasons would remain even in cases where they were at odds with epistemic concerns. There can be ethical, epistemic, and pragmatic reasons to take the epistemic action of suspending judgment. My objective is just to explain what the action is, and the nature of the act should remain the same whether one's reasons for doing it are ethical, epistemic, or pragmatic. A good account of the act should explain how it can serve a variety of purposes.

We consider the epistemic cost and are sometimes willing to pay it when it's low. When we strive for ignorance, we suspend judgment about whether something's true not because doing so is epistemically rational, but because it's not too irrational. When explaining the link to rationality, instead of looking only for cases in which suspending judgment could be epistemically good we should also look for cases where it's not that bad.

The connection between the rationality of suspending judgment and its cost to expected accuracy sheds light on the appeal of the reduction to imprecise credences. Although suspending judgment about a proposition's truth isn't *reducible* to adopting these imprecise credences toward it, suspending judgment is generally more *rational* when one's actual credence is highly imprecise because the cost to expected accuracy is usually lower, since the credences are often less informative. Failing to use a credence costs you less the more imprecise the credence is; this is true at least when the imprecision of the credence results from a lack of evidence. It might not be true if one learned that the objective chance of something was highly imprecise, so maybe suspending judgment in the latter kind of case would be irrational. The main point is that the less informative our credences are the less irrational it is to refrain from using them, and highly imprecise credences are often uninformative, especially when they're wide-ranging and symmetrical around .5. That tendency explains why many cases of suspension of judgment involve those imprecise credences. When we don't know much to begin with, the decision not to apply our knowledge doesn't cost much and is therefore less irrational.

This connection to rationality also explains the relationship between the principle of indifference and suspension of judgment. The principle of indifference tries to select the least informative precise credence in the absence of evidence, which is why it settles on the exact middle, .5. The less informative a credence is, the less the decision not to use it gives up. The

cases where using the principle of indifference seems most appropriate are also the cases where suspending judgment costs the least.[57]

In most situations, suspending judgment about a proposition one has information about is irrational because it amounts to giving up free expected accuracy. This gets at the heart of the criticism that some atheists level against agnostics: the charge appears to be that the agnostic actually thinks it's more likely than not that God doesn't exist, but by suspending judgment about the issue rather than conducting epistemic business-as-usual by having and applying a credence weighted against God's existence the agnostic is giving up free expected accuracy and the epistemic and pragmatic benefits that flow from it. In my experience, people who claim to be agnostic aren't claiming to be perfectly neutral about whether or not God exists—they seem to be deciding to treat their credence weighted against God's existence differently than an atheist would treat the same credence, and differently than the agnostic treats similar credences about other questions.

People who have a credence of roughly .6-.8 that it won't rain later in the day don't claim to be agnostic about whether it will rain; they just report their credence when asked and figure out what implications it has for all their other credences and their decisions. An agnostic might have a credence of roughly .6-.8 that God doesn't exist but be reluctant to report it to others, only

[57] Tang suggests that the natural extension of Joyce's argument for using the Brier score to assess credences (Joyce 1998) would be to treat suspensions like credence .5 and therefore assign "a suspended belief (in either a truth or a falsehood) a score of .25." (Tang, 367) One can also find the idea that suspensions deserve some inaccuracy score in Pettigrew (2016, 255) and Wedgwood (2013, 232).
I think suspensions of judgment and beliefs are such different things that assigning suspensions inaccuracy scores doesn't make sense. Beliefs are the kind of thing that can be accurate; suspensions aren't. Suspending judgment does have an effect on one's accuracy, just as gathering evidence does, but we wouldn't assign an inaccuracy score to the state one is in when they're gathering evidence about something, and we shouldn't assign one to the state one is in when they're suspending about it either. These are states that *affect* accuracy, not ones that *have* it. They affect accuracy because being in them affects one's set of beliefs, which is a proper subject for an inaccuracy score. Instead of assigning a uniform inaccuracy score to all suspensions, my theory explains how different suspensions have different effects on accuracy. The effect declining to use a particular belief has on the accuracy of one's set of beliefs varies depending on what the belief is.

quantifying it and admitting it when pressed, and be reluctant to use the credence to inform their other credences and their actions. An atheist is likely to say things like "God probably doesn't exist, so…" while an agnostic might believe that God probably doesn't exist but is less likely to report their belief, reflect on it, draw inferences from it, or pursue additional evidence to make it more accurate. Agnostics are refusing to play the belief game when it comes to the question of whether God exists, and atheists who object to their agnosticism want them to play the game because they think they'd do the same things as atheists if they played it.

My theory predicts that the decision to suspend judgment about a proposition's truth is more rational not only the less informative one's credence about it is, but the less other propositions' truths depend on its truth, which fits well with the behavior of agnostics—they seem to think not much depends on being right about whether God exists, and that the accuracy of the beliefs they care about and the efficacy of the actions that depend on them won't be significantly affected by suspending judgment about the question. On this theory, someone who had credence .5 in a god who would punish lack of belief severely would be less likely to be agnostic than someone who had credence .5 that some impersonal higher power existed.

Giving up free expected accuracy is irrational in most cases, but there are good reasons for us to ask jurors to do it. From their perspective, we *are* asking them to give up expected accuracy when we ask them to set aside their existing belief about a defendant's guilt before hearing the evidence at the trial. We do this because the law isn't interested in knowing what a juror's actual belief is after hearing all the evidence at a trial; it's interested in knowing what their belief would be if they started with a prior that lay at neither extreme but weighted innocence more heavily. Usually, the problem of bias stems from a prior belief that the defendant is guilty, which is why we typically ask jurors to suspend judgment about the defendant's guilt,

but a high prior credence that the defendant is innocent also prevents their belief at the end of the trial from being determined largely by the evidence presented at the trial, so lawyers can exclude those prospective jurors too.

From a juror's perspective, the mere fact that a defendant has been charged with a crime is highly relevant evidence, but from the judicial system's perspective, it can't be, which is why we ask jurors to suspend judgment about the defendant's guilt in the context of the courtroom but no one would expect them to do so if they encountered the defendant on the street. One way in which suspensions of judgment come in different kinds and degrees is that we can refrain from using our actual credences in some contexts but not others. Jurors suspend judgment in only one context, while agnostics do it in many, generally only using their credence about God's existence when the context of a conversation is manipulated to force them to do so.

One of the questions I began with was the question of why we ever would suspend judgment instead of having and using the best belief possible. I suspend judgment about questions like whether there are spiders in my hair and how much money it would take to make me betray various people in various ways because I'm not interested in knowing those things and not much depends on them. Agnostics suspend judgment about whether God exists for similar reasons, I think. But jurors suspend judgment for a different reason: because we ask them to. This reason generalizes to explain how suspending judgment can serve the pursuit of accuracy—when we ask people with different beliefs than us to suspend judgment about those beliefs, we make their resulting set of beliefs less accurate from their perspective, but more accurate from our own, and they become better sources of evidence for us. The voluntary nature of suspension of judgment entails both that people can choose to do it for their reasons and that we can ask them to do it for our own.

6. Friedman's Theory of Suspension of Judgment

Friedman and I agree that the state of suspended judgment is more than just the absence of belief, and that it's not reducible to the state of having a precise credence. I've added that it's probably not reducible to the state of having an imprecise credence; none of these reductions can explain the act of suspending judgment well. Friedman's positive theory is that one suspends judgment about whether something's true if and only if they're inquiring about its truth. She describes the biconditional in more detail with this:

> The claim that one is inquiring into Q only if one is suspended about Q is the claim that any case in which one has Q on one's research agenda, any case in which one is in the relevant sort of inquiring state of mind (and so any case in which one is genuinely inquiring), is a case in which one is suspended about Q. And the claim that one is suspended about Q only if one is inquiring into Q is the claim that any case in which one is suspended about Q is a case in which one is in an inquiring state of mind with respect to Q (or rather a case in which Q is on one's research agenda, or in which one has an attitude towards Q with the relevant sorts of epistemic satisfaction conditions, or in which one aims to close Q).[58]

I want to note that even if Friedman's biconditional is true it doesn't tell us *what it is* to suspend judgment; it only tells us that suspension of judgment and inquiry always coincide with each other. I don't think Friedman's aim was to say what suspension of judgment is. If she's right that it always happens when inquiry does and vice versa, then that would provide a powerful hint about its nature which would help us figure out what it is—as she puts it, the biconditional "gives us the start of a story about the nature and function of suspended judgment via its connection to inquiry."[59] I agree with her that suspension of judgment often has a close connection to inquiry, and I think I haven't yet shown that my own theory of what suspending judgment is explains how doing so can aid one's own inquiry, but this biconditional seems far too strong to me.

[58] Friedman (2017, 308)
[59] Friedman (2017, 322)

There may be problems with the first claim about necessity, but to me the most clearly objectionable part is the sufficiency claim that any case in which one is suspended about something is a case in which they're in an inquiring state of mind with respect to it.[60] This is straightforwardly at odds with some cases of suspension of judgment that I regard as paradigmatic. Before doing any theorizing, I never would have thought that when one person says "I'm wondering about X" and another responds "I'm suspending judgment about that" the latter is agreeing with the former. It just sounds like a disagreement, or at least an expression of different approaches. Any theory that leads to the conclusion that the two are expressing similar or identical thoughts can't be right. One has thought too deeply if they start out thinking the two are disagreeing, theorize about what suspension of judgment is, and then decide on the basis of the theory that the two must actually be agreeing, or that the second person must be confused about what suspending judgment is.

Let me make this appeal to linguistic intuition more plausible by filling in 'X.' Suppose Person A says "I wonder whether God exists" and Person B responds "I'm suspending judgment about that," or "I'm agnostic." To my ear, B is not saying that they also wonder whether God exists—in fact, they seem to be expressing the thought that they *aren't* wondering or inquiring about whether he does, that it *isn't* on their research agenda. People who say they're agnostic

[60] Friedman presents a few considerations that she thinks aren't decisive, but weigh in favor of the claim that all suspensions involve inquiry. The first is that suspension often becomes inappropriate exactly when inquiry does. The example Friedman focuses on is presupposition failure: "If you know that Jefferson didn't have a Ferrari then suspending judgment about what colour his Ferrari was looks inappropriate." (2017, 316). She then points out that having any inquiring attitude about the color of his Ferrari also looks inappropriate, and concludes by saying "And if we deny that suspension involves this sort of openness, then we'll have to give some other explanation of why it starts to look bad exactly when having a question on the agenda starts to look bad."
But there's already a perfectly good explanation of why suspension looks bad to anyone who shares Friedman's view that it's an attitude. Having *any* kind of attitude with respect to a proposition's truth is inappropriate once you realize the proposition suffers from presupposition failure, because you then believe that the proposition isn't the kind of thing that can be true or false. Once having any attitude about the proposition's truth is unwarranted, it necessarily follows that having an inquiring one is too. The attitude of indifference towards the truth of a proposition one knows contains a failed presupposition is just as unfitting as the attitude of curiosity, and for the same reason. It makes no sense to know Jefferson has no Ferrari and be indifferent about what color his Ferrari is.

about God's existence certainly don't behave the same way that people who say they're inquiring about his existence do—they avoid theological discussions, refrain from gathering evidence, and just generally seem less interested in the issue. Not only does agnosticism fail to be a case where people who suspend judgment inquire, it looks like the exact opposite, where they mean they aren't inquiring.

But agnosticism is one of the few paradigmatic cases where people suspend judgment about something. It makes sense to say that juries suspend judgment to aid inquiry, but agnostics seem to suspend judgment to avoid it; we need a theory of what suspending judgment is that's general enough to explain how it could sometimes be used to enable inquiry and sometimes be used to end it. Anyone who wants to explain the nature of suspension of judgment should grapple with the riddle of how the same kind of act performs such different functions in its paradigmatic cases.

Friedman doesn't say anything about how her theory explains what religious agnostics are doing. But she does make one reference to agnosticism, and it comes at the very beginning: in the first footnote, where she's defining her terms, she says "I take it that [the state of suspended judgment] is also the state people are talking about when they talk about 'withholding belief' although I won't talk about it in their terms. I think that it is also fine to use the term 'agnostic' to describe the suspended subject, but given that it tends to bring to mind discussions of theistic opinions, I'm going to avoid that terminology as well."[61] I agree that it's fine to use the term 'agnostic' to describe people who suspend judgment about something, but I think it's fine precisely because agnosticism about whether God exists is a paradigmatic case where people suspend judgment. Friedman's theory can't explain what religious agnostics or any other kind

[61] Friedman (2017, 322)

appear to be doing. She explains only one facet of suspension of judgment when she says we do so to inquire.[62]

My theory so far has the opposite problem—my theory that suspending judgment amounts to deciding not to use one's belief in the normal way explains how agnostics suspend judgment to avoid inquiry, but I haven't yet fully explained why someone might decline to use their belief about a proposition to pursue the truth about it. I've explained how juries suspend judgment to pursue inquiry, but that's a case of failing to use their actual belief in order to pursue someone *else's* inquiry. I haven't yet said how declining to use their belief about a proposition's truth might lead someone to become more accurate about it from their *own* perspective. I take it this is what Descartes was trying to do. I'll end by taking up this remaining task. If my theory of the act of suspending judgment is general enough, then Descartes' actions should fit in neatly.

What was Descartes claiming to do when he suspended judgment and how did he hope it would help him reach truth? This much is uncontroversial: Descartes wanted to start his inquiry into the world with foundational beliefs he could be completely certain of and derive other beliefs from them. Even beliefs he was very confident of weren't fit to be part of his foundation: "Reason now leads me to think that I should hold back my assent from opinions which are not completely certain and indubitable just as carefully as I do from those which are patently

[62] She acknowledges this in her conclusion: "…one might propose that suspended judgment has many faces or can be done in many ways… perhaps one can think of the project here as one of describing one of those many faces—something like 'active suspension of judgment.' Active suspension involves having a sort of neutral contentful attitude over some interval of time. And as I've now tried to suggest it is intimately tied to inquiring into questions." (322) I think that someone who decides to be agnostic about whether God exists is actively suspending judgment but isn't inquiring whether he exists.
I agree that suspension of judgment has many faces; my goal is to find the body they share. What inquiring suspenders and non-inquiring ones have in common is that they're treating the beliefs they've suspended about differently than the other ones—for instance, by refusing to report them or failing to conditionalize their other beliefs on the suspended ones. They just have different motivations, and the inquiring suspenders tend to plan to stop suspending at some point after they've gained the benefit of looking at things from a perspective different than the one their actual belief recommended.

false."[63] This notion of holding back assent matches well with talk of withholding belief, so this may be the part where the concept of suspending judgment enters Descartes' project. It's important to note that Descartes is withholding *assent* to beliefs he *has*, not withholding *belief itself* from propositions he's *aware of*. Today we speak of withholding belief from propositions, but my theory that the target of a suspension is one's existing belief about a proposition fits better with what Descartes was doing.

Descartes never clearly defines what it means to withhold assent to the beliefs he's confident of, but he does immediately go on to say how he'll withhold assent: "So, for the purpose of rejecting all my opinions, it will be enough if I find in each of them at least some reason for doubt."[64] I take it that rejecting his opinions is a way of managing to withhold his assent from them, and that his plan to reject them is to focus on his reasons to doubt them more than his reasons to believe them. Descartes then does this by showing that even the beliefs he's very confident of are based on his senses and focusing on his reasons to doubt his senses are reliable. He concludes, "So in the future I must withhold my assent from these former beliefs just as carefully as I would from obvious falsehoods, if I want to discover any certainty."[65]

But Descartes then adds the following, which sheds light on what it means to withhold assent: "I shall never get out of the habit of confidently assenting to these opinions, so long as I suppose them to be what in fact they are, namely highly probable opinions… In view of this, I think it will be a good plan to turn my will in completely the opposite direction and deceive myself, by pretending for a time that these former opinions are utterly false and imaginary."[66] It seems to me that when Descartes says he should withhold assent from the opinions he's

[63] Descartes, 12
[64] Id.
[65] Id. at 15
[66] Id.

confident of he means he should avoid considering them and deriving implications from them when pursuing certainty. And it seems that his strategy to avoid using those high credences in his quest for certainty is to temporarily do his best to pretend he doesn't have them. He adds, "I shall consider myself as not having hands or eyes, or flesh, or blood or sense, but as falsely believing that I have all these things… and, even if it is not in my power to know any truth, nevertheless it is in my power to suspend my judgment, that is, resolutely guard against assenting to any falsehoods…"[67]

These passages seem to support my view that, to Descartes, suspending judgment amounted to refusing to use the beliefs one actually has about the truth of a proposition. In particular, Descartes was focused on refusing to draw inferences from any of his actual beliefs that fell short of absolute certainty. The mystery that I've been unable to answer so far is how failing to use a belief one has could ever help bring one closer to the truth from one's own perspective. The key to solving this is that failing to use beliefs he was confident of brought Descartes closer to the truth in a unique way: updating a high credence that falls short of credence 1 with another belief like it can make the resulting belief more accurate, but *can never lead to certainty*. Making full use of our existing beliefs should always make us more accurate from our own perspective, as far as I can tell, but Descartes' immediate goal was not increased accuracy—it was complete certainty. In my terminology, Descartes was initially trying to achieve perfect average accuracy to the detriment of total accuracy, and his hope was that from a foundation with perfect average accuracy he could derive more truths without compromising

[67] Id. Interestingly, the phrase "nevertheless it is in my power to suspend my judgment" appears only in the French version of the First Meditation; the Latin version uses the phrase "I shall at least do what is in my power" instead. Maybe suspension of judgment was a nebulous concept to Descartes too.

average accuracy, leading to high total accuracy too. Descartes *was* turning down free expected total accuracy.

Although Descartes' meditation may seem to have presented a paradigmatic case where suspension of judgment enabled inquiry, his actions were actually a special case where the suspensions only aided inquiry because he was pursuing certainty rather than accuracy. For those of us who don't share his foundationalist project, suspending judgment on our existing beliefs just amounts to taking valuable tools out of our epistemic toolboxes. Doing that might make sense if you suspected that the tool was defective, because, for instance, you were worried you had misleading evidence, and if the cost of being misled was high. Descartes suspended judgment for this reason. But it only makes sense to avoid using an epistemic tool because it might be defective if you can replace it with one you're more confident is not, as Descartes thought he could.

Maybe Friedman's account, where suspending judgment is necessary and sufficient to be engaged in the epistemic project of inquiry, will be more satisfying to epistemologists. I just don't think that theory fits all the facts. On my theory, suspending judgment and deciding not to make full use of the beliefs we have is a rare act that we choose to perform when we don't want to conduct epistemic business-as-usual, either because we don't want to inquire, we're aiding someone else's inquiry, or we're conducting an unusually limited inquiry where we don't want to use all the information we have. Suspending judgment always has epistemic consequences, but we don't always do it for epistemic reasons.

BIBLIOGRAPHY

Atkins, Philip. "A Russellian account of suspended judgment." *Synthese* 194.8 (2017): 3021-3046.

Berker, Selim. "Epistemic teleology and the separateness of propositions." *Philosophical Review* 122.3 (2013): 337-393.

Descartes, René. "The Philosophical Writings of Descartes, vol. 2, translated by John Cottingham, Robert Stoothoff, and Dugald Murdoch." (1984).

Friedman, Jane. "Suspended judgment." *Philosophical studies*162.2 (2013): 165-181.

Friedman, Jane. "Rational agnosticism and degrees of belief." *Oxford studies in epistemology* 4 (2013): 57.

Friedman, Jane. "Why Suspend Judging?." *Noûs* 51.2 (2017): 302-326.

Joyce, James M. "A nonpragmatic vindication of probabilism." *Philosophy of science* 65.4 (1998): 575-603.

Kennedy, Ashley Graham. "Differential diagnosis and the suspension of judgment." *Journal of Medicine and Philosophy*38.5 (2013): 487-500.

Pettigrew, Richard. "Jamesian epistemology formalised: an explication of 'the will to believe'." *Episteme* 13.3 (2016): 253-268.

Perin, Casey. "Skepticism, suspension of judgment, and norms for belief." *International Journal for the Study of Skepticism* 5.2 (2015): 107-125.

Salmon, Nathan. "Being of two minds: Belief with doubt." *Noûs*29.1 (1995): 1-20.

Steup, Matthias. "Belief control and intentionality." *Synthese* 188.2 (2012): 145-163.

Tang, Weng Hong. "Reliabilism and the Suspension of Belief." *Australasian Journal of Philosophy* 94.2 (2016): 362-377.

Sturgeon, Scott. "Confidence and coarse-grained attitudes." *Oxford studies in epistemology* 3 (2010): 126-149.

Wedgwood, Ralph. "The aim of belief." *Noûs* 36.s16 (2002): 267-297.

CHAPTER III

The Problem of Conceptual Learning

1. Introduction

Plato claimed that all learning amounts to remembering. He theorized that we're recalling details about what the ideal forms of things are like, details which we knew as immaterial souls but forgot upon being embodied. This theory naturally raised the question of how our souls ever came to learn about the forms in the first place. Almost two-and-a-half millennia later, we've made a lot of progress in figuring out how learning works, although not as much as one might've hoped. One major theory, the Bayesian one, says that we start out with prior beliefs that generally aren't particularly accurate, a plan about how to update them in response to evidence we might find, and generally get closer to the truth as we encounter more and more evidence. While learning felt like remembering to Plato, to a Bayesian it feels like *correcting*. So the Bayesian explains what the Platonist never attempted to: how we come to learn the truth of a proposition for the first time. To a Bayesian, we learn when we improve our beliefs by conditioning them on evidence.

But although the plan to learn through conditionalization explains how we can learn which propositions are true, it remains silent on the question of how we learn which propositions are candidates for truth to begin with. It explains how we can gradually make our existing beliefs more trustworthy, but not how we could ever acquire a belief about a new possibility and have some initial confidence in it. When we first learn about Einstein's theory of relativity or

Darwin's theory of evolution, for instance, the learning process seems to revolve more around acquiring new thoughts than correcting existing ones, although accomplishing the former often causes the latter. Instead of explaining how we go from knowing nothing about these novel theories to knowing something about them, Bayesians seem to be stuck saying that we're doubting them less. On its own, this account of learning overlooks facts that call for explanation: when we learn a novel idea some set of propositions apparently gets added to the existing set of propositions we have beliefs about, and we apparently follow some method of assessing its initial degree of plausibility. It's hard to see how that way of determining the plausibility of newly recognized possibilities could be derived from our existing beliefs and conditional probabilities, which were only about the possibilities we'd already contemplated, so it seems we must be employing some kind of learning that lets us comprehend new propositions and assign initial degrees of confidence to them that we can then correct through conditioning.

Conditioning explains learning in a certain well-defined set of circumstances: an agent has a set of hypotheses about what the world could be like, a set of potential pieces of evidence they could receive, a set of initial probability assignments about propositions composed of combinations of possible hypotheses and evidence, and a set of conditional probabilities about how those prior confidences should change depending on what evidence they encounter. Updating through conditionalization allows one to adjust one's confidence in a proposition upwards or downwards, but the Bayesian theory of learning doesn't attempt to explain how an agent could go from not having any credence in a proposition to having one. A more complete theory of learning which explains how it can happen in a broader set of circumstances would outline a principled, truth-directed way one could come to understand a new proposition and have an initial degree of confidence in it. This paper describes the main challenges to

understanding how this conceptual learning interacts with Bayesian learning and suggests some answers. I call it conceptual learning mainly because it involves conceiving of new propositions. The task is not just to explain how conceptual learning happens, but how it can work hand-in-hand with conditioning, using some of the same machinery.

The problem is closely related to two well-known obstacles to the Bayesian project, the problem of new hypotheses and the problem of the priors. John Earman describes the former in the context of Thomas Kuhn's theory of scientific revolutions, pointing out that Bayesianism is poorly equipped to explain the introduction of new theories. He says that there are two kinds of revolution, the milder of which "occurs when the new theory articulates a possibility that lay within the boundaries of the space of theories to be taken seriously but that… had previously been unrecognized as an explicit possibility." The second, more radical form "occurs when the space of possibilities itself is significantly altered." Whichever form a new hypothesis takes, Bayesianism doesn't seem to explain how our beliefs adjust to encountering it: "Even the mild form of revolution induces a non-Bayesian shift in belief functions. By 'non-Bayesian' I mean that no form of conditionalization, whether strict of Jeffrey or some natural extension of these, will suffice to explain the change. Conditionalizing (in any recognizable sense of the term) on the information that just now a heretofore unarticulated theory T has been introduced is literally nonsensical, for such a conditionalization presupposes that prior to this time there was a well-defined probability for this information and thus for T…"[68]

The fact that we haven't conceived of a newly-encountered hypothesis before suggests that we can't have a pre-written plan for exactly how to update our beliefs upon hearing of it. It seems like whatever method we ordinarily use at the beginning of Bayesian learning to set priors

[68] Earman, John. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* MIT Press, 1996, p. 196.

and conditional probabilities must be used again partway through the process whenever a new hypothesis is introduced. As Earman puts it, "the problem of the transition from [the old credal function] to [the new one] can be thought of as no more and no less than the familiar Bayesian problem of assigning initial probabilities, only now with a new initial situation involving a new set of possibilities and a new information basis."[69] If the problem of the priors seems troubling, it might be even more troubling that a version of the problem keeps popping up every time an agent learns of a new theory—they have to figure out what priors to assign the claims in the theory and they have to figure out how to adjust their existing beliefs in light of their confidence in the new theory. One difficulty is that it seems like the two tasks depend on each other; we may be tempted to regard the new theory as implausible in light of parts of our pre-existing worldview and vice versa. Following Kuhn, Earman says that weighing the correct redistribution of probabilities in cases of scientific revolution like Einstein's theory of relativity is a matter for persuasion, not proof, meaning that this kind of learning will resist formalization: "The deployment of plausibility arguments is an art form for which there is no taxonomy."[70] There may be something to this, but one of the main goals of this paper is to see how far we can get in formalizing this process instead of giving up, as Earman does, at the very beginning. I argue that we can get considerably further than Kuhn and Earman believed.

It's natural to think that this is a rare problem raised only by revolutionary scientific discoveries, but it's an ordinary problem merely made vivid by those cases. Earman ends his discussion by pointing out that this non-Bayesian learning involving conceiving new hypotheses happens constantly, interfering with the Bayesian kind: "There is little to be salvaged from the Bayesian model of learning as conditionalization by claiming that, although the model fails in

[69] Earman at 197.
[70] Id.

period of scientific revolution, it nevertheless holds for periods of normal science. For normal science defined as the absence of even a weak revolution shrinks to near the vanishing point. New observations, even of familiar scenes; conversations with friends; idle speculations; dreams—all of these and more are constantly introducing heretofore unarticulated possibilities and with them resultant nonconditionalization shifts in our degrees of belief… All that remains of Bayesianism in its present form is the demand that new degrees of belief be distributed in conformity with the probability axioms."[71] Although Earman is right to observe there's no sharp distinction between scientific revolutions and normal science, he overstates the point a bit. Not all previously unarticulated possibilities are new hypotheses in the sense that requires learning through a procedure other than conditionalization—if we realize that we considered only two suspects for a crime and that there's a third, for instance, we can simply assign that suspect relevant conditional probabilities much as we did for the first two, and then condition on the same body of evidence we considered for the first two. We can go back and redo our work, essentially. The kind of new possibility that requires a kind of learning to supplement conditionalization is one that raises concepts different enough from our existing ones that we can't straightforwardly translate our existing conditional probabilities to apply to it. I'll give several examples of these troublesome new concepts throughout the paper.

The great hope Bayesianism offers is that given enough evidence an agent will get closer to the truth through conditionalization despite the inaccuracy of its priors. This process is what learning was supposed to consist of, with the selection of priors being a necessary evil to get the ball rolling. But if Earman is right that we're constantly imagining new possibilities, and that doing so makes us constantly adjust our degrees of confidence using some process other than

[71] Earman at 198.

conditionalization, Bayesianism isn't playing as much of a role in the learning process as one might've originally expected. It starts to look like following whatever process we keep using to select new priors and adjust existing credences isn't just a necessary evil that enables learning, and like applying that process well is part of what it is to learn.

I call this process of finding new ideas possible and plausible conceptual learning: the task is to explain how an agent can comprehend a new proposition for the first time, assign an initial probability to it, create conditional probabilities involving it, and appropriately adjust their existing credences in light of all those additions. For this process to really be learning, and not just the regular shuffling of credences, the method we use to re-evaluate our probability assignments in light of newly recognized propositions has to follow principles which seem likely to lead to truth. Since many of our credences are shaped by our constant conceptual acquisition it's important to understand what principles this non-Bayesian method of adjusting degrees of belief follows.

There are a few reasons I call my focus the problem of conceptual learning instead of the problem of new hypotheses or the problem of the priors. When it comes to explaining how we should change our minds in response to new hypotheses, part of the problem is that we need to select initial probability distributions for new claims in light of a vast amount of background information and belief, and part of the problem is that we need to simultaneously adjust that information and belief in consideration of the new possibility we see. Although the problem of figuring out how conceptual learning works involves some of the same challenges as the problem of the priors, conceptual learning necessarily involves figuring out how to redistribute probabilities between old claims and new ones, meaning we have to adjust our old beliefs without having encountered any evidence anticipated by those old beliefs' updating plan. The

task isn't just to explain the principles governing initial probability assignments to new hypotheses, it's to explain the ones governing probability reassignments to old hypotheses in light of new ones.

This conceptual learning only occurs when a new hypothesis seems relevant to an agent's beliefs about the implications of possible evidence for their existing hypotheses. If a new hypothesis is completely independent of all an agent's old hypotheses, it won't affect their existing conditional probabilities about how potential evidence affects the relative probabilities of those hypotheses—the likelihood ratios of its old hypotheses should stay the same, and their priors for them will stay the same, so it'll still know exactly what to do with its old hypotheses upon encountering evidence. So the only problem raised by learning completely independent new hypotheses is the problem of what priors to assign propositions involving them, which is just the original problem of the priors. Therefore whenever I discuss new hypotheses in this paper, I refer to new hypotheses that are not independent of the agent's existing ones.

The problem of the priors requires figuring out how best to assign probabilities to a partition of the world's basic possibilities, while the problem of conceptual learning requires us to figure out a principled way of moving from one partition to another that divides basic possibilities in a different way due to having a different set of hypotheses. Nothing about the problem of the priors resembles this challenge about how to alter the partitions of hypotheses that our credences are based on, although once we do have revised partitions and revised sets of propositions based on them, we will encounter the problem of how to assign priors to any new propositions. But the way we arrive at these new propositions will likely inform the method we use to decide what initial credences we should have in them.

The problem for a Bayesian is not *that* we learn new hypotheses; the problem is explaining *how* we do it. And, as far as possible, we should explain how conceptual learning works using the same basic ingredients that learning by conditioning uses—partitions of hypotheses and evidence, an algebra of propositions generated from those partitions, priors in those propositions, and conditional probabilities about how to update those priors upon getting evidence. In addition to being parsimonious, this approach would give a unified theory of learning, while explaining conceptual learning using fundamentally different ingredients than conditioning would raise difficult questions about how the two kinds of learning interact.

My main goals in this paper are to describe the problem of conceptual learning with as much specificity as possible and to provide some initial answers. In Section Two I discuss some recent work on modeling growing awareness. I argue that the emerging view called Reverse Bayesianism, which says we should hold all our probability ratios between old hypotheses constant when incorporating new hypotheses, is a useful starting point but too strong. In Section Three I'll raise a potential answer involving something called the catch-all hypothesis, where we open the door to revising our concepts by having a general hypothesis that our worldview might be wrong in some way we can't specify. In Section Four I address the main problems with the catch-all based theory, including the difficulty in assigning the catch-all a prior figuring out how to update credence in it upon encountering evidence. In Section Five I discuss a richer version of the catch-all that can make it easier to learn new hypotheses. Then in Section Six I explain how considering and believing a new hypothesis can make us change our probabilities in existing non-catch-all hypotheses. I give an account of how new hypotheses can make us reconsider the implications of our evidence.

2. Reverse Bayesianism

In recent years a subset of the problem of conceptual learning has been discussed as "the problem of growing awareness." A few economists and philosophers have written on the question of how our existing beliefs should adapt within a Bayesian framework when they're confronted with new possibilities. I am interested in this question, but I'm also interested in the prior question of how our own credal functions could drive us to come up with new possibilities and the further question of how we should assign initial probabilities to them.

The recent line of literature stems from economists Edi Karni and Marie-Louise Vierø's claim that an agent who becomes aware of new possibilities should modify their beliefs "in such a way that the likelihood ratios of events in the original state space remain intact."[72] They call this rule Reverse Bayesianism. Some philosophers have translated Reverse Bayesianism into the epistemic context and advocated it.[73] But more recently, the principle has been challenged in work from Anna Mahtani[74] and Katie Steele and H. Orri Stefánsson.[75] Joe Roussos defends it in a working paper where he formally defines it as follows:[76]

> **Reverse Bayesianism.** Suppose that A and B are maximally-specific propositions the agent was previously aware of, P represents the agent's prior beliefs, and $P+$ represents their extension after a growth of awareness. For any such A, B, where $P(A)>0$ and $P(B)>0$, a rational agent will have:
>
> $$P(A)/P(B) = P+(A)/P+(B)$$

The appeal of Reverse Bayesianism is twofold: it both respects the normative force of our original beliefs and gets us partway to an answer to the question of how to model the

---

[72] Karni, Edi, and Marie-Louise Vierø. "'Reverse Bayesianism': A Choice-Based Theory of Growing Awareness." *American Economic Review*, vol. 103, no. 7, 2013, p. 2801.
[73] Bradley, Richard. *Decision Theory with a Human Face*. Cambridge University Press, 2017.
[74] Mahtani, Anna. "Awareness Growth and Dispositional Attitudes." *Synthese*, vol. 198, no. 9, 2020, pp. 8981–8997.
[75] Steele, Katie, and H. Orri Stefánsson. "Belief Revision for Growing Awareness." *Mind*, vol. 130, no. 520, 2020, pp. 1207–1232., https://doi.org/10.1093/mind/fzaa056.
[76] Roussos, Joe. "Awareness Growth and Belief Revision." 2021, https://doi.org/10.31235/osf.io/bfv56, p. 34.

introduction of new beliefs. This is the beginnings of an effort to take up the task that Kuhn and Earman regarded as hopelessly subjective. The Reverse Bayesians seem to be arguing that we can at least come up with one clear, formal constraint on the credences we possess in our new credal function: the probability ratios between all the old basic possibilities should stay the same. The modern theorists are silent on the larger question Earman would press about what the probability ratios should be between new possibilities and old ones. Reverse Bayesianism therefore cannot deliver a verdict on what probabilities we should assign old possibilities in the new credal function. The Reverse Bayesians seem to be insisting that even if we can't formalize an answer to that question, we can at least impose *some* rational constraint on our new credal function, so the process of moving to it isn't *entirely* subjective. Intuitively, we should try to preserve as much as we can of our original beliefs, since they are our beliefs, and one thing we can preserve is our assessment of their probabilities relative to each other.

Mahtani, Steele, and Stefánsson provide powerful commonsense counterexamples to Reverse Bayesianism (RB). Before I present one of the counterexamples, I want to make a theoretical objection to RB that hasn't been made yet. It appears to hold that it is impossible for awareness of a new possibility to ever cause us to move from non-zero credence in a proposition to credence 0 in it. Suppose your original credal function P assigns non-zero credence to two incompatible basic propositions A and B. Whatever the original ratio is between P(A) and P(B), that ratio cannot be maintained if the new credal function P+ assigns 0 to A or B. Therefore RB forbids assigning A or B credence 0 in P+. So the formal rule RB delivers a surprisingly powerful substantive result: it implies that learning of new possibilities can never make us come to regard old possibilities as impossible. Strikingly, it makes this claim without knowing anything at all about the content of any of the old or new propositions involved. Not only is the

claim empirically dubious, we shouldn't be able to learn such a strong substantive claim by adopting a mere procedural rule like RB. Reverse Bayesianism is born out of an admirable desire to preserve whatever can be preserved from our original beliefs, but this attempt is so inflexible that it entails a substantive claim about the world that our old beliefs likely never asserted.

Steele and Stefánsson convincingly argue that learning of new possibilities can obviously make us reevaluate the evidential relationships between our old possibilities, which would therefore alter the ratio of their probabilities. They build their case around variations on a useful example involving one's beliefs about which foreign movie is playing at the local theater.[77] Suppose you start out with the belief the movie could either be French or German, and either a thriller or comedy. The authors point out that awareness could grow in two distinct ways. If you realize that the language difficulty of the movie matters to you and that difficulty could be either high or low, you will *refine* your partition of possibilities. This would require taking the previously fundamental cell saying, for instance, "The movie is a German thriller" and splitting that into two new basic possibilities saying "The movie is a German thriller with simple language" and "The movie is a German thriller with complex language." On the other hand, say Steele and Stefansson, if you realize that there's a third genre like drama that you haven't considered, you will *expand* your partition of possibilities, keeping the fundamental ones like "The movie is a German thriller" intact but adding "The movie is a German drama" and "The movie is a French drama."

These are good distinctions which I'll apply in my own theory of conceptual learning. The existing literature, whether for or against Reverse Bayesianism, accepts Steele and Stefansson's claim that the two relevant ways an agent's partition of possibilities can change

[77] Steele and Stefánsson, p. 1212.

upon growing awareness are refinement and expansion. Notably, no author entertains the idea that growing awareness could ever cause us to *eliminate* old possibilities from our partitions. Yet this seems like something that could happen. Suppose, for instance, that a movie critic informs you that the French believe all thrillers should be funny and all comedies should be thrilling; they don't distinguish between the two genres. In that case, "French and Thriller" and "French and Comedy" no longer deserve to be separate basic possibilities, since they no longer compete with each other. While you'd maintain the "German and Thriller" and "German and Comedy" basic possibilities, the French ones would be condensed into "French and Comedy/Thriller"—the opposite of refinement. It's not clear what Reverse Bayesianism would recommend when old possibilities are eliminated; it may depend on the reasons for the elimination. In this example I argued that two old basic possibilities could be *condensed* into one new one, but the elimination of old possibilities could take other forms, too, like simple erasure. Recognizing new possibilities might convince us that some of our old possibilities are not coextensive, but incoherent—the opposite of expansion.

Steele and Steffanson raise examples like Einstein's theory of special relativity to point out that paradigmatic cases of awareness growth by expansion can make us reevaluate evidential relationships between old possibilities. Then they revisit their movie example to show that even ordinary cases of refinement can defy Reverse Bayesianism. After supposing you have refined your partition of possibilities to include low and high complexity French and German films, they add "Since it occurs to you that the owner of the cinema is quite simple-minded, you are, after this realisation, much more confident that the movie will have low-level language than high-level language. Moreover, since you associate low-level language with thrillers, this makes you more

confident than you were before that the movie on offer is a thriller as opposed to a comedy."[78] You were previously aware the owner of the cinema was simple-minded, but refining your awareness to include the basic possibilities of simple or complex language makes you draw a new evidential connection between the simple-mindedness of the owner and the kind of movie they'll choose to show.

Thanks to clear counterexamples like this, the tide is turning against Reverse Bayesianism. But to stop here would be an unfortunate result for Bayesianism, since conceptual learning could then easily derail one's plan to learn through conditioning. Someone who plans to learn through conditioning can't allow the slate to be wiped clean each time their awareness grows. Part of the great appeal of the strict conservatism of RB is that it means you might not have to abandon much of your original plan to learn by conditioning; your old work is not wasted and much of your old plan is still valid. Intuitively, that's often the correct result. Of course, a weak version of Reverse Bayesianism still survives all the counterexamples: when new possibilities are not evidentially relevant to the probability ratios between old ones, we should maintain the old ratios in our new credal function. They were, after all, our beliefs, so we should continue to believe them, in the absence of any reason to doubt them. Even Kuhn and Earman would accept that.

In his draft paper, Roussos mounts a defense of the strong version of RB. The heart of his claim is that recognizing the evidential import of new possibilities for old ones is part of the process of belief revision that *follows* awareness growth, and RB is still a rational constraint on awareness growth itself. Essentially, Roussos argues that upon learning of new possibilities, we should first grow our awareness by creating a new algebra that includes the old and new

[78] Steele and Stefánsson, p. 1220.

possibilities. Then, he says, we initially populate the new algebra with beliefs by extending our old beliefs to it wherever possible, obeying RB. Only after this should we consider the evidential import of new possibilities and make necessary revisions to the probability ratios between old ones. He adds "The separation [between stages] is conceptual rather than temporal: the aim is to focus first on the purely awareness related aspects of the experience and then to turn to the attitude of belief. The experience that brings about these three steps might (and often will) be a single, unitary experience."[79]

Roussos' effort to preserve Reverse Bayesianism strikes me as a little like destroying the village to save it. He says that we must maintain our old probability ratios in our new beliefs, but also that at the *very same instant* we extend our old ratios we'll often revise them to respond to the evidence gained from the new possibilities. So in such cases there is no moment in time when the new credal function actually holds all the old probability ratios. I don't know how the new credal function can really be said to maintain all our old probability ratios if there's never a moment in time when it believes them, and I don't see why it must be a rule of rationality that we maintain our old probability ratios if we can in the same instant change them. Roussos says extending our old probability ratios to the new algebra is necessary so that we can update them with the new evidence we've gained—after all, the old credal function didn't contain any updating plan involving the new possibilities.[80] But Roussos seems to ignore the possibility that we could simply start fresh in the new credal function, drawing whatever inspiration we like from our old credal function without being beholden to it. Roussos never really explains why rationality forbids us from creating brand new probability ratios between old possibilities in our new credal function when new possibilities are evidentially relevant. Our new credal functions

[79] Roussos, p. 57.
[80] Id.

very likely would include beliefs about what our old probability ratios were, but it's not clear why they'd necessarily have to include all the old ratios themselves.

So the upshot of the literature on growing awareness is that Reverse Bayesianism is false. But the debate was not wasted. It has delivered us the insight that probability *ratios* between old propositions can be maintained in the new credal function even if we don't know what the probabilities themselves should be. It has delivered the distinction between refining and expanding partitions, to which I would add the eliminative counterparts condensing and erasing. And, in my opinion, we should not throw out the strong version of Reverse Bayesianism entirely, even though the counterexamples are compelling. The motivations for the principle are also compelling. Rather than completely abandoning it, we should downgrade Reverse Bayesianism from a rule to a suggestion. Our old probability ratios should exert some force over us even when new possibilities exert countervailing force modifying them. The question is now how to specify a way of balancing these forces.

To make additional progress on modeling conceptual learning, we need to take up the important questions the growing awareness literature has been silent on. How do our credal functions discover new possibilities? More importantly, how do we assign them priors? How do we come up with probability ratios between incompatible new possibilities and old ones? Instead of modestly picking at isolated aspects of conceptual learning and debating answers to them as if the other aspects are irrelevant, we might be able to make more progress if we try to answer all the questions at once. If the questions are related, the answers probably are too.

3. How a Catch-all Hypothesis can Help with Conceptual Learning

My approach to the problem of conceptual learning revolves around developing an old approach called a "catch-all hypothesis." I'll develop a theory where the catch-all hypothesis plays an important role in finding new possibilities, assigning priors to them, and assigning new probabilities to old beliefs. I'll begin by arguing that to recognize the limitations of our imaginations we ought to have non-zero credence in a catch-all hypothesis saying all our regular hypotheses are wrong. This opens the door for us to discover new possibilities.

Earman thinks of failing to conceive of a possibility as one way for an agent to fail to be logically omniscient. The first requirement of logical omniscience is the ability to recognize all the logical implications of one's beliefs, and satisfying all probability axioms, something computationally limited agents may be unable to do, he writes.[81] He adds "Actual agents also fall short of logical omniscience by being unable to parse all the possibilities, and this inability can skew degrees of belief."[82] The two ways of falling short of logical omniscience are closely linked: the fewer the basic possibilities a computationally limited agent divides the world into, the easier it is for them to recognize all the logical implications of those possibilities. So the better an actual agent does at refining their partitions of hypotheses and evidence to reflect all relevant possibilities, the more difficult it will tend to be for them to recognize all the logical implications of those possibilities.

This failure to be omniscient when recognizing all logical possibilities may be the fate of all actual agents, to some degree, just as having limited computational power prevents them from having beliefs that fully comply with the laws of probability. But to be as rational as it can be, a limited epistemic agent should at least recognize its own limitations as well as it can and have an

[81] Earman at 56.
[82] Id. at 56-57.

appropriate degree of humility about the completeness of the set of propositions its credal function is defined over. The less we expect that we've anticipated all relevant possibilities, the less unanticipated ones will shock us and cast our beliefs into disarray. As Bas van Fraassen observes in an argument against inference to the best explanation, which he defines as believing that our best explanations are probably approximately true, "We can watch no contest of the theories we have so painfully struggled to formulate, with those no one has proposed. So our selection may well be the best of a bad lot… For me to take it that the best of set X will be more likely to be true than not, requires a prior belief that the truth is more likely to be found in X, than not."[83] When a credal function divides all its confidence among a non-exhaustive set of mutually exclusive hypotheses, it commits the sin van Fraassen indicts to the highest degree, because it not only takes it to be more likely than not that the truth is among those hypotheses, it takes it to be certain. Fortunately, there's a way of following Bayesian learning's ordinary plan that can prevent agents from falling into this trap: having some amount of confidence in a catch-all hypothesis that all of one's other hypotheses are wrong.

The idea of a catch-all hypothesis was first raised by Abner Shimony.[84] To explain what exactly the catch-all is and how it might help with understanding conceptual learning we need to go into more detail about what a set of beliefs is to a Bayesian. Typically, a credal function is defined over an algebra of propositions, a comprehensive set of propositions about the world which is closed under the basic Boolean operations, meaning that it contains all the possible negations and countable conjunctions and disjunctions of its elements. To find the elements that form this algebra of propositions, we first create a partition of all the possible hypotheses about

[83] van Fraassen, Laws and Symmetry, p. 143.
[84] Abner Shimony. Scientific inference. In R.G. Colodny ed., *The nature and function of scientific theories*. Pittsburgh: University of Pittsburgh

how the world could be and a partition of all the possible pieces of evidence we could receive about it. A partition of a sample space divides it into mutually exclusive subsets which, taken together, cover the entire sample space. When that sample space is the whole world, creating a partition of hypotheses gives an exhaustive list of hypotheses describing all possible ways the world might be, and creating a partition of evidence creates an exhaustive list of all possible states of total evidence experience could deliver about how the world is. If we take a set composed of all those claims about hypotheses and evidence and generate an algebra from that, we get an algebra of propositions which comprehensively describes all the possibilities we could conceivably learn. These propositions range from basic ones about the individual hypotheses themselves, or individual pieces of evidence themselves, to composite ones like the proposition formed by the disjunction of all possible hypotheses and all possible observations.

When someone divides all their confidence over a non-exhaustive set of hypotheses, if their credal function is defined over an algebra of propositions, they must be assigning a credence of 0 to the proposition formed by the negation of the disjunction of all those hypotheses (and, equivalently, the conjunction of all their negations). For example, if they're a detective considering which employee in a mansion committed a crime and they divide all their confidence among the three mutually exclusive hypotheses "the butler did it," "the maid did it," and "the gardener did it," generating an algebra from that partition of hypotheses delivers the proposition "it's not the case that the butler, maid, or gardener did it," but they can't have any confidence at all in that proposition, so they'll never interpret any observation as incremental evidence for it—one of the defining features of a credence of 0 or 1 in a proposition is that no learning experience can change it. For them to have any non-zero credence in the proposition "it's not the case that

the butler, maid, or gardener did it," they'd have to have a partition of hypotheses that included the hypothesis that none of the three did it. That would be their credal function's catch-all.

Suppose their partition of potential evidence includes things like hearing convincing alibies from the butler, maid, or gardener, seeing videos of them at their homes during the time period the crime was committed, having multiple reliable witnesses vouch for them, and so on, and when they investigate the crime, all of these possibilities turn out to be true for each of them. Because such an agent is committing the sin van Fraassen indicts in his critique of inference to the best explanation to the highest degree, assuming not just that the truth is probably contained within their three hypotheses, but that it must be, they can't react to this overwhelming evidence that each of the three suspects is innocent by concluding that all three are innocent. They would have to update within the confines of their limited partition of possibilities by, for instance, concluding that one of the suspects is a great liar, many of the witnesses must be lying or mistaken, and one of the videos must be doctored, reasoning that although the combination of all those events seemed very unlikely at the outset, it must be true. If we imagine that each of the three hypotheses seemed equally likely at first, and that all the seemingly compelling evidence for the innocence of each suspect is very similar in nature, the agent will be forced to seize upon the slightest differences in the evidence to figure out which suspect is guilty, making mountains out of molehills and coming up with wild tales to distinguish some testimonies from others, asserting, like a bad version of Sherlock Holmes, that once the impossible has been eliminated whatever remains, no matter how improbable, must be true. This is a Bayesian account of how to become a conspiracy theorist. The way for an agent to avoid becoming a conspiracy theorist is to have an open mind, and to do that they need to have non-zero credence in the proposition

expressed by the catch-all hypothesis that all their regular hypotheses are wrong. Having some confidence in that catch-all hypothesis would make it part of their partition of hypotheses.

In theory, an agent could fail to be logically omniscient in Earman's first sense (recognizing all the logical implications of their beliefs) by defining their credences over something other than an algebra out of failure to apply a negation to a proposition, or failure to imagine a conjunction between two propositions, for instance. I think it's reasonable to suppose that a non-ideal agent's credal function can be represented as an algebra. Even if an imperfect memory or impaired reasoning causes me to lack an occurrent or dispositional belief about the proposition ~P to go along with my belief about P, for example, I'm still committed to the principle of having beliefs about the negations of all propositions I have beliefs about. If I reasoned about what to do or believe and I used my credence about P I'd be likely to discover and use my missing credence in ~P along the way. If I reflected and reasoned more carefully, or as part of a group, my beliefs would be likely to better approximate ones defined over an algebra, which makes it plausible to model non-ideal agents like ourselves as having exhaustive beliefs about expressions using logical connectives between the possibilities we're aware of. This idealization is one we can aim for.

It seems far too optimistic to hope we can model non-ideal agents as logically omniscient in Earman's second sense and assume them, and ourselves, to start out with a partition of hypotheses that's refined enough that they'll never need to refine it further and complete enough that they'll never need to expand it. In the detective example the agent is forced to embrace implausible theories about their evidence because there's no room to alter their hypotheses. But if they have non-zero credence in the catch-all hypothesis automatically generated by their algebra of propositions, when they receive compelling evidence that each of the three suspects

isn't guilty they can conclude that none of the three are guilty, which amounts to becoming very confident in the catch-all. Shimony argues that we should have non-zero credence in all coherent possibilities, which would include the catch-all.[85] Crucially, this automatic catch-all hypothesis supplied by their algebra states only that none of the three suspects they're aware of is guilty—it says nothing positive about who is guilty, or even if anyone at all is guilty. Those answers can't come from within the agent's current partition of hypotheses.

But going from a credence close to zero in the automatic catch-all to a high credence in it should focus the agent's attention on their general, automatic catch-all hypothesis, and, now that they know the right answer is contained within it, the natural next step for them is to refine that general catch-all hypothesis into more specific hypotheses. The process of refining the vague, general catch-all into explicit hypotheses equips them with new hypotheses to incorporate into a new partition, either through refining, expanding, condensing, or erasing old basic possibilities. Writing in support of a catch-all hypothesis, Sylvia Wenmackers and Jan-Willem Romeijin borrow Earman's language to say that in cases of growing awareness we "shave off" new hypotheses from the catch-all. As I'll elaborate in the next section, Earman was entertaining the idea that we shave off *confidence* in the catch-all to obtain confidence for new hypotheses, but Wenmackers and Romeijin seem to understand him as saying that we shave off the new hypotheses themselves from the catch-all, which amounts to refining the catch-all hypothesis.[86] When we refine the catch-all hypothesis, we create new regular hypotheses which we can in turn use to modify our partition of regular hypotheses.

[85] Shimony at 92.
[86] Wenmackers, Sylvia, and Jan-Willem Romeijn. "New Theory about Old Evidence." *Synthese*, vol. 193, no. 4, 2015, p. 1236.

For instance, once the detective knows the solution to the case is contained within their general catch-all "it's not the case that the butler, gardener, or maid did it," they might consider a variety of possible refinements to the catch-all, like "there's some employee other than those three," or "there wasn't actually a crime," or "multiple employees did it," or "it wasn't an employee who did it." They'll be forced to question assumptions they treated as given when creating their original partition of hypotheses. When they consciously refine their automatic catch-all hypothesis into more specific hypotheses, the natural next step is to think about things like what kinds of other employees there might be. As they refine the automatic catch-all hypothesis and specify it, that amounts to generating new regular hypotheses—for instance, their new partition might expand to become "the butler did it, the maid did it, the gardener did it, the accountant did it, or none of those four did it." Or it could be refined to include the possibility that there are two gardeners. Or, in a more complex change, the detective could significantly expand their basic possibilities to include the hypotheses that various combinations of employees working together committed the murder.

In the easiest cases, like just adding the hypothesis that the accountant did it, it would be simple for the agent to figure out how to adjust their priors and conditional probabilities. Assuming that they had a prior of slightly below 1/3 in each suspect being guilty in their original credal function, in the new one they might have, for instance, credence slightly below ¼ in each. These straightforward adjustments in the priors of their old hypotheses would naturally entail straightforward adjustments in conditional probabilities involving them. For instance, in the original credal function, learning the evidence that the gardener had the perfect alibi would make the agent update to a credence close to 0 in the gardener being guilty, and about .5 each in the butler and maid being guilty. In the new credal function, learning about the gardener's perfect

alibi would lead to credences of about 1/3 each in the butler, maid, and accountant being guilty. The easy cases of conceptual learning are solved by making simple modifications to a partition and applying Reverse Bayesianism.[87]

The more difficult cases can arise in two separate ways. The first is the category the literature on growing awareness has come to recognize: even simple alterations to a partition of hypotheses can sometimes challenge RB because the presence of new possibilities can make us reconceive the evidential relationships between old ones. Imagining the accountant as a possible suspect, for instance, or two gardeners, could potentially change some of our old probability ratios between the original suspects. The second category of difficult case arises when changes to a partition of hypotheses are themselves so complex that RB doesn't give much guidance. If the detective realizes any combination of the employees could've been working together, their new partition will be so different from their old one that Reverse Bayesianism doesn't help much. Maybe a few old probability ratios can be maintained, but what little Reverse Bayesianism can say about the new credal function is not much. The second difficulty may tend to be accompanied by the first: the more radical the changes we make to our partition, the more likely it is we will reconceive the evidential relationships between whatever old possibilities remain. The more complex the changes to a partition, the less Reverse Bayesianism can tell us about what the new credal function should look like. The answer depends more on how we assign priors to new possibilities and create new conditional probabilities involving them. I consider the

[87] When the Reverse Bayesians speak of holding old probability ratios between hypotheses constant, that's usually another way of saying conditional probabilities involving only old hypotheses and observations should be held constant, and that any changes to the expected probabilities of the inputs should be symmetrical. One of the weaknesses of RB is that it ignores the fact that new hypotheses can create new possible observations, and that conditional probabilities connecting the new observations to old hypotheses can then easily change the probability ratios of old hypotheses. I discuss this more in Section Six.

details of this process in Section Six, but for now I want to make a more general point about how having credence in the catch-all hypothesis could help in the difficult cases.

It seems to me that an agent who's deciding how to modify their partition of hypotheses would naturally prefer to start with the simplest adjustments, trying out things like just adding the hypothesis "the accountant did it," hoping that would work before trying more complicated modifications like "the accountant helped the butler do it." That would be a way to acknowledge both their original credal function and their newfound high confidence in its catch-all hypothesis that it's missing something: they start by assuming it was missing as little as possible. They'd modify their partition of hypotheses in the most minimal way that seemed like it might work, then take their new credal function and update on the same body of evidence that led to high confidence in the original credal function's catch-all, hoping it wouldn't lead to high confidence in the new credal function's catch-all. They'd hope that adding a hypothesis like "the accountant did it" and conditioning on the same body of evidence that confounded them before would now just lead to high credence that the accountant did it. But if that simple expansion still led to high credence in the new catch-all "none of those four employees did it" they'd know they had to try more radical adjustments, like considering that multiple employees could've worked together.

In other words, we start out by hoping we're dealing with an easy case of conceptual learning. But when treating the case as easy doesn't work, our high credence in our new catch-alls drives us to make more radical changes to either our probability ratios, our partitions, or both. This is what I mean when I say that we should treat Reverse Bayesianism as a suggestion instead of a rule. Psychologically, our old partitions and probability ratios both will and should exert a strong grip on us. We will start out by making the most minimal changes possible, driven to make ever more complex changes until we finally solve the case and arrive at a credal

function that isn't stumped by our total evidence, one with low credence in its catch-all. If adding the hypothesis "the accountant did it," adjusting their old priors to fit their new partition, and conditioning on the same evidence that confounded their old credal function still confounded their new one, before questioning the assumptions underlying the way they partitioned all their regular hypotheses, the detective might prefer to try something like refining the hypothesis "the gardener did it" to account for the possibility there are two gardeners. That would leave many of their conditional probabilities involving hypotheses about the maid and the butler intact, but rethinking the assumption that a single suspect committed the crime would require them to alter the way they drew their entire partition of hypotheses, and would therefore require rethinking far more of their old priors and conditional probabilities.

The general principle I suggest is that an agent ought to respond to high credence in its catch-all hypothesis by altering its probability ratios and partition of hypotheses in the way that lowers credence in the catch-all while requiring the least change to its existing credal function. I'll call this the principle of minimal change. This process of trying minimal changes, watching them fail to solve the problem (or make only slight progress), and then making more radical changes to our partition of hypotheses, is an account of how we gradually refine our concepts. On my theory, the difference between the easiest cases of conceptual learning and the hardest is, like belief itself, just a matter of degree.

In hard cases of conceptual learning, we're trying to modify our credal functions to balance three competing goals: we want to preserve as much of our old probability ratios as possible, preserve as much of our old partitions as possible, and lower credence in our new catch-all hypothesis as much as possible. The most important of the goals is the third— if a credal function makes minor changes and they only manage to slightly lower credence in its

catch-all hypothesis, then by its own lights the new credal function is still missing something important, and it ought to make further adjustments.

On my theory, we ought to choose the probability ratios for old possibilities and the priors for new ones that best achieve these three goals. After we create a new partition and obey Reverse Bayesianism as far as possible, there are many possible combinations of priors we could choose to assign new possibilities, and those would in turn entail claims about the probability ratios between new possibilities and old ones. The fundamental goal of lowering credence in the catch-all hypothesis provides a powerful constraint on which new priors we select: in the absence of any other constraints, we might as well choose the combination of priors which minimizes credence in our new catch-all hypothesis. When the detective has tried all the easy changes to their partition and none of them have meaningfully lowered credence in the catch-all, for instance, once they reluctantly dramatically change their partition to include all the combinations of employees that might've committed the crime together, they face a question: how do they assign priors to all these new possibilities? My answer is that they should choose the set of priors that minimizes credence in their new credal function's catch-all. If the set of priors that best accounts for the total evidence assigns the highest credence to the possibility that the accountant and butler worked together, that's the best solution to the case.

To sum up, I think the principle of minimal change generally recommends proceeding in several steps:

1. Make the most minimal plausible change to your partition, trying to keep as many old basic possibilities as possible and add as few new ones as possible. I'll call the algebra generated by the old partition P and the algebra generated by the new one P+.

2. Apply Reverse Bayesianism within P+, extending all old probability ratios to the new algebra. If any new possibilities within P+ are obviously evidentially relevant to old probability ratios, use the old ones as a foundation, but adjust them.[88]
3. To complete the credal function C+ defined over P+, assign all new basic possibilities the set of priors which minimizes credence in P+'s catch-all hypothesis. Combined with the old probability ratios extended by RB (or adjusted versions of those), these priors will determine the final credences for old possibilities within P+. In easy cases of conceptual learning, the minimal change of adopting C+ should result in credence near 0 in C+'s catch-all. If so, no further changes are necessary.
4. If not, the next step is to consider whether the new possibilities in P+ have changed the evidential relationships between the old possibilities in P in ways that aren't obvious. Loosen adherence to RB. See if the most minimal tweaks, combined with any set of priors for new possibilities, are sufficient to lower credence in the catch-all, gradually making larger tweaks to old probability ratios if not.
5. If after step 4 the best credal function defined over P+ still has high credence in P+'s catch-all hypothesis, a more radical change in partition is necessary. Return to step 1, this time making the most minimal plausible change to P+, generating P++. Repeat these steps until you arrive at a credal function that has near 0 credence in its catch-all, or, failing that, until it's impossible to find any credal function that has lower credence in its catch-all than the last one did.

[88] This process is easier said than done. It involves creating conditional probabilities relating new and old observations to old hypotheses, assigning probabilities to those observations, and then combining the two to arrive at new probabilities for old hypotheses. I elaborate in Section Six.

There is a substantive commitment in this method that may be false. In this way of fleshing out minimal change, I have assumed that preserving as much as possible of old partitions is significantly more important than preserving old probability ratios. This assumption is in the spirit of Bayesianism: acquiring new evidence can make us radically change the probability ratios of hypotheses all the time. If we can get close to credence 0 in the catch-all by making a small change to our partition plus a large change to some old probability ratios, that just amounts to thinking that the new possibilities have large evidential import for some old ones. The old probability ratios can ultimately be traced back to our original priors for hypotheses anyway, so they may not have been based on much, and may not be worth clinging to in the face of unanticipated possibilities. If the detective can account for all the evidence by adding the hypothesis the accountant did it and having high credence in that, but doing so dramatically alters the relative probabilities between the butler and the gardener, maybe that's just the way things should be.

But maybe not. Here's where some subjectivity could enter the process. What if making a more radical change to one's partition allows one to adhere more closely to Reverse Bayesianism? What if, for instance, adding the possibility of the accountant *and* simultaneously refining to include the possibility of two gardeners allows one to maintain all old probability ratios, while merely adding the accountant requires a change to old probability ratios? What should we be more attached to, our old probability ratios or our old partition? This may be a matter of personal preference. One could, for instance, use a version of my procedure where they place a limited range on how much they're allowed to tweak old probability ratios during Step 4 before being forced to make a new change to their partition. The scope of that range would express how attached one is to their old probability ratios.

So there can be multiple valid solutions to the optimization problem of conceptual learning, where we try to minimize credence in the catch-all by making the most minimal changes possible to our old partitions and our old probability ratios. There will often be many Pareto-optimal approaches to how to revise our concepts. For instance, what if new credal functions Y and Z each manage to account for one's total evidence far better than original credal function X, but Y does a better job preserving old probability ratios while Z does a better job preserving the old partition? The subjectivity that troubled Kuhn and Earman rears its head again: there seems to be no rationally mandated way to choose between Y and Z. I think it's in the spirit of Bayesianism to prefer Z, prioritizing keeping a similar partition and allowing the new possibilities to have greater evidential weight. Others might disagree. But what matters most for my purposes is that Y and Z are both better accounts of the total evidence than X, and are each strictly better than many other potential modifications to X. By more precisely specifying how the process of conceptual learning is subjective, we make it less subjective.

It would be a mistake to assume that there must be a single, precise answer to the question of how our credal functions should change during conceptual learning. An objective method of choosing a new credal function doesn't necessarily have to deliver a precise result; it's more important that it rule out many wrong answers than that it deliver a single right one. Perhaps the single most correct thing to do in these cases is to adopt a complicated, highly imprecise credal function which treats all Pareto-optimal solutions to the optimization problem as plausible. Roussos reaches a similar conclusion about accepting high imprecision after delivering his own theory of how to model growing awareness, though he doesn't use a catch-all hypothesis and doesn't address how we might determine priors for new possibilities.[89]

[89] Roussos at 62.

4. Problems for the Catch-all Approach

Wenmackers and Romeijin, unlike other recent thinkers on growing awareness, do support having a catch-all hypothesis and flesh out a proposal for how it would work. But there's a big difference between our approaches. At every step mine revolves around adjusting credence in the catch-all. Confounding evidence can start the process of conceptual learning by driving an agent's credence in their catch-all up near 1, compelling them to create new hypotheses. Then during the process of modifying their partitions, adjusting probability ratios between old hypotheses, and assigning priors to new ones, at every step of the way the goal is to make the most minimal changes necessary to drive credence in the catch-all down near 0. High credence in the catch-all starts the process of conceptual learning, and low credence in the catch-all ends it. Wenmackers and Romeijin, on the other hand, support having non-zero credence in a catch-all, but don't think one can find reasons to assign any determinate credence to it, accepting a similar claim made by both Shimony and Earman. They therefore don't discuss adjusting credence in the catch-all up or down.

So this is the first common objection to a catch-all hypothesis: it's impossible to know what credence to assign it. Why not assign a particular prior to the catch-all? Wenmackers and Romeijin write, "Since the catch-all is not based on a scientific theory, the usual 'arational' considerations (to employ the terminology of Earman 1992, p. 197) for assigning it a prior, namely by comparing it to hypotheses produced by other theories, do not come into play here."[90] The view they ultimately arrive at is that most of the time, we behave as if our credence in the catch-all is 0—what they call "silent open-minded Bayesianism." When this leads us to hold implausible beliefs, they say our theoretical context has changed, and we become vocal open-

[90] Wenmackers and Romeijin at 1234.

minded Bayesians, acknowledging the possibility of new hypotheses and changing our partitions and algebras to include them.[91]

I don't think the mere existence of a catch-all hypothesis with some indeterminate non-zero credence can do all the theoretical work Wenmackers and Romeijin want it to. When we're in the silent phase of open-minded Bayesianism, pretending credence in our catch-all is 0, we should *never* encounter evidence we regard as confounding or hold beliefs we regard as implausible—instead we should always confidently contort our beliefs to make our evidence fit into our limited partition of possibilities. Even if we're vocal open-minded Bayesians, merely acknowledging that the catch-all hypothesis is a theoretical *possibility* is insufficient to make us introduce new hypotheses—to justify changes to our algebras, we need to believe that the catch-all hypothesis is probably right. Wenmackers and Romeijin try to create a theory of conceptual learning where the mere existence of the catch-all does important work, but you need to be able to quantify credence in the catch-all to really make use of it. Their own central example illustrates the point that credence in the catch-all sometimes seems to increase toward 1 to kick off conceptual learning, and that this increase can be gradual:

> A food safety inspector wants to determine whether or not a restaurant is taking the legally required precautions against food poisoning. She enters the restaurant anonymously and orders a number of dishes. She uses food testing strips to determine for each of the dishes whether or not it is infected by a particularly harmful strain of Salmonella. She assumes that these tests work perfectly, interpreting a positive test result as a Salmonella-infected dish and a negative result as an uninfected one. She also assumes that in kitchens that implement the precautionary practices each dish has a probability of 1% of being infected, whereas this probability rises to 20% in kitchens that do not implement the practices. She orders five dishes from the kitchen and they all turn out to be infected. This prompts her to consider a third hypothesis: the test strips may have been contaminated, rendering all test results positive, irrespective of whether the dish is infected or not.[92]

[91] Id. at 1245.
[92] Id. 1226.

The way Wenmackers and Romeijin present the case, the possibility of the new hypothesis that the test strips are contaminated hits the inspector all at once. They phrase it as if she suddenly acquires the observation "All five dishes are infected" and finds that surprising, jolting her into questioning her assumptions. But in reality, she could be experiencing a sequence of five separate observations, one of each dish, finding each one increasingly surprising. What really seems to be going on is that with each observation of an infected dish, the inspector becomes increasingly confident in her catch-all hypothesis, gradually but exponentially raising her credence in it. Or, if she sees all the results simultaneously, she suddenly becomes very confident in her catch-all. But Wenmackers and Romeijin ignore this psychological reality because of their theoretical commitment to avoid ever assigning a credence to the catch-all. Their theory seems to revolve around some unstated meta-belief system outside of an agent's regular credal function, where the inspector never becomes confident in her catch-all, but becomes confident in some meta-belief telling her that her catch-all is now relevant and that she ought to revise her regular credal function's algebra. Why not just skip the middleman and say she becomes confident in her catch-all? Any problem for assigning a credence to the catch-all will probably also be a problem for assigning a credence to the meta-belief telling an agent their catch-all has become relevant.

Wenmackers and Romeijin, along with Earman, seem reluctant to ever commit any determinate credence to the catch-all because they're persuaded by an argument raised by Shimony, the originator of the idea. Shimony wrote about the catch-all in the context of scientific theories, as Earman and Wenmackers and Romeijin do. His argument is complex, and really only relevant when referring to hypotheses in the strict scientific sense, not the very ordinary sense in which philosophers discussing growing awareness mean the term. Shimony himself notes that

his concern is only applicable when discussing hypotheses in the scientific sense, not when using "hypothesis" to refer to a singular proposition.[93] But it may be worth briefly summarizing.

The gist of Shimony's point seems to be this: the history of human knowledge shows us that virtually all scientific theories are wrong, so strictly speaking, our current theories' probabilities are probably close to 0. But the best old theories often contain seeds that are refined and developed in the best new theories. We can still meaningfully compare the probability ratios of scientific theories (or hypotheses within them) to each other to find which of them best explain a set of observations. But once you accept the idea that what we're really doing is comparing the relative success of scientific hypotheses in a particular domain, it's not clear how to construct a probability ratio between scientific hypotheses and the catch-all hypothesis, which is just the negation of the disjunction of all one's scientific hypotheses. We can, for instance, evaluate how many correct predictions one scientific hypothesis makes relative to another in explaining a set of observations, or compare the percentage of correct predictions they make. But the catch-all hypothesis can't be evaluated in those terms; it doesn't make concrete predictions about any particular set of observations. We're comparing like-to-like when we construct probability ratios between scientific hypotheses, but not when we attempt to construct probability ratios between scientific hypotheses and the catch-all hypothesis.[94]

Although Wenmackers and Romeijin cite Shimony's reasoning to explain why they never assign any credences to the catch-all,[95] simple hypotheses like "The first dish has Salmonella" and "The tests are contaminated" are more prosaic than the scientific ones his argument concerns. The hypotheses he's concerned with are sweeping ones like "Everything is made of

[93] Shimony at 95.
[94] Id. at 94-96.
[95] Wenmackers and Romeijin at 1228.

water" and "Everything is made of atoms." Scientific progress wouldn't invalidate an ancient Greek's high credence in the hypothesis "This dish is poisoned," for instance. Shimony wouldn't doubt our ability to come up with probabilities for everyday hypotheses, so his argument about constructing probability ratios wouldn't apply. Likewise, Shimony's inductive argument about the fallibility of all scientific theories raises no doubt for our ability to assign probabilities to ordinary hypotheses like "The movie is a French thriller," and therefore raises no doubt for our ability to assign credence to a catch-all hypothesis for that credal function's algebra. As the title of his article says, he focuses specifically on scientific inference.

So here's one easy way we could come up with a credence for the catch-all hypothesis: assign probabilities that don't sum to 1 to all regular hypotheses. The catch-all would get whatever scraps remain. Shimony's problem about needing to compare like-to-like to create probability ratios only gets off the ground when we can't imagine how to assign probabilities to our regular hypotheses and can therefore rely on nothing but probability ratios. In the absence of that difficulty, he'd say we should have non-zero credence in all coherent possibilities, which would include the catch-all.[96] Recognizing the mere possibility of the catch-all would be enough to decrease our confidence at least marginally in all our regular hypotheses.

Earman raises a similar concern about the difficulty of assigning priors involving the catch-all, although his argument doesn't involve the special nature of scientific hypotheses. It's not credence in the catch-all itself that worries Earman; his problem centers around the difficulty of creating conditional probabilities about how likely various observations are *given* the catch-all hypothesis. It's similar to the problem Shimony runs into when saying it's hard to create probability ratios between the catch-all and regular hypotheses because the catch-all doesn't

[96] Shimony at 92.

make the same kind of predictions, but Earman reaches that difficulty through a different path. Consider the following formulation of Bayes' Theorem, where H is a particular hypothesis and E is a particular observation:

$P(H|E) = (P(E|H)*P(H))/P(E)$

Although it's fairly common to represent the probability of an observation as P(E), Earman's argument revolves around the fact that we way we actually arrive at a value for the term is to take the sum of the probabilities of E given each hypothesis in our partition times the probability of that hypothesis. So to find the probability of an observation, one of the terms we need to add is $P(E|H_c)*P(H_c)$, where H_c is the catch-all hypothesis. But, asks Earman, how could we possibly know the value of $P(E|H_c)$? As he puts it, "…such an evaluation would of necessity seem to be ill informed, since H_c stands for unexplored territory."[97]

This particular way of putting the problem can actually be dealt with, to a degree, if we simply say that our prior in the catch-all is very low, near 0. Even if we can't find any principled way to assign a value to $P(E|H_c)$, if we set $P(H_c)$ near 0 multiplying the two will result in a product near 0, and the value of P(E) will ultimately be determined almost entirely by its probability under all of our regular hypotheses. The lower our credence in the catch-all is, the less troubling Earman's problem about figuring out P(E) is.

Earman's point about the ambiguity of $P(E|H_c)$ is actually much more troubling if we view it as a problem for how to update credence in the catch-all given any particular observation. Earman phrases it as if the problem comes when $P(E|H_c)$ appears in the denominator of the right-hand side of Bayes' Theorem, but I think the bigger problem occurs when it appears in the *numerator* there because $P(H_c|E)$ appears on the left-hand side. Given a theory like mine which

[97] Earman at 229.

relies on credence in the catch-all rising in response to evidence, for every potential observation, an agent will need to have a conditional probability for how likely the catch-all is given that observation:

$P(H_c|E) = (P(E|H_c)*P(H_c))/P(E)$

And so here, Earman's point about the ambiguity of $P(E|H_c)$ resurfaces in a new, more important context, and this time it can't be mitigated by supposing $P(H_c)$ is near 0. The ambiguous term is relevant each time we want to update credence in the catch-all upon making an observation. How can we know $P(H_c|E)$ without knowing $P(E|H_c)$?

Earman makes another objection to use of a catch-all when he introduces the idea: it can't be a sustainable source of confidence for new hypotheses because there's no obvious way for credence in it to increase. In my view, then, Earman's two problems are related. If we could get clearer ideas on how to assign values to conditional probabilities involving the catch-all, we could have a better idea of how to increase credence in it upon encountering evidence. He writes,

> …we can try to acknowledge the failure of logical omniscience (L02) by means of Abner Shimony's (1970) device of a catch-all hypothesis H_c, which asserts in effect that something, we know not what, beyond the previously formulated theories $T_1, T_2, \ldots, T_q$ is true. Now suppose that a new theory T is introduced and that as a result the old degree-of-belief function Pr is changed to Pr'. The most conservative way the shift from Pr to Pr' could take place is by the process I will call *shaving off*; namely, $Pr(T_i) = Pr'(T_i)$ for i = 1, 2, ... , q, $Pr'(T) = r > 0$, and $Pr'(H_c) = Pr(H_c) - r$. That is, under shaving off, H_c serves as a well for initial probabilities for as yet unborn theories, and the actual introduction of new theories results only in drawing upon this well without disturbing the probabilities of previously formulated theories. Unfortunately, such a conservatism eventually leads to the assignment of ever smaller initial probabilities to successive waves of new theories until a point is reached where the new theory has such a low initial probability as to stand not much of a fighting chance.[98]

Earman seems to have given up on the shaving-off approach partly because of the inadequacy of positing an ever-shrinking credence in a catch-all and therefore ever-shrinking

[98] Earman at 196.

priors for new hypotheses. A different problem would remain even if we took a less conservative approach than he considers and always shaved confidence off of the catch-all but took some from regular hypotheses too to fund initial priors in new hypotheses. On this less conservative version of shaving-off, credence in the catch-all would still converge toward the limit of 0 as more new hypotheses are introduced over time, and eventually virtually all of the initial confidence we had in new hypotheses would come from our non-catch-all ones. Although an ever-shrinking credence in the catch-all would no longer entail ever-shrinking priors for new hypotheses, it would be very strange if it were always true that the more unanticipated new possibilities we learned, the more confident we became that we knew all relevant possibilities.

Wenmacker and Romeijin's example involving the food inspector can help us get an answer about one way credence in the catch-all can increase. Supposing, for simplicity, that she tests five dishes simultaneously and observes that all five are infected, it seems obvious her credence in her catch-all should suddenly jump near 1. Let's say E_5 refers to the observation that all five dishes are infected. Suppose, for simplicity, that the inspector thought that $P(E_5)=.000001$—one in a million. What's going on seems clear: the inspector can either be confident a one-in-a-million event has happened or she can be confident her catch-all is right, and most people will probably have the intuition that she should choose the latter, driving her to come up with the hypothesis the tests are contaminated. The question is how to spell out how she gets from observing E_5 to having high credence in H_c.

Here is one unpromising path worth briefly exploring. In this case, it seems intuitive that the inspector's value for $P(H_c|E_5)$ is near 1; let's call it around .99. And we can assume that her prior for H_c was near 0, say .01. We know that her value for $P(E_5)$ was .000001. So now we can

plug all of those into Bayes' Theorem and see what our intuitions say the value of $P(E_5|H_c)$ is, calling that term X.

.99 = .01X/.000001

Working backward then would tell us that to reach the right intuitive end result, $P(E_5|H_c)$ must have been around .000099, or around one in 100,000. Could anyone have known that in advance? Is it really plausible to say we've figured out that $P(E_5|H_c)$ was .000099? Probably not. That's why I say this line of reasoning is ultimately unenlightening, at least as a way of deducing values for Earman's troubling term $P(E|H_c)$.

What cases like the inspector example reveal is that it's much easier for us to find intuitive values for $P(H_c|E)$ than for $P(E|H_c)$. When we reason through Bayes' Theorem with normal hypotheses, P(H|E) is often the murkier term, and it's easier for us to figure out P(E|H)—with regular hypotheses, it's generally easier to figure out how well a hypothesis predicts an observation than how well the observation predicts the hypothesis. That's because regular hypotheses make clear predictions about the world. But the catch-all doesn't, as Shimony noted. With regular hypotheses, it's usually easier to figure out the values on the right-hand side of Bayes Theorem than the one on the left-hand side. With the catch-all hypothesis, it's hard to figure out a value for the term $P(E|H_c)$, but somehow sometimes very easy to know what $P(H_c|E)$ should be. I don't think the inspector ever knows what $P(E_5|H_c)$ is.

The one observation H_c can be said to predict is that because our partitions are missing the right hypothesis, we will eventually be forced to believe implausible things, including hypotheses we initially regarded as unlikely. It may also predict that we will encounter observations we initially regarded as unlikely—since we lack the right hypothesis, we might lack belief about the observations it would predict, so we will forced to fit that evidence into our

impoverished partition of observations. The detective example I began with illustrates one set of unlikely observations predicted by the catch-all: he observes strong evidence against each of his regular hypotheses. But in my example, because he has no credence in the catch-all, he is forced to deny that he has evidence all of his regular hypotheses are false, leading him to embrace increasingly implausible beliefs to make sense of his evidence. When you have credence 0 in the catch-all and haven't imagined the right answer to a problem, you'll have to interpret every piece of evidence in a way that makes it support one of the wrong answers, which leads to increasingly poor beliefs about both your evidence and your hypotheses.

In other words, our catch-all hypothesis predicts that we will eventually be forced to become conspiracy theorists, contorting our evidence to support our hypotheses, small stretches adding up and eventually compounding to make us believe crazy theories. Imagine the theories the inspector will be driven to embrace if she concludes that every dish in every restaurant she visits has Salmonella. Regular hypotheses make predictions about the behavior of the *world*, but the catch-all hypothesis makes predictions about the behavior of *the credal function it belongs to*. Shimony and Earman and their followers have been holding the catch-all to an impossible standard when they ask how to quantify its predictions about our regular observations; those aren't the entities it makes predictions about. The catch-all hypothesis predicts our regular hypotheses will be inadequate, and that our beliefs will therefore behave increasingly strangely.

The catch-all is incrementally confirmed by the observations that one's credal function is witnessing improbable evidence and believing improbable hypotheses. This could happen gradually or all at once, as illustrated by whether the inspector views dishes one by one or simultaneously. As one is led into holding beliefs that they initially regarded as very unlikely, they become more confident that the catch-all is true. One way to flesh this out is to think that

anytime we come to believe a proposition we thought was very unlikely, that offers a degree of confirmation for the catch-all. H_c does not specifically predict observation E_5 that the inspector will observe five positive test results—not until she assigns a low prior to E_5. H_c predicts that she will eventually observe some evidence that she thought was very unlikely, and after she assigns a low prior to E_5, it fits the bill. H_c also predicts that she will eventually come to believe in a hypothesis she thought was very unlikely, and her low prior in the hypothesis H_5 that all five dishes are infected fits that prediction. Being forced to believe unlikely things confirms the catch-all, and in general, the more unlikely they are, the more believing them confirms it. This point is even clearer when we look past one-off events—after all, unlikely things do happen. The catch-all predicts we will become locked into a pattern of conspiratorial behavior where our predictions increasingly go awry and we consistently start believing stranger and stranger things to justify our mistaken predictions. It's like an epistemic circuit-breaker that tells us we're becoming conspiracy theorists and we ought to stop and reassess our assumptions.

Because the probabilities of regular hypotheses and the catch-all must add to 1, I don't think the inspector gains high credence that all five dishes are infected *and* simultaneously gains high credence that her catch-all is right. What happens, I think, is that first, for a moment, she gains credence near 1 in the hypothesis H_5 that all five dishes are infected. Then the observation she's believed a very unlikely hypothesis confirms the catch-all, and she then decides to transfer most of her high confidence from H_5 to H_c, though she still believes H_5 more than other regular hypotheses. She'd rather believe she missed a relevant hypothesis than that a one-in-a-million event has happened.

So this is my outline of how to assign a prior to the catch-all hypothesis and increase it upon encountering evidence, two tasks which others think can't be done. First, considerations

raised by Shimony and Earman support assigning a low initial credence to the catch-all. Next, that low credence rises as we start to witness improbable evidence and believe improbable regular hypotheses, and we transfer confidence from the improbable regular hypotheses we've come to believe to the catch-all. Then high credence in our catch-all makes us revise our partitions of hypotheses, following the principle of minimal change which I fleshed out at the end of the previous section. Once we're done, our new credal function will have lower credence in its catch-all.

The process of moving from old credal functions to new ones when revising our concepts is not as subjective as critics fear—following rules like "make the most minimal changes to your partition of hypotheses possible, try out the new credal function on the body of evidence that confounded the old one, then try out more radical changes if the new credal function is still confounded" provides some guidance about how to proceed. Doubts about formalizing conceptual learning seem to be inspired by Kuhn's model of scientific revolution, where proponents of radical new theories simply overpower the adherents to the old ones by persuading younger generations. My account of the process doesn't involve persuading anyone; the agent just runs credal functions defined over increasingly different algebras through their body of evidence until they arrive at a set of beliefs that isn't stumped by it.

5. Learning About New Hypotheses Directly

Cases like the detective example, where an agent is begrudgingly driven toward coming up with new hypotheses to explain their puzzling evidence, are actually not the most common kind of conceptual learning. I used that case to show how the bare bones of Bayesian machinery already supports having non-zero credence in a catch-all hypothesis, and to show how high

credence in that automatic catch-all hypothesis could compel an agent to come up with new hypotheses. But there's a far more ordinary type of case that ought to lead us to consider new hypotheses: being told them.

Back at the beginning of the paper I pointed out that when we learn about new theories like evolution by natural selection we don't learn them purely through conditioning. The detective in my case is actually doing a primitive version of what Darwin did, painstakingly formulating new hypotheses, rather than doing what the student in a biology class does and just learning about them directly. In a much simpler version of the detective case, instead of raising their credence in their automatic catch-all hypothesis after encountering a body of evidence that disconfirms each of their regular hypotheses, they could simply talk to another detective who says something like "what if there's an accountant who helped the butler do it?" Or maybe, like Earman suggests, no one tells the detective the new hypothesis but it just comes to them in a dream, like the chemist who came up with a hypothesis about the shape of benzene after dreaming of a snake eating its tail. We could even come up with new hypotheses with no real effort by mishearing or misunderstanding what someone says, or misremembering or misstating our own thoughts. Sometimes we painstakingly formulate new hypotheses like Einstein or Darwin did, but we often just stumble upon fully formed new hypotheses in various ways. Experience delivers them to us directly, and our task is just to figure out how to adjust our existing beliefs to fit them.

Whether we become consciously aware of new hypotheses through reading about them in a book, hearing about them in class, dreaming them up, or just misunderstanding someone, it's worth distinguishing cases where we're forced to invent new hypotheses from ones where experience of the world delivers them to us directly, already formed. One big difference between

the cases where we formulate new hypotheses and the ones where we learn about them directly is that in the former, compelling evidence against our old hypotheses drives us to come up with new ones, while in the latter, becoming consciously aware of the new hypotheses seems to provide us evidence that affects our old hypotheses. A second difference is that the question of what priors to assign new hypotheses is sometimes easier when experience delivers them to us directly, because the way we're told the new hypothesis often indicates the degree of confidence we ought to have in it. If a trusted authority tells us a new hypothesis is right, we probably ought to have high initial credence in it. If it came to us in a dream or we read it from a random person on Twitter, we probably ought to have low initial credence in it.

The automatic catch-all hypothesis every Bayesian learner ought to have can get the ball rolling on conceptual learning, but it isn't essential to the process. In this section I'll argue that if an agent just starts out with a little more than the bare-bones account says they must have and consciously expects that they might come across new hypotheses, they can easily create a sustainable system for learning them.

One interesting connection between the bare-bones way of introducing new hypotheses and this richer one is that the former will naturally lead to the latter—if an agent starts off without a conscious expectation that they'll learn about new hypotheses, but still has minimal but non-zero credence in the automatic catch-all that all their regular hypotheses are false, overwhelming evidence against each of their regular hypotheses should drive them into the laborious process of formulating new hypotheses. Like a child who touches a hot stove for the first time, they'll only need to learn this painful epistemic lesson that there might be relevant hypotheses they haven't considered once. From then on, they can avoid much of that trouble by adopting credal functions that include conscious expectations that they'll learn about new

hypotheses, though they very well might have low credence in that possibility. This means they can move past just having non-zero confidence in the automatic catch-all hypothesis formed by the negation of the disjunction of all their regular hypotheses and they can have a more specific catch-all hypothesis that makes positive claims, and they can have corresponding expectations about specific evidence confirming it.

For instance, if the detective originally had credence .0001 in "it's not the case that the gardener, butler, or maid did it," on their second attempt, after introducing the hypothesis that the accountant did it, the detective might be more cautious, since they've already been burned once, and adopt a credal function that contains a specific catch-all hypothesis that says "maybe there's some other employee I haven't considered as a suspect." They'd then also have to partition their potential evidence to include the possibility that they learn about another employee. Modifying one's partition of hypotheses requires modifying one's partition of observations to account for evidence confirming the new hypotheses; there's not much point to adding a new hypothesis if you can't recognize evidence that would confirm it.

This is a point that the growing awareness literature glosses over—discussion focuses on partitions of hypotheses, not observations. This may be because we assume that changes to our partition of observations will generally mirror the ones we make to our partition of hypotheses: if we refine our hypotheses to include the possibility of two gardeners, we will refine our potential observations to give evidence for the guilt of two separate gardeners. I will follow others in assuming that figuring out how to modify our partitions of hypotheses will give us insights that translate into modifying our partitions of observations, although I wouldn't be at all surprised if questions about how to modify partitions of observations raised unique problems worthy of separate papers. For instance, a new hypothesis might make some of the same predictions as an

old one, so one's new credal function ought to assign higher priors to those observations, all else equal, and they ought to lower their conditional probability for how much those observations confirm the old hypothesis, if it competes with the new one. This will be relevant in Section Six.

Anytime I refer to the *automatic* catch-all hypothesis I mean the most general one logically possible, the one automatically generated by an agent's algebra by having non-zero credence in the negation of the disjunction of all their other hypotheses. The automatic catch-all only makes a negative claim. The bare-bones account from the last section illustrates how an agent who starts out with only the automatic catch-all could easily be led to realize they should have a more specific one which makes a positive claim. This positive catch-all about hypotheses would say "there's at least one relevant hypothesis I haven't considered," and the automatic catch-all would say "it's not the case that any of my hypotheses are true." I'll call that general positive catch-all H_{c+}. So far in this section I've argued that it's not hard for a Bayesian agent to basically pull itself up by its bootstraps and move from the automatic catch-all's mere negative claim to one like H_{c+} that makes a positive one.[99] I'll argue a positive general catch-all like H_{c+} makes conceptual learning easier.

The most natural candidate for a piece of evidence that an agent can learn to incrementally confirm H_{c+}, increasing their total evidence for it, is not the particular new hypothesis they become aware of itself, but the fact that they have become aware of some new hypothesis. A hypothesis is not evidence, but the observation that we have become aware of a new hypothesis can be. For that observation to be evidence for any of the agent's existing hypotheses, including H_{c+}, they must have included the possibility they would encounter it in the

[99] Once an agent has a positive catch-all like H_{c+}, they'll still have the automatic catch-all but it won't be nearly as important, because now it will make the claim that all their other hypotheses including H_{c+} are false. So the automatic catch-all H_c would say that all their other hypotheses are false, including the hypothesis that there's at least one relevant hypothesis they haven't considered.

partition of potential pieces of evidence that they used to form an algebra of propositions in combination with their partition of hypotheses. That's why the particular details of the new hypothesis the agent acquires can't be part of the evidence they learn—since they were unaware of the details of the new hypothesis when they formed their initial credal function, it can't include prior probability assignments about possible data that refer to those details. If the new hypothesis is "Light behaves like a wave and a particle," for instance, it makes no sense to say that the agent had a non-zero prior credence in the potential evidence "At some point I become aware of the hypothesis that light behaves like a wave and a particle."

They can, however, have a prior probability that they will introduce *some* unspecified new hypothesis, and their initial credal function could account for that possibility by regarding the potential introduction of any new hypothesis as possible evidence. So the evidence E that they learn with certainty upon introducing a new hypothesis H can be "At some point in time I consider a relevant new hypothesis." They could also potentially refine observation E to account for different ways they come to consider the new hypothesis, some of which would be more credible than others. In that case, they would have multiple basic possible observations within E, such as "At some point in time I dream of a relevant new hypothesis" and "At some point in time a teacher tells me a relevant new hypothesis."

Instead of saying someone learns E when they become aware of a new hypothesis, it might be more accurate to say that the potential evidence they learn with certainty upon introducing the hypothesis "Light behaves like a wave and a particle" is "At some point in time I become aware of a relevant new hypothesis about the nature of light" or "At some point in time I become aware of a relevant new hypothesis about whether light is a wave or a particle." Our expectations about the evidence we could receive from experience can be more or less general,

and this is reflected by how coarsely or finely we draw the partition of possible states of total evidence we generate our algebra of propositions from, which depends on how coarsely or finely we draw our partition of hypotheses. If a high school physics student's partition of hypotheses after the first day of the section on light includes hypotheses like "Light is a wave," "Light is a particle," and H_{c+}, "There's at least one relevant hypothesis I haven't considered," creating an algebra will result in having hypotheses like "Light is neither a wave nor a particle and there's at least one relevant hypothesis I haven't considered." They'll have a correspondingly fine partition of possible evidence that expects they might receive evidence that light is neither a wave nor a particle and there's at least one relevant hypothesis they haven't considered. Then when they go to the next day of class and hear the new hypothesis that light is neither but behaves like both they can interpret that as evidence about whether light is a wave or a particle. But if a different student skipped the earlier class and didn't start out with the hypotheses "Light is a wave" and "Light is a particle," when they attend the later class and hear the new hypothesis that light behaves like a wave and a particle they can at best take that as evidence about the nature of light.

It's possible for someone to have all three of the expectations about possible evidence I mentioned, "At some point I consider a relevant new hypothesis," "At some point I consider a relevant new hypothesis about the nature of light," and "At some point I consider a relevant new hypothesis about whether light is a wave or a particle." If they did, their prior probability for each of the three statements should be equal or higher the more general the statement is. Any time we become aware of a new hypothesis about whether light is a wave or a particle we become aware of a new hypothesis about the nature of light, but not vice versa, and any time we become aware of a new hypothesis about the nature of light we become aware of a new hypothesis, but not vice versa.

The agent's initial estimate of the probability they will acquire the evidence "At some point in time I become aware of a relevant new hypothesis" might be very low, if their prior probability in the positive catch-all hypothesis "There's at least one relevant hypothesis I haven't considered" is low. But being told a new hypothesis is an experience that arguably requires an agent to become certain in the observation that they have become aware of a new hypothesis. We have direct access to our own awareness, so it's hard to be wrong about whether we've become aware of a new hypothesis. Still, I can imagine someone thinking they've become aware of a new hypothesis and being wrong, perhaps because they didn't realize it amounted to a restatement of an old hypothesis, or because it ended up being contradictory or unintelligible upon careful consideration. The main reason I discuss becoming certain in E and using simple conditionalization to raise credence in H_{c+} in this section is that for many purposes modeling a credence close to 1 as 1 will be good enough, and if the difference is relevant for some reason we can just use Jeffrey conditionalization instead.

Jeffrey conditionalization generalizes simple conditionalization to apply to all cases where an agent changes their degrees of belief in an evidential statement and its negation from initial probabilities between 0 and 1 to new ones. According to the principle of Jeffrey conditionalization, when there's some body of evidence E and some hypothesis H which the agent assigns prior probabilities between 0 and 1 to, an observation that causes a non-inferential change in the probability of E should result in changing the probability of H in the following way: $P_f(H) = P_i(H|E) \times P_f(E) + P_i(H|{\sim}E) \times P_f({\sim}E)$. $P_f({\sim}E)$ can also be described as $(1 - P_f(E))$. When $P_f(E)$ is 1 and $P_f({\sim}E)$ is therefore 0, Jeffrey conditionalization reduces to simple conditionalization.[100] So in some context where we wanted to account for the possibility that a

[100] Jeffrey, Richard, *The Logic of Decision*, 2nd ed., Chicago: University of Chicago Press, 1983.

seemingly new hypothesis someone encountered might not really be one, we could say that the agent directly learns the truth of the potential evidence E "At some point I consider a relevant new hypothesis" with credence .99 or some other credence close to 1 and then uses Jeffrey conditioning to raise credence in H_{c+}.

For conditioning on E to result in a posterior probability in H_{c+} higher than the agent's prior probability in it H_{c+} must predict E, so the catch-all hypothesis "There's at least one relevant hypothesis I haven't considered" must predict the evidence "At some point in time I consider a relevant new hypothesis." The degree to which two things predict each other comes down to the degree to which they're more likely to occur together than they are to occur separately. An agent's probability of eventually considering a relevant new hypothesis given that there's at least one relevant hypothesis they haven't considered, $P(E|H_{c+})$, generally ought to be higher than their unconditional probability that they'll eventually become aware of a relevant new hypothesis, P(E), since assuming there *is* a relevant hypothesis they haven't considered should make it strictly more probable that they'll eventually consider it. If there were no relevant new hypothesis for them to consider they definitely wouldn't ever consider one, while if there were one, they might. Additionally, the fact that there's a *relevant* hypothesis they haven't considered makes it likelier that their current credal function will encounter evidence that can't be explained by any of their non-catchall hypotheses, which would raise credence in H_{c+}. So $P(E|H_{c+})$ ought to be higher than P(E) for most people, barring unusual cases like an agent who knows they're about to die or be lobotomized.

$P(H_{c+}|E)$ clearly ought to be higher than $P(H_{c+})$. The probability that there's at least one relevant hypothesis the agent hasn't considered in their current credal function given that they eventually become aware of a relevant new hypothesis ought to be higher than their prior

probability that there's at least one relevant hypothesis they haven't considered in their current credal function; it should be 1. The positive catch-all hypothesis H_{c+} does predict the evidence E that one will eventually become aware of a new hypothesis, so learning of the existence of a new hypothesis can provide incremental confirmation of H_{c+}, justifying raising credence in it to reflect greater total evidence for it.

The fact that H_{c+} says "there's *at least* one relevant hypothesis I haven't considered" is important. Upon becoming aware of a new hypothesis H, depending on what it is, we might have more reason than before to think that there's some unknown *second* new hypothesis H' that we hadn't considered, and still haven't considered because we can't specify it. We can acknowledge that possibility by assigning a higher prior than we otherwise would to our new credal function's positive catch-all hypothesis after we create a new algebra that includes H. I'll call the new credal function's positive catch-all $H_{c+'}$.

Whether introducing H gives us additional reason to expect some unknown H' depends on the particular hypothesis H represents—becoming aware of the theory of special relativity seems to give us significant reason to expect some additional new hypothesis is out there, while becoming aware of evolution by natural selection doesn't seem to do so nearly as much. In many cases, like when a doctor realizes a possible diagnosis no one's thought of yet, or a detective thinks of a new suspect in a crime, becoming aware of a new hypothesis can sometimes make someone become less confident that there's some additional new hypothesis H' that they also missed. When we're very confident in the new hypothesis we learn of and it resolves mysteries without raising additional ones, that will justify assigning a lower prior than we otherwise would to our new catch-all, $H_{c+'}$.

A new hypothesis could suggest additional new hypotheses will eventually arise, instead of suggesting no more will be necessary, if it seemed well-supported but incomplete, or contradicted another hypothesis which seemed just as plausible, suggesting the need for some unifying theory to be discovered. Or you could just directly learn of new hypothesis H in a way that makes you think the same source could inform you of other new hypotheses. Here's a simple example: suppose a trusted source tells you "Off the top of my head, here's one possibility you haven't considered," and the way they say it makes you suspect that if they thought for a moment, they'd be able to come up with a second. Or suppose you're reading a textbook with ten chapters, you get halfway through, and you realize you've learned of one relevant new hypothesis in each chapter.

6. How New Hypotheses Change the Probabilities of Old Ones

Most of this paper has focused on the role of some version of a catch-all hypothesis in explaining how we could come up with a new hypothesis or learn of one. In Section Three, I provided a roadmap for how we might follow a principle of minimal change to revise our partitions of hypotheses, determine priors for new hypotheses, and modify probability ratios between old ones. The general idea is that as we try to adopt a credal function that minimizes credence in its catch-all, our old partition of hypotheses constrains us greatly, our old probability ratios between them constrain us somewhat, and we're hardly constrained at all when it comes to assigning priors to the new hypotheses we create, although when we're directly told new hypotheses our method of learning of them can influence the prior we should have. In conceptual learning, we try to solve an optimization problem where we adopt the nearest variation of our

original credal function that manages to account for the evidence that stumped the original.[101] This procedure of minimal change is what allows us to trust that our old learning by conditioning has not been wasted and that much of our original plan to learn by conditioning remains valid.

Although this is the broad outline of my theory, I have said relatively little about exactly *how* we adjust probability ratios between old hypotheses or select priors for new ones. That process revolves around creating new conditional probabilities between new observations and old hypotheses, between new observations and new hypotheses, and between old observations and new hypotheses. Selecting new conditional probabilities and selecting priors for new observations and old ones are the mechanisms by which we achieve the general goals of adjusting old probability ratios and setting priors for new hypotheses. What I've outlined is a *strategy* for conceptual learning; now I turn to the tactics.

This may be the most complex question of all, and though I have some answers about how the process works, I doubt I have all of them; it's a broad topic. I'll begin with an example.

Suppose Darwin is visiting a new island just before he's formulated the hypothesis that species evolve through natural selection. I'll use the name Darwin in this example, but I could just as well consider some biology student who visits a new island just before hearing about Darwin's theory—my focus in this section is not on the formulation of the theory of evolution by natural selection, but what to do with the rest of one's beliefs after considering it. Suppose Darwin's only seen the birds on the forested half of the island, and observed most of them have long beaks. He knows the other half of the island is unforested, but doesn't think that's relevant to the length of the birds' beaks, and is pretty sure most of the birds there have long beaks too,

[101] Interestingly, solving optimization problems like this is something machine learning can often accomplish better than humans. Once we frame conceptual learning as a particular kind of optimization problem, it moves into the territory of something artificial intelligence could potentially accomplish well rather than a task that seems hopeless for it.

because most other conditions on the two halves of the island are the same. Suppose he starts out with probability .8 that most of the birds on the unforested side have long beaks and probability .2 that they don't. Then he finishes formulating the hypothesis H_n that species evolve through natural selection. Not only must Darwin now add H_n to his partition of hypotheses and come up with an initial probability for it, he must reconsider his opinion about the probability most of the birds he'll observe on the unforested side of the island will have long beaks.

Suppose he concludes, in light of the hypothesis of evolution by natural selection, that forested environments favor long beaks because they help birds open the fruit on the trees. He should now think it's less probable than before that he'll encounter the data that the birds on the unforested side of the island have long beaks, which I'll call E_b. Introducing H_n and finding it probable should make Darwin change his estimate of the probability that he'll acquire E_b, which can in turn change his estimate of the probabilities of various hypotheses. Suppose his new probability in E_b is .3 and his probability in its negation is .7. If he had some hypothesis H_d that predicted the birds on the unforested side of the island had long beaks, finding that data less probable in light of evolution by natural selection should result in thinking the hypothesis that predicts it is less probable, holding everything else fixed. If H_d were "The birds on the unforested side of the island are recent descendants of the ones on the forested side," for instance, Darwin might have thought that H_d predicted the evidence E_b that the birds on the unforested side would mostly have long beaks. Now that introducing H_n has made him think E_b is less probable, he should also think H_d is less probable than he thought it was before considering H_n.

Last section I argued that when we become aware of a relevant new hypothesis we become certain of the evidential statement "At some point I consider a relevant new hypothesis," which I'll now call E_h. That certainty allows us to use simple conditionalization to update the

catch-all hypothesis H_{c+} based on the evidence. That evidence had great import for H_{c+}, leading to a posterior probability of 1 in it, but told us little about what to think of the probabilities of non-catch-all hypotheses. For instance, Darwin's probability for the hypothesis that the birds on the unforested side are descendants of the ones on the forested side given that at some point he considers some new hypothesis shouldn't be much different from his unconditional probability for the hypothesis that the birds on the unforested side are recent descendants of the ones on the forested one; $P(H_d|E_h)$ is roughly equal to $P(H_d)$. Assuming only that he becomes aware of some unspecified new hypothesis should do little to change his confidence that the birds on the unforested side descend from the ones on the forested one, although it's highly relevant evidence to the catch-all hypothesis H_{c+}. That means changes in $P(H_d)$ upon becoming aware of H_n must be explained by changes in the probability of some evidential statement other than E_h. In my example, Darwin did become certain in E_h upon considering H_n, assuming it was contained in his partition of possible states of evidence, but the more relevant fact was that he also changed his assessment of the probability of E_b, that most of the birds on the unforested side have long beaks. Considering H_n and assigning high credence to it makes him change his assessment of $P(E_b)$. To update on changes in the probability of E_b and its negation, he would have to use Jeffrey conditionalization.

I think the process of going from considering H_n to changing $P(H_d)$ involves four basic steps: first, Darwin becomes certain in E_h and uses simple conditionalization to become confident in H_{c+}. Second, his high credence in H_{c+} tells him he needs to adopt a new credal function that modifies his old one to include H_n in its partition. I explained how these two steps might work in the last few sections, and I've argued that an agent would want to make the most minimal changes possible to its old partition of hypotheses and priors, so it would hold them

constant until something forced it to modify them. Now I introduce an important third step: when Darwin modifies his old credal function to include H_n, he's forced to come up with initial probability assignments for H_n and $P(E_b|H_n)$. His old credal function anticipated the possibility of H_n through its catch-all H_{c+}, but it couldn't possibly have contained any prior for H_n or conditional probabilities referring to it because he hadn't considered it yet. So even when making the most minimal changes possible he must come up with initial probability assignments for H_n and $P(E_b|H_n)$. This necessary modification will force him to have a different value for $P(E_b)$ than he did before he considered H_n. This leads to a fourth step, where his new value for $P(E_b)$ will then force him to rethink the probabilities of old hypotheses like H_d. To do this, he can take whichever conditional probabilities from his original credal function he's currently holding fixed, including $P(H_d|E_b)$, and update on the change in $P(E_b)$ using Jeffrey conditionalization. This is my basic picture of how he goes from considering H_n to changing $P(H_d)$. Now I'll elaborate on steps three and four.

When I speak of holding our original conditional probabilities fixed until something forces us to change them, I'm fleshing out my theory that Reverse Bayesianism should be treated as a suggestion rather than a rule. RB holds that probability ratios between our old hypotheses should be strictly maintained in our new credal function. Presumably, it also holds that probability ratios between old observations should be held constant, although Reverse Bayesians speak of hypotheses far more often than they speak of observations. As I said earlier, it seems to me that RB is really a claim that either we should hold old conditional probabilities and probability assignments for old observations and hypotheses constant, or that any changes to those entities should affect the probabilities of old hypotheses symmetrically. As Roussos put it,

RB holds that for any maximally specific propositions A and B from our old credal function, P, our new credal function P+ ought to obey the following rule:

$$P(A)/P(B) = P+(A)/P+(B)$$

Simply put, although we now have ample counterexamples to RB, my proposal is that for any two of our old basic possibilities, whether hypotheses or observations, we adhere to RB and our old conditional probabilities until we are forced to depart from them to respond to new hypotheses, new observations, priors for them, and new conditional probabilities involving them. We *have* to come up with those things. Probability assignments we create for these new terms will then ripple out into our beliefs through Jeffrey conditioning as we plug them into the old conditional probabilities we're holding constant. Creating necessary probability assignments for these new terms has an effect very similar to observing evidence.

When we introduce new hypotheses, they can entail changes to both our partition of possible observations and the probabilities of old observations. First I'll talk about the implications of creating new observations. When you refine your partition of hypotheses about possible movies to include ones like "The movie is a French comedy with simple language," and "The movie is a French comedy with complex language," you have to revise your possible observations to include evidence about whether the language is simple or complex. For instance, while before you might have imagined figuring out what the movie was by hearing people who saw it complaining "It wasn't very funny," or "It wasn't very exciting," now you might also imagine them complaining "The language was too hard to follow." You'd then assign a probability to that possible observation, which I'll call O_n.[102]

[102] Interestingly, in this kind of case, the probability for this new observation wouldn't necessarily be taken from probabilities for old observations. One important difference between revising partitions of hypotheses and revising partitions of observations is that the probability of all our hypotheses ought to sum to 1, but the probability of all our observations doesn't have to.

You also have to create conditional probabilities linking the new observation O_n to both the new hypotheses *and* the old ones. When a possible new observation would lend greater support to one old hypothesis than another, we have a likely counterexample to Reverse Bayesianism. Suppose, for instance, that the observation that you hear people complain the language was too hard to follow is more likely if the movie is a comedy than if it's a thriller, since jokes in one language often don't translate easily into another (assuming the foreign movies have English subtitles). Simple puns might be lost in translation, and sophisticated wordplay and irony are especially likely not to translate well.[103] If you think that's true, and you think the theater manager's decisions on which movies to show are influenced by a desire to avoid disappointing customers, that's a reason to assign a higher probability to $\sim O_n$ than O_n—you think the manager has taken steps to avoid customers complaining that the movie's language was too hard to follow, and that involves showing foreign comedies less frequently than other genres. This would justify assigning a low probability to the new hypotheses that the movie is a French or German comedy with complex language. But the new observation is relevant to all hypotheses about foreign comedies. The old hypothesis of a foreign comedy being shown weakly predicts the new observation O_n, and assigning a high prior to $\sim O_n$ then decreases confidence in that old hypothesis relative to other old hypotheses.

Specific examples aside, the general point here is that some old hypotheses very well might predict new observations better or worse than others; that's probably to be expected. When that's true the introduction of new observations predicted by new hypotheses, and the assignment of priors to them, will alter the probability ratios of at least some old hypotheses. Whenever a new hypothesis raises a new possible observation, it would be pretty surprising if every old

[103] Hoffman, Jascha. "Me Translate Funny One Day." The New York Times, The New York Times, 19 Oct. 2012, https://www.nytimes.com/2012/10/21/books/review/the-challenges-of-translating-humor.html.

hypothesis predicted that observation to the exact same degree. So the recognition of possible new observations, and assignment of probabilities to them and creation of conditional probabilities linking them to old hypotheses, has an effect similar to observing evidence.

Alternatively, a new hypothesis might not introduce any new observations; in that case, it necessarily must change the probabilities of old ones. When a new hypothesis makes the same prediction as an old one, our probability for the old observation increases, but the degree of confirmation it offers for the old hypothesis decreases. The fact that the new hypothesis predicts an old observation changes our old conditional probability about how much the occurrence of that observation supports the old hypothesis. This is what's going on in the food inspector case: the new hypothesis H_t "All five tests are contaminated" is created to explain the *existing* observation E_5 "All five dishes test positive for Salmonella," so it doesn't necessarily require the inspector to revise her partition of observations. The old hypothesis H_5 that all five dishes were infected predicted E_5, but the new hypothesis H_t also predicts it, and does so more strongly. In a credal function with only H_5, observing E_5 would therefore be strong evidence for H_5, but in one with both H_5 and H_t, observing E_5 would more strongly confirm H_t than H_5. When H_t is an option, $P(H_5|E_5)$ is lower than it otherwise would be. New hypotheses can change conditional probabilities involving old hypotheses and old observations.

So introducing new hypotheses can cause us to create new observations, change old conditional probabilities, and change our probability assignments for old observations, and each of these changes has implications for the probabilities of our old hypotheses when we combine them with all the beliefs we *haven't* changed. Perhaps the simplest of these is when a new hypothesis makes us find an old observation more or less likely. Once a new hypothesis has forced us to reevaluate our probability for an old observation, that simple change radiates

outward into our credal function as we take our new probability for the old observation and plug it into all the old conditional probabilities we have held constant. After our probability assignment for our old observation E_1 changes, we plug that into our old conditional probability $P(H_1|E_1)$ and use Jeffrey conditioning to arrive at a new value for $P(H_1)$. Then our new value for $P(H_1)$, combined with our old conditional probability $P(E_2|H_1)$, through Jeffrey conditioning delivers us a new value for $P(E_2)$. Then our new value for $P(E_2)$, combined with our old conditional probability $P(H_2|E_2)$, delivers us a new value for $P(H_2)$. And so on until there are no changes left to make. Following this procedure could very easily change the ratio of $P(H_1)$ to $P(H_2)$. The moment a new hypothesis changes the probability of an old observation, Reverse Bayesianism is on shaky ground.

Jeffrey and simple conditionalization are the inferential stages of Bayesian confirmation that are generated from a first stage involving non-inferential changes in probabilities of evidential statements and their negations. Those non-inferential changes are supposed to be the products of observation, direct experience of the world that gives us data about it. Our conditional probabilities outline how we should change our probabilities in hypotheses upon interacting with the world to find out through experience which possible states of total evidence are actual. Although Darwin directly observes the data that he becomes aware of some new hypothesis when he formulates H_n, he doesn't seem to directly observe the data relevant to E_b that it's less probable than before that the birds on the unforested side of the island have long beaks. This change in his probability of E_b is the product of inference about E_b's new probability conditional on H_n, combined with his unconditional probability in H_n. But he didn't have a prior about H_n or $P(E_b|H_n)$. I'll argue that when he assigns an initial probability to H_n and recognizes

that H_n makes E_b less probable than he originally thought it's as if he's observed evidence against E_b.

When experience of the world directly gave him confidence in H_n and he understood it made E_b less probable than he thought, experience changed $P(E_b)$. It's not that *considering* H_n or becoming *aware* of it directly changes $P(E_b)$; it's that *assigning a high initial probability* to H_n directly changes $P(E_b)$. This is even clearer in cases where confidence in a new hypothesis makes evidence that used to seem possible now seem impossible (or vice versa); I give an example on the next page. Nothing in the priors in Darwin's original credal function gave him high initial credence in H_n. That could only have come from experience, and so if high credence in H_n lowers his credence in E_b, that change in $P(E_b)$ could only have come from experience, not his priors.

In this case, Darwin's initial probability assignment to H_n should be high, perhaps even close to 1, considering that he carefully formulated the hypothesis that species evolve through natural selection to best account for his total evidence,[104] so the question of what he should now think of $P(E_b)$ comes down mostly to what he thinks $P(E_b|H_n)$ should be, although he could multiply that by $P(H_n)$ and add $(P(E_b|\sim H_n)*P(\sim H_n))$ to account for the possibility his new hypothesis of evolution by natural selection might be wrong.[105] Instead of directly observing data that causes a non-inferential shift in $P(E_b)$, he is forced to infer a new value for $P(E_b)$ when his addition of H_n to his credal function's algebra forces him to come up with assignments for $P(H_n)$

[104] Or, if we replace Darwin with the biology student who visits the island then hears about his theory, the student could have a high initial probability assignment for Hn because their teacher assures them it's the best account of all the evidence.

[105] $P(E_b|\sim H_n)$ is a somewhat murky term: how should we know what the probability of E_b is given that H_n is false? Taking inspiration from the Reverse Bayesians, we could start with the assumption that $P(E_b|\sim H_n)$ is pretty close to the probability our old credal function assigned E_b before we'd ever considered H_n.

and $P(E_b|H_n)$. If he were certain of H_n and it entailed that some evidential statement E_x were impossible, for instance, his assignment for $P(E_x)$ should change from whatever it was to 0.

Consider, for instance, a devout 1800s priest who has credence .9 in Lamarck's theory of evolution by acquired characteristics (H_a) and has credence .1 in the possible observation E_a that if they pray about it God will tell them H_a is false. Then suppose the priest hears about Darwin's new theory and becomes certain that H_n is right. Just hearing about the theory of evolution by natural selection won't necessarily raise their credence that if they pray God will tell them Lamarck is wrong, but if they assign initial credence 1 to H_n, they should simultaneously conclude it's certain God will tell them Lamarck's theory is false if they pray about it; they create the initial probability assignment $P(E_a|H_n)=1$. So when the priest assigns credence 1 to H_n, they also become certain E_a is true. This change in $P(E_a)$ doesn't come from the priest's priors in their original credal function; it comes from experience revealing H_n and $P(E_a|H_n)$ to them and giving them credence 1 in H_n. The devout priest started out with credence .9 in Lamarck's theory H_a, but suppose they also had in their original credal function a conditional probability saying that if they pray to God and he tells them H_a is false (observation E_a), they should lower credence in H_a to 0; $P(H_a|Ea)=0$. Even after the priest moves from their old credal function to their new one that includes H_n they *still* ought to believe $P(H_a|E_a)=0$. They just originally thought it was very unlikely God would tell them H_a was false, so they had a low prior probability in observing E_a. But now considering H_n and assigning credence 1 to it makes them certain in E_a, so they can combine that with their old probability $P(H_a|E_a)$, use simple conditionalization, and conclude $P(H_a)=0$. Lamarck is wrong.

In the birds case, Darwin's initial assignments for $P(H_n)$ and $P(E_b|H_n)$ entail changes in $P(E_b)$. As with acquiring perceptual data or expert testimony, the reason he changes $P(E_b)$ is that

he comes to possess information he didn't have when he defined his original credal function, although this information comes in the form of a hypothesis he didn't know when he defined his original credal function instead of evidence he didn't know at the time. To compare the two ways of changing $P(E_b)$, and see whether there's any difference in the way he should adjust $P(H_d)$ to respond to them, we can compare my first case to a new one where Darwin uses direct observation to support increasing $P(E_b)$ and then uses Jeffrey conditionalization to find a posterior probability for H_d, $P_{final}(H_d)$. Suppose, for instance, that he goes to a canyon separating the two sides of the island and sees large flocks of a variety of different birds on the unforested side, and they generally appear to have short beaks, although he isn't sure the beaks are actually short since he's viewing the birds from across the canyon and doesn't fully trust his vision or sense of scale at that distance. Because of that, he thinks there's now a probability of about .3 that the birds on the unforested side of the island have long beaks and .7 that they don't, where before going to the canyon he thought there was a probability of .8 that most of the birds on the unforested side had long beaks and .2 that they didn't. His observation makes him less confident in E_b and more confident in $\sim E_b$. This is an ordinary case where perceptual data makes him increase his confidence in an evidential statement but, due to its fallibility, doesn't allow him to become certain of it, so Darwin should update his credal function using Jeffrey conditionalization instead of simple conditionalization.

In the original case, he becomes less confident that the birds on the unforested side have long beaks without ever going there to look at them. Instead the decrease in $P(E_b)$ and increase in $P(\sim E_b)$ come from his initial assignments for $P(H_n)$ and $P(E_b|H_n)$, because he's confident in the new hypothesis that species evolve through natural selection and thinks that if that's the case, the probability that the birds on the unseen unforested side have long beaks is lower than he thought;

$P_{initial}(E_b|H_n) < P_{initial}(E_b)$ and $P_{initial}(\sim E_b|H_n) > P_{initial}(\sim E_b)$. The reason Darwin reached this conclusion was that before formulating H_n he thought that the data E_f that the birds on the forested side mostly had long beaks was good reason to expect the data E_b that the birds on the unforested side did as well, since the two sides had similar conditions other than their degree of forestation and he didn't think that wasn't relevant. Knowing that species evolve new traits by having those traits selected for by their environments, and realizing that long beaks help birds open the fruit on trees, makes him realize that his observation E_f of long beaks on the forested side was partly explained by the forests there. Understanding this should then make him think the fact that the birds on the forested side have long beaks doesn't as strongly imply that the ones on the unforested side do, so observing E_f doesn't give him as much reason as he thought to expect to observe E_b. It could still give him some reason to expect the birds on the other side to have long beaks too, since the forests might not be the only important feature of the environment to select for them and the two sides of the island have similar conditions other than forestation.

Although Darwin lowers $P(E_b)$ and raises $P(\sim E_b)$ upon adding initial probability assignments for $P(H_n)$ and $P(E_b|H_n)$ to his credal function without actually going to the canyon and seeing that most of the birds there don't have long beaks, he now expects that if he did go, it's less probable than he originally thought it was that he would observe that most of the birds had long beaks. Consequently, since H_d, the hypothesis that the birds on the unforested side recently descend from the ones on the forested one, predicts (to some degree) that they'll have long beaks, his increased confidence that they don't disconfirms H_d, justifying lower confidence in it. In a normal case of Jeffrey conditioning he observes that the birds on the unforested side seem to have short beaks; in the case of conceptual learning he observes a new hypothesis which makes him more confident he'd observe they seemed to have short beaks if he bothered to look.

One way to think of this move from believing H_n to disbelieving H_d is that considering the implications of evidence in light of the hypothesis of evolution by natural selection ultimately made Darwin reassess how probable H_d was given E_f; the question I've been answering is how it did so. At first, Darwin thought the observation that the birds on the forested side had long beaks had no significant impact on the probability that the ones on the unforested side recently descended from them; $P(H_d|E_f)$ was roughly equal to $P(H_d)$. Formulating H_n makes him see the significance of forests to beak length, so he should now think $P(H_d|E_f)$ is lower than $P(H_d)$. My account of how he goes from considering H_n to changing $P(H_d)$ revolves around how assigning high confidence to H_n made him change his belief in E_b; this lowering of $P(E_b)$ then has ripple effects for the rest of his beliefs, especially when combined with some beliefs that *didn't* change.

Since he still believes offspring tend to inherit the characteristics of their parents, observing long beaks on the birds on the forested side means that if the ones on the unforested side recently descended from them, they probably have long beaks. Even after introducing H_n and believing it, and becoming less confident the birds on the unforested side have long beaks, he should still remain confident that if the birds there recently descend from the ones on the forested side they're likely to have long beaks—the belief that children tend to inherit the characteristics of their parents persists through his old credal function and the new one. What has changed is that high credence in H_n has made him less confident the birds on the unforested side have long beaks; he in turn becomes less confident they descended from the ones on the forested side the moment he observes those ones have long beaks, or the moment he even recalls observing it. Since believing H_n gives him reason he didn't have before to think the birds on the unforested side don't have long beaks, the observation that the ones on the forested side do have

long beaks gives him more evidence than he originally thought that the ones on the other side aren't their recent descendants.

7. Conclusion

In Section One of this paper I raised two distinct but related questions: first, how do we go from being entirely unaware from a hypothesis to conceiving it and having some belief in it? Second, how does believing in a new hypothesis end up changing our beliefs about old ones? My goal was to show that a Bayesian has considerably more resources to answer these questions than one might expect, so the project of explaining conceptual learning using the same basic elements as conditioning isn't hopeless. The task is also to explain how one can maintain a long-term plan to learn through conditioning that isn't derailed when conceptual learning interrupts.

I started out by showing that a Bayesian agent already has the raw material to anticipate the possibility of new hypotheses by using what I called the automatic catch-all hypothesis. I argued that if an agent merely has non-zero credence in the hypothesis that all their other hypotheses are wrong, experience can make them confident in that automatic catch-all by disconfirming each of their regular hypotheses. I then argued that this newfound confidence in the automatic catch-all would force the agent into the laborious process of redrawing their partition of hypotheses, and I suggested they would attempt to make the most minimal changes to their credal function that would account for their confounding evidence. I argued that there are three main variables to be adjusted during this process: one's algebra, their probability ratios between hypotheses, and their priors for new propositions. I argued that in general, making the most minimal change that accounts for the confounding evidence requires treating one's original algebra as a strong constraint and their original probability ratios as a weaker constraint, and that

they are largely free to select whichever priors for new propositions best lower credence in their catch-all. Then I argued that it isn't as impossible a task as some have thought to assign a credence to the catch-all, and I gave a theory for how we might have a fuller version of it that makes it easier to learn of new hypotheses.

I ended with an account of how assigning initial confidence to a new hypothesis has ripple effects that ought to lead an agent to change their expectations that they'll eventually observe some evidence, and I argued that this in turn would make them change their expectations about how likely their old hypotheses were. This account showed how learning new hypotheses can have the same kind of effect that observing evidence does during conditioning. In conceptual learning, we observe new hypotheses that make predictions about new and old possible observations, changing our overall assessment about what evidence we'll encounter. We can use Jeffrey conditioning on those new probabilities for observations. Revising one's prediction of what evidence they'll observe has an effect very similar to observing evidence.

Taken together, although these arguments don't amount to a complete explanation of how conceptual learning occurs, such a goal seems too ambitious for one paper. The biggest area for exploration involves the general topic I touched upon toward the end of the paper: the way new hypotheses change not our partition of hypotheses, but our partition of observations, and the way the predictions of new hypotheses make us assign probabilities to new observations, adjust probabilities of old ones, and revise conditional probabilities linking observations to hypotheses. I have illustrated a variety of different ways this can occur and raised relevant issues; for instance, the probabilities of possible observations don't have to sum to one. The topic is broad enough that I don't think I've given a complete account of it. But there's no doubt that both conditioning and conceptual learning frequently occur within the same agent; my goal has been

to show how the two systems can be linked in many ways, to make it more plausible that they work together instead of separately. One underappreciated connection between the two is that conceptual learning leads to conditioning.

BIBLIOGRAPHY

Bradley, Richard. *Decision Theory with a Human Face.* Cambridge University Press, 2017.

Earman, John. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* MIT Press, 1996

Jeffrey, Richard, *The Logic of Decision*, 2nd ed., Chicago: University of Chicago Press, 1983.

Karni, Edi, and Marie-Louise Vierø. "'Reverse Bayesianism': A Choice-Based Theory of Growing Awareness." *American Economic Review*, vol. 103, no. 7, 2013

Mahtani, Anna. "Awareness Growth and Dispositional Attitudes." *Synthese*, vol. 198, no. 9, 2020, pp. 8981–8997.

Roussos, Joe. "Awareness Growth and Belief Revision." 2021, https://doi.org/10.31235/osf.io/bfv56

Shimony, Abner. Scientific inference. In R.G. Colodny ed., *The nature and function of scientific theories*. Pittsburgh: University of Pittsburgh

Steele, Katie, and H. Orri Stefánsson. "Belief Revision for Growing Awareness." *Mind*, vol. 130, no. 520, 2020, pp. 1207–1232.

Van Fraassen, Bas C. *Laws and Symmetry*. Clarendon, 1989.

Wenmackers, Sylvia, and Jan-Willem Romeijn. "New Theory about Old Evidence." *Synthese*, vol. 193, no. 4, 2015, p. 1236.