NEW BOUNDARY LINES

Alejandro Pérez Carballo
University of Massachusetts, Amherst

According to Bayesian orthodoxy, a rational agent's epistemic state is essentially a probability function—her *credence* function. At any point in time, an agent's credence function is defined over a *hypothesis space*—typically a Boolean algebra of propositions or events. As an agent receives new evidence, her credence function is updated by *conditionalizing* on her evidence in the following sense: if an agent's credence function at a given point is C and she receives evidence e, then (assuming $C(e) \neq 0$) her updated probability function C' is set to:

$$C'(x) = C(x \mid e) = \frac{C(xe)}{C(e)}.$$

Typically, conditionalization is taken to be the only rational epistemic change. And this entails that a rational agent's hypothesis space will remain constant throughout time. This is because for any e, if $C(p)$ is not well-defined, neither is $C(p \mid e)$.[1]

1 The same holds of Jeffrey-conditionalization (Jeffrey 1983), for exactly the same reasons. Matters are less straightforward if we take conditional probabilities to be primitive—as in e.g. Hájek 2003, Popper 1959, Rényi 1970. For if we allow $C(x \mid e)$ to be well-defined even when the ratio is not, we can have $C(p \mid e)$ be well-defined even when $C(p)$ is not. But in any case, no 'radical' changes in the hypothesis space are allowed. More precisely, let the *generalized hypothesis space* consist of those propositions x for which $C(x \mid y)$ is well-defined for some y that is logically independent of x. Even if we allow for primitive conditional probabilities, it is a consequence of Bayesian orthodoxy that an agent's generalized hypothesis space will remain constant throughout time. Furthermore, as will become clear later on, while capturing the phenomenon of rational conceptual change requires the possibility of changes in the hypothesis space, it is not obviously sufficient. The kind of change we will be interested in is rarely (if at all) triggered by new evidence.

Is this a feature or a bug?

Surely, some important changes in our epistemic states are left out of the orthodox picture. Typical examples involve the introduction of new theories, usually via the introduction of new concepts. In acquiring the concept of a gene, say, we enlarged our hypothesis space—e.g. we introduced new competing hypotheses to explain inheritance patterns of certain phenotypic traits.

Still, one might think such changes fall outside of the scope of a theory of epistemic rationality. After all, epistemic rationality arguably requires a certain amount of immodesty: if you are rational, then you should take yourself to be doing as well as you can, *epistemically*, given your available evidence. And it certainly seems as if changing your epistemic state without new evidence would involve moving to an epistemic state that is *worse*, by your own lights, than your current state. So it is tempting to conclude that a change in one's hypothesis space—at least when it is not the result of acquiring new evidence—can never be epistemically rational. (A good inquirer might be required to *act* in certain ways—perhaps she will be required not to shield herself from relevant evidence. Yet as far as what her epistemic state should be, one might think, the only sensible thing to say is whatever orthodoxy has to say.)

But one can consistently maintain that epistemic rationality is immodest and that some rational epistemic changes need not be triggered by new evidence. To see what I have in mind, consider the following analogy.

Lars is a *fun-maximizer*. At any point in time, he always takes himself to be doing the most fun thing he can do. While having lunch, Lars' perspective on the world is quite striking: I could be doing a number of things right now, but nothing would be as fun as having lunch. Of course, Lars is not stuck having lunch all day long. As he eats, his evidence changes—he acquires evidence that he is no longer hungry. He thus proceeds to do what he takes to be the most fun activity he could do, in light of his new evidence: sit quietly under an oak tree.

Now, if you looked at the specific choices Lars makes throughout his life, you would find them incredibly boring—imagine he spends his waking hours alternating between having lunch and sitting quietly under an oak tree. But this is not because his preferences are much different from yours. Rather, the problem is that Lars lacks imagination. If you could only get him to see that there are many things he could do with his day other than spend it quietly under an oak tree, you know he would

be thankful. And this can be so even though Lars is doing as well as he can in order to have fun. Among all the options that ever occur to him as things he could be doing at any particular time, he always takes himself to be doing the one he finds most fun. But this is not because he is always having that much fun: he is just not able to imagine all the fun things he could do instead.

Now consider an agent, Tom, who always takes himself to be doing as well as he can, epistemically. He only has beliefs about a particular issue: whether he is tired. He is very good at responding to the evidence he receives, and at any time, he takes himself to be doing as well as he can with regards to the issue of whether he is tired. To some extent, Tom is doing quite well, epistemically. But he could be doing much better: he could be asking questions about many things, including issues that have little to do with how tired he is.

If Tom lacks imagination, he could take himself to be doing as well as he can, epistemically, because he only considers a limited range of options. If he came to see that he could be in an epistemic state he had not considered, he would pick it in a heartbeat. So if a new issue occurred to him, he could in principle come to have a view on the matter without having acquired any new evidence. And crucially, without taking his earlier self to have been at fault.

We can think of the introduction of new concepts as an increase in our capacity for *epistemic* imagination. Once new concepts are introduced, epistemic states that were not previously available to us become available. The question arises as to how this affects what epistemic state we ought to be in: is it ever rational for us to move to a new epistemic state just because new concepts have been introduced?

This a type of epistemic change that orthodox Bayesian theories of epistemic rationality have little to say about.[2] Indeed, the question cannot even be formulated within the orthodox picture. Here I want to sketch a way of thinking about rational epistemic change that gives us the flexibility

2 Let me flag here an important distinction, one we will return to, between two distinct if related questions. First, assuming new concepts will be introduced, what is the rational way of distributing credence among the resulting set of propositions? Second, assuming new concepts could be introduced, when is it rational to do so? My concern here will mainly be with the second of these two questions. The first, of course, is one I will need to have something to say about. But I will remain largely neutral on what the best way to answer it is. See e.g. Romeijn 2005, Williamson 2003 for discussion.

to ask that question. The hope is to show how epistemic rationality could constrain expansions of the range of hypothesis we rely on in inquiry.

There will be many moving parts. It will thus be helpful, before moving on, to give a detailed road map of what is to follow. I start (§1) by making explicit my main working assumptions. I then spell out in §2 a minimal way of modeling conceptual change within a Bayesian picture of our epistemic states. Next (§3), I situate the question of the rationality of conceptual change within the broadly decision-theoretic approach to epistemic rationality sometimes labeled 'epistemic utility theory'. Doing so requires extending existing the epistemic utility framework to allow for comparisons of epistemic states that assign credence to different sets of propositions. I then motivate a richer account of epistemic utility (§4), one that goes beyond accuracy considerations and incorporates explanatory considerations. I offer a particular framework for doing so, one where accuracy with respect to a proposition is valued in a way that is proportional to its explanatory worth. After motivating a particular way of understanding the explanatory worth of a given proposition, I offer in §5 a way of measuring explanatory worth so as to generate the weights needed to define epistemic utility functions that are sensitive to explanatory considerations. Before closing and listing a few issues left outstanding (§7), I turn back in §6 to some of the motivating toy examples and show how the strategy I develop can be used to vindicate some of our intuitive judgments about the epistemic gain that comes from certain instances of conceptual innovation.

1 METHODOLOGICAL ASSUMPTIONS

I will identify propositions with sets of possible worlds. I will speak interchangeably of the conjunction and the intersection of two propositions—similarly for the disjunction or union of two propositions, and the negation or the complement of a propositions. I will also speak interchangeably of a proposition entailing another and of the former being a subset of the latter.

I will be relying on a familiar, broadly Bayesian picture of our cognitive states. An epistemic agent, for our purposes, is essentially a *credence*

function—a function C that assigns real numbers to propositions, subject to the following constraints:[3]

> BAYESIAN CORE: C is a probability function defined over a finite Boolean algebra $\mathcal{B}_C$ of subsets of W, the collection of possible worlds. That is, C is a real-valued function whose domain is a collection of propositions that is closed under conjunction and negation, that satisfies the following constraints:
>
> a. $C(x) \geq 0$
> b. $C(W) = 1$.
> c. $C(x \vee y) = C(x) + C(y)$, whenever x and y are incompatible propositions (i.e., whenever their intersection is empty) in the domain of C.

Whenever C satisfies the constraints in BAYESIAN CORE, I will say that C is *probabilistically coherent*. I will often refer to the domain of C as its *hypothesis space*

Fix such a function C. Since $\mathcal{B}_C$ is finite, there is a unique subset π_C of $\mathcal{B}_C$ of *atoms*: for all $s \in \pi_C$ and $x \in \mathcal{B}_C$, either s entails x or s and x are incompatible. Any nonempty member of $\mathcal{B}_C$ can thus be written out as the disjunction of elements of π_C. I will call π_C the *state space* of C. I will sometimes refer to the atoms of $\mathcal{B}_C$ as *C-atoms*.

If C is defined over the entire collection of propositions, its state space will be the collection of singleton propositions, viz. propositions of the form $\{w\}$ for $w \in W$. But in general, π_C will be a *partition* of W—a collection of pairwise exclusive and jointly exhaustive subsets of W. When there is no risk of ambiguity, I will drop the subscripts and simply use π (and $\mathcal{B}$) to talk about the state space (resp. the domain) of C.

By design, $\mathcal{B}$ is the *Boolean closure* of π—the smallest Boolean algebra that contains all the members of π. For any $x \in \mathcal{B}$, we have

$$C(x) = \sum_{s \in \pi} C(sx) = \sum_{s \subseteq x} C(s).$$

3 Throughout, I will be assuming we are dealing with credence functions defined over a finite collection of proposition. As far as I can see, the bulk of what I say survives if we allow for credence functions defined over countably many propositions. The gain in generality, however, would come at the cost of simplicity. (I do not know what would happen if we took into account credence functions defined for uncountably many propositions.)

Thus, in order to specify a given credence function, all we need to know is what values it assigns to elements of its state space. The elements of π—the C-atoms—can be thus thought of as the possible states of the world that play the role of possible worlds for C.[4]

I will call C_e the credence function that results from updating C on evidence e. When $e \in \mathcal{B}_C$, I will assume:

BAYESIAN UPDATE: For $e, x \in \mathcal{B}$, if $C(e) \neq 0$,

$$C_e(x) = C(x \mid e) = \frac{C(xe)}{C(e)}$$

This completes the sketch of our background framework.

Now, our concern here is, ultimately, with the rationality of conceptual change. So before we can ask whether conceptual change can ever be epistemically rational, we need a way of thinking about conceptual change within this background picture of our cognitive economy.

2 MODELING CONCEPTUAL CHANGE IN A BAYESIAN FRAMEWORK

Concepts, we are often told, are 'constituents of thoughts'. This presupposes a certain picture of thoughts as having structure—thoughts are 'built out of' concepts, as it were. But even those who want to reject this presupposition can agree that, to the extent that concepts have a role to play in a theory of the mind, what concepts one has constrains what thoughts one can have. Conceptually impoverished creatures, we can say, can only entertain a limited range of propositions. And acquiring new concepts, among other things, increases the range of thoughts an agent can have.[5]

This minimal characterization of the role of concepts in our cognitive economy suggests a straightforward way of thinking about conceptual change in a Bayesian framework.[6]

4 Elements of the state space can be thought of as 'small worlds', in the terminology of Savage 1972.

5 Whether the relation between having a concept and having the ability to entertain the relevant thoughts is one of constitution or not is something I want to remain neutral on for the purposes of this paper.

6 For related discussion of how this approach can model talk of concepts in a broadly Bayesian picture, see Yalcin 2018, as well as [REDACTED FOR BLIND REVIEW].

2.1 *Epistemic transitions*

Think of an *epistemic transition* as a change from one credence function to another. A transition is *trivial* iff the initial and final credence functions are identical. An epistemic transition is an *update* if the initial credence function and the final credence function of the transition have the same domain.[7] An epistemic transition is an *expansion* if the domain of the initial credence function is a subset of the domain of the final credence function. A credence function C' is an *extension* of C iff C' is an expansion of C that is conservative in the following sense: for all x in the domain of C, $C'(x) = C(x)$ (see Figure 1). Other types of transitions are possible.

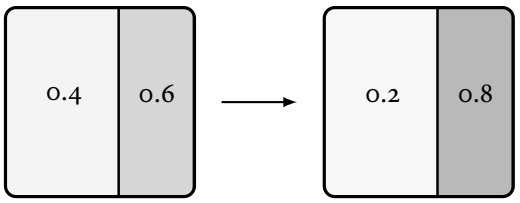
It is natural to think of the domain of an agent's credence function as the collection of propositions the agent can entertain.[8] To fix terminology, then, I will say that an epistemic transition is an instance of *conceptual change* only if the domain of the final credence function is distinct from of the domain of the initial credence function—if there are propositions in the domain of one of the two credence functions that are not in the domain of the other. So expansions, as I understand them, count as instances of conceptual change.

I should emphasize that this is not meant as an analysis of the intuitive notion of conceptual change. Instead, it is a modeling suggestion: if we model epistemic states using credence functions, we can model changes in an agent's conceptual resources as changes in the domain of her credence function.

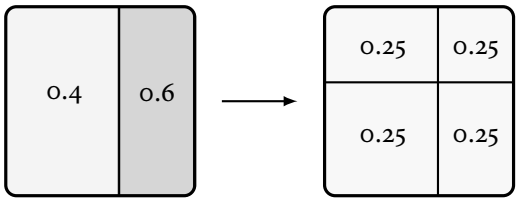
This modeling choice has the advantage of requiring a minimal adjustment to the orthodox Bayesian picture. Admittedly, it leaves many

7 Hence, not all updates satisfy the constraints in BAYESIAN UPDATE.

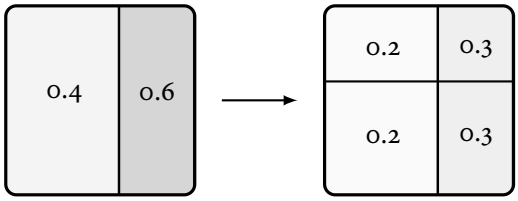
8 More precisely, my proposal is that we think of the domain of the agent's credence function There are other interpretations. Some have identified the domain of an agent's credence function (at a time t) as the collection of propositions the agent is attending to (at t). See e.g. Franke & de Jager 2011, Swanson 2006, Yalcin 2007. For our purposes here, however, I want to ignore the difference between the propositions an agent is attending to and those she is not.



(a) An update.



(b) An expansion



(c) An extension

Figure 1: Three kinds of epistemic transitions. Think of each square as representing the collection of possible worlds, partitioned along the lines of the state space of a credence function. We can use shades of gray to represent the amount of probability assigned to elements of the corresponding partition: the darker the cell, the more credence it gets. Updates can thus be represented by changes in the degree of gray in the cells of the partition. Expansions may also involve moving to a finer partition. Extensions are extensions that leave the distribution of probability over the state space of the prior function unchanged.

questions unanswered.[9] But the resulting picture is flexible enough to model some important features of conceptual change.

A couple of choice examples illustrate this point. (These examples will prove to be helpful later on, so I will go through them in a bit more detail than might seem necessary at this point.)

2.2 *Toy example 1: The Reds*

You are studying an unfamiliar type of organism, call them 'reds', and their reactions to certain stimuli.[10] You keep your reds inside dark boxes for a little while and then proceed to flash different colored lights on them to see how they react.

On Monday, you notice that some reds start moving faster when exposed to blue lights; others show no change in behavior when exposed to blue light. You also notice that some reds start moving faster after being exposed to red lights; others show no change in behavior when exposed to red light. To your surprise, you notice too that that most of the reds that move faster after exposure to red light are among those that move faster after being exposed to blue light, and that most of those that move faster after being exposed to blue light are among those that move faster after exposure to red light.

On Tuesday, it occurs to you that there could be an internal state R such that being in state R makes a red respond to blue and red light by moving faster than normal. Once you bring R into the picture, you can formulate a hypothesis that could be used to explain why some reds

9 For example, take an epistemic transition that, according to our terminology, counts as an instance of conceptual change. Which are the new concepts available in the final stage? Nothing in what I have said thus far allows us to answer that question.

To be sure, our modeling suggestion could be refined in many ways. For example, it may be that not all changes in the domain of an agent's credence function correspond to genuine instances of *conceptual* change. Perhaps only those transitions where the final credence function satisfies some kind of closure condition—say, something like the so-called generality constraint (Evans 1982)—should be classified as instances of conceptual change. Further elaboration of these ideas, however, is beyond the scope of this paper.

10 This example is based on a series of cases discussed in great detail in Sober 1998. See also Forster 1999, for related discussion of how conceptual innovation can be motivated by epistemic considerations.

respond to blue and red lights the way they do—e.g., that they, unlike the rest, are in state R.[11]

In order to model your epistemic state on Monday we need a credence function whose domain is the Boolean closure of propositions of the form: *c moves faster after exposure to blue light*, *c moves faster after exposure to red light*, *c's behavior is unaffected by exposure to blue light*, *c's behavior is unaffected by exposure to red light*.

To model your epistemic state on Tuesday, we need an expansion of your Monday credence function. We need a credence function whose domain includes the domain of your Monday credence function together with propositions of the form *c is in state R*. It is largely because of the addition of these new propositions that you can formulate the hypotheses that reds in state R move fast when exposed to red light and that a red in state R stops moving while exposed to blue light.

2.3 *Toy example 2: Cubes and spheres*

On Wednesday you discover some critters.[12] They come in two shapes, spheres and cubes. You soon realize that by squeezing two of them together, a new, slightly larger critter appears. You start gathering data about how the shape of the generated critter depends on the shape of the two you pressed against one another, as depicted on Table 1.

	Second parent	
First parent	*cube*	*sphere*
cube	cube	cube
sphere	cube	sphere

Table 1: Shape of second-generation critters as a function of parents' shape.

Once you press the new, slightly larger critters against one another, things get messier. Spheres pressed together yield even larger spheres. But for every other combination of the slightly larger critters you notice that the shape of the created critter depends on the shapes of the ones

11 I am not taking a stand on whether this is a *good* explanation. Whether this amounts to anything beyond a 'dormitive virtue' explanation is not relevant to our present purposes.

12 The example that follows is essentially due to Arntzenius 1995.

you pressed together and of the shapes of the ones *those* came from (see Table 2).

	Second parent		
First parent	*cube, cube parents*	*cube, mixed parents*	*sphere*
cube, cube parents	100% cube	100% cube	100% cube
cube, mixed parents	100% cube	75% cube	50% cube
sphere	100% cube	50% cube	100% sphere

Table 2: Shape of third-generation critters as a function of ancestors' shape.

Things get even messier when you try to gather data about third generation critters. (As it happens, squeezing together critters of different sizes does nothing to them.)

On Thursday morning it occurs to you that cube critters could be of two kinds, call them *pure* and *hybrid*. Once you factor in hybrid cubes, you can accommodate your data using the hypothesis in Table 3.

	Second parent		
First parent	*pure cube*	*hybrid cube*	*sphere*
pure cube	100% pure cube	50% pure cube 50% hybrid cube	100% hybrid cube
hybrid cube	50% pure cube 50% hybrid cube	25% pure cube 50% hybrid cube 25% sphere	50% hybrid cube 50% sphere
sphere	100% hybrid cube	50% hybrid cube 50% sphere	100% sphere

Table 3: A new hypothesis about the shape inheritance pattern.

To model your epistemic state on Wednesday, we need a credence function that is defined over the Boolean closure of all propositions of the form: *c is a cube*, *c is a sphere*, *c is an* n^{th}*-generation critter*, and *a and b are c's parents*.

To model your epistemic state on Thursday, we need an expansion of your Wednesday credence function. Its domain should be the Boolean closure of the domain of your Wednesday credence function together with propositions of the form *c is a pure cube* and *c is a hybrid cube*. Crucially, the hypothesis that the inheritance pattern of shapes is given

by Table 3 is in the domain of your Thursday credence function, but not in the domain of your Wednesday credence function. This allows us to think of the change in your epistemic state that took place on Thursday as an instance of conceptual change.

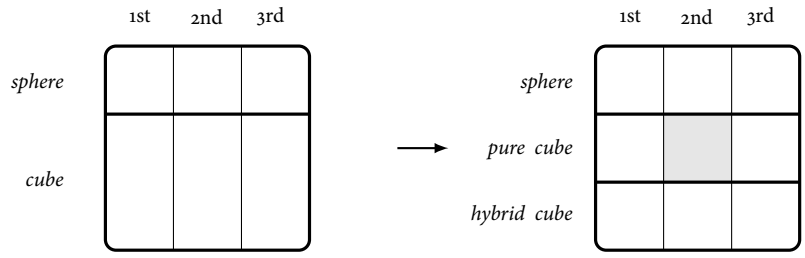


Figure 2: Each of the figures above represents a different space of hypotheses about the properties of a given critter. The space of hypotheses on the left only contains propositions of the form 'x is an n^{th} generation critter of shape S', for $n \in \{1, 2, 3\}$ and S one of 'cube' or 'sphere'. The output of the expansion contains a larger space of hypotheses, also generated by propositions of the form 'x is an n^{th} generation critter of shape S', but where S is allowed to range over 'pure cube', 'sphere', and 'hybrid cube'. Note that we're identifying the hypothesis 'x is a cube', in the first hypothesis space, with the hypothesis 'x is a hybrid cube or x is a pure cube' in the second hypothesis space. The proposition that the critter is a second-generation pure cube (in gray) is one that is in the domain of the output credence function, but not in that of the input credence function.

3 EPISTEMIC UTILITY AND EPISTEMIC RATIONALITY

Which epistemic transitions are rational?

On a standard Bayesian picture, the only rational epistemic transitions are updates. This is not because Bayesian dynamics has a criterion that is applicable to all epistemic transitions and which rules out as non-rational anything other than updates. Rather, it is because the scope of Bayesian dynamics is limited to updates: the question Bayesian dynamics aims to answer is whether an update is a rational response to the acquisition of a new piece of evidence. If we are to tackle the question of the rationality of conceptual change, we need a framework in which the general question of the rationality of epistemic transitions can be formulated.

Fortunately, we need not look too far. We can adapt standard decision-theoretic tools to build such a framework.[13]

13 There is a growing body of literature on so-called *epistemic utility theory*. See Greaves 2013, Greaves & Wallace 2006, Joyce 1998, 2009, Leitgeb & Pettigrew 2010, Moss 2011,

3.1 *The epistemic decision-theoretic framework*

Start by focusing on a simple form of expected utility theory. We evaluate different alternatives relative to a credence function and a *utility function*—an assignment of numerical values to each alternative relative to a given state of the world. A particular alternative is *better* than another if it has a higher expected utility—again, relative to a credence and utility functions. More precisely, suppose C is a credence function with state space π and U assigns a numerical value to each alternative $a \in \mathcal{A}$ relative to each $s \in \pi$. Then, the *expected utility* of a, relative to C, is given by

$$\mathbb{E}_C[U(a)] = \sum_{s \in \pi} C(s) \cdot U(a, s).$$

The canonical application of this framework is to give an account of rational decision theory—to solve *decision problems*. Typically, a decision problem is just a set of alternative courses of action. We evaluate these relative to a given utility function and a credence function. On one view, an agent's actions are *rational* just in case they have the highest expected utility (among a relevant set of alternatives) relative to her own credence function and her own utility function.[14]

But we can apply this conception of rationality to other situations. Whenever we have a range of options and an assignment of utility to each option relative to each possible state of the world, we can apply expected utility theory to evaluate each of the relevant options. In particular, we can think of decision problems where the alternatives are possible epistemic states one could be in. So long as we have a credence function (defined over possible states of the world) and a utility function defined over the

---

Pettigrew 2012, 2013. As will become apparent below, I will be building upon some of this work. The framework I will develop can be seen as a generalization of the more traditional epistemic utility theory to allow for variability in the domain of the credence functions that are the object of evaluation. For some critical discussion of the particular modeling assumptions of most work in epistemic utility theory, see Caie 2013. For critical discussion of the underlying 'consequentialist' framework, see Berker 2013. For insight on how to understand the epistemic utility framework without the seemingly problematic teleological assumptions of most interpretations of the framework, see Stalnaker 2002.

14 Since the debate over whether to formulate decision theory in evidential or causal terms is orthogonal to our present purposes, I am skating over some subtle issues here. Everything I go on to say would survive defining expected utility in more evidentialist-friendly ways, since I will be restricting my attention to cases where the different formulations of decision theory give the same verdict.

relevant alternatives (relative to a possible state of the world), we can use expected utility to compare different possible epistemic states.

This way of evaluating alternatives is relativistic in an important sense: it only makes sense to ask whether a given option is better than another relative to a particular credence function and utility function.[15] But this does not prevent us from using it to capture thicker notions of value: we only need to make more restrictions on what is to count as an admissible utility function (ditto for credence functions). In particular, we can use it to capture a notion of *epistemic value*. Given a credence function C, an 'epistemic utility function' u, and a set of epistemic states, we say that an epistemic state is better than another, epistemically, iff it has higher expected utility relative to C and u.

Now, for this to be of any help, we need to specify what a utility function must be like if it is to count as an epistemic utility function—a utility function that corresponds to an epistemic dimension of evaluation. And you might worry that there is not much content to the notion of *purely* epistemic value: perhaps any epistemic dimension of evaluation will be somewhat entangled with pragmatic considerations.[16] Still, we can aim to evaluate our beliefs so as to minimize the interference of pragmatic considerations. We can set aside particular idiosyncrasies of our judgments of practical value, and focus instead on how some beliefs more than others help us make sense of the world.

To give you a sense of the kind of evaluation I'm after, consider the following case.[17]

> An Oracle (which you know to be perfectly reliable) tells you the truth-value of every proposition. She then tells you that you will be put to sleep and your memory will be erased. Fortunately, you can now pick which credence function you will wake up with.

If you could pick *any* credence function, I suppose you would know what to choose: the one that assigns 1 to all and only the true propositions. But here is the catch: you cannot pick any credence function. You will be

15 Cf. Stalnaker 2002, p. 158.

16 For more on skepticism about the notion of a purely epistemic notion of value, see Gibbard 2008, as well as Arntzenius 2008.

17 A similar case is used for essentially the same purpose in Moss 2011, p. 1063f. Thanks to [REDACTED FOR BLIND REVIEW] for bringing this to my attention.

given a choice among a small set of credence functions which does not include the one you currently have.

I suspect you have a rough idea of how you will choose, if you set aside your practical interests for a moment. You will be able to compare at least *some* epistemic states with each other, in a way that corresponds to an epistemic dimension of evaluation. Intuitively, a utility function will count as an *epistemic* utility function just in case it corresponds to the way a fully informed agent would rank epistemic states from an epistemic perspective.

To be sure, this cannot be a *definition* of an epistemic utility function, unless we have a clear enough notion of what an epistemic dimension of evaluation is. But it is a useful heuristic, one that can help motivate plausible conditions on epistemic utility functions.

3.2 *Accuracy-based measures of epistemic value*

One way in which we can evaluate epistemic states along an epistemic dimension is in terms of accuracy. My credence in p is accurate, relative to the actual world, to the extent that it is close to p's actual truth-value. If p is true, there is a relevant sense in which, again relative to the actual world, a credence function that assigns .9 to p is better, all else equal, than one that assigns .8 to p.

This intuitive thought has been made precise in the literature on so-called *scoring rules* or *accuracy measures*. An accuracy measure, as I will use the term, is a function that assigns to each credence function and possible state of the world, a numerical value—a measure of how accurate the given credence function is relative to that state of the world. An accuracy measure is thus a plausible example of an epistemic utility function.

There is a lively, ongoing debate about how best to characterize measures of accuracy.[18] For now, and for the sake of concreteness, let us stick to a particular working example of a measure of accuracy, viz. the *Brier score* β^*, defined as:

$$\beta^*(C, s) = -\sum_{t \in \pi} (C(t) - \mathbb{1}\{s = t\})^2,$$

18 See, e.g. Joyce 2009, Leitgeb & Pettigrew 2010.

where $\mathbb{1}\{s = t\}$ equals 1 if $s = t$ and 0 otherwise. This utility function assigns, to each credence function C and world w, the sum over $s \in \pi_C$ of the negative square of the distance between $C(s)$ and s's truth-value at w. Since we can think of the negative square of the distance between $C(s)$ and s's truth value at w as a measure of the accuracy of assigning $C(s)$ to s in world w, we can think of $\beta^*(C, w)$ as telling us how accurate C is at world w.

What happens if we assume that epistemic rationality is a matter of maximizing the expected Brier score of our credence function? Remarkably, we can derive a large part of the tenets of a Bayesian picture of epistemic rationality. There is an interesting sense in which, if rationality is a matter of maximizing the expected Brier score of our credal state, then rationality requires that we have a probabilistically coherent credal state and that we update our degrees of belief by conditionalizing on new evidence—in short, that our credence functions satify BAYESIAN CORE and BAYESIAN UPDATE.[19,20]

3.3 *Expansions and epistemic utility*

We can formulate the Bayesian picture of epistemic rationality within the framework of epistemic utility theory. But the scope of the epistemic utility framework is broader than that of the orthodox Bayesian picture.

So far, the focus in the literature has been on epistemic utility functions defined over a collection of credence functions with a fixed domain. But in principle, an epistemic utility function could be defined so as to compare, relative to a given world, credence functions with different domains.

19 Cf. Greaves & Wallace 2006, Joyce 2009, Leitgeb & Pettigrew 2010. One needs to tread carefully if one is to argue that probabilistic coherence is a requirement of rationality. So far, I have only talked of credal states that are probabilistically coherent, and defined the notion of expected epistemic utility for probability functions. Joyce offers a more general definition, where it makes sense to talk about the expected epistemic utility of an assignment of degrees of belief relative to another assignment, even if neither of them are probabilistically coherent. Since those issues are orthogonal to our present concerns, I will simply refer the reader to Joyce's paper for the details.

20 Certain complications arise if we look at credence functions defined over propositions about one's own credal assignments. For present purposes, I will simply assume those complications away—any credence function that will be relevant here is one that is undefined for any such 'higher-order' proposition.

Return to the example from §2.3, involving cubes and spheres. Consider your Wednesday credence function, one whose domain does not include propositions about whether a critter is a pure cube. Consider now your Thursday credence function, one whose domain does include those propositions. Here are a few things about it worth noting:

- The hypothesis—call it T1—that the inheritance pattern of shapes among the critters is given by Table 3 is generation independent. It applies equally well to critters of any generation, and so is *more general* than the hypotheses available to you on Wednesday.
- Our hypothesis T1—which, again, is in the domain of your Thursday credence function, but not in the domain of your Wednesday credence function—accurately predicts the inheritance pattern of shapes among the critters. And T1 is *simpler* than any reasonably accurate hypothesis that you could have formulated with your Wednesday credence function.
- On Thursday you are in a position to explain why a particular pair of cubes have some spheres as offspring by appealing to the fact that they are both hybrid cubes. And *this explanation is better*, I submit, than any one you could have formulated on Wednesday.[21]

In short, there is something to be said, epistemically, and relative to the actual world, for your Thursday credence function over your Wednesday credence function. A theory of epistemic utility should be sensitive to such facts.

Our first order of business is to spell out a way of comparing credence functions with different hypothesis spaces relative to the same state of the world. This will require introducing a few definitions.

Let $\mathcal{P}$ be the collection of all probability functions with a finite domain. To fix terminology, let us stipulate that a *utility function* is a function u that associates, to each probability function $P \in \mathcal{P}$ and world $w \in W$ a real number $u(P, w)$, which I will call the *utility of P in w*.

DEFINITION 3.1. A *utility function* is a real-valued function defined over $\mathcal{P} \times W$.

Intuitively, any such function must satisfy the following desideratum: if P does not distinguish between w and w'—that is, if for all x in the

21 For further discussion of this issue, see §4.3.

domain of P, $w \in x$ iff $w' \in x$—then $u(P, w) = u(P, w')$. Thus, for any world w, $u(P, \cdot)$ should be constant throughout the π_P cell of w—that is, the unique element of π_P containing w. This amounts to the following constraint:

DEFINITION 3.2. A utility function u is *nice* iff for each P, the function $u(P, \cdot) : W \to \mathbb{R}$ is constant throughout any $s \in \pi_P$. In other words, for each $s \in \pi_P$ and any $w, w' \in s$, $u(P, w) = u(P, w')$.

Equivalently, u is nice iff for each P, $u(P, \cdot)$ P-measurable—i.e. if for each $r \in \mathbb{R}$,

$$\{w \in W : u(P, w) = r\}$$

is in the domain of P.[22]

From now on, I will assume that all utility functions are nice.

If u is a nice utility function, it makes sense to talk of the utility of P 'at s'. More generally:

REMARK 3.3. If u is nice and t is a subset of a given P-atom s, we can set $u(P, t)$ to $u(P, w)$ for any $w \in t$, so that $u(P, t) = u(P, s) = u(P, w)$.

To see what a nice utility function looks like, consider the following utility function β:

$$\beta(C, x) = -\sum_{t \in \pi_C} (C(t) - \mathbb{1}\{[x]_C = t\})^2,$$

where $[x]_C$ is the unique element of π_C containing x. Note that for all C, if $[x]_C = [x']_C$, then $\beta(C, x) = \beta(C, x')$. Thus, β is a nice utility function. Note moreover that, when restricted to comparing probability functions of a given domain, our utility function β is just the Breier score, i.e.:

$$\beta(C, x) = \beta^*(C, [x]_C).$$

Nice utility functions can be used to make comparisons among credence functions with different domains. To see how, consider a simple example.

22 In the terminology of Lewis 1981, q P-measurability is just the requirement that all value-level propositions be in the domain of P.

EXAMPLE 3.4. Let P_0 be the only probability function defined over the trivial algebra—that is, $\pi_{P_0} = \{W\}$. Let P_1 be such that $\pi_{P_1} = \{s_0, s_1\}$, with $P_1(s_0) = P_1(s_1) = .5$. You can think of the state space of P_1 as what you get from the state space of P_0 if you simply partition W into two cells, s_0 and s_1—P_1 is just the uniform probability distribution over the finer state space.

Now, for all w,

$$\beta(P_1, w) = -((.5-1)^2 + (.5-0)^2) = -.5.$$

Since for all w,

$$\beta(P_0, w) = 0,$$

we have that P_0 is doing better than P_1 relative to β and any $w \in W$.

In this case, it makes sense to ask for the expected score of each of P_0 and P_1 relative to P_1 and β, where as before:

$$\mathbb{E}_{P_1}[\beta(P)] = \sum_{s \in \pi_{P_1}} P_1(s) \cdot \beta(P, s)$$

In particular:

$$\mathbb{E}_{P_1}[\beta(P_0)] = 0,$$
$$\mathbb{E}_{P_1}[\beta(P_1)] = -.5.$$

Hence, relative to P_1 and β, P_0 is doing better than P_1.[23] □

Since we are dealing with probability functions with different domains, we need to be careful when defining the expected utility of a probability function relative to another. For ease of reference, let us introduce a familiar definition:

23 This might seem surprising to those familiar with epistemic utility theory. After all, one of the notable features of the Brier score is that it is *strictly proper* (in a sense to be defined in §3.4). But recall that the claim that β is strictly proper essentially amounts to the claim that any credence function should take itself to be doing better than any other credence function *with the same domain*. I discuss generalizations of the more familiar notion of propriety in the context of comparing credence functions with different domains below. In the terminology to be introduced in §3.4, we will say that β is not 'downwards proper', even though it is strictly proper.

DEFINITION 3.5. Say that a partition π of W is a *refinement* of π' iff for any $s' \in \pi'$ there is $s \in \pi$ such that $s \subseteq s'$. The refinement relation induces a partial ordering on the set of partitions of a set, where $\pi \preceq \pi'$ iff π' is a refinement of π.[24] If π is a refinement of π', we will say that π' is a *coarsening* of π.

Now note that if P' is an expansion of P then $\pi_{P'}$ is a refinement of π_P. And a straightforward consequence of Remark 3.3 is that if $\pi_{P'}$ is a refinement of π_P, then for all $s' \in \pi_{P'}$, $u(P, s)$ is well-defined. Hence, whenever P' is an expansion of P, we can define the expected utility of P relative to P' and any nice utility function in the usual way:

$$\mathbb{E}_{P'}[u(P)] = \sum_{s \in \pi_{P'}} P'(s) \cdot u(P', s).$$

The difficulty arises when $\pi_{P'}$ is not a refinement of π_P, since we cannot guarantee that for all $s \in \pi_{P'}$, $u(P, s)$ is well-defined. In particular, if P' is an expansion of P, we cannot guarantee that $u(P', s)$ will be well-defined for all $s \in \pi_P$, since $u(P', \cdot)$ may not be constant throughout all $s \in \pi_P$.

We will be particularly interested in evaluating, relative to a probability function P and utility function u, different expansions of P. Although we cannot assume that $\mathbb{E}_P[u(P')]$ will be defined for an arbitrary expansion P' of P, there are two different but related well-defined quantities that will come in handy.[25]

If π is a refinement of π_P, let $\mathcal{P}_\pi\,/_P$ be the extensions of P to π—the collection of all probability functions whose state space is π that agree with P on P's domain. If P' is an expansion of π_P, we will let $\mathcal{P}_{P'}\,/_P$ stand for $\mathcal{P}_{\pi_P}\,/_P$—so $\mathcal{P}_{P'}\,/_P$ is the collection of extensions of P to the domain of P'.[26]

24 For the *cognoscenti*: the partial ordering I will be relying on is the inverse of the partition lattice of W (see e.g. Ellerman 2010.) In the present context, the inverse order seems easier to work with, for $\pi \preceq \pi'$ iff $\mathcal{B}_\pi \subseteq \mathcal{B}_{\pi'}$, so that 'moving forward' along the refinement relation amounts to an increase in the set of propositions in the corresponding algebra.

25 Cf. Goldstein 1984, Manski 1981.

26 Note that we can think of $\mathcal{P}_{P'}\,/_P$ as an imprecise probability function, most naturally identified with a *representor* in the sense of van Fraassen n.d., viz. a set of probability functions. In this case, we can think of $\mathcal{P}_{P'}\,/_P$ as an imprecise probability function that assigns precise values to each member of $\mathcal{B}_P$ and imprecise values to any other member of $\mathcal{B}_{P'}$. The definitions to follow can thus be seen as the familiar definition of upper and lower expectation for imprecise probabilities (Gilboa 1987, Satia & Lave 1973). See Troffaes 2007 for a recent of these and related works.

DEFINITION 3.6. Let P' be an expansion of P, and let $\mathcal{P}_{P'}/_P$ be the collection of extensions of P to the domain of P'—that is, the collection of probability functions with the same domain as P' and which assign the same value as P does to every proposition in P's domain (recall that if P' is an expansion of P then the domain of P' includes that of P). The *upper expected value* of P', relative to P and u is

$$\begin{aligned}\overline{\mathbb{E}}_P[u(P')] &= \sup_{P^+ \in \mathcal{P}_{P'}/_P} \mathbb{E}_{P^+}[u(P')] \\ &= \sup_{P^+ \in \mathcal{P}_{P'}/_P} \sum_{\pi_{P'}} P^+(s) \cdot u(P', s)\end{aligned}$$

The *lower expected value* of P', relative to P and u, is

$$\begin{aligned}\underline{\mathbb{E}}_P[u(P')] &= \inf_{P^+ \in \mathcal{P}_{P'}/_P} \mathbb{E}_{P^+}[u(P')] \\ &= \inf_{P^+ \in \mathcal{P}_{P'}/_P} \sum_{\pi_{P'}} P^+(s) \cdot u(P', s)\end{aligned}$$

Note that $\overline{\mathbb{E}}_P[u(P')] \geq \underline{\mathbb{E}}_P[u(P')]$, with equality if $\pi_P = \pi_{P'}$.

EXAMPLE 3.7. Let P_0 be as in Example 3.4, viz. the unique probability function defined over the trivial algebra. Let P_2 be such that $\pi_{P_2} = \{s_0, s_1\}$, with $P_2(s_0) = 1$. Then:

$$\beta(P_2, w) = \begin{cases} -((1-1)^2 + (0-0)^2) = 0 & \text{if } w \in s_0 \\ -((0-1)^2 + (1-0)^2) = -2 & \text{otherwise}, \end{cases}$$

which means $\beta(P_2, s_0) = 0$ and $\beta(P_2, s_1) = -2$. Thus,

$$\begin{aligned}\overline{\mathbb{E}}_{P_0}[\beta(P_2)] &= \sup_{P^+ \in \mathcal{P}_{P_2}/_{P_0}} \sum_{\pi_{P_2}} P^+(s) \cdot \beta(P_2, s) \\ &= \sup_{P^+ \in \mathcal{P}_{P_2}/_{P_0}} P^+(s_1) \cdot -2 \\ &= 0.\end{aligned}$$

Since $\mathbb{E}_{P_0}[\beta(P_0)] = 0$, we have

$$\mathbb{E}_{P_0}[\beta(P_0)] = \overline{\mathbb{E}}_{P_0}[\beta(P_2)].$$

Note further that

$$\begin{aligned}\underline{\mathbb{E}}_{P_0}[\beta(P_2)] &= \inf_{P^+ \in \mathcal{P}_{P_2}/_{P_0}} \sum_{\pi_{P_2}} P^+(s) \cdot \beta(P_2, s) \\ &= \inf_{P^+ \in \mathcal{P}_{P_2}/_{P_0}} P^+(s_1) \cdot -2 \\ &= -2.\end{aligned}$$

Hence

$$\underline{\mathbb{E}}_{P_0}[\beta(P_2)] < \mathbb{E}_{P_0}[\beta(P_0)] = \overline{\mathbb{E}}_{P_0}[\beta(P_2)].$$

□

Given a credence function C and a utility function u, we can now compare different expansions of C in different ways. Importantly, as shall become clearer, we can compare different expansions of C defined over different collections of propositions. For example, we can ask which one maximizes $\overline{\mathbb{E}}_P[u(\cdot)]$ (maximax), which one maximizes $\underline{\mathbb{E}}_P[u(\cdot)]$ (maximin), or perhaps which one maximizes some weighted average of the two. Presumably there will be things to be said in favor of each of these decision rules. Here, though, I hope to avoid getting into that debate. My focus will be first and foremost on the preliminary question how to motivate epistemic utility functions that allow for such comparisons in the first place. And, fortunately, some of the applications I will briefly review towards the end allow us to remain neutral on exactly how best to choose different expansions.

3.4 Propriety and expansions

Not all utility functions, in the sense defined above, are admissible as *epistemic* utility functions. A function that assigns the same value to any probability function relative to any world, for example, cannot count as an epistemic utility function.

A minimal requirement on epistemic utility functions, one that has been proposed in a slightly different form in the literature, is this. Suppose P and P' are defined over the same state space π and for all $s \in \pi$, the distance between $P(s)$ and the actual truth-value of s is as close, and sometimes closer, than that between $P'(s)$ and the actual truth-value of s.

Then the utility of P at the actual world is greater than that of P'. More generally, for any two P and P' defined over a given state space π, we will say that P is *as close to the truth* as P', relative to $s^* \in \pi$, iff for all $s \in \pi$,

$$|P(s) - \mathbb{1}\{s^* = s\}| \leq |P'(s) - \mathbb{1}\{s^* = s\}|.$$

We will say that P is *closer to the truth* than P', relative to $s^* \in \pi$, if P is as close to the truth as P', relative to $s^* \in S$, and for some $s_0 \in \pi_{P'}$,

$$|P(s_0) - \mathbb{1}\{s^* = s_0\}| < |P'(s_0) - \mathbb{1}\{s^* = s_0\}|.$$

DEFINITION 3.8. A utility function is *truth-directed* iff for any two P and P' defined over a given state space π, (i) if P is as close to the truth as P', relative to $s^* \in \pi$, then

$$u(P, s^*) \geq u(P', s^*),$$

and (ii) if P is closer to the truth than P', relative to $s^* \in \pi$, then

$$u(P, s^*) > u(P', s^*).$$

Plausibly, any reasonable epistemic utility function will be truth-directed. The Brier score, widely taken to be a reasonable epistemic utility function, is a truth-directed utility function.

There is broad consensus that not all truth-directed utility functions are reasonable epistemic utility functions. For example, consider the *absolute distance* measure α, where

$$\alpha(P, w) = -\sum_{s \in \pi_P} |P(s) - \mathbb{1}\{[w]_P = s\}|.$$

This is plausibly the simplest truth-directed utility function. But it has the following property, which many consider to be a bug: if you are slightly more confident in p than in its negation, you should expect that being fully confident in p is better, by the lights of α, than having your current degrees of belief.

For example, suppose you have a credence function C whose state space consists of two propositions, p and $\neg p$, and suppose that $C(p) > .5$. Then, with a bit of algebra, we can see that the expected utility of C, relative to C and α, will be lower than that of the function C', where $C'(p) = 1$. If epistemic rationality is a matter of maximizing expected

epistemic utility, and if α is a reasonable measure of epistemic utility, then it would be rational to jump to extremes. Since jumping to extremes is not epistemically rational, once we accept that epistemic rationality is a matter of maximizing expected epistemic utility, we must conclude that α is not a reasonable measure of epistemic utility.

We can draw a general lesson from this. Suppose it is epistemically rational to have credence function C defined over a state space π. Then someone with that credence function cannot expect to be doing better, epistemically, by assigning *different* credences to the propositions in π.[27] In general, suppose any probabilistically coherent credence function can sometimes be epistemically rational. Then all reasonable epistemic utility functions must be *partition-wise proper* in the following sense:

DEFINITION 3.9. A utility function u is *partition-wise proper* iff for all $P \neq P'$ defined over the same state space,

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_P[u(P')].$$

A utility function is *partition-wise strictly proper* iff for all $P \neq P'$ defined over the same state space,

$$\mathbb{E}_P[u(P)] > \mathbb{E}_P[u(P')].$$

Note the restriction to comparisons of probability functions with the same domain. Since most work on epistemic utility functions has implicitly made this restriction, our definition of partition-wise propriety coincides with the more familiar definition of propriety. The Brier score, for example, is a paradigmatic example of a (strictly) proper utility function, and it is as a partition-wise (strictly) proper utility function. In what follows, and when there is no risk of confusion, I will sometimes use 'proper' (resp. 'strictly proper') as shorthand for 'partition-wise proper' (resp. 'partition-wise strictly proper').

We could introduce a stronger constraint: if someone with credence function C is rational, she cannot expect to be doing better, epistemically, by switching to any other credence function whose domain is a subset

27 What is wrong with α, then, on this picture, is that it is sometimes rational to have a credence function C such that $C(p) > C(\neg p) > 0$. And if α were a reasonable measure of epistemic utility, someone with that credence function C would expect to be doing better, epistemically, by becoming certain of p.

of the domain of C. In other words, one should not be able to do better either by changing one's assignment of credence in p or by no longer assigning any credence (not even 0) to p. If again we suppose that any probability function can sometimes be a rational credence function, we would require that all reasonable epistemic utility functions be *strictly downwards proper* in the following sense:

DEFINITION 3.10. A utility function u is *downwards proper* iff for all $P \neq P'$, if $\pi_{P'}$ is a coarsening[28] of π_P ($\pi_{P'} \preceq \pi_P$), then

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_P[u(P')].$$

A utility function is *strictly downwards proper* iff for all $P \neq P'$ if $\pi_{P'} \preceq \pi_P$, then

$$\mathbb{E}_P[u(P)] > \mathbb{E}_P[u(P')].$$

As we saw in Example 3.4, β is not downwards proper. If we require that all epistemic utility functions be downwards proper, then the Brier score will not count as an epistemic utility function.[29]

A related but different constraint we could impose is this. If someone with credence function C is rational, she cannot expect to be doing better, epistemically, by switching to any other credence function whose domain extends the domain of C. In other words, one should not be able to do better either by changing one's assignment of credence to those propositions one currently assigns credence to or by assigning credence to propositions not in the domain of one's current credence function. If again we suppose that any probability function can sometimes be a rational credence function, we would require that all reasonable epistemic utility functions be *strictly upwards proper* in the following sense:[30]

28 See Definition 3.5.

29 Sometimes the Brier score is defined in a different way, viz.

$$\beta^*(P, w) = -\frac{1}{N}\sum_{\pi_P}(P(s) - \mathbb{1}\{s_w = s\})^2,$$

where N is the size of π_P. The same example can be used to illustrate that β^* is not downwards proper.

30 As I show in Appendix A, in the presence of partition-wise propriety, upwards propriety is equivalent to the following condition: for each $P \neq P'$, if $\pi_P \preceq \pi_{P'}$, then

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_{P'}[u(P')].$$

DEFINITION 3.11. A utility function u is *upwards proper* iff for all $P \neq P'$, if $\pi_{P'}$ is a refinement of π_P ($\pi_P \preceq \pi_{P'}$), then

$$\mathbb{E}_P[u(P)] \geq \overline{\mathbb{E}}_P[u(P')].$$

A utility function is *strictly upwards proper* iff for all $P \neq P'$, if $\pi_P \preceq \pi_{P'}$, then

$$\mathbb{E}_P[u(P)] > \overline{\mathbb{E}}_P[u(P')].$$

As we saw in Example 3.7, the Brier score is not strictly upwards proper. If we require that all epistemic utility functions be strictly upwards proper, then the Brier score will not count as an epistemic utility function.

Say that a theory of epistemic rationality requires *strong immodesty* if the only rational epistemic transitions are the result of updating on new evidence. If epistemic rationality is a matter of maximizing expected epistemic utility, strong immodesty requires that any rational credence function judge itself to be doing better, epistemically, than *any* of its alternatives, regardless of their domain. If we think any probability function can sometimes be a rational credence function, strong immodesty would entail that all epistemic utility functions be *strictly universally proper* in the following sense:

DEFINITION 3.12. A utility function u is *universally proper* iff it is upwards proper and downwards proper. A utility function is *strictly universally proper* iff it is strictly downwards proper and strictly downwards proper.

As it turns out, strong immodesty is too strong a requirement. For it rules out all utility functions as reasonable epistemic utility functions:[31]

FACT 3.13. *There are no strictly universally proper utility functions.*

See Fact A.7.

31 Note that we could have formulated upwards propriety using lower expected value rather than upper expected value (see Definition 3.6). The corresponding requirement would have been that for each P and each $P' \neq P$, if $\pi_P \preceq \pi_{P'}$, then

$$\mathbb{E}_P[u(P)] \geq \underline{\mathbb{E}}_P[u(P')].$$

As we will see, however, this requirement is also incompatible with strict downwards propriety, at least given some fairly harmless assumptions—see Corollary A.10.

What is more, say that a utility function u is *partition-wise strictly proper* iff for each P and any Q defined over the same domain, if $P \neq Q$ then $\mathbb{E}_P[u(P)] > \mathbb{E}_P[u(Q)]$. (Most work on the literature so far uses the term 'strict propriety' to mean what I'm calling 'partition-wise strict propriety'.) If we assume all epistemic utility functions are partition-wise strictly proper, then not only do we know there are no strictly universally proper epistemic utility functions. We can also show that there are no universally proper epistemic utility functions:

FACT 3.14. *If u is universally proper, it is not partition-wise proper.*[32]

These results raise a couple of questions. Should all epistemic utility functions be downwards proper? If not, should they all be upwards proper? (Examples of upwards proper and downwards proper epistemic utility functions can be found in Appendix B.)

I think an argument can be made for requiring downwards propriety, much along the lines of a familiar argument for (partition-wise) strict propriety. Suppose u is not downwards proper. Then there are credence functions such that they take themselves to be doing strictly worse, epistemically, than one of their restrictions. This means that anyone with such a credence function should think, of some of issues she has an opinion on, that she would better off simply having no view whatsoever on the matter. Thus, if we assume that any credence function can sometimes be epistemically rational, we must require that all epistemic utility functions be downwards proper.[33]

32 See Theorem A.3.

33 Note that this argument does not straightforwardly generalize to an argument for requiring upwards propriety, for reasons spelled out in the introduction (see the discussion of immodesty and epistemic imagination).

Richard Pettigrew has shown—building on Carr 2015—that any epistemic utility function that satisfies some plausible constraints will give rise to a dilemma: either (a) for all $\epsilon > 0$, there is a proposition p such that assigning credence $< \epsilon$ is equally epistemically good relative to worlds in which p is true as it is relative to worlds in which p is false, or (b) some credence functions are strictly dominated by some of their expansions (Pettigrew 2016). Ultimately, Pettigrew thinks we should learn to live with the second horn of the dilemma (see his discussion in section 3.4). I agree that this is the horn we should plump for, but my reasons are slightly different: as long as the agent is not entertaining the propositions in the expanded domain, it may be rational for her to stick to her current credence function *even if* there is an expansion of her credence function defined over the larger domain that dominates it.

It is not my goal here, however, to make a case for a requirement of downwards propriety. Instead, I will tentatively assume that all ways of measuring *accuracy* should be downwards proper. From this and Fact 3.14 we can now conclude that no accuracy measure will be upwards proper, and thus that expansions can sometimes be epistemically rational. There is more to be said, though. In the remainder of the paper, I want to find a way of vindicating the intuitive judgments about our motivating examples, by relying on a principled way of comparing different expansions of a given credence function along an epistemic dimension of evaluation.

Recall, for example, our toy example from §2.3: there you moved from a credence function (your Wednesday credence function) that was not defined over propositions of the form *c is a pure cube* or *c is a hybrid cube* to one (your Thursday one) which was. Incorporating such propositions into your posterior, I claimed, was an epistemic improvement, and this is from the perspective of your prior credence function. Before incorporating the new propositions into your credence function—before assigning a particular credence to it and before deploying that proposition in your theorizing—you could have expected that doing so will have some non-trivial epistemic benefits. To capture these epistemic judgments, we need more than just the assumption that there are epistemic utility functions that are downwards proper. We need a principled way of defining epistemic utility functions which are not just downwards proper, but also which get the judgments in question right.

In the next section, I offer a way of doing just that. In particular, I will offer a family of epistemic utility functions according to which, from the point of view of your Wednesday credence function, Thursday's credence function had a higher expected epistemic value than your Wednesday credence function itself, but also than other possible expansions of your Wednesday credence function. In other words, if the utility functions in question do in fact correspond to a genuine epistemic dimension of evaluation, as I will suggest they do, we will have vindicated the intuition that switching to your Thursday credence function was a Good Thing.

4 A RICHER THEORY OF EPISTEMIC UTILITY

We could try to appeal to accuracy considerations, broadly understood, in order to account for why the move to your Thursday credence function was epistemically beneficial. After all, by introducing new distinctions

you increased the number of propositions about which you have beliefs. And this could turn out to increase the amount of overall accuracy, on at least some ways of measuring it, of your body of beliefs.

For example, we could modify the Brier score so that the score of C at w depends on how large its state space is.[34] And we could do this in a way that would vindicate the thought that, relative to the actual world, your Thursday credence function is doing better than your Wednesday credence function. But this would not vindicate the similarly plausible thought that the epistemic gain from incorporating those propositions into your Thursday credence function exceeds that of incorporating (say) propositions of the form *c is a pure sphere* and *c is a hybrid sphere*.[35] There is epistemic work to be done by the former set of propositions, but not (as much) by the latter. We want our theory of epistemic utility to be sensitive to that fact.

The epistemic gain corresponding to the transition from your Wednesday credence function to your Thursday credence function is (at least in part) due to the *content* of those propositions themselves. You reaped epistemic benefits from the transition to your Thursday credence function because the added propositions gave you explanatory resources you did not have before, and because you gained the ability to formulate a simpler, more general hypothesis that accurately predicts the inheritance pattern of the traits you were interested in. And this suggests that our epistemic utility function should not treat accuracy with respect to a given proposition to be as important as accuracy with respect to any other proposition. It should instead weigh accuracy with respect to a given proposition in a way that is proportional to the explanatory benefits it would provide (if true).[36]

More precisely, the proposal is this:

34 Cf. Corollary B.4 in the appendix.

35 This is because, like many measures of accuracy, the modified Brier score would be subject to the following constraint (cf. Joyce 2009, p. 273):

> EXTENSIONALITY: Let P and Q be defined, respectively, over $\pi_P = \{p_i : 0 \leq i < m\}$ and $\pi_Q = \{q_i : 0 \leq i < m\}$. If $P(p_i) = Q(q_i)$ and $\mathbb{1}\{[w]_P = p_i\} = \mathbb{1}\{[w']_Q = q_i\}$ for all i, then $\mu(P, w) = \mu(Q, w')$.

36 Cf. fn. 42.

EXPLANATION SENSITIVITY: Relative to the goal of explaining e, accuracy with respect to p matters more, epistemically, to the extent that p would contribute towards explaining e.

Our first order of business, then, is to define an epistemic utility function that is sensitive to accuracy considerations, but which gives different weight to accuracy with respect to different propositions. Our second and more challenging order of business is to justify a particular assignment of weights—one that assigns to each proposition a number that measures its explanatory benefits. Let us take each in turn.

4.1 *Additive epistemic utility functions*

Recall our working example of an accuracy measure, the Brier score:

$$\beta(C,x) = -\sum_{t \in \pi_C} (C(t) - \mathbb{1}\{[x]_C = t\})^2.$$

This utility function treats accuracy about s at w exactly like it treats accuracy about s' at w: in each case what matters is the distance between the credence assignment in a proposition and that proposition's truth-value. But we can easily define a variant of our epistemic utility function that treats accuracy about s at w differently from accuracy about s' at w. Given any real-valued function λ of s, where each $\lambda(s) > 0$, we can define:

$$\beta^\lambda(C,x) = -\sum_{s \in \pi_C} \lambda(s) \cdot (C(s) - \mathbb{1}\{x \in s\})^2.$$

Such a utility function would weigh accuracy with respect to s as a function of $\lambda(s)$.

This is not, of course, the only way of obtaining utility functions that treat accuracy with respect to s in a way that sensitive to certain features of s. Any *additive* function, in the following sense, will do:[37]

DEFINITION 4.1. A utility function u is *additive* iff for each partition π, each $s \in \pi$, there is a function $\delta_s^\pi : [0,1] \times \{0,1\} \longrightarrow \mathbb{R}$, such that (i) $\delta_s^\pi(x,1)$ is continuous, twice differentiable, and strictly increasing, (ii)

37 The definition to follow is a slight variant of the one given in Joyce 2009, p. 272. Note that I'm building in niceness (something Joyce need not worry about) into the definition of additive utility functions. I'm also building in truth-directedness, much like Joyce does.

$\delta_s^\pi(x,0)$ is continuous, twice differentiable, and strictly decreasing, and (iii) for all C with $\pi = \pi_C$

$$u(C,w) = \sum_{s \in \pi} \delta_s^\pi(C(s), \mathbb{1}\{w \in s\}).$$

Additive utility functions are epistemic utility functions that can treat accuracy with respect to different propositions quite differently.

An additive utility function will be partition-wise strictly proper so long as each δ_s^π is strictly proper, in the sense that, for all $r \neq t \in [0,1]$:[38]

$$r \cdot \delta_s^\pi(r,1) + (1-r) \cdot \delta_s^\pi(r,0) > r \cdot \delta_s^\pi(t,1) + (1-r) \cdot \delta_s^\pi(t,0).$$

In particular, for any assignment λ of weight functions $\lambda_\pi : \pi \to \mathbb{R}^+$ to each state space π, the function β^λ is a strictly proper utility function, one which weighs accuracy with respect to s, in a credence function defined over π, in a way that is proportional to $\lambda_\pi(s)$.

Of particular relevance to the purposes of this paper is a family of additive utility functions that we will call *weighted accuracy measures*. Say that a function $\delta : [0,1] \times \{0,1\} \longrightarrow \mathbb{R}$ is a *local accuracy measure* iff for each $i \in [0,1]$, $\delta(x,i)$ is a continuous, twice differentiable, strictly decreasing function of $|i-x|$. (Thus, $\delta(x,1)$ will be strictly increasing and $\delta(x,0)$ strictly decreasing.) We can think of a local accuracy measure as a way of evaluating an assignment of credence in a proposition in a way that is sensitive to the truth-value of the proposition and the distance between the credence assignment and that truth-value.[39]

DEFINITION 4.2. A utility function δ is a *weighted accuracy measure* iff there is a local accuracy measure $\delta : [0,1] \times \{0,1\} \longrightarrow \mathbb{R}$ and, for each

38 This result is stated without proof in Joyce 2009, p. 276, but it is a straightforward consequence of the additivity of expectation, which ensures that (for C' defined over π_C):

$$\sum_{s \in \pi_C} C(s) \sum_{s' \in \pi_C} \delta_s^\pi(C'(s'), \mathbb{1}\{s = s\}) = \sum_{s \in \pi_C} C(s) \cdot \delta_s^\pi(C'(s),1) + (1-C(s)) \cdot \delta_s^\pi(C'(s),0).$$

On the assumption that each δ_s^π is proper, we can infer that this sum will be maximized at $C' = C$.

39 Compare the definition of a *local epistemic utility function* in Pettigrew 2016, §1.1.3. Note that, unlike Pettigrew, I do not build into the definition either the requirement that δ be proper or the requirements that $\delta(0,1) < 0 < \delta(1,1)$ and $\delta(1,0) < 0 < \delta(0,0)$. Indeed, as we will see below (cf. Corollary C.2) assuming $\delta(0,0) > 0$ imposes substantive constraints on epistemic utility functions.

partition π, a weight function $\lambda_\pi : \pi \longrightarrow \mathbb{R}^+$ such that for all P with $\pi_P = \pi$,

$$u(P, w) = \sum_{p \in \pi} \lambda_\pi(p)\delta(P(p), \mathbb{1}\{w \in p\}),$$

The weighted version of the Brier score, β^λ, is a weighted accuracy measure in this sense.

We can use weighted accuracy measures in order to capture a way of measuring epistemic value that is sensitive to explanatory considerations. Doing so requires that we identify an assignment of weights to different propositions that measure the extent to which they contribute to an explanation of that explanandum. Given such an assignment, we can define epistemic utility functions that treat accuracy with respect to a proposition in a way that depends on how much it contributes to a given explanandum.

Presumably, a theory of explanation should tell us how much a particular proposition contributes to a (good) explanation of any given explanandum. Now, I do not have a general theory of explanation to offer. But there is a common core to many theories of explanation that we can appeal to in order to illustrate how to go about measuring the explanatory strength of a given proposition (relative to a particular explanatory goal). This will allow us to vindicate the thought that, in an interesting sense, the move from your Wednesday credence function to your Thursday credence function (and the same goes for the transition from Monday to Tuesday) was epistemically rational.

It bears repeating. In order to build in explanatory considerations into a theory of epistemic utility, I will need to take a stand on difficult questions about the nature of explanation. But it is not my goal here to defend those answers, and the main conclusions of this paper should not depend on those being the correct answers. Rather, I will rely on them in order to offer a proof of concept: I want to explore how we can extend the Bayesian framework if we enrich our theory of epistemic utility.

Before moving on, I should make explicit two methodological assumptions. The first is that I will presuppose, in order to keep things somewhat simpler, that whenever an agent is interested in explaining e, she is certain that e is true. This is no doubt an oversimplification. For our purposes, however, it will do—none of the main conclusions of the paper hinge on it.

The second is that I will be mostly concerned not with whether (and to what extent) p is a good explanation of e *simpliciter*. Rather, I will be concerned with whether (and to what extent) p is a good explanation of e relative to some particular body of beliefs. This is so for two reasons. First: I want to remain neutral on the debate over whether there is such a thing as an 'objective' notion of explanation. I thus want to rely only on judgments of explanatory relevance, which would make sense against a background body of belief even if there is no such thing as 'the' explanation of a given explanandum. Second: in line with the the broadly 'internalist' perspective that underlies much work on epistemic decision theory, I think that what epistemic utility function we use to evaluate an agent's epistemic rationality should reflect the agent's judgments about epistemic value.[40] To the extent that an epistemic utility function is meant to be sensitive to explanatory considerations, those considerations will reflect the agents own judgments about what explains what. And since, again, those judgments may well be sensitive to the agent's background body of beliefs, we are better off focusing on whether p is a good explanation of e relative to a given body of beliefs.

4.2 *Explanation, invariance, and stability*

You flip a coin ten times in a row. To your surprise, it lands tails nearly every single time. Here is a possible explanation of what happened:

BIAS: The coin is heavily biased toward tails.

Another possible explanation consists of a specification of each of the starting positions of the coin and your hand, together with a specification of the force with which you flipped it and the wind conditions which, together with the laws of physics, make it extremely likely that the coin landed tails nearly every time. Call this explanation INITIAL, and sup-

40 I suspect those who think that explanations are, as Woodward puts it, ''a matter of exhibiting systematic patterns of counterfactual dependence'' (Woodward 2005, p. 191) will agree that explanatory judgments like 'the glass broke because it fell from the table' only deserve to be so-called because they take place against a background set of assumptions which, together with the specific judgment in question, entail facts about about the relations of counterfactual dependence among the relevant events.

pose it is incompatible with BIAS, perhaps because we can derive from the facts cited in INITIAL that the coin is fair.[41]

To some extent, the first explanation is more satisfying than the second. This is not because it is more or less likely: it may be highly unlikely that the coin you got from the bank was heavily biased towards tails. Rather, it is because *if true* it would be more satisfying as an explanation than the second one would be, if *it* happened to be true.[42]

Why would an explanation in terms of BIAS be more satisfying than one in terms of INITIAL? It is not because BIAS makes the explanandum more probable— we would prefer BIAS even if we modified INITIAL so that it entailed the truth of the explanandum, and therefore raised its probability to one. Rather, it is because BIAS has some familiar explanatory virtues that INITIAL lacks.

For example, BIAS is simpler than INITIAL. We want our theories to be simple—we want them to involve no more detail than it is necessary—partly because theorizing has cognitive costs, and we rather not spend cognitive resources on details that promise little in terms of theoretical payoff.[43]

Another reason for preferring BIAS over INITIAL is that it is more general—it can be applied to many different circumstances. Generality tends to make for good explanations. This is why appealing to beliefs and desires to explain my behavior can be more satisfying than giving a full description of the state of my brain.[44]

Consider this example, essentially due to Alan Garfinkel.[45] Tom is running late for a meeting, because he had a leisurely breakfast. He gets in

41 This example is based on a slightly different example in White 2005, where it's used to illustrate an explanatory virtue called *stability*. Although the point White goes on to make is different from the one I will make, and although the characterization of stability he provides is not quite the notion of counterfactual resilience that I introduce in this paper, there is much in common to the spirit of both proposals.

42 The distinction I'm drawing here is of course reminiscent of Peter Lipton's well-known distinction between the *loveliness* and *likelihood* of an explanation (Lipton 2004).

43 Admittedly, it is tempting to think that simplicity is a virtue not just because of our cognitive limitations, but we needn't take a stance on that issue. For related discussion, see e.g. Baker 2003, Nolan 1997. See also Baker 2011 for a helpful overview of some of the relevant issues.

44 Cf. Jackson & Pettit 1988, Strevens 2004, and the discussion of causal relevance in Yablo 1992 for related discussion.

45 Garfinkel 1981, p. 30.

his car and drives somewhat recklessly—so much so that he loses control of the car at some point and gets into an accident. A natural explanation of this unfortunate event is that Tom was driving recklessly—that he was speeding, say. Given background assumptions, his speeding makes it very likely that he got into an accident. But this cannot be all it takes for something to be a good explanation of the accident. After all, a fuller description of Tom's morning would also make it highly likely that he got into an accident. And this, I submit, would not be as good an explanation of the accident.

The reason—or at least, a reason that many have offered to account for similar cases—is that, unlike the first explanation, the second one is not very portable.[46] Had Tom not had a leisurely breakfast, we couldn't have used this second one as an explanation for why he got into the accident. The explanation in terms of his reckless driving, in contrast, is applicable to many other situations—there are many other ways Tom's morning could have been that, given his reckless driving, would have ended up in a car accident.

I do not intend to argue that simplicity and generality are explanatory virtues. I will simply take it for granted for the purposes of this paper. What I want to suggest is a way of capturing these explanatory virtues in a way that is more amenable to the framework of epistemic utility theory.

To see how, start by noting that both simplicity and generality have one thing in common. Having a simple, or a very general, explanation, makes the explanandum very stable.[47] The simpler the explanation, the fewer stars had to align in just the right way to make the explanandum occur. The same goes for more general explanations—the more circumstances it applies to, the easier it is that the explanandum occurs, given the explanans. This suggests a strategy for coming up with a diagnostic tool for good explanations: a way of assessing how well a given claim contributes to explaining another.

46 I have in mind a number of authors who have argued for the explanatory relevance of higher-level, or 'more abstract', properties—e.g. Garfinkel 1981, Jackson & Pettit 1988, 1990, Strevens 2004, 2008, Weslake 2010.

47 The notion of stability I am after is related to, although distinct from, the notion of resilience discussed in Jeffrey 1983 (when discussing the so-called paradox of ideal evidence), or Skyrms 1977, 1980. Their focus is on stability under conditionalization—or 'indicative supposition'. Mine is on stability under counterfactual supposition.

My suggestion, in a nutshell, is this: the contribution of p to having an explanation of e is proportional to the extent to which the truth of p would make e stable.[48]

One way to see that satisfying explanations increase the counterfactual stability of the explanandum is to think about laws of nature. Laws of nature have a high degree of counterfactual stability.[49] They are also some of the best candidates for explanatory bedrock. We all know the explanatory buck has to stop somewhere. We all agree that stopping at the laws of nature is as good a place as any. I say it is no coincidence that their high counterfactual robustness goes hand in hand with their not being in need of an explanation. It is because laws of nature are so counterfactually robust—because they would have obtained (almost) no matter what—that they do not cry out for explanation.[50]

Another way of motivating the connection between counterfactual stability and explanation is to reflect on the plausibility of so-called *contrastive* accounts of explanation.[51] The idea is simple: any request for explanation takes place against the backdrop of a contrast class. What we want out of an explanation of an event e is a story as to why e *rather than* some other member of the contrast class occurred. Now, the harder it is to find a natural contrast class, the harder it is to take e to be in need of explanation, on this way of thinking. And if e has a high degree of counterfactual stability, then the harder it is to think of e as crying out for an explanation.

Counterfactual stability is also a helpful diagnostic tool for simplicity, a highly plausible candidate for an explanatory virtue. The fewer variables are involved in an explanation, the more robust will the explanandum

48 Of course this will not do, as it stands, when it comes to low-probability events. But these are vexed issues far beyond the scope of this paper. See Woodward 2010 for discussion and references. For all I say in this paper, there may be other explanatory virtues that are not captured by the notion of counterfactual resilience. For my purposes, however, all I need is that there be an important dimension of explanatory value that is captured by the notion of counterfactual resilience.

49 Indeed, some would go so far as to use counterfactual stability in order to characterize what laws of nature are. See, e.g. Lange 2005, 2009.

50 This is not to say that we cannot explain a given law of nature. The same caveat listed in fn. 48 applies here.

51 See Garfinkel 1981, as well as van Fraassen 1980, Lipton 1990, *inter alia*.

be, and vice-versa. The fewer variables we need to fix for the explanation to go through, the more variables we can modify consistent with the explanandum obtaining. And every aspect of the situation that we can counterfactually modify without affecting the explanandum will plausibly correspond to a variable that is not involved in the explanation. Seeking explanations that make the explanandum counterfactually robust is thus likely to lead to simpler explanations. And explanations that make the explanandum counterfactually robust can be applied, *mutatis mutandis*, to many different circumstances: they are highly portable.

Consider again the explanation of the sequence of nearly ten tails in a row in terms of the coin's initial conditions (together with specification of the forces involved, wind conditions, etc.)—what I called INITIAL. Slight variations in the initial conditions would have made this explanation inapplicable: there are many ways things could have been—ways similar to the way things actually are—where the explanandum might have been false.

For example, for all the explanation tells you, if you had held the coin in a slightly different way in one of the tosses, the coin might have easily landed heads. Had someone sneezed nearby, altering the wind conditions, the outcome might have been different. In contrast, if you had held the coin slightly differently, then according to BIAS the coin would have still landed tails nearly every time. BIAS is applicable to many situations—it wears its portability on its sleeve—not just involving different coins and different initial conditions, but different processes involving binary random variables.[52]

It is hard to cash out the notion of counterfactual stability in a more precise way. The number of ways things might have turned out such that, for all that INITIAL says, the explanandum might have been false is infinite. But so is the number of ways things might have turned out such that, for all BIAS says, the explanandum might have been true. We cannot just *count* the relevant possibilities. And while it is in principle possible to provide a measure that would differentiate between the relevant infinite

52 There are some tricky issues I'm skating over. For example, one might think that counterfactuals of the form *If p had been false, the coin would have landed tails* cannot be true, since BIAS does not rule out entirely the possibility of the coin landing heads—cf. Hájek n.d. For our purposes, however, these complications are best set aside.

sets of possibilities, it is not obvious how to motivate one measure over another in a way that will work for all cases.

But assume we can agree on a finite set of relevant suppositions.[53] If the explanandum is more robust under counterfactual suppositions in that set relative to one body of beliefs than another, I submit, that would give us an epistemic reason (albeit a *pro tanto* one) for preferring the first body of beliefs over the other. This is not to say this is a reason for taking the first body of beliefs to be more accurate than the second one.[54] But it is surely a reason for favoring the one over the other when both are (expected to be) equally accurate.

For example, suppose you are interested in explaining why the outcome of the ten coin tosses is what it is. You are told your memory will be erased, but you will have some say on what credence function you will have afterwards. In particular, you are given the choice of waking up with a credence function that gets very close to the truth about the bias of the coin, and a credence function that gets very close to the truth about the initial position of each of the coin tosses.

If all else is equal, you will prefer the former over the latter. If you can only have a view on one of the two questions, you would rather know what the bias of the coin is than what the particular initial conditions of those ten coin tosses were. After all, you can expect to have more stable beliefs about the outcomes of the tosses of that coin if you know what its bias is than if you only know what the initial conditions of one particular sequence of tosses were. Holding fixed a class of explananda, having a counterfactually robust body of beliefs is better than not—and this, I submit, from an epistemic point of view.[55]

53 Assume further that all such suppositions should be treated equally, an assumption that we might want to relax at some point.

54 Although see White 2005, especially §3.

55 Note that the notion of robustness at issue is different from the one that figures in discussions of the value of knowledge inspired by the *Meno*, or in theories of knowledge that impose a so-called 'safety condition'. Your belief in p is counterfactually robust, in the sense that is relevant for our present purposes, if you think p would have happened no matter what. This can be so even though it is just a fluke that you happened to have that belief in the first place.

5.1 *Measuring explanatory potential*

We want to compare how much a given proposition would contribute to making e well-explained relative to a given body of beliefs—a credence function. This, I have suggested, can be done by measuring how learning that proposition would increase the counterfactual robustness of e relative to that credence function.

Relative to a particular set of suppositions B, we can measure the counterfactual robustness of e relative to a credence function C as a function of the average amount of variation between $C(e)$ and $C(b \mathrel{\square\!\!\rightarrow} e)$ for $b \in B$ (as long as $C(b \mathrel{\square\!\!\rightarrow} e)$ is well-defined for each $b \in B$).[56]

There are different ways of measuring the relevant variation. To fix ideas, let us use the most straightforward option: the counterfactual resilience of e, relative to a credence function C and a finite set of suppositions B is given by the average difference between $C(e)$ and $C(b \mathrel{\square\!\!\rightarrow} e)$, for $b \in B$.[57]

DEFINITION 5.1. The *counterfactual resilience* of e relative to a credence function C (and a set of suppositions B), $\mathbf{r}_B(C, e)$, is given by:

$$\mathbf{r}_B(C, e) = 1 - \frac{1}{|B|} \sum_{b \in B} \left| C(e) - C(b \mathrel{\square\!\!\rightarrow} e) \right|.$$

56 One must tread carefully here. Williams 2012 argues that one cannot identify the credence in $\phi \mathrel{\square\!\!\rightarrow} \psi$ with the credence one assigns to ψ on the counterfactual supposition that ϕ. For reasons carefully laid out in Schwarz 2016, I think the argument fails. So I will proceed henceforth on the assumption that whenever a and b are propositions, $a \mathrel{\square\!\!\rightarrow} b$ is a proposition. I will postpone the question of how to assign credence to counterfactuals until Appendix D, and note here that everything I say here can be reformulated in terms of *imaging* (cf. Gärdenfors 1982, Joyce 1999, Lewis 1976) rather than in terms of credences in counterfactual propositions.

57 This measure is thus simimlar to Skyrm's notion of resilience (Skyrms 1977, 1980, *inter alia*), which is defined in terms of conditional probabilities instead of probabilities of counterfactuals, and is essentially the same as

$$1 - \max_{b \in B} |C(e) - C(e \mid b)|.$$

An alternative measure of resilience in Skyrm's sense would equal one minus the *average* difference between $C(e)$ and $C(e \mid b)$ as opposed to one minus the *maximum* difference between $C(e)$ and $C(b \mid e)$.

Thus, $\mathbf{r}_B(C, e)$ is a number between 0 and 1 that increases to the extent that the average amount of variation between $C(e)$ and $C(b \mathrel{\square\!\!\rightarrow} e)$ decreases. How resilient e is, relative to your credence function, is thus meant to capture how well-explained you take e to be. (In what follows, I will drop explicit relativization to the set of suppositions, and assume throughout that a given explanatory context fixes such a set.)

Note that the resilience of e relative to your credence function says nothing about what you take the explanation of e to be. In other words, the claim

> EXPLANATION PROVIDES RESILIENCE (EPR): $\mathbf{r}(C, e)$ measures how well-explained an agent whose credence function is C takes e to be.

is fairly neutral as to what the correct theory of explanation is. After all, EPR is a claim about how well-explained someone with a given credence function takes e to be. So unless we grant some additional, and by no means obvious, assumptions—say that someone takes e to be well-explained to the extent that they assign high credence to the true answer to the question *what explains e?*—EPR is compatible with any theory of what, if any, is the correct explanation of e. Furthermore, EPR does not by itself entail anything about what is the explanation of e, even by the lights of an agent who takes e to be well-explained: for all EPR says, an agent can take e to be well-explained even if there is no answer to the question *what explains e according to her?*

At the same time, EPR is not completely neutral on some well-known questions about explanation. For instance, according to EPR, there is a sense in which high-level explanations are preferable, all else equal, to low-level explanations. This can be seen more clearly if we focus on causal explanations.

A high-level explanation will abstract away from more details of the causal history of a particular explanandum. As such, it will be applicable to a wider variety of events that differ from the target explanandum in the details of its causal history. So, all else equal, a high-level explanation will make the explanandum more stable than a lower-level one.[58]

58 I am making some nontrivial assumptions here on how the truth-value of certain counterfactuals is affected by the features of a particular explanatory context. In particular, I am assuming that in a context where we are given every single detail of the causal history

Still, as stated EPR is highly schematic. For any credence function C and any proposition e we can find a set of suppositions relative to which $\mathbf{r}(C, e)$ takes its maximum value (i.e. 1). If my purpose here were to provide an account (reductive or not) of what makes for a good explanation, I would be in urgent need of a way of specifying a set of suppositions for any given explanandum. But that is not my purpose here. Rather, I want to offer a way of building in some of our views on explanatory value into the framework of epistemic utility theory in order to see how a more flexible theory of Bayesian dynamics could go. Thus, my hope is that we can agree, in any particular case, on a suitable class of suppositions, and look at the consequences that would have for questions about the rationality of conceptual change.

5.2 *Combining accuracy and resilience*

Suppose we fix a particular explanandum e and a given credence function C. Further suppose λ^e_C is a function which assigns to each s a positive number that increases as the amount of change in resilience of e that would come from updating C on s increases. Then, for any local accuracy measure δ,

$$\delta_C(P, x) = \sum_{s \in S_P} \lambda^e_C(s) \cdot \delta(P(s), \mathbb{1}\{x \in s\})$$

is a weighted accuracy measure, whose weight function measures the extent to which s contributes to explaining e relative to C. Moreover, if δ is (strictly) proper, then δ_C will be partition-wise (strictly) proper.[59]

---

of some event e, counterfactuals of the form 'if d had not obtained, e would still have obtained', where d is some particular detail that figures in the putative explanation in question, are not true. Thus, even if a low-level explanans e logically entails a high-level explanans E, it could be that the explanatory context set up by e, the closest not-d worlds (where d is incompatible with e but compatible with E) will include not-E worlds. This is by no means uncontroversial, but it will not be needed in what follows. If there are readings of the relevant counterfactuals on which lower-level explanations come out as providing more stability than higher-level ones, that makes EPR less controversial than I'm taking it to be.

59 There is a small wrinkle here I'm ignoring for now. We want perfect accuracy with respect to more important propositions to count for more than perfect accuracy with respect to less important propositions. If our local accuracy measure is always non-positive, then we will want the weight assigned to a proposition to be *smaller* the more

For concreteness, let us define, for each $s \in \pi$,

$$\chi^e_C(s) = 1 + \mathbf{r}(C_s, e),$$

where $C_s = C(\cdot \mid s)$. For a given s, then, $\chi^e_C(s)$ will be a number between 1 and 2 that increases as the stability of e relative to C_s increases. For a fixed C and e, we can now define an epistemic utility function—one which is sensitive to both accuracy and explanatory considerations—by combining our χ^e_C function with any strictly downwards proper local accuracy measure δ:

$$\varepsilon^\delta_C(P, x) = \sum_{s \in \pi_P} \chi^e_C(s)\delta(P(s), \mathbb{1}\{x \in s\})$$

Note that, unlike any other epistemic utility function we've discussed thus far, ε^δ_C is defined in terms of a particular probability function C. This is to be expected. After all, the weights given to each of the propositions in question are supposed to measure how much learning the truth about them contributes to the explanatory closure *of a particular body of beliefs*. If we think of an agent's epistemic utility function as something determined by her epistemic values, we can think of an agent with credence function C who uses ε^δ_C as her epistemic utility function as someone who values accuracy as well as the resilience of her body of beliefs—which, given EPR, would be tantamount to valuing both accuracy and explanatory closure.[60]

It is also worth noting that, because ε^δ_C is defined in terms of a fixed probability function, it can be partial: it may not be defined for all pairs

---

important the proposition. If instead our local accuracy measure is always non-negative, we will want the weight assigned to a proposition to be greater the more important the proposition. (Things get even trickier if our local accuracy measure is sometimes positive and sometimes negative.) Fortunately, if we assume that our local accuracy measure is downwards proper—in a sense to be made precise—and that it is 0/1-symmetric—in that $\delta(x, 1) = \delta(1 - x, 0)$ for all $x \in [0, 1]$, our local accuracy measure will be non-negative. See the Remark C.4.

60 We could, of course, find some fully objective way of assigning to each proposition a measure of its explanatory potential—cf. the discussion of weighted accuracy measures in REDACTED FOR BLIND REVIEW. Doing so would allow us to define an epistemic utility function that is sensitive to explanatory considerations in a way that does not depend on the agent's credence function. Further investigation may result in one such measure of explanatory worth, but for our purposes, I will set that strategy aside.

consisting of an arbitrary probability function and a possible world. Specifically, $\varepsilon^{\delta}_{C}(P, x)$ will be well-defined only if $\mathbf{r}(C_s, e)$ is well-defined for all $s \in \pi_P$.

Now, $\mathbf{r}(C_s, e)$ may not be well-defined for one of two reasons. First, it may be that C_s is well-defined, but that $C_s(b \boxdot\!\!\rightarrow e)$ is not, for some relevant b. Second, it may be that C_s is not well-defined, perhaps because $C(s)$ is not well-defined. Fortunately, in the cases that will be of interest—when using ε^{δ}_{C} to compare expansions of C—our epistemic utility function will always be well-defined. In the remainder of this section, I briefly explain why, leaving the details to an appendix.

I start by assuming that counterfactual conditionals are only understood against the backdrop of a *similarity function*—a function σ that assigns, to each proposition a and possible world w, the set $\sigma(a, w)$ of those a-worlds that are most similar to w. The counterfactual conditional $a \boxdot\!\!\rightarrow_{\sigma} b$ is thus the set of all worlds w such that all worlds in $\sigma(a, w)$ are b-worlds. (Henceforth, I will drop the relativization to σ and assume one such function is fixed throughout.)

Now, recall that from the perspective of an agent whose state space is π, we can think of each $s \in \pi$ as playing the role of a possible world—the elements of π are each maximally consistent relative to the agent's credence function, in that each proposition in the agent's credence function is entailed by or inconsistent with some member of π. Thus, from the perspective of such an agent, we can think of the counterfactual conditional as the proposition that is true in all and only those $s \in \pi$ such that their 'closest worlds' that make a true, also make b true. Intuitively, we can think of $\sigma_{\pi}(a, s)$ as the set of those elements of π that entail a and which are most like s *in all similarity respects the agent can entertain*. For a given partition π, $(a \boxdot\!\!\rightarrow b)^{\pi}$ is thus the union of the set of atoms $s \in \pi$ such that their most similar a-atoms all entail b. In other words, $(a \boxdot\!\!\rightarrow b)^{\pi}$ will be the weakest proposition in the Boolean closure of π that entails $a \boxdot\!\!\rightarrow b$.

In general, $(a \boxdot\!\!\rightarrow b)^{\pi}$ and $(a \boxdot\!\!\rightarrow b)$ are different propositions.[61] After all, $(a \boxdot\!\!\rightarrow b)^{\pi}$ is a proposition that is 'visible' in the algebra generated by π, even though $(a \boxdot\!\!\rightarrow b)$ may not be.

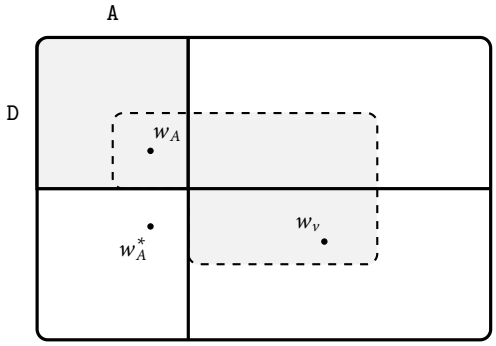


Figure 3: Think of the solid lines as tracing the partition π that generates the algebra containing the proposition that Alice enters the room on Thursday (here labeled as A) and the proposition that Alice had the disease on Friday (here labeled as D). Relative to π, the counterfactual $\mathtt{A} \boxdot\!\!\rightarrow \mathtt{D}$ ends up being equivalent to $\mathtt{A} \cap \mathtt{D}$. The area traced by the dashed line corresponds to the proposition V that there was a virus in the room. Relative to the refinement π' of the initial partition that contains this new proposition, the counterfactual $\mathtt{A} \boxdot\!\!\rightarrow \mathtt{D}$ ends up expressing the proposition corresponded to the grayed out region—i.e. it turns out to be equivalent to $(\mathtt{A} \cap \mathtt{D}) \cup \mathtt{V}$. Consider now w_v—a $\neg\mathtt{A} \cap \neg\mathtt{D} \cap \mathtt{V}$ world. If we only consider similarity respects available in π, then w_A and w_A^* are both among the closest A-worlds to w_0. If instead we consider similarity respects available in π', w_A is a closest A-world to w_v, but w_A^* is not.

This is at it should be. We can think of $(a \boxdot\!\!\rightarrow b)^{\pi}$ as the proposition that plays the role of $(a \boxdot\!\!\rightarrow b)$ in an agent whose conceptual resources are given by π. This proposition will be true at a world just in case the a-worlds that most agree with that world in all similarity making respects *that the agent can entertain* are b-worlds. And it may not be true at a world even though the a-world that most agrees with that world in *all* similarity making respects *simpliciter* is a b-world. (See Figure 3.)

Take, for example, the following counterfactual:

> CAUGHT: If Alice had entered the room, she would have caught the disease.

61 To avoid clutter, I will no longer make explicit the relativization to σ in what follows, unless it matters. For our purposes, we can assume that we have fixed a selection function—perhaps the one corresponding to what Lewis 1979 calls the 'standard resolution' of the 'vagueness' of counterfactuals.

And suppose π is a partition that does not distinguish between worlds that differ only at the microscopic level (see Figure 3). Consider now some possible world w_v in which there is some virus in the air inside the room. Presumably, CAUGHT is true at that world: on a standard interpretation, CAUGHT is true at w_v just in case the world most similar to w_v in which Alice enters the room is one in which she catches the disease. Since such a world—call it w_A—will agree with w_v as to the presence of the virus, CAUGHT is true in w_v (at least if we assume that Alice's absence is causally independent of the presence of the virus). But consider now all worlds that are closest to w_v in all respects that are visible in π in which Alice enters the room. Among those worlds is w_A. But so is w_A^*, a world that differs from w_A before the time at which Alice enters the room only at the microscopic level, and in which there is no virus in the room. Since the difference between w_A and w_A^* is not visible in π, both worlds will count as being equally close *in all respects visible in* π. As a result, since presumably in w_A^* Alice does not catch the disease, if we restrict ourselves to similarity respects that are available in π, CAUGHT comes out false.

In short, there is a principled sense in which, for each agent with state space π and a and b in their algebra $\mathcal{B}$, there will be a proposition $(a \boxdot\!\!\rightarrow b)^\pi$ that plays the role of $a \boxdot\!\!\rightarrow b$ for that agent. It's the proposition that will be at a world $w \in s$, with $s \in \pi$, just in case the closest $s' \in \pi$ that entails a also entails b.

The second reason $\varepsilon_C^\delta(P, x)$ could fail to be well-defined, recall, is that $\mathbf{r}(C_s, e)$ may not well-defined if $C(a \boxdot\!\!\rightarrow e \mid s)$ is not well-defined for some $s \in \pi_P$. In assessing the epistemic value of an extension P, you need to assign weights to propositions in π_P. These weights in turn depend on your credence, given $s \in \pi_P$, of some counterfactuals. Since P is an extension of your credence function, though, we run into a problem. For whenever q is not in the domain of your credence function, your credence conditional on q may not be well-defined—not if we assume, as I have been throughout, that the conditional probability of q given p is defined in terms of the ratio of your credence in qp and your credence in p. Fortunately, if p is an element of π_P, any proposition in the domain of P is either entailed by p or it is incompatible with p. Thus, for any probability function defined over π_P and any q, the probability of q given p, for $p \in \pi$, will be 0 or 1, depending on whether $p \subseteq q$.

The upshot of these two claims—which I argue for in more detail in Appendix D—is this. Say that P is *fully opinionated* iff it assigns 0

or 1 to every proposition in its domain. If P is fully opinionated, there is a unique $s \in \pi_P$ such that $P(s) = 1$—in that case, we will say that P is *concentrated* on s. As I show in Appendix D, our two observations ensure that whenever an agent with credence function C is considering an extension C' of C, $\chi^e_C(s')$ will in fact be well-defined, for each $s' \in \pi'$, since $\mathbf{r}(C_{s'}, e)$ will be. In particular, we can establish the following remark, which will come in handy in what follows:

REMARK 5.2. Suppose P is fully opinionated and concentrated on s. If $P(e) = 1$ then

$$\mathbf{r}(P, e) = 1 - \frac{1}{|B|} \sum_{b \in B} \left|1 - \mathbb{1}\{\sigma(b, s') \subseteq e\}\right|$$

In other words, the resilience of e relative to P) is determined by the proportion of $b \in B$ such that s_P entails $b \boxdot\!\!\rightarrow e$.

It is also worth highlighting the following consequence of Remark 5.2:

COROLLARY 5.3. *Suppose P and Q are fully opinionated and defined over the same partition π. If $P(e) = Q(e) = 1$, then $\mathbf{r}(P, e) = \mathbf{r}(Q, e)$.*

6 THE RATIONALITY OF CONCEPTUAL CHANGE

Let me take stock. I have taken for granted a particular picture of epistemic rationality. On this picture, epistemic rationality is a matter of maximizing expected epistemic value. It is well known that, if we impose some minimal constraints on the notion of epistemic value, this picture allows to recover many of the norms of a broadly Bayesian picture of rationality.

I have argued that we can generalize familiar ways of thinking about epistemic value so as to allow for comparisons of credence functions defined over different hypothesis spaces. The strategy I recommended—if only as a proof of concept—was this. First, use *weighted accuracy* measures as epistemic utility functions—where these are functions that treat accuracy with respect to a proposition in a way that depends on what a given weight function assigns to that proposition (§4.1). Second, use weight functions that assign to the relevant proposition something that measures its *explanatory value*—where this is something that is meant to capture

how much it would contribute, if true, to having a good explanation of an antecedently given explanandum (determined by the relevant agent's explanatory goals). Third, measure the explanatory value of a given proposition using the notion of *counterfactual resilience*—the explanatory value of a given proposition is determined by how much learning that proposition would increase the counterfactual stability of the explanandum (§5.1). Crucially, how much a given proposition would increase the stability of another proposition depends on what other propositions are available to the agent—for it is those propositions that determine what the similarity respects used to evaluate counterfactuals are (§5.2). The resulting strategy allows us to compare the epistemic value of different expansions of an agent's function in terms of how much the added proposition would contribute to the agent's explanatory goals.

My ultimate goal, however, is to show that this generalization provides us with the resources to make sense of questions about the rationality of conceptual change. And to do this, we need to specify some *bridge principle* yielding norms of epistemic rationality from facts about epistemic value.

In the simple case when we're comparing choices among credence functions with the same domain, the familiar injunction to *maximize expected value* is a reasonable bridge principle. As we saw in §3.3, though, once we allow for options whose domains are refinements of the domain of our prior credence function, the notion of expected value is not well-defined. Instead, we have two related but distinct quantities—the *upper expected value* and the *lower expected value* (see Definition 3.6)—and at least two distinct but related bridge principles—*maximize upper expected value* (maximax) and *maximize lower expected value* (maximin).

To make things slightly more concrete, suppose we could argue that the right way to pick an expansion of P to some refinement π of π_P is to pick the credence function $\overline{P}_\pi$ with domain π that maximizes $\overline{\mathbb{E}}_P[u(\cdot)]$ for a given utility function u. In order to compare two distinct refinements π and π' of π_P from P's perspective, we could then simply compare the values of $\overline{\mathbb{E}}_P[u(\overline{P}_\pi)]$ and $\overline{\mathbb{E}}_P[u(\overline{P}_{\pi'})]$. Suppose instead we agreed that the right way to pick an expansion of P to π is to pick the credence function $\underline{P}_\pi$

that maximizes $\underline{\mathbb{E}}_P[u(\cdot)]$ for a given u. Then we could compare distinct refinements π and π' by comparing $\underline{\mathbb{E}}_P[u(\underline{P}_\pi)]$ and $\underline{\mathbb{E}}_P[u(\underline{P}_{\pi'})]$.[62]

I do not here have a case to offer in favor of one or another such bridge principle. To that extent, at least, this is just a progress report. Fortunately, sometimes it will turn out that no matter which way we go, one partition comes out ahead of another one. And in those cases, we can avoid settling the vexed question how to assign credence to new propositions and still declare a choice of one partition over another to be rational. Our two toy examples can illustrate this point.

As we will see, our strategy for both examples will rely on two observations. The first is that, for each C and $e \in \mathcal{B}_C$, χ^e_C has the following monotonicity property relative to any extension of C:

REMARK 6.1. Suppose C' is an extension of C, $s \in \pi_C$ and $s' \in \pi_{C'}$. If $s' \subseteq s$ then $\chi^e_C(s) \leq \chi^e_C(s')$.[63]

This ensures that, as long as our local accuracy measure is downwards proper, so will be ε^e_C.[64]

Our second observation requires a little more setup. Suppose your credence function C is defined over π and suppose π_1 and π_2 are two refinements of π. Fix a weight function λ and assume that for each $s \in \pi$ that is partitioned by π_1, there is an equally likely (by your own lights) $s' \in \pi$ that is partitioned by π_2 into the same number of cells. Further suppose that π_1 splits s into propositions whose weight is at least as high as those into which π_2 splits s', in the sense that you can send each member t of π_1 that is included in s into exactly one member t' of π_2 that is included in s', so that the λ weight of t is at least as great, and sometimes greater, than that of t'. Then, we say that C and λ rank π_1 over π_2. More precisely:

62 Again, these are not the only possible ways to go about choosing a distribution over a refinement of a prior credence function. One could assign, for example, to each such distribution a weighted average of its upper and lower expected values, for fixed weights α and β (this is the so-called Hurwicz criterion, after Hurwicz 1951). Here, however, I will restrict my attention to the two simpler bridge principles, and simply note that in the cases we will be concerned with, the Hurwicz criterion would yield exactly the same results. I do not know what would happen if we relied on other bridge principles instead (e.g. minimax regret).

63 Cf. Remark D.10.

64 See Fact B.9.

DEFINITION 6.2. If π_1 and π_2 are two refinements of π_P we say that λ and P *rank* π_1 over π_2 iff there is a bijection $f : \pi_1 \longrightarrow \pi_2$ such that:

- for all $s \in \pi_1$, $\lambda_{\pi_1}(s) \geq \lambda_{\pi_2}(f(s))$,
- for some $p \in \pi_P$,

$$\max_{s \subseteq p, s \in \pi_1} \lambda_{\pi_1}(s) > \max_{s \subseteq p, s \in \pi_1} \lambda_{\pi_2}(f(s)),$$

- for all $p \in \pi_P$,

$$\bigcup_{\substack{s \subseteq p,\\ s \in \pi_1}} f(q) \in \pi_2,$$

- for all $p \in \pi_P$,

$$P(\bigcup_{\substack{s \subseteq p,\\ s \in \pi_1}} q) \geq P(\bigcup_{\substack{s \subseteq p,\\ s \in \pi_1}} f(q)).$$

Our final observation is that whenever λ and P rank π_1 over π_2, it makes sense to adopt π_1 rather than π_2 no matter which of the bridge principles under consideration turns out to be the correct bridge principle. In other words:[65]

FACT 6.3. *Suppose u is a weighted accuracy measure with weight function λ and local accuracy measure δ that is a downwards proper, affine transformation of the Brier local accuracy measure. For any two refinements π_1 and π_2 of P, if λ and P rank π_1 over π_2, then*

1. $\max_{C \in \mathcal{P}_{\pi_1}/_P} \overline{\mathbb{E}}_P[u(C)] > \max_{C \in \mathcal{P}_{\pi_2}/_P} \overline{\mathbb{E}}_P[u(C)]$, *and*
2. $\max_{C \in \mathcal{P}_{\pi_1}/_P} \underline{\mathbb{E}}_P[u(C)] > \max_{C \in \mathcal{P}_{\pi_2}/_P} \underline{\mathbb{E}}_P[u(C)]$.

□

What we want to show, then, is that in each of our toy examples, introducing the new propositions you did made epistemic sense. This is

65 Cf. Corollary C.16. In the terminology of the appendix, λ and P ranks π_1 over π_2 iff there is some π that is π_P-equivalent to π_1 and λ-P-equivalent to π_2 such that λ favors π_1 over π.

because, regardless of how exactly we think you should assign probabilities over new propositions, refining the way your did was epistemically rational.

Our strategy will be to show that, in each case, one of the atoms in the posterior credence function gets a strictly greater weight—has more explanatory value than—the atom of the prior credence function that it entails. This will suffice to show that your epistemic transition was rational, at least when your options were your posterior credence function and any refinement π^* of your prior such that (a) for each atom of your prior that is split by your posterior, there is an equally probable atom in π^* that is split into the same number of elements, and (b) none of the atoms of π^* get greater weight than the atom of your prior that they entail. This is because of the following simple fact, which follows from Remark D.10 in Appendix D.

REMARK 6.4. Fix a credence function C with state space π, fix some explanandum E, and let π_1 and π_2 be two refinements of π such that there is a bijection $f : \pi_1 \longrightarrow \pi_2$ such that, for all $s \in \pi$ and all $t \in \pi_1$,

$$t \subseteq s \Leftrightarrow f(t) \subseteq s.$$

Suppose for each $s \in \pi$ and each $t \in \pi_2$ with $t \subseteq s$, $\chi^E_C(s) = \chi^E_C(t)$ and suppose for some $s \in \pi$ there is $t \in \pi_1$ with $t \subseteq s$ and $\chi^E_C(t) > \chi^E_C(s)$. Then C and χ^E_C rank π_1 over π_2.

6.1 *The Reds, revisited*

Your Monday credence function, recall, was defined over the Boolean closure of propositions of the form: *c moves faster after exposure to blue light*, *c moves faster after exposure to red light*, *c's behavior is unaffected by exposure to blue light*, *c's behavior is unaffected by exposure to blue light*. Your Tuesday credence function was one whose domain is the result of adding propositions of the form *c is in state R* and (again) closing under Boolean operations. We will only focus on propositions involving a particular red, call it c_0.

Call $\mathtt{F_r}$ (resp. $\mathtt{F_b}$) the proposition that c_0 moves faster after exposure to red (resp. blue) light, and let $\mathtt{R}$ denote the proposition that c_0 is in state R. For simplicity, let us assume that the only change in behavior we are interested in is whether c_0 moves fast, so that $\mathtt{F_b}$ is equivalent to

the negation of c_0*'s behavior is unaffected by blue light* and similarly for $\mathtt{F_r}$. Let us further assume that, together with everything else you believe, $\mathtt{R}$ entails (but is not entailed by) $\mathtt{F_b} \cup \mathtt{F_r}$ (see Figure 4). Finally, assume $\mathtt{E} := \mathtt{F}_b$ is the proposition you are interested in explaining (the reasoning below can be reused, *mutatis mutandis*, to establish the same conclusion on the assumption that it is $\mathtt{F_r}$ instead you are interested in explaining).

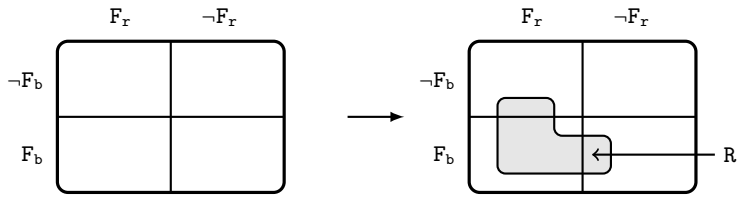


Figure 4: The transition from your Monday credence function to your Tuesday credence function. After adding the proposition $\mathtt{R}$ and closing under Boolean operations, the result is the algebra generated by the state space that results from taking the state space of your Monday credence function and replacing each of $\mathtt{F_b} \cap \mathtt{F_r}$, $\mathtt{F_b} \cap \neg\mathtt{F_r}$, $\neg\mathtt{F_b} \cap \mathtt{F_r}$, and $\neg\mathtt{F_b} \cap \neg\mathtt{F_r}$, with their intersections with $\mathtt{R}$ and with $\neg\mathtt{R}$. Note that for worlds in $\mathtt{F_{br}} \cap \mathtt{R}$, their closest $\neg\mathtt{F_r}$ worlds are also $\mathtt{R}$ worlds. This means that the atom $\mathtt{F_{br}} \cap \mathtt{R}$ entails the counterfactual $\neg\mathtt{F_r} \boxdot\!\!\rightarrow \mathtt{F_b}$, which $\mathtt{F_{br}}$ does not (since there are $\mathtt{F_{br}}$ worlds whose closest $\neg\mathtt{F_r}$) are $\neg\mathtt{F}_b$ worlds.

Call C your Monday credence function, let π denote C's state space, and let π' denote the coarsest refinement of π whose Boolean closure includes $\mathtt{R}$. Assume C is concentrated on the proposition $\mathtt{F_{br}} = \mathtt{F_b} \cap \mathtt{F_r}$, reflecting the fact that you are certain that c_0 moves faster after exposure to blue and red lights.

We want to show that for some $s' \in \pi'$, $s \in \pi$, $s' \subseteq s$, $\chi_C^{\mathtt{E}}(s') > \chi_C^{\mathtt{E}}(s)$. In other words, that some proposition in π' has greater weight than the proposition in π that encloses it—or, in yet other words, that some proposition in π' has more explanatory value than the proposition in π it entails.

Let $\mathtt{X} = \mathtt{F_{br}} \cap \mathtt{R}$. We want to show that $\mathtt{X}$ has greater weight than $\mathtt{F_{br}}$: roughly, that upon being certain that $\mathtt{F_{br}}$ is true, learning further that $\mathtt{R}$ is true would put you in a better position to explain $\mathtt{E}$ (that is, $\mathtt{F_b}$). Since $\mathtt{X} \subseteq \mathtt{F_{br}}$, $\mathtt{X}$ entails any counterfactual that is entailed by $\mathtt{F_{br}}$. So in order to show that $\chi_C^{\mathtt{E}}(s') > \chi_C^{\mathtt{E}}(s')$, we need to show that there is some relevant counterfactual of the form $b \boxdot\!\!\rightarrow \mathtt{E}$ that is entailed by $\mathtt{X}$ but not by $\mathtt{F_{br}}$. For this would show that the proportion of relevant counterfactuals entailed by $\mathtt{X}$ is greater than the proportion of those entailed by $\mathtt{F_{br}}$, which given Remark 5.2, is enough to show that $\chi_C^{\mathtt{E}}(\mathtt{X}) > \chi_C^{\mathtt{E}}(\mathtt{F_{br}})$.

Take an off-the-shelf theory of similarity, like the one in Lewis 1979. Clearly, $\neg \mathtt{F_r}$-worlds in which $\mathtt{R}$ is true—worlds in which c_0 does not move faster when exposed to red light but in which c_0 is in state R—are closer to any $\mathtt{X}$-world than $\neg \mathtt{F_r}$-worlds in which $\mathtt{R}$ is not true. After all, $\mathtt{R}$ concerns matters of particular fact prior to the time at which $\mathtt{F_r}$ is true.[66] Thus, $\mathtt{X}$ entails $\neg \mathtt{F_r} \boxdot\!\!\rightarrow \mathtt{E}$, since $\mathtt{R}$ entails $\mathtt{F_b} \cup \mathtt{F_r}$ and thus any $\mathtt{R}$-world in which $\neg \mathtt{F_r}$ is true is an $\mathtt{F_b}$-world (since c_0 is in state R, worlds in which c_0 is in state R are closer to the actual world than worlds in which it isn't; thus, the closest worlds in which c_0 does not move faster when exposed to red light are those in which he is also in state R, which means that they are worlds in which he moves faster when exposed to blue light). In contrast, $\mathtt{F_{br}}$ does not entail $\neg \mathtt{F_r} \boxdot\!\!\rightarrow \mathtt{E}$, for there are $\mathtt{F_{br}}$-worlds (namely $\neg \mathtt{R}$-worlds) such that their closest $\neg \mathtt{F_r}$-worlds are not $\mathtt{F_b}$-worlds.

Consider now any other refinement of π which splits the same cells as π' does, and in the same number of ways, but whose weights do not exceed the weights of their corresponding C-cells. For example, consider the refinement of C that results from adding the proposition, call it $\mathtt{T}$, that c_0 was tired after moving faster from exposure to either blue light or red light. Like $\mathtt{R}$, $\mathtt{T}$ entails (given everything else you believe) $\mathtt{F_b} \cup \mathtt{F_r}$.[67] And it is compatible with but logically independent of any of $\mathtt{F_{br}}$, $\mathtt{F_r} \cap \neg \mathtt{F_b}$, and $\neg \mathtt{F_r} \cap \mathtt{F_b}$. Plausibly, however, for any $\mathtt{X}$ of the form $\mathtt{T} \cap s$, with $s \in \{\mathtt{F_{br}}, \mathtt{F_r} \cap \neg \mathtt{F_b}, \neg \mathtt{F_r} \cap \mathtt{F_b}\}$, the closest $\neg F_r$ worlds to $\mathtt{X}$-worlds include both worlds in which $\mathtt{E}$ is true and worlds in which it isn't. (After all, had $\neg \mathtt{F_r}$ been true, $\mathtt{T}$ might have been false, perhaps because c_0 wouldn't have moved at all.) Similar reasoning shows that, for any relevant supposition b and any $s \in \pi_C$, if s does not entail $b \boxdot\!\!\rightarrow \mathtt{E}$, neither does $\mathtt{T} \cap s$. As a result, for each atom t of this new refinement (call it π^*), $\chi_C^{\mathtt{E}}(t) = \chi_C^{\mathtt{E}}(s)$, where $s \in \pi_C$ and $t \subseteq s$. This shows that $\chi_C^{\mathtt{E}}$ ranks π' over π^*, and thus that regardless of which bridge principle we use, expanding to π' rather than π^* would be rational.

66 I'm relying on context to provide the relevant time indices here. Strictly speaking, we should let $\mathtt{R}$ be the proposition that c_0 is in state R at t_0, $\mathtt{F_r}$ the proposition that c_0's behavior at t_1 is unaffected by red light, etc.

67 I am ignoring the possibility that $\mathtt{F_b} \cup \mathtt{F_r}$ is not entailed but merely presupposed by $\mathtt{T}$. Adjust accordingly if you think that's something we should not ignore.

6.2 *Cubes and spheres, revisited*

Let C denote your Wednesday credence function, which was defined over the Boolean closure of propositions of the form: *c is a cube*, *c is a sphere*, *c is an n^{th}-generation critter*, and *a and b are c's parents*. Let $\pi = \pi_C$ and let π' denote the state space of your Thursday credence function, whose domain results from enlarging $\mathcal{B}_C$ by adding propositions of the form *c is a hybrid cube* and *c is a pure cube* and closing under Boolean operations.

Take two second-generation cubes with mixed parents. Let $\mathtt{E}$ be the proposition that their offspring is 75% cube and 25% sphere and assume $C(e) = 1$. Again, we want to show there is one $s \in \pi$ and some $s' \in \pi'$ with $s' \subseteq s$ such that $\chi^{\mathtt{E}}_C(s') > \chi^{\mathtt{E}}_C(s)$—that some proposition in π' has more explanatory value than the proposition in π it entails.

As before, this requires showing that there is $s' \in \pi_{C'}$, $s \in \pi_C$, with $s' \subseteq s$, and some relevant proposition b such that s' entails $b \boxdot\!\!\rightarrow \mathtt{E}$ but s does not. In other words, a proposition which, were you to update on it, the resilience of $\mathtt{E}$ relative to your credence function would increase.

Let $\mathtt{M}$ be the proposition that the cubes are second generation cubes with mixed parents and let $\mathtt{H}$ be the proposition that they are second generation hybrid cubes. (Note that, by construction, $\mathtt{M}$ is an atom of π and $\mathtt{H}$ is an atom of π'.) Let $\mathtt{T}$ be the proposition that the cubes are third-generation cubes. We want to show that, while $\mathtt{H}$ entails $\mathtt{T} \boxdot\!\!\rightarrow \mathtt{E}$, $\mathtt{M}$ does not. On the assumption that $\mathtt{T}$ is a relevant proposition, this would establish that $\chi^{\mathtt{E}}_C(\mathtt{H}) > \chi^{\mathtt{E}}_C(\mathtt{M})$—roughly, that learning $\mathtt{H}$ even after being certain of $\mathtt{M}$ would increase the resilience of $\mathtt{E}$, and thus leave you with a better explanation of $\mathtt{E}$.

Note now that there are $\mathtt{M}$-worlds in which one of our cubes is not a hybrid cube, and thus there are $\mathtt{T}$-worlds closest to them in which their offspring is not 75% cube—which is to say that $\mathtt{M}$ does not entail $\mathtt{T} \boxdot\!\!\rightarrow \mathtt{E}$. In contrast, $\mathtt{T} \cap \mathtt{M}$-worlds are closer to any $\mathtt{H}$-world than any $\mathtt{T} \cap \neg\mathtt{M}$-worlds. But given everything else you believe, $\mathtt{M}$ entails $\mathtt{E}$, which means that the closest $\mathtt{T}$ worlds—which will be $\mathtt{T} \cap \mathtt{M}$ worlds—will also be $\mathtt{E}$ worlds. We thus have that $\mathtt{H}$ entails $\mathtt{T} \boxdot\!\!\rightarrow \mathtt{E}$, as desired.

Contrast now π' with another refinement of π, call it π^*: this is the result of adding to $\mathcal{B}$ the proposition $\mathtt{B}$ that one of each cube's parents is a blue sphere. Let $\mathtt{S}$ be the proposition that one of each cube's parents is a blue sphere. Plausibly, for any relevant supposition b and any $s \in \pi$, if $\mathtt{E} \cap s$ entails $b \boxdot\!\!\rightarrow \mathtt{E}$, so does $s \boxdot\!\!\rightarrow \mathtt{E}$. Thus, for each $s \in \pi$, $\chi^{\mathtt{E}}_C(s) = \chi^{\mathtt{E}}_C(s \cap \mathtt{B})$.

Again, this shows that expanding to π' is better than expanding to π^*, regardless of which bridge principle we use.

7 CLOSING

Bayesian epistemology has long remained silent on questions about how best to carve up the space of hypotheses we use in theorizing about the world. But it need not: there is a natural way of generalizing the classical Bayesian framework so as to formulate and perhaps answer this question. Doing so requires thinking of epistemic rationality as having a broadly decision theoretic structure. The key is to allow for a more expansive notion of epistemic value to play the role of 'epistemic utility'—more expansive, that is, than the accuracy-centered approach that has dominated work on epistemic utility theory thus far. I suggested a simple way to do this: we should use *weighted* accuracy measures, where the weight assigned to a proposition corresponds to how epistemically important it is. And I offered a proof of concept: a way of assigning weights to propositions that measures their *explanatory value*. The resulting picture is conservative with respect to standard, Bayesian epistemology. But it offers answers to questions that cannot even be formulated in the classical framework.

I have only told the beginning of the story. Before concluding, I want to highlight just a few questions for future work to address.

Perhaps the most pressing one is whether we can drop the relativization to particular explananda in our characterization of epistemic utility functions. One could avoid this by making the choice of epistemic utility function be even more dependent on the particular agent whose credences we are evaluating. On this view, whatever the agent herself seeks to explain gives rise to the particular weighted accuracy measure, which we should use to assess the rationality of her epistemic transitions. But a more ambitious strategy would be to identify which explanada cry out for explanation relative to a body of beliefs, and find a way of aggregating the different weight functions generated by each explanandum. The resulting function would measure the extent to which a given proposition contributes to explaining *that which ought to be explained*. Aside from concerns about the possibility of aggregating different weight functions,

the main obstacle I foresee for this strategy is that of characterizing what it is for a proposition to be in need of explanation.[68]

Another issue left outstanding is the status of some of the constraints on epistemic utility functions that I have taken for granted. In particular, it might be worth considering the possibility that downwards propriety is too strong a demand. On the resulting picture, neither downwards propriety nor upwards propriety should be taken as constraints on epistemic utility functions. Rather, we should think that only those credence functions that are 'stable' are fully rational—only those credence functions which take themselves to be doing better than any of their extensions and any of their restrictions. An interesting question would be how to characterize the class of such probability functions relative to a particular epistemic utility function. But a more pressing concern would be to motivate the choice of one such epistemic utility function on independent grounds.

Finally, assuming we abandon both downwards propriety and upwards propriety, the question arises as to whether it can be rational to move from a given probability function to an expansion of it that is not an extension of it—whether, in other words, enlarging the domain of propositions one assigns credence to requires revising one's prior credences. An affirmative answer to this question would promise to shed light on the so-called problem of new theories, one of the big open questions for Bayesian epistemology. At this stage, however, I cannot tell which answer will turn out to be right.[69]

I have argued here that if we enrich the framework of epistemic utility theory with a more expansive notion of epistemic value, we can better understand how our hypothesis spaces should change, and vindicate the plausible idea that conceptual innovation is rationally constrained. There is still a significant role for epistemic imagination: there may be little we can do but wait for new distinctions to occur to us. But we need to know which such distinctions to take seriously and which to ignore as mere clutter. The framework outlined in this paper can help us do just that.

68 Some suggestive remarks in White 2005 might be used to characterize what being in need of explanation amounts to in a way that is amenable to the present framework. I discuss this and related ideas in my [REDACTED FOR BLIND REVIEW].

69 [ACKNOWLEDGMENTS OMITTED FOR BLIND REVIEW.]

I stated, without proof, that no utility function can be strictly universally proper (Fact 3.13), and that no strictly proper utility utility function can be universally proper (Fact 3.14). The purpose of this appendix is to provide proofs of these and related results.

First, we establish the following lemma:

LEMMA A.1. *Fix a nice utility function u. Suppose there are P and Q such that Q is an extension of P and $\mathbb{E}_Q[u(Q)] > \mathbb{E}_Q[u(P)]$. Then u is not upwards proper.*

Proof. Suppose there are such P and Q, for a given u. Let q and p range over π_Q and π_P, respectively. Since $\pi_P \preceq \pi_Q$, we know from Remark 3.3 that for each q, $u(P,q)$ is well-defined, with $u(P,q) = u(P,q')$ whenever q and q' are in the same cell of S_P. It thus follows from the probability calculus that[70]

$$\sum_q Q(q) \cdot u(P,q) = \sum_p P(p) \cdot u(P,p).$$

Now, by definition:

$$\overline{\mathbb{E}}_P[u(Q)] \geq \sum_q Q(q) \cdot u(Q,q).$$

And by assumption,

$$\sum_q Q(q) \cdot u(Q,q) > \sum_q Q(q) \cdot u(P,q) = \sum_p P(p) \cdot u(P,p).$$

Thus,

$$\overline{\mathbb{E}}_P[u(Q)] > \mathbb{E}_P[u(P)],$$

which means u is not upwards proper. □

70 Since π_P is a partition of W,

$$\sum_q Q(q) \cdot u(P,q) = \sum_q \sum_p Q(q \mid p) Q(p) \cdot u(P,q) = \sum_p Q(p) \sum_q Q(q \mid p) \cdot u(P,q).$$

And for each p, since u is nice and Q extends P,

$$Q(p) \sum_q Q(q \mid p) \cdot u(P,q) = P(p) \sum_{q \subseteq p} Q(q \mid p) \cdot u(P,q) = P(p) \cdot u(P,p).$$

The result stated in Fact 3.13 follows immediately.

THEOREM A.2. *There are no strictly universally proper epistemic utility functions.* □

What is more, if we assume that epistemic utility functions must be partition-wise strictly proper, we can show that there are no universally proper epistemic utility functions.

THEOREM A.3. *If u is universally proper, it is not partition-wise strictly proper.*

Before proceeding with the proof of Theorem A.3, let me introduce one more definition, which will come in handy in what follows:

DEFINITION A.4. A probability function Q is an *opinionated extension* of a probability function P iff Q is an extension of P and for each $p \in \pi_P$ there is $q_p \in \pi_Q$ with $q_p \subseteq p$ and $Q(q_p) = P(p)$. [71]

Proof of Theorem A.3. Our result follows immediately from the following lemma:

LEMMA A.5. *Let u be a strictly proper utility function that is downwards proper. Suppose P is a probability function and π is a refinement of π_P such that there is $p \in \pi_P \setminus \pi$ with $P(p) \neq 0$. Then there is an extension P^* of P to π such that*

$$\overline{\mathbb{E}}_P[u(P^*)] > \mathbb{E}_P[u(P)]$$

The proof of Lemma A.5 relies on an observation.

REMARK A.6. Suppose Q is an extension of P. For each utility function u there is an opinionated extension Q^*_P of P such that

$$\overline{\mathbb{E}}_P[u(Q)] = \mathbb{E}_{Q^*_P}[u(Q)].$$

71 Note that any probability function is an opinionated extension of itself. Note also that not all opinionated extensions of a probability function are fully opinionated in the sense defined in §5.2. Indeed, only those opinionated extensions of fully opinionated probability functions are themselves fully opinionated.

Proof. For each $p \in \pi_P$, pick $q_p \in \pi_Q$ with $q_p \subseteq p$ and such that $u(Q, q_p) \geq u(Q, q)$ whenever $q \subseteq p$. Let Q^*_P be the unique extension of P such that $Q^*_P(q_p) = P(p)$. Clearly, Q^*_P is an opinionated extension of P. Furthermore, for any extension Q' of P, we have

$$\mathbb{E}_{Q'}[u(Q)] \leq \mathbb{E}_{Q^*_p}[u(Q)],$$

and thus

$$\overline{\mathbb{E}}_P[u(Q)] \leq \mathbb{E}_{Q^*_p}[u(Q)] \leq \overline{\mathbb{E}}_P[u(Q)],$$

as desired. □

Proof of Lemma A.5. Suppose u is partition-wise strictly proper and downwards proper. Fix a probability function P, let π be a refinement of π_P, and suppose there is $p_\pi \in \pi_P \setminus \pi$ with $P(p_\pi) \neq 0$. Let Q be an extension of P with $Q(q) \neq P(p_\pi)$ for each $q \subseteq p_\pi$. We know from Remark A.6 that there is an opinionated extension Q^*_P of P to π such that

$$\overline{\mathbb{E}}_P[u(Q)] = \mathbb{E}_{Q^*_p}[u(Q)],$$

with $Q^*_P(q) \in \{P(p), 0\}$, whenever $q \subseteq p$. By construction, $Q^*_P \neq Q$. Since u is partition-wise strictly proper, we have

$$\mathbb{E}_{Q^*_p}[u(Q)] < \mathbb{E}_{Q^*_p}[u(Q^*_P)].$$

And by definition, we have $\mathbb{E}_Q[u(Q)] \leq \overline{\mathbb{E}}_P[u(Q)]$. Since u is downwards proper, we have $\mathbb{E}_Q[u(Q)] \geq \mathbb{E}_Q[u(P)] = \mathbb{E}_P[u(P)]$. Thus,

$$\mathbb{E}_P[u(P)] \leq \overline{\mathbb{E}}_P[u(Q)] = \mathbb{E}_{Q^*_p}[u(Q)] < \mathbb{E}_{Q^*_p}[u(Q^*_P)] \leq \overline{\mathbb{E}}_P[u(Q^*_P)],$$

which entails that u is not upwards proper. □

□

It might seem too demanding to require upwards propriety of epistemic utility functions. So it might be tempting to require instead that for each P and each extension Q of P,

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_Q[u(Q)].$$

The thought behind this is that, whereas in order to compare an extension of one's credence function one needs to use some credence function with

the same domain as that extension, you would be stacking the deck against your credence function if you evaluate extensions using the credence function that gives the most favorable assessment of it. Perhaps, then, the thing to do is to evaluate each extension not with whoever gives it the most positive evaluation, but with itself. The problem is, in the presence of partition-wise propriety, this requirement turns out to be equivalent to the requirement of upwards propriety, as made clear by the following fact:

FACT A.7. *A partition-wise proper utility function u is upwards proper if and only if, for each P and each opinionated extension Q^* of P,*

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_{Q^*}[u(Q^*)].$$

Proof. The left to right direction is straightforward, since by definition

$$\overline{\mathbb{E}}_P[u(Q^*)] \geq \mathbb{E}_{Q^*}[u(Q^*)].$$

For the right to left direction, we rely again on Remark A.6. For take P, let Q be an extension of P, and let Q^*_P be such that

$$\overline{\mathbb{E}}_P[u(Q)] = \mathbb{E}_{Q^*_P}[u(Q)].$$

Since u is partition-wise proper, we have

$$\mathbb{E}_{Q^*_P}[u(Q^*_P)] \geq \mathbb{E}_{Q^*_P}[u(Q)].$$

Which gives us what we want, since by assumption,

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_{Q^*_P}[u(Q^*_P)].$$

□

COROLLARY A.8. *Suppose that for each P and any extension Q of P,*

$$\mathbb{E}_P[u(P)] \geq \mathbb{E}_Q[u(Q)].$$

Then u is not downwards proper. □

Let me conclude by noting that, if we make some restrictions on u, we can strengthen Theorem A.2 in an interesting way.[72] We say that a function $\delta : [0,1] \longrightarrow \mathbb{R}$ is *concave* iff for each $x, y, \alpha \in [0,1]$,

$$\delta(\lambda x + (1-\lambda)y) \geq \lambda\delta(x) + (1-\lambda)\delta(y).$$

We say that δ is *strictly concave* iff the above inequality is always strict.

Recall now (see Definition 4.1) that a utility function u is *additive* iff for each partition π and each $s \in \pi$ there is $\delta_s^\pi : [0,1] \times \{0,1\} \longrightarrow \mathbb{R}$, with $\delta_s^\pi(x,0)$ (resp. $\delta_s^\pi(x,1)$) continuous, twice differentiable, and strictly increasing (resp. decreasing), such that for all C with $\pi_C = \pi$,

$$u(C,w) = \sum_{s \in \pi} \delta_s^\pi(C(s), \mathbb{1}\{w \in s\}).$$

We say that an additive utility function u is (*strictly*) *concave* iff each *component function* $\delta_s^\pi(x,i)$ ($i \in \{0,1\}$) is (strictly) concave. Any utility function which, restricted to probability functions with a given domain, is equivalent to an affine transformation of the Brier score is concave.[73] We can now formulate the following theorem (recall that $\mathcal{P}_{C'}\!/_C$ is the set of extensions of C to the domain of C'):

THEOREM A.9. *Suppose u is an additive, concave, strictly partition-wise proper utility function. Fix C and let C' be an expansion of C. Then:*

$$\max_{P \in \mathcal{P}_{C'}} \underline{\mathbb{E}}_C[u(P)] = \min_{\hat{C} \in \mathcal{P}_{C'}\!/_C} \mathbb{E}_{\hat{C}}[u(\hat{C})],$$

where $\mathcal{P}_{C'}$ is the set of probability functions defined over $\pi_{C'}$.

72 The restrictions aren't all strictly necessary—an even more general result for pretty much any natural partition-wise proper utility function follows as a corollary of Theorem 6.2 in Grünwald & Dawid 2004—but they allow for a more self-contained presentation of our results.

73 The requirement that a utility function be concave is a generalization of what Joyce calls CONVEXITY (Joyce 2009, p. 282). This unfortunate terminological difference is probably due to the fact that, when Joyce formulated a similar requirement in Joyce 1998, he was focused on *dis*utility functions (or *in*accuracy measures), and a utility function u is concave iff the corresponding disutility function $-u$ is convex. Requiring concavity rules out any affine transformation of the log score, and while I believe we can extend our results below so as to include the log score and its affine transformations if we allow for local accuracy measures to take on the value $-\infty$ for $(0,1)$ and $(1,0)$, I will not explore this issue further here.

COROLLARY A.10. *Suppose u is an additive, concave, strictly partition-wise proper utility function. Suppose for each P and each extension Q of P,*

$$\mathbb{E}_P[u(P)] \geq \underline{\mathbb{E}}_P[u(Q)].$$

Then u is not strictly downwards proper.

Proof. Pick any such u. Fix P and let P' be an extension of P. By assumption, we know that

$$\mathbb{E}_P[u(P)] \geq \max_{Q \in \mathcal{P}_{P'}/_P} \underline{\mathbb{E}}_P[u(Q)].$$

Theorem A.9 thus entails that

$$\mathbb{E}_P[u(P)] \geq \min_{Q \in \mathcal{P}_{P'}/_P} \mathbb{E}_Q[u(Q)].$$

Letting Q^* be such that

$$\mathbb{E}_{Q^*}[u(Q^*)] = \min_{Q \in \mathcal{P}_{P'}/_P} \mathbb{E}_Q[u(Q)],$$

we thus have

$$\mathbb{E}_{Q*}[u(Q*)] \not> \mathbb{E}_P[u(P)] = \mathbb{E}_{Q^*}[u(P)],$$

which means u is not strictly downwards proper. □

In order to prove Theorem A.9, we will rely on the following well-known result:[74]

THEOREM A.11 (VON NEUMANN'S MINIMAX THEOREM). *Suppose $X \subseteq \mathbb{R}^n$ and $Y \subseteq R^m$ are compact and convex and let*

$$f : X \times Y \longrightarrow \mathbb{R}$$

be concave for a fixed $y \in Y$ and convex for a fixed $x \in X$. Then

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y).$$

□

74 The canonical reference here is von Neumann 1928. For a more accessible proof of this result, see Binmore 2004.

Proof of Theorem A.9. Fix C and let C' be an expansion of C to some partition π. As before, let $\mathcal{P}_{pi}$ be the set of probability functions defined over π and let $\mathcal{P}_{C'}/_{C}$ be the set of extensions of C to π. Fix an enumeration $\{s_i : i \leq n\}$ of π, with $n = \|\pi\|$ and identify each credence function $P \in \mathcal{P}_{pi}$ defined over π with $\mathbf{x}_P = \langle P(s_1), \ldots, P(s_n)\rangle$. Set $X = \{\mathbf{x}_P : P \in \mathcal{P}_{C'}/_{C}\}$ and $Y = \{\mathbf{x}_P : P \in \mathcal{P}_{pi}\}$. Note that both X and Y are compact (since they are each subsets of $[0,1]^n$) and convex (Y trivially so, and X because any convex combination of extensions of C is itself an extension of C). Define $f : X \times Y \longrightarrow \mathbb{R}$ by

$$f(\mathbf{x}_P, \mathbf{x}_Q) = -\mathbb{E}_P[u(Q)].$$

Since u is concave, $f(\mathbf{x}_P, \mathbf{x}_Q)$ is convex for a fixed $\mathbf{x}_P$. And clearly, for a fixed $\mathbf{x}_Q$, $f(\mathbf{x}_P, \mathbf{x}_Q)$ is a linear function of $\mathbf{x}_P$, and thus concave. From Theorem A.11 we can thus conclude that

$$\max_{\mathbf{x}_P \in X} \min_{\mathbf{x}_Q \in Y} f(\mathbf{x}_P, \mathbf{x}_Q) = \min_{\mathbf{x}_Q \in Y} \max_{\mathbf{x}_P \in X} f(\mathbf{x}_P, \mathbf{x}_Q).$$

Note now that

$$\min_{\mathbf{x}_Q \in Y} \max_{\mathbf{x}_P \in X} f(\mathbf{x}_P, \mathbf{x}_Q) = \max_{\mathbf{x}_Q \in Y} \min_{\mathbf{x}_P \in X} -f(\mathbf{x}_P, \mathbf{x}_Q) = \max_{Q \in \mathcal{P}_{pi}} \underline{\mathbb{E}}_C[u(Q)].$$

And since u is proper, we have

$$\max_{\mathbf{x}_P \in X} \min_{\mathbf{x}_Q \in Y} f(\mathbf{x}_P, \mathbf{x}_Q) = \min_{P \in \mathcal{P}_{C'}/_{C}} \max_{Q \in \mathcal{P}_{pi}} \mathbb{E}_P[u(Q)] = \min_{P \in \mathcal{P}_{C'}/_{C}} \mathbb{E}_P[u(P)].$$

Putting these two observations together, we thus get

$$\max_{Q \in \mathcal{P}_{\pi}} \underline{\mathbb{E}}_P[u(Q)] = \min_{P \in \mathcal{P}_{C'}/_{C}} \mathbb{E}_P[u(P)].$$

□

APPENDIX B

The purpose of this appendix is to provide examples of downwards proper and of upwards proper epistemic utility functions and to prove two characterization theorems (Theorem B.3 and Theorem B.5) for a simple class of weighted accuracy measures.

Recall that an epistemic utility function is a *weighted accuracy measure* iff for each partition π there is a weight function $\lambda_\pi : \pi \longrightarrow \mathbb{R}^+$ and a local accuracy measure $\delta : [0,1] \times \{0,1\} \longrightarrow \mathbb{R}$ such that

$$u(P,w) = \sum_{p \in \pi_P} \lambda_{\pi_P}(p)\delta(P(p), \mathbb{1}\{w \in p\}),$$

We will say that u is a *simple accuracy measure* iff it is a weighted accuracy measure with constant weight function. Note that if u is a simple accuracy measure, then there is a local accuracy measure δ_u such that

$$u(P,w) = \sum_{p \in \pi} \delta_u(P(p), \mathbb{1}\{w \in P\}).$$

If we restrict our attention to simple accuracy measures, we can provide a general characterization of all downwards proper utility functions. To do that, it will be helpful to have at our disposal the following definition.

DEFINITION B.1. For any $x, y \in [0,1]$, let

$$\mathbb{E}_x\, \delta(y) := x \cdot \delta(y,1) + (1-x) \cdot \delta(y,0)$$

This gives us a slightly more convenient way of rewriting the definition of expected utility.

REMARK B.2. Suppose u is a weighted accuracy measure with weight function λ and local accuracy measure δ. Then for any partition π and any probability functions P and P' defined over π

$$\mathbb{E}_P[u(P')] = \sum_{p \in \pi} \lambda_\pi(p)\, \mathbb{E}_{P(p)}\, \delta(P'(p)).$$

□

We can now establish the following result:

THEOREM B.3. *A simple accuracy measure u is downwards proper iff for all $x, y \geq 0$, if $x + y \leq 1$ then*

$$\mathbb{E}_x\, \delta_u(x) + \mathbb{E}_y\, \delta_u(y) \geq \mathbb{E}_{x+y}\, \delta_u(x+y).$$

It is strictly downwards proper iff for all $x, y \geq 0$, if $x + y \leq 1$, then

$$\mathbb{E}_x\, \delta_u(x) + \mathbb{E}_y\, \delta_u(y) > \mathbb{E}_{x+y}\, \delta_u(x+y)$$

Proof. The left to right direction of each biconditional follows immediately from the definitions. For the right to left direction of the first biconditional, suppose u is not downwards proper. Then there are P and Q such that P is an extension of Q with

$$\mathbb{E}_P[u(Q)] > \mathbb{E}_P[u(P)].$$

In other words,

$$\sum_{q \in \pi_Q} \mathbb{E}_{P(q)}\, \delta_u(Q(q)) > \sum_{q \in \pi_Q} \sum_{p \subseteq q} \mathbb{E}_{P(p)}\, \delta_u(P(p)).$$

But this entails that there is $q \in \pi_Q$ such that

$$\mathbb{E}_{Q(q)}\, \delta_u(Q(q)) > \sum_{p \subseteq q} \mathbb{E}_{P(p)}\, \delta_u(P(p)),$$

or equivalently that there are non-negative $x_1, \ldots x_k$ such that

$$\mathbb{E}_{x^*}[\delta]_u(x^*) > \sum_i \mathbb{E}_{x_i}[\delta]_u(x_i),$$

where $x^* = \sum_i x_i \leq 1$ and $k = |\{p : p \subseteq q\}|$. And this in turn entails there are $x, y \geq 0$ with $x + y \leq 1$ such that

$$\mathbb{E}_x\, \delta_u(x) + \mathbb{E}_y\, \delta_u(y) < \mathbb{E}_{x+y}\, \delta_u(x + y)$$

Analogous reasoning can be used to establish the right to left direction of the second biconditional. □

COROLLARY B.4. *Let*

$$\beta_\theta(P, w) = \sum_{p \in \pi_P} \theta - (P(p) - \mathbb{1}\{w \in p\})^2.$$

β_θ is downwards proper iff $\theta \geq 1/2$.

Proof. Note that β_θ is a simple accuracy measure with local accuracy measure

$$b_\theta(x, i) = \theta - (x - i)^2,$$

with

$$\mathbb{E}_x\, b_\theta(x) = \theta - (x - x^2),$$

From Theorem B.3, we know that β_θ is downwards proper iff for all $x, y \geq 0$ with $x + y \leq 1$,

$$\mathbb{E}_x\, b_\theta(x) + \mathbb{E}_y\, b_\theta(y) \geq \mathbb{E}_{x+y}\, b_\theta(x+y).$$

Now,

$$\mathbb{E}_{x+y}\, b_\theta(x+y) = \theta - (x+y-(x+y)^2) = \mathbb{E}_x\, b_\theta(x) + \mathbb{E}_y\, b_\theta(y) - \theta + 2xy.$$

So β_θ is downwards proper iff $\theta \geq 2x(r-x)$ for all $x \geq 0$, $r \leq 1$. Since $2x(r-x)$ is maximized at $x = r/2$, we can conclude that β_θ is downwards proper iff $\theta \geq 1/2$. □

We can also provide a characterization of all upwards proper simple accuracy measures, one that includes the Brier score β.

THEOREM B.5. *If u is a simple accuracy measure, then u is upwards proper (resp. strictly upwards proper) iff $\delta(0,0) \leq 0$ (resp. $\delta(0,0) < 0$).*

Proof. We start with the following fact.

LEMMA B.6. *Suppose u is a proper, weighted accuracy measure. Then for each P and any opinionated extension Q of P, if $\lambda_{\pi_Q}(q) = \lambda_{\pi_P}(p)$ whenever $q \subseteq p$, then*

$$\mathbb{E}_Q[u(Q)] = \mathbb{E}_P[u(P)] + \sum_{p \in \pi_P} \lambda_{\pi_P}(p) \cdot (\|p\|_{\pi_Q} - 1) \cdot \delta(0,0).$$

Proof. Since Q is an opinionated extension of P, we know that for each $p \in \pi_P$, $Q(q_p) = P(p)$ and that for $q \subseteq p$, $q \neq q_p$ implies $Q(q) = 0$. And since $\mathbb{E}_0\, \delta(0) = \delta(0,0)$, we have

$$\mathbb{E}_Q[u(Q)] = \sum_{p \in \pi_P} \left(\lambda_Q(q_p)\, \mathbb{E}_{P(p)}\, \delta(P(p)) + \sum_{\substack{q \subseteq p, \\ q \neq q_p}} \lambda_Q(q) \delta(0,0) \right).$$

Now, by assumption, we have that for each $p \in \pi_P$, $q \in \pi_Q$, $q \subseteq p$ entails $\lambda_Q(q) = \lambda_P(p)$. Thus, we can conclude

$$\mathbb{E}_Q[u(Q)] = \sum_{p \in \pi_P} \left(\lambda_P(p)\, \mathbb{E}_{P(p)}\, \delta(P(p)) + \lambda_P(p) \cdot (\|p\|_{\pi_Q} - 1) \cdot \delta(0,0) \right).$$

as desired. □

COROLLARY B.7. *If u is a simple accuracy measure, then for each P and any non-trivial opinionated extension Q of P there is $k > 0$*

$$\mathbb{E}_Q[u(Q)] = \mathbb{E}_P[u(P)] + k \cdot \delta(0,0).$$

□

Theorem B.5 follows immediately from Corollary B.7 and Remark A.6.
□

COROLLARY B.8. *For each $\theta \leq 0$,*

$$\beta_\theta(P,w) = \sum_{p \in \pi_P} \theta - (P(p) - \mathbb{1}\{w \in p\})^2$$

is upwards proper. In particular, the Brier score $\beta = \beta_0$ is upwards proper.
□

Before concluding this appendix, let me note the following easy fact, which allows to generate a class of downwards proper weighted accuracy measures from any downwards proper local accuracy measure.

FACT B.9. *Suppose u is a weighted accuracy measure with a downwards proper local accuracy measure. Suppose for each $\pi \preceq \pi'$, $p \in \pi$ and $p' \in \pi'$, $p \subseteq p'$ entails $\lambda_\pi(p) \leq \lambda_\pi(p')$. Then u is downwards proper.*

APPENDIX C

The purpose of this appendix is to establish Corollary C.16, a straightforward consequence of which is Fact 6.3.

Start by first generalizing the notion of downwards propriety to apply to local accuracy measures:

DEFINITION C.1. A local accuracy measure $\delta : [0,1] \longrightarrow \mathbb{R}$ is *(strictly) downwards proper* iff the corresponding simple accuracy measure u_δ, defined by

$$u_\delta(P,w) = \sum_{p \in \pi_P} \delta(P(p), \mathbb{1}\{w \in p\}),$$

is (strictly) downwards proper.

The following is a straightforward consequence of Theorem B.3:

COROLLARY C.2. *If δ is downwards proper, then $\delta(0,0) \geq 0$. If δ is strictly downwards proper, then $\delta(0,0) > 0$.* □

Some local accuracy measures care only about the distance between a credal assignment and truth-value. We will say that such local accuracy measures are *normal*.

DEFINITION C.3. A local accuracy measure is *normal*[75] iff for each $x \in [0,1]$,

$$\delta(x,1) = \delta(1-x,0).$$

The following observation—an immediate consequence of the definitions—will come in handy:

REMARK C.4. If δ is normal, then for each $x \in [0,1]$,

$$\mathbb{E}_x\, \delta(x) = \mathbb{E}_{1-x}\, \delta(1-x).$$

□

The combination of normality and downwards-propriety ensures that $\mathbb{E}_x\, \delta(x)$ is non-negative.

FACT C.5. *Suppose δ is normal and downwards proper. Then for each $x \in [0,1]$, $\mathbb{E}_x\, \delta(x) \geq 0$. If in addition, δ is strictly downwards proper, then for each $x \in [0,1]$, $\mathbb{E}_x\, \delta(x) > 0$.*

Proof. Suppose δ is normal and strictly downwards proper, and suppose for *reductio* that there is $x \in [0,1]$ such that $\mathbb{E}_x\, \delta(x) \leq 0$. Since δ is normal, we can apply Remark C.4 we can conclude that

$$\mathbb{E}_x\, \delta(x) + \mathbb{E}_{1-x}\, \delta(1-x) = \mathbb{E}_x\, \delta(x) + \mathbb{E}_x\, \delta(x) \leq 0.$$

From the normality of δ we also know that $\delta(0,0) = \delta(1,1)$. Thus, from Theorem B.3 and the fact that δ is strictly downwards proper we know that

$$\mathbb{E}_x\, \delta(x) + \mathbb{E}_{1-x}\, \delta(1-x) > \mathbb{E}_1\, \delta(1) = \delta(1,1) = \delta(0,0).$$

75 Cf. Joyce 2009, p. 274. Note that a weighted accuracy measure with a normal local accuracy measure need not be normal in Joyce's sense. In Joyce's terminology, what I'm calling normality best corresponds to 0/1-symmetry. But when restricted to local accuracy measures, 0/1-symmetry and normality are equivalent.

Putting all of this together with Corollary C.2, we get

$$0 \geq \mathbb{E}_x\, \delta(x) + \mathbb{E}_{1-x}\, \delta(1-x) > \delta(0,0) \geq 0,$$

a contradiction. Perfectly analogous reasoning (*mutatis mutandis*) shows that if u is downwards proper then $\mathbb{E}_x\, \delta(x) \geq 0$.

□

For the remainder of this appendix, we will be working with a fixed weighted accuracy measure u whose local accuracy measure δ is normal and downwards proper. (Note that u may fail to be downwards proper even if its local accuracy measure δ is.)

Our ultimate goal is to establish Fact 6.3. Before getting there, let me introduce a few definitions.

DEFINITION C.6. Suppose π_1, π_2 are two refinements of π. Say that π_1 and π_2 are *π-equivalent* iff there is a bijection $f : \pi_1 \longrightarrow \pi_2$ such that, for all $s \in \pi$ and all $t \in \pi_1$,

$$t \subseteq s \Leftrightarrow f(t) \subseteq s.$$

REMARK C.7. If π_1 and π_2 are π-equivalent, then for each $s \in \pi$,

$$\left|\{t \in \pi_1 : t \subseteq s\}\right| = \left|\{t \in \pi_2 : t \subseteq s\}\right|.$$

□

DEFINITION C.8. Suppose π_1 and π_2 are π-equivalent. Say that a weight function λ *π-favors* π_1 over π_2 iff there is a bijection $f : \pi_1 \longrightarrow \pi_2$ such that for all $s \in \pi$ and all $t \in \pi_1$,

1. $t \subseteq s \Leftrightarrow f(t) \subseteq s$ and
2. $\lambda_{\pi_1}(t) \geq \lambda_{\pi_2}(f(t))$.

REMARK C.9. If λ π-favors π_1 over π_2, then for each $s \in \pi$,

$$\max_{\substack{t \in \pi_1,\\ t \subseteq p}} \lambda_{\pi_1}(t) \geq \max_{\substack{t \in \pi_2,\\ t \subseteq p}} \lambda_{\pi_2}(t).$$

□

THEOREM C.10. *Suppose u is a weighted accuracy measure with a strictly downwards proper and normal local accuracy measure. Fix P and let π_1 and π_2 be two π_P-equivalent refinements of π_P. If λ π-favors π_1 over π_2 and for some $p \in \pi$ with $P(p) \neq 0$*

$$\max_{\substack{q \in \pi_1, \\ q \subseteq p}} \lambda_{\pi_1}(q) > \max_{\substack{q \in \pi_2, \\ q \subseteq p}} \lambda_{\pi_2}(q),$$

then

1. $\max_{Q \in \mathcal{P}_{\pi_1}/_P} \mathbb{E}_Q[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_2}/_P} \mathbb{E}_Q[u(Q)]$, *and*
2. $\min_{Q \in \mathcal{P}_{\pi_1}/_P} \mathbb{E}_Q[u(Q)] > \min_{Q \in \mathcal{P}_{\pi_2}/_P} \mathbb{E}_Q[u(Q)]$.

Proof. Fix $f : \pi_1 \longrightarrow \pi_2$ witnessing that λ π-favors π_1 over π_2 and fix Q and R such that:

$$\mathbb{E}_Q[u(Q)] = \min_{P' \in \mathcal{P}_{\pi_1}/_P} \mathbb{E}_{P'}[u(P')],$$
$$\mathbb{E}_R[u(R)] = \min_{P' \in \mathcal{P}_{\pi_2}/_P} \mathbb{E}_{P'}[u(P')].$$

Let p, q, and r range over elements of π_P, π_1, and π_2, respectively. For each p, fix an enumeration q_i^p $(1 \leq i \leq n_p)$ of all subsets of p in π_1 and let $r_i^p = f(q_i^p)$. To reduce clutter, for each p and $1 \leq i \leq n_p$, let $\mathbf{q}_i^p = Q(q_i^p)$. Finally, define R_Q over π_2 by letting $R_Q(r_i^p) = \mathbf{q}_i^p$.

From Fact C.5 we know that for each x, $\mathbb{E}_x\, \delta(x) > 0$. Since λ π-favors π_1 over π_2, we thus have

$$\mathbb{E}_Q[u(Q)] = \sum_p \sum_{q \subseteq p} \lambda_{\pi_1}(q_i^p)\, \mathbb{E}_{\mathbf{q}_i^p}\, \delta(\mathbf{q}_i^p) > \sum_p \sum_{1 \leq i \leq n_p} \lambda_{\pi_2}(r_i^p)\, \mathbb{E}_{\mathbf{q}_i^p}\, \delta(\mathbf{q}_i^p).$$

But

$$\sum_p \sum_{1 \leq i \leq n_p} \lambda_{\pi_2}(r_i^p)\, \mathbb{E}_{\mathbf{q}_i^p}\, \delta(\mathbf{q}_i^p) = \mathbb{E}_{R_Q}[u(R_Q)],$$

and by construction

$$\mathbb{E}_{R_Q}[u(R_Q)] \geq \mathbb{E}_R[u(R)],$$

which establishes the second claim. To establish the first claim we will use the following observation, which follows immediately from the definitions and Fact C.5:

REMARK C.11. Suppose u is a weighted accuracy measure with a normal, downwards proper local accuracy measure. Then for each P and each refinement π of π_P

$$\max_{Q \in \mathcal{P}_\pi /_P} \mathbb{E}_Q[u(Q)] = \sum_p \left(\max_{q \subseteq p} \lambda_\pi(q) \, \mathbb{E}_{P(p)} \, \delta(P(p)) \right) + \sum_p \left(k_p^\pi \cdot \delta(0,0) \right),$$

where for each p

$$k_p^\pi = \left(\sum_{q \subseteq p} \lambda_\pi(q) \right) - \max_{q \subseteq p} \lambda_\pi(q).$$

□

Note now that our assumptions ensure that

$$\sum_p \max_{q \subseteq p} \lambda_{\pi_1}(q) \, \mathbb{E}_{P(p)} \, \delta(P(p)) > \sum_p \max_{q \subseteq p} \lambda_{\pi_2}(q) \, \mathbb{E}_{P(p)} \, \delta(P(p)),$$

and that

$$\sum_p k_p^{\pi_1} > \sum_p k_p^{\pi_2}.$$

Given Remark C.11 and Corollary C.2, we can conclude

$$\max_{Q \in \mathcal{P}_{\pi_1} /_P} \mathbb{E}_Q[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_2} /_P} \mathbb{E}_Q[u(Q)],$$

as desired. □

COROLLARY C.12. *Suppose u is weighted, accuracy measure with a strictly downwards proper local accuracy measure that is normal and concave. Fix P and let π_1 and π_2 be two π_P-equivalent refinements of π_P. If λ π-favors π_1 over π_2 and for some $p \in \pi$ with $P(p) \neq 0$*

$$\max_{\substack{q \in \pi_1, \\ q \subseteq p}} \lambda_{\pi_1}(q) > \max_{\substack{q \in \pi_2, \\ q \subseteq p}} \lambda_{\pi_2}(q),$$

then

1. $\max_{Q \in \mathcal{P}_{\pi_1} /_P} \overline{\mathbb{E}}_P[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_2} /_P} \overline{\mathbb{E}}_P[u(Q)]$, *and*
2. $\max_{Q \in \mathcal{P}_{\pi_1} /_P} \underline{\mathbb{E}}_P[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_2} /_P} \underline{\mathbb{E}}_P[u(Q)]$.

Proof. In light of Theorem C.10 and Theorem A.9, it suffices to establish the following lemma:

LEMMA C.13. *Suppose u is partition-wise proper. Then for each P and any refinement π of π_P,*

$$\max_{Q \in \mathcal{P}_\pi /_P} \overline{\mathbb{E}}_P[u(Q)] = \max_{Q \in \mathcal{P}_\pi /_P} \mathbb{E}_Q[u(Q)].$$

Proof. Fix $\hat{Q}$ such that

$$\overline{\mathbb{E}}_P[u(\hat{Q})] = \max_{Q \in \mathcal{P}_\pi /_P} \overline{\mathbb{E}}_P[u(Q)],$$

and fix $\dot{Q}$ such that

$$\mathbb{E}_{\dot{Q}}[u(\dot{Q})] = \max_{Q \in \mathcal{P}_\pi /_P} \mathbb{E}_Q[u(Q)].$$

From Remark A.6, we know that there is an opinionated extension $\hat{Q}^*_P$ of P such that

$$\overline{\mathbb{E}}_P[u(\hat{Q})] = \mathbb{E}_{\hat{Q}^*_P}[u(\hat{Q})].$$

By definition, we have

$$\max_{Q \in \mathcal{P}_\pi /_P} \overline{\mathbb{E}}_P[u(Q)] \geq \overline{\mathbb{E}}_P[u(\dot{Q})] \geq \mathbb{E}_{\dot{Q}}[u(\dot{Q})] = \max_{Q \in \mathcal{P}_\pi /_P} \mathbb{E}_Q[u(Q)].$$

And since u is partition-wise proper, we have

$$\begin{aligned} \max_{Q \in \mathcal{P}_\pi /_P} \mathbb{E}_Q[u(Q)] &\geq \mathbb{E}_{\hat{Q}^*_P}[u(\hat{Q}^*_P)] \\ &\geq \mathbb{E}_{\hat{Q}^*_P}[u(\hat{Q})] \\ &= \overline{\mathbb{E}}_P[u(\hat{Q})] \\ &= \max_{Q \in \mathcal{P}_\pi /_P} \overline{\mathbb{E}}_P[u(Q)]. \end{aligned}$$

□

□

To conclude, let me state a slight generalization of Theorem C.10.

DEFINITION C.14. Fix P and let π_1 and π_2 be two refinements of π_P. Say that π_1 and π_2 are *λ-P-equivalent* iff there is a bijection $f : \pi_1 \longrightarrow \pi_2$ such that:

- for all $p \in \pi_P$,
$$\bigcup_{q \subseteq p} f(q) \in \pi_P,$$
- for all $p \in \pi_P$,
$$\sum_{\substack{q \subseteq p,\\ q \in \pi_1}} P(f(q)) = \sum_{\substack{q \subseteq p,\\ q \in \pi_1}} P(q),$$
- for all $q \in \pi_1$,
$$\lambda_{\pi_1}(q) = \lambda_{\pi_2}(f(q)).$$

REMARK C.15. Suppose π_1 and π_2 are two λ-P-equivalent refinements of π_P. Then:

1. $\max_{Q \in \mathcal{P}_{\pi_1}/_P} \mathbb{E}_Q[u(Q)] = \max_{Q \in \mathcal{P}_{\pi_2}/_P} \mathbb{E}_Q[u(Q)]$, and
2. $\min_{Q \in \mathcal{P}_{\pi_1}/_P} \mathbb{E}_Q[u(Q)] = \min_{Q \in \mathcal{P}_{\pi_2}/_P} \mathbb{E}_Q[u(Q)]$.

Proof. Fix f witnessing that π_1 and π_2 are λ-P-equivalent. Define $\phi : \mathcal{P}_{\pi_1}/_P \longrightarrow \mathcal{P}_{\pi_2}/_P$ by letting
$$\phi(Q)(q) = Q(f^{-1}(q)).$$
Clearly, ϕ is a bijection from $\mathcal{P}_{\pi_1}/_P$ to $\mathcal{P}_{\pi_1}/_P$. And for each $Q \in \mathcal{P}_{\pi_1}/_P$,
$$\mathbb{E}_Q[u(Q)] = \mathbb{E}_{\phi(Q)}[u(\phi(Q))].$$
□

COROLLARY C.16. *Suppose u is weighted, accuracy measure with a strictly downwards proper local accuracy measure that is normal and concave. Fix P, let π_1 and π_2 be two π_P-equivalent refinements of π_P. If λ π-favors π_1 over π_2 and for some $p \in \pi$ with $P(p) \neq 0$*
$$\max_{\substack{q \in \pi_1,\\ q \subseteq p}} \lambda_{\pi_1}(q) > \max_{\substack{q \in \pi_2,\\ q \subseteq p}} \lambda_{\pi_2}(q),$$
then, whenever π_3 and π_2 are λ-P-equivalent,

1. $\max_{Q \in \mathcal{P}_{\pi_1}/_P} \overline{\mathbb{E}}_P[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_3}/_P} \overline{\mathbb{E}}_P[u(Q)]$, *and*
2. $\max_{Q \in \mathcal{P}_{\pi_1}/_P} \underline{\mathbb{E}}_P[u(Q)] > \max_{Q \in \mathcal{P}_{\pi_3}/_P} \underline{\mathbb{E}}_P[u(Q)]$.

□

APPENDIX D

The purpose of this appendix is to argue in more detail for the two claims I relied on in §5.2 in making a case for Remark 5.2, as well as to provide a proof of Remark 5.2 starting from those two claims.

Recall that we are seeking to assess, for a given credence function C, any expansion C' of C. The proposal was to use a weighted accuracy measure whose weights were sensitive to explanatory considerations. The construction proceeded in two steps. First, we defined the *counterfactual resilience* of a given explanandum e relative to a function C as follows.

DEFINITION D.1. The *counterfactual resilience* of e relative to a credence function C (and a set of suppositions B), $\mathbf{r}_B(C, e)$, is given by:

$$\mathbf{r}(C, e) = 1 - \frac{1}{|B|} \sum_{b \in B} |C(e) - C(b \boxdot\!\!\rightarrow e)| .$$

As before, I will drop the relativization to the set of suppositions in what follows. We then defined a weighted accuracy measure using a weight function defined in terms of $\mathbf{r}$.

DEFINITION D.2. Let δ be a local accuracy measure. Fix P and $e \in \mathcal{B}_P$. For any extension Q of P,

$$\varepsilon_P^\delta(Q, x) = \sum_{s \in \pi_Q} \chi_C^e(s) \delta(Q(s), \mathbb{1}\{x \in s\})$$

where for each $s \in \pi_Q$,

$$\chi_C^e(s) = \mathbf{r}(C_s, e),$$

with $C_s = C(\cdot \mid s)$.

This function will be well-defined, for a given extension Q of P, if and only if for all $q \in \pi_Q$, $\mathbf{r}(P_q, e)$ is well-defined. And $\mathbf{r}(P_q, e)$ will be well-defined if and only if for each b in the relevant set of suppositions B,

$P_q(b \boxdot\!\!\rightarrow e)$ are well-defined. Since $b \boxdot\!\!\rightarrow e$ is not a Boolean combination of b and e, we have no guarantee that $b \boxdot\!\!\rightarrow e$ is in the domain P even if b and e are. So we need to say something about how to assign credences to counterfactuals. Further, since for any non-trivial extension Q of P there will be $q \in \pi_Q$ that is not in π_P, we need to say something about how to understand P_q. After all, using the standard ratio definition (as before), the fact that $q \notin \mathcal{B}_P$ entails that P_q is not well-defined.

In what follows, I want to offer a principled way of addressing each of these concerns.

D.1 *Credences in counterfactuals*

For our purposes, we will rely on the familiar selection function semantics associated with Stalnaker 1968, which is a particular case of the similarity based semantics associated with Lewis 1973.

DEFINITION D.3. A *selection function* is a function

$$\sigma : \wp(W) \times W \longrightarrow \wp(W)$$

such that (i) for each $a \in \wp(W)$, $\sigma(a, w) \subseteq a$; for each $a \in \wp(W)$, if $w \in a$ then $\sigma(a, w) = \{w\}$; (iii) for all $a, b \in \wp(W)$, if $\sigma(a, w) \subseteq b$ and $\sigma(b, w) \subseteq a$, then $\sigma(a, w) = \sigma(b, w)$; (iv) $\sigma(a, w) = \varnothing$ only if $a = \varnothing$.

As usual, we think of $\sigma(a, w)$ as the closest worlds to w in which a is true, and we introduce the usual definition:

DEFINITION D.4. The *counterfactual conditional* relative to σ, $\boxdot\!\!\rightarrow_\sigma$, is a binary propositional operator defined as

$$a \boxdot\!\!\rightarrow_\sigma b := \{w : \sigma(a, w) \subseteq b\}$$

Thus, $a \boxdot\!\!\rightarrow_\sigma b$ contains all and only those worlds such that all of their closest a-worlds are b-worlds.

Now, recall that from the perspective of an agent whose state space is π, we can think of each $s \in \pi$ as playing the role of a possible world—the elements of π are each maximally consistent relative to the agent's credence function, in that each proposition in the agent's credence function is entailed by or inconsistent with some member of π. Thus, from the perspective of such an agent, we can think of the counterfactual conditional

as the proposition that is true in all and only those $s \in \pi$ such that their 'closest worlds' that make a true, also make b true.

For this to make sense, however, we need a function that assigns a set of closest worlds not to a member of W and a proposition, but rather to a member of π and a proposition. Fortunately, doing so is straightforward.

DEFINITION D.5. Given a selection function σ we can define a function

$$\sigma^* : \wp(W) \times \wp(W) \longrightarrow \wp(W)$$

by letting

$$\sigma^*(a, s) := \{\sigma(a, w) : w \in s\}.$$

Slightly abusing notation, we will identify σ^* and σ, and write $\sigma(a, s)$ instead of $\sigma^*(a, s)$. (Note that even if $\sigma(a, w)$ contains at most one world for all w, $\sigma^*(a, s)$ will typically contain more than one world.)

DEFINITION D.6. Fix σ and π. The *projection* of $a \mathrel{\Box\!\!\rightarrow_\sigma} b$ onto π, written $(a \mathrel{\Box\!\!\rightarrow_\sigma} b)^\pi$, is defined as

$$(a \mathrel{\Box\!\!\rightarrow_\sigma} b)^\pi := \bigcup\{s \in \pi : \sigma(a, s) \subseteq b\}.$$

We can think of $(a \mathrel{\Box\!\!\rightarrow_\sigma} b)^\pi$ as the proposition that contains all and only those cells in π such that their closest cells that entail a also entail b.[76] More precisely, for each a, b in $\mathcal{B}$, $(a \mathrel{\Box\!\!\rightarrow_\sigma} b)^\pi$ is the weakest proposition in $\mathcal{B}$ that entails $a \mathrel{\Box\!\!\rightarrow_\sigma} b$.

REMARK D.7. Suppose $\pi' \preceq \pi$, and let a and b be in the Boolean closure of π'. Then

$$(a \mathrel{\Box\!\!\rightarrow} b)^{\pi'} \subseteq (a \mathrel{\Box\!\!\rightarrow} b)^\pi.$$

□

Suppose now C is a credence function with state space π. Suppose $a, b \in \mathcal{B}_C$ but $a \mathrel{\Box\!\!\rightarrow} b \notin \mathcal{B}_C$. I submit that $(a \mathrel{\Box\!\!\rightarrow} b)^\pi$ has a very strong claim to being the proposition that plays the role of the counterfactual conditional in an agent whose conceptual resources are given by π. For

76 Note that, for a given $s \in \pi$, $\sigma(a, s)$ may be the union of more than one member of π. Thus, even if the underlying selection function satisfies the so-called uniqueness assumption—so that for all $a \subseteq W$ and $w \in W$, $\sigma(a, w)$ contains at most one world—the selection function filtered through an agent's state space π may not.

$(a \boxdot\!\!\rightarrow b)^{\pi}$ will be the proposition that is true at a 'world'—where this is just an atom of the agent's state space—if and only if in all 'worlds' 'most similar' that 'world' in which a is true, b is true—where similarity is understood in the only terms the agent can grasp. Thus, whenever $a, b \in \mathcal{B}_C$, I will henceforth use $C(a \boxdot\!\!\rightarrow_{\sigma} b)$ to denote the value C assigns to $(a \boxdot\!\!\rightarrow_{\sigma} b)^{\pi_C}$, a proposition which, by definition, is in $\mathcal{B}_C$.

We should accordingly understand $\mathbf{r}(C, e)$ as a function of the value that C assigns to $(b \boxdot\!\!\rightarrow e)^{\pi_C}$. This addresses our first concern, at least on the assumption that C_s is well-defined. We turn now to the second concern, viz. that C_s may not be well-defined, perhaps because $C(s)$ is not.

D.2 *Conditioning beyond one's state space*

Given everything we've said thus far, for $C_s = C(\cdot \mid s)$ to be well-defined, we need $C(s)$ to be well-defined (and non-zero)—this is because we have defined $C(x \mid y)$ using the so-called ratio formula

$$C(x \mid y) = \frac{C(xy)}{C(y)}.$$

But if C' is a non-trivial extension of C, there will be a C'-atom s' that such that $C(s')$ is not well-defined. We thus need some alternative way of defining C'_s, then, in order for $\varepsilon_C(C', w)$ to be well-defined.

Suppose C is a credence function with state space π and suppose s' is not in π. Consider the set $\mathcal{C}'$ of all possible extensions of C whose domain include s' and x. If for all $P, P' \in \mathcal{C}'$ we have $P(x \mid s') = P'(x \mid s') = \alpha$, it makes good sense to assign $C(x \mid s') = \alpha$. After all, an agent with credence function C would be committed to assigning α to $C(x \mid s')$—assuming every extension of C whose domain includes x and s' assigns α to x conditional on s'. (That said, we may wish to restrict our extension of the definition of $C(x \mid y)$ so that it is only defined over pairs of propositions that the agent whose credence function is C can grasp.)

Consider now an agent with credence function C that is assessing an extension C' of her credence function—an agent perhaps who just underwent an expansion of her conceptual resources. For most a and b in the domain of C', different extensions of C will assign different values to b conditional on a. But for our purposes this turns out not to matter.

REMARK D.8. Suppose s is an atom of P and let a be any proposition in the domain of of C. Then for any probability function Q whose domain includes s and a, with $Q(s) \neq 0$, $Q(a \mid s)$ will be either 1 or 0, depending on whether $s \subseteq a$ or not.

□

We can now see that, whenever an agent with credence function C is considering an extension C' of C, $\chi^e_C(s')$ will be well-defined, for each $s' \in \pi'$, since $\mathbf{r}(C_{s'}, e)$ will be, with

$$\begin{aligned}\mathbf{r}(C_{s'}, e) &= 1 - \frac{1}{|B|} \sum_{b \in B} \left| C(e) - \mathbb{1}\{s' \subseteq (b \boxdot\!\!\rightarrow e)^{\pi'}\} \right| \\ &= 1 - \frac{1}{|B|} \sum_{b \in B} \left| C(e) - \mathbb{1}\{\sigma(b, s') \subseteq e\} \right|.\end{aligned}$$

This suffices to establish the following remark, which has Remark 5.2 as an immediate consequence:

REMARK D.9. If $C(e) = 1$, s is a C-atom, π' is a refinement of C and $s' \subseteq s$, with $s' \in \pi'$, then

$$\mathbf{r}(C_{s'}, e) = \frac{\left|\{b \in B : s' \subseteq (b \boxdot\!\!\rightarrow e)^{\pi'}\}\right|}{|B|}$$

□

And from Remark D.7 we can then establish the following observation, from which Remark 6.4 follows immediately.

REMARK D.10. Suppose C' is an extension of C to $\pi' \geq \pi = \pi_C$. Then for each $s' \in \pi'$, $s \in \pi$, if $s' \subseteq s$ then

$$\chi^e_C(s) \leq \chi^e_C(s').$$

□

REFERENCES

Arntzenius, Frank. 1995. A Heuristic for Conceptual Change. *Philosophy of Science* 62(3). 357–369.

Arntzenius, Frank. 2008. Rationality and Self-Confidence. In Tamar Szábo Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology*, vol. 2, 165–178. Oxford: Oxford University Press.

Baker, Alan. 2003. Quantitative Parsimony and Explanatory Power. *The British Journal for the Philosophy of Science* 54(2). 245–259.

Baker, Alan. 2011. Simplicity. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2011.

Berker, Selim. 2013. Epistemic Teleology and the Separateness of Propositions. *Philosophical Review* 122(3). 337–393.

Binmore, Ken. 2004. Guillermo Owen's Proof Of The Minimax Theorem. *Theory and Decision* 56(1). 19–23.

Caie, Michael. 2013. Rational Probabilistic Incoherence. *Philosophical Review* 122(4). 527–575.

Carr, Jennifer. 2015. Epistemic Expansions. *Res Philosophica* 92(2). 217–236.

Ellerman, David. 2010. The Logic of Partitions: Introduction to the Dual of the Logic of Subsets. *The Review of Symbolic Logic* 3(2). 287–350.

Evans, Gareth. 1982. *The Varieties of Reference*. John McDowell (ed.). Oxford: Clarendon Press.

Forster, Malcolm R. 1999. How Do Simple Rules 'Fit to Reality' in a Complex World? *Minds and Machines* 9(4). 543–564.

van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford University Press.

van Fraassen, Bas C. N.d. FIgures in a Probability Landscape. In, 345–356.

Franke, Michael & Tikitu de Jager. 2011. Now That You Mention It: Awareness Dynamics in Discourse and Decisions. In Anton Benz, Christian Ebert, Gerhard Jäger & Robert van Rooij (eds.), *Language, Games, and Evolution* (LNAI 6207), 60–91. Heidelberg: Springer.

Gärdenfors, Peter. 1982. Imaging and Conditionalization. *Journal of Philosophy* 79(12). 747–760.

Garfinkel, Alan. 1981. *Forms of Explanation*. New Haven: Yale University Press.

Gibbard, Allan. 2008. Rational Credence and the Value of Truth. In Tamar Szábo Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology*, vol. 2, 143–164. Oxford: Oxford University Press.

Gilboa, Itzhak. 1987. Expected Utility with Purely Subjective Non-Additive Probabilities. *Journal of Mathematical Economics* 16(1). 65–88.

Goldstein, Michael. 1984. Turning Probabilities Into Expectations. *The Annals of Statistics* 12(4). 1551–1557.

Greaves, Hilary. 2013. Epistemic Decision Theory. *Mind* 122(488). 915–952.

Greaves, Hilary & David Wallace. 2006. Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind* 115(459). 607–632.

Grünwald, Peter D. & A. Philip Dawid. 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics* 32(4). 1367–1433.

Hájek, Alan. n.d. Most Counterfactuals Are False. Unpublished ms., Australian National University.

Hájek, Alan. 2003. What Conditional Probability Could Not Be. *Synthese* 137(3). 273–323.

Hurwicz, Leonid. 1951. The generalized Bayes minimax principle: a criterion for decision making under uncertainty. *Cowles Commission Discussion Paper* 335. 1950.

Jackson, Frank & Philip Pettit. 1988. Functionalism and Broad Content. *Mind* 96(387). 381–400.

Jackson, Frank & Philip Pettit. 1990. Program Explanation: A General Perspective. *Analysis* 50(2). 107–117.

Jeffrey, Richard C. 1983. *The Logic of Decision*. Chicago: University of Chicago Press.

Joyce, James M. 1998. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science* 65(4). 575–603.

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. New York: Cambridge Univ Press.

Joyce, James M. 2009. Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Franz Huber & Christoph Schmidt-Petri (eds.), *Degrees of Belief*, vol. 342 (Synthese Library), chap. 10, 263–297. Dordrecht: Springer Netherlands.

Lange, Marc. 2005. Laws and Their Stability. *Synthese* 144(3). 415–432.

Lange, Marc. 2009. *Laws and Lawmakers*. New York: Oxford University Press.

Leitgeb, Hannes & Richard Pettigrew. 2010. An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science* 77(2). 201–235.

Lewis, David. 1973. *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Lewis, David. 1976. Probabilities of Conditionals and Conditional Probabilities. *Philosophical Review* 85(3). 297–315.

Lewis, David. 1979. Counterfactual Dependence and Time's Arrow. *Noûs* 13(4). 455–476.

Lewis, David. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59(1). 5–30.

Lipton, Peter. 1990. Contrastive Explanation. *Royal Institute of Philosophy Supplement* 27(1). 247–266.

Lipton, Peter. 2004. *Inference to the Best Explanation*. 2nd edn. London: Routledge.

Manski, Charles F. 1981. Learning and Decision Making When Subjective Probabilities Have Subjective Domains. *The Annals of Statistics* 9(1). 59–65.
Moss, Sarah. 2011. Scoring Rules and Epistemic Compromise. *Mind* 120(480). 1053–1069.
von Neumann, John. 1928. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100(1). 295–320.
Nolan, Daniel. 1997. Quantitative Parsimony. *The British Journal for the Philosophy of Science* 48(3). 329–343.
Pettigrew, Richard. 2012. Accuracy, Chance, and the Principal Principle. *Philosophical Review* 121(2). 241–275.
Pettigrew, Richard. 2013. A New Epistemic Utility Argument for the Principal Principle. *Episteme* 10(01). 19–35.
Pettigrew, Richard. 2016. The Population Ethics of Belief: In Search of an Epistemic Theory X. *Noûs*. Forthcoming.
Popper, Karl. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
Rényi, Alfréd. 1970. *Foundations of Probability*. San Francisco: Holden-Day.
Romeijn, Jan-Willem. 2005. Theory Change and Bayesian Statistical Inference. *Philosophy of Science* 72(5). 1174–1186.
Satia, Jay K. & Roy E. Lave. 1973. Markovian Decision Processes with Uncertain Transition Probabilities. *Operations Research* 21(3). 728–740.
Savage, Leonard J. 1972. *The Foundations of Statistics*. Second. New York: Dover.
Schwarz, Wolfgang. 2016. Subjunctive Conditional Probability. *Journal of Philosophical Logic*.
Skyrms, Brian. 1977. Resiliency, Propensities, and Causal Necessity. *Journal of Philosophy* 74(11). 704–713.
Skyrms, Brian. 1980. *Causal Necessity*. New Haven: Yale University Press.
Sober, Elliott. 1998. Black Box Inference: When Should Intervening Variables Be Postulated? *The British Journal for the Philosophy of Science* 49(3). 469–498.
Stalnaker, Robert C. 1968. A Theory of Conditionals. *Studies in Logical Theory* 2. 98–112.
Stalnaker, Robert C. 2002. Epistemic Consequentialism. *Aristotelian Society Supplementary Volume* 76(1). 153–168.
Strevens, Michael. 2004. The Causal and Unification Approaches to Explanation Unified—Causally. *Noûs* 38(1). 154–176.
Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, Mass.: Harvard University Press.
Swanson, Eric. 2006. *Interactions With Context*. PhD dissertation, Massachusetts Institute of Technology.
Troffaes, Matthias C.M. 2007. Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning* 45(1). 17–29.
Weslake, Brad. 2010. Explanatory Depth. *Philosophy of Science* 77(2). 273–294.

White, Roger. 2005. Explanation as a Guide to Induction. *Philosopher's Imprint* 5(2). 1–29.
Williams, J. Robert G. 2012. Counterfactual Triviality: A Lewis-Impossibility Proof for Counterfactuals. *Philosophy and Phenomenological Research* 85(3). 648–670.
Williamson, Jon. 2003. Bayesianism and Language Change. *Journal of Logic, Language, and Information* 12. 53–97.
Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Woodward, James. 2010. Scientific Explanation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2010.
Yablo, Stephen. 1992. Mental Causation. *Philosophical Review* 101(2). 245–280.
Yalcin, Seth. 2007. Epistemic Modals. *Mind* 116(464). 983–1026.
Yalcin, Seth. 2018. Belief as Question-Sensitive. *Philosophy and Phenomenological Research* 97(1). 23–47. Early View, 2016.