Version 2.3. Last major revision: 12.02.2012 at 11:25:45 -0500

GOOD QUESTIONS

Alejandro Pérez Carballo
University of Southern California
aperezca@usc.edu

We care about the truth. We want to believe what is true, and avoid believing what is false. But not all truths are created equal. Having a true botanical theory is more valuable than having true beliefs about the number of plants in North Dakota. To some extent this is fixed by our practical interests. We may want to keep our plants looking healthy, and doing botany is more likely to help us do that than counting blades of grass. (Of course, if you find yourself out of luck, your main goal in life might be to keep tally of blades of grass.) But setting our practical interests aside, there is something more valuable, *epistemically*, about our botanical beliefs than about those we get out of counting blades of grass.

That, at least, is the intuition driving this paper. I think it is a powerful intuition. But it remains to be cashed out without relying too heavily on metaphors. The first task of the paper will be to do just that. More specifically, I want to offer a way of evaluating different courses of inquiry—different research agendas, as it were—from a purely epistemic perspective. There will be some additional benefits from looking at things the way I suggest. First, we will be better placed to give an account of epistemic value that is sensitive both to considerations of accuracy and to explanatory considerations. Having a more accurate belief about the answer to a good question, I will suggest, should count for more, epistemically, than having a more accurate belief about a lesser one. Second, we will be have the beginnings of an account of conceptual evaluation. A set of conceptual tools is better than another, I will suggest, to the extent that it allows us to formulate better questions. The abstract issues about questions I will discuss thus turn out to be relevant to giving an account of epistemic value, as well as to the question of the rationality of conceptual change.

A word about terminology is in order. I will speak interchangeably of 'beliefs' and 'credences', or 'degrees of confidence'. I realize there are some diffi-

cult questions about how those two notions relate to one another, if at all.[1] For our purposes, however, we can set them aside. Most of what I will say can be recast purely in terms of all-out beliefs—only towards the end will the notion of a credence function play a major role. For now, those details do not matter.

1 EVALUATING QUESTIONS RELATIVE TO A DECISION PROBLEM

We want to evaluate truths.[2] But better first to evaluate lines of inquiry, or research agendas. To think that p would be good to believe, if true, is at least to think it is worth engaging in finding out *whether* p—at least if we set aside, as I will for now, the cost of inquiry, and if we assume that our attempt to find out whether p will succeed. To the extent that you think it is worth trying to answer the question whether p, you will think that at least one of its answers would be good to believe, if true.

We cannot just identify the value of p with the value of answering the question whether or not p. This would have as a consequence that every proposition will be as valuable as its negation. And while a sophisticated scientific theory might be very valuable, epistemically, its negation may not be worth much.

Keeping this in mind, let us start by evaluating lines of inquiry. Connecting issues about the value of truths to issues about the value of *actions* gives us some much needed tractability. We have some idea about how to decide whether to engage in finding out whether p. So this gives us a good place to start giving an account of the value of p.

You can think of a course of inquiry as a collection of questions. Indeed, you can think of it as a single question: what is the answer to each of these questions? Any way of evaluating questions will thus correspond to a way of evaluating courses of inquiry.

We can devise a framework for evaluating questions using familiar decision-theoretic tools. We need only assume that we can identify the value of a question with the expected value of *learning* the (true) answer to that question. For whenever you are facing a choice among a set of options, you can evaluate questions according to how likely, and to what extent, learning its true answer will help you make the right choice.

[1] For discussion, see e.g. Foley 2009; Sturgeon 2008.

[2] For the sake of variety and legibility, I will speak interchangeably of the value of p, the value of believing p if true, and the value of learning p.

Let's cash this out a bit more explicitly. Two coins will be tossed. You are told that the first coin is fair. The second one is biased: there is a 70% chance it will land heads. Consequently, you assign credence .5 to the first coin landing heads, and .7 to the second one landing heads. You are then asked to predict a particular outcome: you will be rewarded only if you predict the actual outcome. The reward will depend on what the prediction is, according to this table (where, e.g. 'HT' stands for the act of predicting that the first coin lands heads and the second coin lands tails):

	HH	HT	TH	TT
Reward if correct (in $)	0	5	10	15
Penalty if incorrect (in $)	0	0	0	0

After computing the expected utility of each possible action, you realize that TH is the action that maximizes expected utility.[3]

Before you state your choice, however, you are told that an Oracle you take to be fully reliable will answer for you one of these two questions:

(Q1) Did the first coin land heads?

(Q2) Did the second coin land heads?

Clearly, if you have nothing to lose, you should ask one of these questions.[4] The question is: which one?

To answer this, we need to consider two different issues. First, all things being equal, we prefer to ask a question Q over another Q' if we are less opinionated about the answer to Q than of the answer to Q'. If we have good evidence that the answer to Q is a, but no evidence pointing to what the right answer to Q' is, we have a *pro tanto* reason for asking Q' rather than Q. At the same time, if we expect that having an answer to one question will have little

[3] Your credence assignment is as follows: $C(\text{HH}) = C(\text{TH}) = .35$, $C(\text{HT}) = C(\text{TT}) = .15$. Thus, the expected utility of TH is $3.5, that of TT is $2.25. The expected utility of HH is $0, and that of HT is $.75. (Note that I'm being sloppy in using e.g. 'HH' to stand both for the proposition that both coins land heads and for the action of predicting that both coins land heads. But context should have resolved the ambiguity.)

[4] We know from a result by I. J. Good that for any Q (and any decision problem) the value of asking Q is never negative, *so long as asking Q is cost-free*. See Good 1967. Good attributes the result to Raiffa and Schlaifer 1961. For a general discussion of Good's theorem, see Skyrms 1990.

impact on our choice—perhaps we would choose the same action no matter what the answer to that question is—we may have reason to ask a different question instead. We need a way of arbitrating between these potentially conflicting considerations.

Following I. J. Good (1967), we can do so in the following way. We set the value of a question as the weighted average of the value of (learning) its answers. The value of each answer p is obtained as follows. First, let a be the alternative that maximizes expected value relative to your current credence function. Now let a' be the alternative that maximizes expected value relative to the result of updating your credence function with the proposition p. The value of (learning) p is the difference in the *posterior* expected value (i.e. the expected value calculated using the result of updating your credence function with p) between a and a'.[5]

Return to the coin example. Relative to your prior credence function, TH was the action that maximized expected utility. But if you learned that the first coin landed heads (henceforth, 'H1') you would no longer pick TH. For assuming you update your credence function by conditionalizing on your evidence, that would be a sure loss. The sensible thing to do if you learned H1 would be to pick HT, since the expected utility (relative to your new credence function) of each other option is \$0. Now, the expected value of HT relative to the result of updating your credence function with the information at hand is \$1.5. Since upon learning H1 the expected value of TH would be \$0, the net gain in utility from learning H1 is $V(\text{H1}) = \$1.5 - \$0 = \$1.5$.

Similarly, we can compute the expected gain in utility from learning that the first coin landed tails (i.e. T1): it is the expected value of TH minus the expected value of TH, both calculated using the posterior. Since T1 would not affect your choice, we have that $V(\text{T1}) = 0$.

We can then set the value of Q1 to equal the weighted average of the values of its answers, so that $V(\text{Q1}) = \$.75$. [6] It is easy to verify that $V(\text{H2}) = \$0$, and $V(\text{T2}) = \$7.5$. This allows us to assign a value to Q2, viz. the weighted average of the value of H2 and T2, i.e. $V(\text{Q2}) = \$2.25$.[7] The upshot is that the value of

[5] As it turns out, one could also assign value to a proposition p by looking at the difference between the *prior* expected values of the action that maximizes expected value relative to the result of conditionalizing on p and the action that maximizes expected value relative to your prior. The value of p will of course be different if we do things this way, but the resulting $V(Q)$ will be the same. See van Rooy 2004, p. 397.

[6] Since $C(\text{H1}) \times V(\text{H1}) + C(\text{T1}) \times V(\text{T1}) = .5 \times \$1.5 + .5 \times \$0$.

[7] Since $C(\text{H2}) \times V(\text{H2}) + C(\text{T2}) \times V(\text{T2}) = .7 \times \$0 + .3 \times \$7.5$.

Q2 is higher than that of Q1, so that Good's strategy recommends you ask Q2, as we would expect.

I want to use this strategy to spell out a way of evaluating questions from a purely epistemic perspective. But first we need to find the right decision problem.

2 EPISTEMIC DECISION PROBLEMS

Let me step back for a moment and review some of the background assumptions I have been relying on thus far. I have implicitly appealed to a very minimal form of *expected utility theory*. On this framework, different alternatives are evaluated relative to a credence function and a *utility function*—an assignment of numerical values to each alternative in any possible world. A particular alternative is *better* than another if it has a higher expected utility—again, relative to a credence and utility functions.

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 actions. 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.

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 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.[8] But this does not prevent us from using it to capture thicker notions of value: we only need to make

[8] Cf. Stalnaker 2002, p. 158.

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, an 'epistemic utility function', 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 that credence and utility functions. But 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.

Now, 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.[9] But we can surely 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.

> An Oracle 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 given a choice among a small set of credence functions which does not include the one you currently have. If you set aside your practical interests for a moment, how will you choose?

I suspect you have a rough idea of how you will choose. 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.

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

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 conditions that must plausibly be met by anything that could count as an epistemic utility function.

2.1 *Truth-directedness*

Here is the first thing we can say about what an epistemic utility function must be like. Suppose you are comparing two credence functions C and C', defined over a body of propositions $\mathcal{B}$. If you know which propositions in $\mathcal{B}$ are true, then you will probably think that C is better than C', epistemically, if for each $p \in \mathcal{B}$, $C(p)$ is closer to p's truth-value than $C'(p)$ is.

We can extract a minimal constraint on epistemic utility functions from this. Recall that utility functions are assignments of numerical values to each pair consisting of an alternative—from the relevant decision problem—and a possible world. If alternatives are just credence functions, our utility functions in question assign numerical values to pairs of the form (C, w), where C is a probability function defined over a fixed $\mathcal{B}$, and w is a possible world. Say that C is *uniformly closer to the truth* than C', relative to w, if for each proposition p in $\mathcal{B}$ (i) $C(p)$ is at least as close to p's truth value in world w than $C'(p)$ is, and (ii) for at least some $p \in \mathcal{B}$, $C(p)$ is closer than $C'(p)$ to p's truth-value in w. The constraint that must be satisfied by any epistemic utility function can now be stated as:[10]

> TRUTH-DIRECTEDNESS: If C is uniformly closer to the truth than C', then $u(C, w) > u(C', w)$,

To illustrate, suppose you are only interested in one question: whether Secretariat won the Kentucky Derby in 1973. You are told by a reliable source that he did. If you are given the choice between waking up with a credence function that assigns .6 to the proposition that Secretariat won that race, and waking up with a credence function that assigns .9 to the proposition that Secretariat won that race, you would presumably choose the latter. After all, it gets closer to the truth of the proposition you are interested in, and you have nothing to lose. TRUTH-DIRECTEDNESS constrains epistemic utility functions to agree with your judgment: from an epistemic point of view, in a world in

[10] Cf. e.g. Gibbard 2008; Horwich 1982; Joyce 2009.

which Secretariat did win the race in 1973, it is better to have a credence function that assigns .9 to that proposition than one that assigns .6 to it.

Note that TRUTH-DIRECTEDNESS only tells you to rank C above C', relative to a world w, if C is closer than C' to the truth-value, at w, of *every* proposition. But suppose you were given two credence functions C and C' to choose from when you wake up. You now know the truth-value of p and q. If C is closer to the truth about p but C' is closer to the truth about q, TRUTH-DIRECTEDNESS will tell you nothing about how to chose, even if C is much closer to the truth about p than C' is about the truth about q.

To evaluate different lines of inquiry we need a principled way of going beyond TRUTH-DIRECTEDNESS in precisely those contexts. Doing that will be the object of §4.[11] Before moving on, however, I want to consider an additional constraint on epistemic utility functions, one that is almost as uncontroversial as TRUTH-DIRECTEDNESS.[12]

2.2 *Propriety*

Suppose you are rational in having credence function C. Further suppose you are evaluating alternative credence functions relative to your own credence function and an epistemic utility function u. Could it be that you take some *other* credence function to be better, epistemically, than your own?

The question is not whether you could reasonably think that some of your beliefs could be false. The question is whether you could reasonably think: even though my actual credence function is C, it would be better, from a purely epistemic perspective, to have $C' \neq C$ as my credence function. If you are rational, it seems, you could not. So if you have rationally arrived at your credence function C, then the expected epistemic value of C' should not be greater than the expected epistemic value of C.

If we further assume that any credence function could be rationally held by some agent at some point, then we will be tempted to endorse the following condition on epistemic utility functions:

[11] Thus, I will be suggesting that epistemic utility functions need not be *normal* nor *extensional*, in the sense defined in Joyce 2009. As I understand him, while he endorsed normality as a condition on epistemic utility functions in Joyce 1998, he seems to have changed his mind and now endorses normality and extensionality 'to the extent that we see epistemic utility as reflecting considerations of accuracy alone' (Joyce 2009, p. 275).

[12] Other constraints have been proposed. For example, Joyce 1998 discusses a constraint he calls LOCALITY, which makes $u(C, w)$ be sensitive only to the value of $C(p_w)$, where p_w is the strongest $p \in \mathcal{B}$ such that $p(w) = 1$.

PROPRIETY: For any credence function C, the expected value of C relative to u and C must be greater than or equal to the expected value of $C' \neq C$ relative to u and C.

PROPRIETY ensures that a probability function will always evaluate itself, relative to u, as having maximum expected value.[13] In fact, one might want to impose a slightly stronger constraint on epistemic utility functions, viz. that the expected value of C relative to C and u must be *higher* than that of any other C'. But we need not argue over that. What matters is that we have some idea of what an epistemic utility function must be like. It is time to put this to use to start answering our initial question.

3 THE EPISTEMIC VALUE OF QUESTIONS

I set out to find a decision problem that could allow us to evaluate questions from an epistemic perspective. So far, all I have is a sketch of such a decision problem: a decision problem where options—credence functions—are assessed relative to an epistemic utility function.

To be sure, I have said very little about what epistemic utility functions are. But assuming that all epistemic utility functions are proper, we can at least say this much: the alternative that maximizes expected value relative to your prior C is C. The alternative that maximizes expected value relative to the result of updating your prior with p is $C_p = C(\cdot \mid p)$.[14] So the value you will assign to (learning) p is the difference in expected value, relative to C_p, between C and C_p. And the value you will assign to a question Q is the weighted average of the value you will assign to (learning) each of its answers.

We can simplify things even further, if what we are after is a way of *comparing* different questions. For the expected value of Q will be higher than the expected value of Q' relative to a proper scoring rule u and a credence function C if and only if the weighted average of the expected values of C_p relative

[13] By itself, TRUTH-DIRECTEDNESS is not enough to guarantee PROPRIETY. For instance, take the following value function defined over all credence functions whose domain is a given finite set X of size N:

$$\varphi(C, w) = (-1)\frac{1}{N}\sum_{p \in X} |C(p) - p(w))|.$$

Clearly, φ satisfies TRUTH-DIRECTEDNESS. But for most credence functions C, φ will sanction moving to the extremes. See Gibbard 2008 for discussion.

[14] Cf. the discussion in Greaves and Wallace 2006 of why conditionalizing maximizes expected epistemic utility, assuming that all epistemic utility functions are proper.

to u and C_p, where p is an answer to Q, is higher than the expected value of C_r relative to u and C_r, where r is an answer to Q'.

But just assuming that an epistemic utility function is proper and truth-directed is not enough to get an answer to our starting question. To see that, suppose you only assign credence to two independent propositions, p and q, as well as their Boolean combinations. Suppose we have an epistemic utility function that is truth-directed and proper, such as[15]

$$u_B(C, w) = -1/2((C(p) - p(w))^2 + (C(q) - q(w))^2),$$

where we identify a proposition p (thought of as a set) with its characteristic function. You are trying to determine which of $?p$ (whether p) and $?q$ (whether q) to ask. If you assign .5 credence to each of p and q, then the expected value of $?p$ will be exactly that of $?q$.[16] So if all we know of epistemic utility functions is that they satisfy TRUTH-DIRECTEDNESS and PROPRIETY, this way of cashing out a notion of epistemic value will not allow for non-trivial comparisons between different lines of inquiry.

The reason our utility function u_B fails to distinguish between $?p$ and $?q$ in this particular situation is that it is insensitive to the *content* of the propositions being assessed. If we want to compare different lines of inquiry from an epistemic perspective, we need a different utility function: one that *is* sensitive to the content of the propositions being assessed.

Now, we could assign different weights to the terms in the sum above. We could define a scoring rule that would differ from u_B only in that rather than calculating the average distance from the truth about p and the truth about q, we take a *weighted* average distance, say by multiplying $(C(p) - p(w))^2$ by some factor greater than 1. This would take deviations from the truth about p to be worse than deviations from the truth about q. But we need a good reason for favoring one proposition over the other, and we need to do so in a way that corresponds to an epistemic benefit to be accrued from being closer to the truth about p rather than closer to the truth about q. In the next section, I want to propose a way of doing just that.

To anticipate, the guiding thought will be this: we should favor those questions whose answers we expect to increase the explanatory closure of our body

[15] This is the so-called *Brier score*—cf. Brier 1950.

[16] To see that, note that C_p and C_q are perfectly symmetric, so that the expected value of C_p relative to C_p is equal to that of C_q relative to C_q, and the same goes for $C_{\neg p}$ and $C_{\neg q}$.

of beliefs. Fleshing this out however requires a way of incorporating explanatory considerations into the framework of epistemic utility functions that we have been working with. I propose to do that by first looking at the notion of *counterfactual resilience*.

4 COUNTERFACTUAL RESILIENCE AND EXPLANATION

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 would consist 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 suppose it is incompatible with BIAS, perhaps because we can derive from the facts cited in INITIAL that the coin is fair.[17]

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.

4.1 *Some platitutdes*

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

[17] 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.

because theorizing has cognitive costs, and we rather not spend cognitive resources on details that promise little in terms of theoretical payoff.

But not all reasons for preferring BIAS over INITIAL have to do with our particular cognitive limitations.[18] 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 account of my brain state.[19]

Consider this example, essentially due to Alan Garfinkel.[20] Tom is running late for a meeting, because he had a leisurely breakfast. He gets in 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 a good explanation of the accident.

The reason—or at least, a reason that has been offered by many to account for similar cases—is that, unlike the first explanation, the second one is not very portable. Had Tom not driven recklessly, we couldn't have used that 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

[18] 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 as well as Baker 2011 for a helpful overview of some of the relevant issues.

[19] Cf. Jackson and Pettit 1988, Strevens 2004, and the discussion of causal relevance in Yablo 1992 for related discussion.

[20] Garfinkel 1981, p. 30.

the explanandum very stable.[21] 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 putative explanation does in helping us understand the explanandum.

4.2 *Counterfactual resilience*

Rather than focusing on properties of the explanans, however, let us focus on what the explanans does to the explanandum, relative to a given body of beliefs. Let us first ask how well-explained a given proposition is relative to a body of beliefs. We can then use this to tell how much learning a particular proposition—a putative explanation—can contribute to having the explanandum be well-explained.

Like subjective Bayesians who think of questions of evidential support as making sense only against the backdrop of background beliefs, I think of questions of explanation as making sense only against such a background. Thus, I think it is best to ask how well explained e is relative to a given body of beliefs. Questions about what is 'the' explanation of e are, on my view, less pressing. But we can still ask whether p contributes more to an explanation of e than q does, by looking at how well-explained e would be conditional on p vs. how well-explained it would be conditional on q. And we can use this to explain why simpler, or more general, explanations are better.

My suggestion, in a nutshell, is this: the more counterfactually robust a particular claim is, the more well-explained it is (relative to a given body of beliefs). Explanations that make the explanandum more counterfactually robust tend to be more satisfying than those that do not. [22]

One way to see that increasing the counterfactual stability of the explanandum makes for satisfying explanations is to think about laws of nature. Laws

[21] The notion of stability I am after is related to, although distinct from, the notion of resilience discussed in Jeffrey 1983 (when discussing the 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.

[22] 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.

of nature have a high degree of counterfactual stability.[23] 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 high counterfactual robustness goes hand in hand with 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.[24]

Another way of motivating the connection between counterfactual stability and explanation is to reflect on the plausibility of so-called *contrastive* accounts of explanation.[25] 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 reasonably expect an explanation of *e*, 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.

Moreover, counterfactual stability is 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 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 likely to lead to simpler explanations. And explanations that make the explanandum counterfactually robust can be applied to many different circumstances. They are 'portable', in that they can be used, *mutatis mutandis*, to explain many different phenomena.

Consider again the explanation of the sequence of nearly ten tails in a row

[23] 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.

[24] This is not to say that we cannot explain a given law of nature. 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 (the same applies to the worries about low probability events mentioned in fn. 22.)

[25] See Garfinkel 1981, as well as van Fraassen 1980; Lipton 1990, *inter alia*.

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.[26]

It is hard to cash 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 false. 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 sets of possibilities, it is not obvious how to motivate one measure over another that will work for all cases.

But assume we can agree on a finite set of relevant suppositions. If the explanandum is made more robust under counterfactual suppositions in that set by one explanation than another, I submit, that would give us an *epistemic* reason (albeit a *pro tanto* one) for preferring the one explanation over the other. This is not to say this is a reason for taking the first explanation to be more likely than the second one.[27] But it is surely a reason for favoring inquiry into the truth of the one over the other.

For example, suppose you are interested in explaining why the outcome of the ten coin tosses is what it is, and in general you want to be able to explain facts about the outcome of coin tosses involving that coin. You are told your memory will be erased, but you will have some say on what credence function

[26] 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.

[27] Although see White 2005.

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. You would rather know the bias of the coin than the particular initial conditions of those ten coin tosses. After all, you can expect that having true beliefs about the bias of the coin will be more likely to explain other features of the coin. As I will put it, claims about the bias of a coin plausibly have more *explanatory potential* than claims about the particular initial conditions of some arbitrary sequence of ten coin tosses—at least relative to the sort of things we tend to want to explain. Holding fixed a class of explananda, having true beliefs that have a higher explanatory potential is, all else equal, better than not—and this, I submit, from an epistemic point of view.

Once we have an account of what the explanatory potential of a proposition is, we can make sense of the explanatory potential of a *question* as the expected explanatory potential of its correct answer. Other things being equal, the more explanatory potential a particular question has, the better it is, epistemically, to engage in finding out its correct answer. By making our epistemic utility functions sensitive to the explanatory potential of the relevant propositions, we can finally spell out a framework for evaluating different lines of inquiry from an epistemic point of view.

5 MEASURING EXPLANATORY POTENTIAL

Suppose we fix on a particular explanandum e. In comparing different explanations of e, I have relied on instances of following schema:

> STILL: Given the relevant explanation, if p had been false, e would have still obtained.

I want to say more about what exactly these instances are supposed to mean.

It is important to note that STILL is *not* equivalent to: if the explanans E were true and p were false, e would have still obtained. Assuming the truth of E, the closest possible world in which p is false may not be a world in which E is also true. In other words, the logical form of STILL, at the relevant level of abstraction, should be understood as

$$E \rightarrow (\neg p \mathrel{\Box\!\!\rightarrow} e),$$

which is not equivalent to:

$$(E \wedge \neg p) \boxdot\!\!\rightarrow e.$$[28]

This is at it should be. For one thing that motivates the idea that BIAS is a better explanation than INITIAL is that the following is true:

> SPIN: Given BIAS, *if* the spin velocity of the coin had been different, the coin would have still landed tails nearly every time.

whereas the instance corresponding to INITIAL is not.

> SPIN-INITIAL: Given INITIAL, *if* the spin velocity of the coin had been different, the coin would have still landed tails nearly every time.

And if we understand SPIN-INITIAL as

> had the spin velocity been slightly different in any one toss and INITIAL been correct, the coin would have still landed tails nearly every time,

then we get a counterfactual with a logically impossible antecedent. After all, INITIAL specifies, inter alia, the spin velocity that each coin toss actually had. The reason SPIN-INITIAL is not true is that, even if all the details are in fact as INITIAL states them to be, slightly altering the initial conditions of the coin-tosses we could have easily yielded a sequence of nearly ten coin-tosses. But if we assume the truth of INITIAL, when we counterfactually suppose that the initial conditions are different from what INITIAL states them to be, we are not supposing that some logical contradiction is true.

We can refine this criterion some more, by defining a degreed version of counterfactual resilience. Other things being equal, the more resilient an explanandum is according to an explanation, the better the explanation is—where resilience is understood as relative to a fixed set of candidate suppositions. Intuitively, we want to measure how stable e is under counterfactual supposition conditional on a particular explanation E. Thus, relative to a credence function C, we want to measure the amount of variation between $C(e \mid E)$ and $C(s \boxdot\!\!\rightarrow_\sigma e \mid E)$, where (i) $\boxdot\!\!\rightarrow_\sigma$ denotes the counterfactual condi-

[28] I am assuming a semantics for counterfactuals broadly along the lines of Stalnaker 1968 and Lewis 1973, on which counterfactuals are sensitive to a contextually determined similarity ordering.

tional relativized to the similarity function σ,[29] and (ii) s is an element of the fixed set of candidate suppositions.

There are different ways of measuring the relevant variation. To fix ideas, let us use a generalization of Euclidean distance. Thus, the counterfactual resilience of e, relative to a credence function C, a similarity function σ, and a finite set of suppositions S is given by the average of the square of the difference between $C(e)$ and $C(s \mathrel{\square\!\!\rightarrow_\sigma} e)$, for $s \in S$. We can give a more explicit definition if we appeal to *imaging*[30] as a way of understanding counterfactual supposition, but those technicalities can wait for the appendix (§A.4).

We can now define the counterfactual resilience of e relative to a credence function C as follows:

$$CR_{\sigma,C}(e) = 1 - \frac{1}{|S|}\sum_{s\in S}(C(e) - C(s \mathrel{\square\!\!\rightarrow_\sigma} e))^2.$$

For a given explanation E, we can now think of $CR_{\sigma,C_E}(e)$ as a measure of how robust E makes the explanandum e.

In order to build these considerations into a notion of epistemic value, we need to specify what an epistemic utility function that is sensitive to the expected degree of resilience given to an explanandum e would look like. Presumably there are different ways of bringing the expected degree of resilience to bear into a measure of epistemic utility. We have two criteria for ranking credence functions. On the one hand, we can rank them in terms of their expected accuracy. On the other, we can rank them in terms of how resilient they make a given explanandum. How exactly to weigh each of these criteria is a question for some other day.[31]

---

[29] One must tread carefully here. Williams 2012 shows that, under certain assumptions, one cannot identify the credence in $\varphi \mathrel{\square\!\!\rightarrow} \psi$ with the credence one assigns to ψ on the counterfactual supposition that φ. For our purposes, however, we can treat $C(\varphi \mathrel{\square\!\!\rightarrow} \psi)$ as merely shorthand for the value $C^\varphi(\psi)$—in the notation introduced below. I need not assume that there is anything as the proposition expressed by a counterfactual conditional.

[30] Cf. Lewis 1976.

[31] As I gestured at in §3, one option would be to use an additive scoring rule (Joyce 2009, p. 272), i.e. a scoring rule of the form:

$$u(P, w) = \sum_{s\in\pi_P} \lambda_{\pi_P}(s) f(P(s), s(w)).$$

We just need to have the *weight* $\lambda_{\pi_P}(s)$ of each s be proportional to the degree to which s is expected to contribute to the resiliency of a given explanandum. (Here, π_P is the collection of atoms of the algebra over which P is defined, and $f(P(s), s(w))$ measures the utility of having credence $P(s)$ in s in a world w where s has $s(w)$ as its truth-value. Cf. §A.1 for more details

Fortunately, we can already say something about how to compare different questions from an epistemic perspective. Recall from §3 that we had an epistemic utility function—the Brier score—on which $?p$ and $?q$ were on a par. Now, the Brier score is just a measure of expected accuracy. So it is no surprise that, from the point of view of maximizing accuracy, inquiry into p and inquiry into q are taken to be on a par.[32] But we can use counterfactual resilience as a tie breaker. In other words, in addition to PROPRIETY and TRUTH-DIRECTEDNESS, we have the following constraint on epistemic utility functions:

> RESILIENCE: If all else is equal, and the expected resilience of e relative to a credence function C is higher than the expected resilience of e relative to a credence function C', then the expected epistemic utility of C should be higher than the expected epistemic utility of C'.

Go back to the example of the explanations of the sequence of ten tails. From the point of view of maximizing accuracy, inquiry into BIAS and inquiry into INITIAL may well be on a par. It will depend on the prior degrees of belief you assign to the relevant propositions. But from the point of view of explanatory potential, I have argued, inquiry into BIAS is to be preferred.

6 IMMODESTY AND EPISTEMIC IMAGINATION

What we have so far is a way of comparing different lines of inquiry from an epistemic point of view. We have, in other words, an account that would allow us to tell which questions we should try to answer from an epistemic perspective. If you expect that answering Q is more likely to give you an explanation of what you want to explain than answering Q', then you should prefer to gather evidence that bears on Q rather than Q'.

Thus, nothing in what I've said so far suggests that Q having a higher expected epistemic value than Q' should in any way affect your epistemic state.[33] Yet if assessments of the epistemic value of different lines of inquiry are to be incorporated into an account of rational inquiry, it seems like such assessments *should* be able to change, in some cases, your epistemic state.

---

on the notation used here.)

[32] This brings out the fact that we have not taken into account how likely we take finding out the answer to question to be.

[33] At least when it comes to propositions that are not about the difference in expected epistemic value of the relevant questions.

One reason to think this cannot be is that realizing that Q has higher expected epistemic value than Q' does not involve acquiring new evidence. And in order for a change in our epistemic state to be *rational*, it would seem, it must be triggered by the acquisition of new evidence. After all, epistemic rationality 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. To change your epistemic state without new evidence would involve moving to an epistemic state that is *worse* than the one you currently are.

Appearances are misleading, however. One can consistently maintain that epistemic rationality requires that we take ourselves to be doing as well as we can, epistemically, 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, and proceeds to do what he takes to be the most fun activity he could do, in light of his evidence.

If you looked at the life Lars lives, you would find it incredibly boring. 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. Rather, it's because his lack of imagination precludes him from seeing 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.

The point is a rather simple one. If Tom, like Lars, 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.

This is a type of epistemic change that orthodox Bayesian theories of epistemic rationality have little to say about. I want to sketch a way of thinking about epistemic rationality that is conservative, in that it allows us to capture some basic principles behind traditional Bayesian accounts of epistemic rationality. But it gives us the flexibility to ask questions about the rationality of conceptual change: about how epistemic rationality could constrain expansions of the range of hypothesis we rely on in inquiry.

7 RATIONAL DYNAMICS AND EPISTEMIC VALUE

If we make some plausible assumptions, claims about epistemic value can be used to motivate the standard norms of Bayesian rationality. For instance, Joyce (1998, 2009) shows that on certain plausible assumptions about epistemic utility functions, probabilistically incoherent credences are dominated by probabilistically coherent ones: if your credences are not probabilistically coherent, there will be some probability functions that you take to have higher expected epistemic value. And Greaves and Wallace (2006) show that, once an agent receives evidence E, updating by conditionalization on E is what maximizes expected epistemic value—again, assuming that epistemic utility functions are proper, and that your prior degrees of belief are probabilistically coherent.

But an account of epistemic rationality in terms of epistemic value gives us more flexibility. Suppose our agent, Tom, has never considered a particular question Q. Further suppose that, once the question occurs to him, he can estimate a particular expected epistemic value to learning the answer to Q. Then it might be rational for him to gather evidence that bears on Q: and this, I submit, will require a certain change in Tom's epistemic state, one that is not recognized by most broadly Bayesian theories of epistemic rationality. Let me explain.

I have been thinking of an agent's epistemic state as a credence function: an assignment of numerical values to a set of propositions that satisfies the standard Kolmogorov axioms of the probability calculus. Crucially, a credence function assigns values to some propositions: it needn't assign numerical values to *all* propositions. Why not take an agent's epistemic state to assign a numerical value to all propositions? Of course, if we are not interested in what value an agent assigns to a particular collection of propositions, we could simply have them drop out of the picture. But in that case, a proposition not being in the domain of an agent's credence function would not tell us anything of substance about the agent's epistemic state. It would simply tell us that we are not interested in what value the agent assigns to that proposition. Is there anything we could model about the agent by leaving a proposition out of the domain of her credence function?

Start out with an easy case. Tom is a much more primitive version of Jackson's Mary.[34] He lives in a black and white world but he knows very little physics. Further, he has not even heard of color words. He is unable to distinguish red things from the rest not just visually, but in any other way.

It is quite tempting to say that Tom has no doxastic attitude towards propositions about the colors of things. In particular, the proposition that he is wearing all blue is one he has no view on. But it is also tempting to say something stronger, viz. that Tom is not even aware of that proposition: given his current epistemic state, there is little sense to be made of him suspending judgment on that proposition. His epistemic state is blind to that proposition, as it were. In general, when an agent is unable to entertain a given proposition, there is a principled reason for leaving it out of the domain of her credence function.[35]

I suspect you are still on board. Nothing I've said about what partial credences can be used to model is terribly controversial. But it is roughly a psychological question what sort of distinctions an agent is able to make. And whereas there surely are some interesting questions about the way in which Tom could come to acquire the relevant distinctions, they are better left for those with enough budget to run a lab.

But we can use partial credences to model something of more evident bearing to our present purposes. Tom, we suppose, has no doxastic attitudes to-

[34] Jackson 1986.

[35] Note that the suggestion so far is not to leave out a proposition p whenever the agent is not attending to the proposition. We may have good grounds for thinking that the agent assigns a particular credence to p even if the agent is not currently entertaining the issue of whether p.

wards propositions about the color of things. The reason is that, as described, Tom lacks the conceptual resources to make the relevant distinctions. But suppose Tom acquires the ability to make these distinctions. Should we insist that Tom's epistemic state be modeled by a credence function that assigns value to all propositions about the color of things?

From our point of view, as modelers, there is little to recommend this. It is hard to see what reasons we would have for using some real number or other to represent Tom's attitude to the proposition that some potatoes are blue. But even from *Tom's* point of view, the proposition that some potatoes are blue is not even part of his epistemic landscape. There is a sense in which Tom will, and in my view *should*, ignore this proposition. Having his credence function be undefined on that proposition is a way of modeling just that.[36]

Of course, things can change. It can be that it will become advantageous for Tom to gather evidence bearing on the question whether some potatoes are blue. And in order to make room for this in our model, we need to allow for changes in his epistemic state that involve the *expansion* of the domain of his credence function.

Bayesian epistemologists tend to focus on a distinctive type of epistemic change: when an agent's credence function comes to assign different values to a given body of propositions. Some of these changes are said to be rational—those resulting from conditionalization on the evidence available to the agent, say. Some are not: my moving from uncertainty as to whether p to full certainty as to whether p without having acquired any new evidence. But Bayesians have little to say about expansions: the orthodox Bayesian machinery lacks the resources to even ask whether some expansions could be epistemically rational.[37, 38]

[36] Cf. Rayo 2011. See also the discussion of digital encoding of information in Dretske 1981, ch. 7. In Dretske's terminology, what I'm suggesting is that we have $C(p)$ be undefined whenever p is not part of what is digitally encoded by an agent's belief system.

[37] You might think that PROPRIETY should be enough to rule out this possibility, once we move to a framework in which rational dynamics is a matter of maximizing expected epistemic value. But as I show in the Appendix (§A.2, Fact A.5), PROPRIETY is only a reasonable constraint when we are focusing on credence functions that assign values to the same collection of propositions.

[38] There is a related but distinct problem that has received some discussion in the literature—e.g. Earman 1992; Maher 1995—the so-called problem of new theories. The problem is that of explaining, once a new hypothesis is added to the domain of a credence function, how to assign a probability to that hypothesis. The question of rational expansions is in some sense prior to the question of new theories—the former is the question of which hypotheses to add, the latter the question of how to assign credence to such hypotheses.

In contrast, by moving to a framework in which rational dynamics involves maximizing expected epistemic value, we can ask the question whether a particular expansion is epistemically rational. In particular, we can ask the question whether introducing new conceptual machinery, or postulating new hypotheses, are likely to be beneficial, epistemically.

8 THE RATIONALITY OF CONCEPTUAL CHANGE

Start with a simple toy example.[39] You are studying an unfamiliar type of organism, call them 'Reds', and their reactions to certain stimuli. 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.

You notice that Reds that were exposed to red light tend to stop moving altogether until the lights are switched off. In contrast, exposing Reds to blue or green light seems to have little or no effect on their behavior.

However, if you take a Red that was previously exposed to red light, you observe that exposing it to blue light tends to make it move significantly faster than normal. It occurs to you that there could be an internal state R such that a Red could get in state R as a result of being exposed to red light, and once in state R it would respond to blue light differently than it would had it not been in state R.

Once you bring R into the picture, you can formulate a hypothesis about Reds that could be used to explain why Reds would respond to blue light by moving faster after being exposed to red light. For example, that Reds in state R tend to get excited by blue lights, and that exposure to red light tends to cause Reds to be in state R. Under this hypothesis, the claim that a particular Red will move faster when exposed to blue light is made more counterfactually robust than it would be otherwise. For given that they are in state R, had they not been exposed to Red light, they would have still responded to blue light the way they did (given the new hypothesis). In other words, the introduction of state R allows for the formulation of a hypothesis that would, *if true*, increase the counterfactual resilience of the claim that a given Red would respond the way it did to blue light.

[39] 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.

Now, nothing here suggests that you should conclude that Reds do *in fact* get to be in state *R* when exposed to red light. But the expansion of your hypothesis space to include that particular hypothesis—which will have to be assessed in light of the data at hand[40]—can be motivated by the considerations of epistemic utility above. After all, some claims involving state *R* have a high explanatory potential, so the expected epistemic value of expanding your credence function to allow for such propositions is relatively high.[41]

Here is a slightly more complex example, due in its essentials to Frank Arntzenius.[42] You arrive in a strange land, where you find a collection of round critters, each of about 1 inch in diameter. Each critter is either red or white. You pick some up and discover that, if you press two critters against each other, they combine to form a larger critter, of about 2-inches in diameter, uniform in color—red or white. You try combining two 2-inch critters and discover that they too combine to form a larger critter, of about 3-inches in diameter, uniform in color.

You set out to understand how the colors of smaller critters relate to the color of the larger critters they turn into when combined. After gathering data for a while, you have the following observations. First, if you combine two 1-inch critters, they will turn into a 2-inch *red* critter unless both the 1-inch critters were white. If you combine two 2-inch critters, however, things get slightly trickier.

If you combine two 2-inch white critters, they combine to form a white 3-inch critter. But if at least one of the two 2-inch critters is red, the color of the resulting 3-inch critter is sometimes white and sometimes red. After trying this out with a large number of 2-inch critters, you get the following frequencies. First, if one of the 2-inch critters is red, and it came from two 1-inch red critters, then no-matter what other 2-inch critter it's combined with, the result will be a 3-inch critter. But if you only look at combinations of 'mixed' 2-inch red critters with other 2-inch critters—either mixed red or white—the color of the resulting 3-inch critter will have the following distribution:

[40] On which, again, see Sober 1998. Specifically, Sober discusses ways in which the introduction of intervening variables can sometimes be motivated by frequency data.

[41] See §A.6 for an illustration of how one can compute the expected epistemic value of a given expansion even though one hasn't assigned credence to the 'new' propositions.

[42] Arntzenius 1995.

	3-inch red (%)	*3-inch white* (%)
mixed 2-inch red	75	25
2-inch white	50	50

Now, imagine you want to explain why a particular 3-inch critter is red. You look at your records and notice that it came from two 2-inch red critters. How good of an explanation is this? Granted, on the basis of your observation, your credence that a 3-inch critter will be red given that it came from two 2-inch red critters is relatively high. But this explanation does not make your explanans particularly robust. For example had one of the red 2-inch critters come from different colored parents, they might have combined to form a white critter instead. So the color of the 2-inch critters does not, by itself, suffice to make the explanandum counterfactually robust.

Suppose now it occurs to you that the red critters could come in two varieties—strong-red and weak-red.[43] If a 2-inch red critter comes from two 1-inch red critters, it is strong-red. If it comes from a 'mixed' pair of critters, it is weak-red.[44] You now note that a 2-inch strong-red critter combined with a 2-inch white critter yields a 3-inch red critter, and that a 2-inch weak-red critter combined with a 3-inch white critter yields a 3-inch red critter 50% of the time, and a 3-inch white critter 50% of the time. You can now formulate the following hypothesis—a hypothesis that was not part of your hypothesis state before you considered that red critters could come in two types:

	strong-red	*weak-red*	*white*
strong-red	100% strong-red	50% strong-red 50% weak-red	100% weak-red
weak-red	50% strong-red 50% weak-red	25% strong-red 50% weak-red 25% white	50% weak-red 50% white
white	100% weak-red	50% weak-red 50% white	100% white

[43] Of course, this is just a simplified version of the conceptual innovation behind Mendelian genetics.

[44] The sense in which these are 'new' properties is this: these properties are causally dependent on, but they are not reducible to, facts about the critters' lineage.

If you were to add this hypothesis to your body of beliefs, you could make claims about the heredity of color features among the critters more robust: assuming that a red critter is strong-red you could now explain why it would, when combined with a white critter, yield a red critter most of the time. Knowing what type of red a critter is would allow you to explain things about the distribution of color among its offspring independently of what generation the critter happens to be—you've thereby made your explanandum resilient under the counterfactual supposition that the given red critter is a third-generation, say, rather than a second-generation one—and also independently of what its 'parents' were—you've made your explanandum resilient under the assumption that your critter had different colored ancestors, say.[45]

Now, in both these cases, conceptual innovation allowed for the formulation of hypotheses that were not part of the starting hypothesis space. There is still a question to be asked, viz. how is it that a given hypothesis gets entertained for the first time? But perhaps there isn't much to be explained here.

Imagine a machine that is generating possible new hypotheses at random—new ways of partitioning the state space it is working with. Allow for the machine to evaluate each possible hypothesis in terms of its expected epistemic value. By constraining the process of crafting its hypothesis space by considerations of epistemic value the machine is more likely to yield better theories, rather than undertaking avenues of inquiry that would lead to nowhere. Not because theories with high expected epistemic value are more likely to be true—but because *if true*, they are more likely to be more satisfying, from an epistemic point of view.[46]

9 CONCLUSION

Let me take stock. I started out with an intuition: that some lines of inquiry are better, *epistemically*, than others. I proposed a way of cashing out this intuition: by assigning epistemic value to different bodies of belief, we can evaluate a given line of inquiry on the basis of the value of the body of belief that this line of inquiry is expected to yield. I outlined a framework for understanding this notion of epistemic value, by looking at the extent to which a given body

[45] See §A.6 for a detailed example.

[46] Cf. Bromberger 1992 for more on the role of questions in, and the importance of formulating new questions for, inquiry.

of beliefs was explanatorily closed. I then conjectured that counterfactual resilience could be a tractable guide to explanatory closure.

This strategy has two additional benefits. First, it opens the door to using scoring rules to assess credence functions in ways that go beyond accuracy considerations. The notion of counterfactual resilience can be seen as the first step toward better understanding explanatory value within the framework of epistemic utility theory. Second, it allows us to assess *expansions* of our hypothesis space before setting off to gather evidence for the new hypothesis. We should only spend cognitive resources on new lines of inquiry that promise to be epistemically valuable. Of course, having a high expected epistemic value is no guarantee that the given line of inquiry will prove to be helpful. But it gives us an epistemic reason to look into it—to take the relevant hypotheses seriously in inquiry. This strategy thus provides us with a model of how conceptual change and theoretical innovation could fall under the scope of a theory of epistemic rationally. How much this is so remains to be seen. But at the very least, it gives us a framework for asking questions about the rationality of a type of epistemic change that was ruled out by default from an orthodox Bayesian framework.

A APPENDIX

I will outline a formal framework for investigating how epistemic utility functions can be used to assess different expansions of a credence function. But first, some definitions are in order.

A.1 *Basic definitions*

Given any probability function P, let $\mathcal{A}_P$ denote the domain of P. To simplify our discussion, we will restrict our attention to atomic algebras. Thus, for each P, we can define π_P as the smallest subset of $\mathcal{A}_P$ whose Boolean closure is $\mathcal{A}_P$. Note that π_P is a partition of $\mathcal{W}$.

For our purposes, a *utility function* is a function u that associates, to each probability function P and world $w \in \mathcal{W}$ a real number $u(P, w)$, which is the *score* of P in w.

Intuitively, any such function must satisfy the following desideratum: if P does not distinguish between w and w', then $u(P, w) = u(P, w')$. Thus, if $\{w\}$ is not in the domain of P, then $u(P, \cdot)$ will be constant throughout the π_P cell of w—that is, the unique element of π_P containing w. The following definition captures this intuition:

Definition A.1: A utility function u is *nice* iff for each P, $u(P, \cdot) : \mathcal{W} \to \mathbb{R}$ is P-measurable.

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

If a probability function P is defined over the entire power set of $\mathcal{W}$, then we can define the *expected score* of any probability Q relative to u in the usual way, viz.

$$EU_{u,P}(Q) = \sum_{w \in \mathcal{W}} P(\{w\})u(Q, w).$$

But we need a different definition in order to allow for π_P to be coarser than the set of singletons of $\mathcal{W}$. The above definition is of no help, since there will be some $w \in \mathcal{W}$ such that $\{w\} \notin \pi_P$, and thus $P(w)$ will be undefined.[47]

The best we can do is to approximate the expected value of Q relative to u and any *extension* of P to the entire power set (at least to π_Q—as we will see, this makes no difference under the assumption that u is nice).

[47] Whenever there's no risk of ambiguity, I will abuse notation and write '$P(w)$' instead of '$P(\{w\})$', for $w \in \mathcal{W}$.

For any algebra $\mathcal{A}$ and any probability function P such that $\pi_P \subset \mathcal{A}$, let $\mathbb{P}_P(\mathcal{A})$ denote the set of all extensions of P to $\mathcal{A}$. If Q is a probability function, I will write $\mathbb{P}_P(Q)$ to denote $\mathbb{P}_P(\pi_Q)$. I will use $\mathbb{P}_P$ to denote the set of all extensions of P to the entire power set.

Fix a probability function P. We can now define, for each probability function Q and each $p \subset \mathcal{W}$:

Definition A.2:

$$\overline{EU}(Q,p) = \sup_{P' \in \mathbb{P}_P} \sum P'(w \mid p) u(Q,w)$$
$$\underline{EU}(Q,p) = \inf_{P' \in \mathbb{P}_P} \sum P'(w \mid p) u(Q,w)$$

Since u is nice, whenever $w \in q \in \pi_Q$ we have $\overline{EU}(Q,q) = \underline{EU}(Q,q) = u(Q,w)$, so we can extend our function so that $u(Q,q)$ is well-defined. This has as an immediate consequence the following easy fact:

Fact A.3: For any $P' \in \mathbb{P}_P$, and any Q,

$$\sum P'(w \mid p) u(Q,w) = \sum_{q \in \pi_Q} P'(q \mid p) u(Q,q).$$

We can now define the upper and lower expected values of an extension Q of P as follows (cf. Manski 1981):

Definition A.4:

$$\overline{EU}(Q) = \overline{EU}(Q,\mathcal{W}) = \sup_{P' \in \mathbb{P}_P(Q)} \sum_q P'(q) u(Q,q).$$
$$\underline{EU}(Q) = \underline{EU}(Q,\mathcal{W}) = \inf_{P' \in \mathbb{P}_P(Q)} \sum_q P'(q) u(Q,q).$$

Clearly, $\overline{EU}(Q) \geq \underline{EU}(Q)$, with equality iff $\pi_P = \pi_Q$.[48]

Now, given a utility function u, we can compare two extensions of Q, Q' of P in many ways. For example, we can ask which one maximizes $\overline{EU}$ (maximax), which one maximizes $\underline{EU}$ (maximin), etc. Presumably there will be things to be said in favor of each of these decision rules. But we need a better understanding of what utility functions are like if these decision rules are not to degenerate into triviality. Further, we need to see whether there are any useful generalizations to be made given a set of constrains on utility functions.

[48] Note that if Q is a coarsening of P, then $EU_{P,u}(Q) = \sum_p P(p) u(Q,p)$ is well-defined.

A.2 *Epistemic utility functions*

There is a substantial body of literature on so-called *scoring rules* or *epistemic utility functions*.[49] All extant discussions, however, restrict their attention to functions of the form

$$u : \mathbb{P}(\mathcal{A}) \times \mathcal{W} \to \mathbb{R},$$

where $\mathbb{P}(\mathcal{A})$ is the set of all probability distributions over a fixed algebra $\mathcal{A}$. In this context, epistemic utility functions are utility functions that satisfy a number of constraints, like TRUTH-DIRECTEDNESS, or PROPRIETY. How far can we generalize these constraints to the case at hand? In other words, how should we state versions of these constraints for epistemic utility functions whose domain includes pairs of the form (P, w) and (Q, w) with $\pi_P \neq \pi_Q$?

The weakest extension of these principles would just require that an epistemic utility function satisfies TRUTH-DIRECTEDNESS and PROPRIETY when restricted to a given algebra. On my view, this is the only plausible generalization to epistemic utility functions that can evaluate probability distributions over different domains. Let me explain

Let us first consider TRUTH-DIRECTEDNESS. The only candidate extension that seems to make sense would be this. First, assume that π_P and π_Q have the same cardinality. Fix a bijection $f : \pi_P \to \pi_Q$. Then the generalized version of TRUTH-DIRECTEDNESS would require that if for all $s \in \pi_P$, $|P(s) - s(w)| \leq |Q(f(s)) - f(s)(w)|$, then $u(P, w) \geq u(Q, w)$. The problem is to find a principled way of fixing the bijection. For a given $w \in \mathcal{W}$, we can require that $s(w) = f(s)(w)$, but this does not give us that much traction. If we had some version of EXTENSIONALITY (Joyce 2009) then perhaps we could motivate this way of generalizing the principle.

Now, the case against generalizing PROPRIETY is more straightforward. Again, the minimal change we need to make to PROPRIETY is what I'll call PARTITION-WISE PROPRIETY, which essentially amounts to the claim that $u \restriction \mathbb{P}(\mathcal{A}) \times \mathcal{W}$ must be proper. But beyond this, the only generalization that can be motivated is this:

UNIVERSAL PROPRIETY: For any $P \neq Q$, $EU_{P,u}(P) > \overline{EU}_{P,u}(Q)$.

However, UNIVERSAL PROPRIETY cannot be satisfied by *any* epistemic utility

[49] E.g. Greaves and Wallace 2006; Joyce 1998, 2009.

function:

Fact A.5: No nice epistemic utility function can be universally proper.

Proof. Assume otherwise: let u be nice and universally proper, and let Q be a non-trivial extension of P. Since $Q \neq P$, we have

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

where the last equality follows from the probability calculus.[50]

Now, since Q is an extension of Q, we have $P(p) = Q(p)$ for $p \in \pi_P$. Thus, since u is universally proper, we have:

$$\sum_q Q(q)u(Q,q) > \sum_p P(p)u(P,p) >$$
$$\sup_{P' \in \mathbb{P}_P(Q)} \sum_q P'(q)u(Q,q) \geq \sum_q Q(q)u(Q,q).$$

a contradiction. □

Can we specify additional constraints that must be met by all epistemic utility functions, other than partition-wise truth-directedness and partition-wise propriety? The notion of counterfactual resilience promises to give us a way to do so—in particular, a way of extending our framework in order to compare probability functions with different domains.

A.3 *Counterfactual resilience*

Let us fix a class S of *potential suppositions*. The degree of counterfactual resilience of e, relative to P and S, is given by:

$$CR_{S,P}(e) = 1 - \frac{1}{|S|}\sum_{s \in S}(P(e) - P(s \mathrel{\square\!\!\rightarrow} e))^2.$$

In other words the more counterfactually resilient e is, relative to P and S, the more robust the value of $P(e)$ is under counterfactual suppositions with elements of S.

Now, suppose we have a class $\mathcal{E}$ of *explananda*. Other things being equal, a probability function that assigns to each $e \in \mathcal{E}$ a high degree of counterfactual

[50] Since u is nice, $u(P,q) = u(P, \pi_P(q))$, where π_P is the projection onto π_P of q, so $u(P,q)$ is well-defined.

resilience relative to S is better, epistemically, than one that does not. What we want is for our epistemic utility functions to track these differences.

To get there, however, we need to say something about how $P(s \boxrightarrow e)$ is related to $P(s)$ and $P(e)$. (This is because we have no guarantee that if s and e have a well-defined value under P, so does $s \boxrightarrow e$.) In other words, we need to say something about the probabilities on counterfactuals: better yet, about credences in counterfactuals.[51]

A.4 *Generalized imaging*

Fix $P : \mathcal{A} \to [0, 1]$ a probability function, and for each $c \in \mathcal{A}$ let $\mu_c : \mathcal{A} \times \pi_P \to [0, 1]$ assign to each $p \in \pi_P$ a probability function $\mu_c(\cdot, p)$ over $\mathcal{A}$.

Definition A.6: The *image* of P on c relative to μ is defined as:

$$P_\mu(x \setminus q) = \sum_{p \in \pi_P} P(p)\mu_c(x, p)$$

Intuitively, μ_c is supposed to correspond to a measure of similarity among worlds: $\mu_c(x, p)$ tells you how x worlds fare in terms of closeness to p worlds among worlds in which c holds.

Following Joyce and Lewis,[52] I will take as primitive a similarity function such that $p[c]$ consists of those c-worlds that are most similar to p-worlds.[53] Given this similarity function, we can define our similarity measure as follows

$$\mu_c(x, p) = P(x \mid p[c]),$$

so that

$$P_\mu(x \setminus c) = \sum_{p \in \pi_P} P(p)P(x \mid p[c])$$

[51] There is a long tradition in philosophy linking beliefs in conditionals with conditional probabilities. The idea goes back to Ramsey, and the pipe dream is to find a semantics for the indicative conditional '>' such that $P(a > b) = P(b \mid a)$. It is well-known that this is just that: a pipe dream. In light of a number of triviality results, we now know that there is no interesting way of designing such a semantics. Despite all this, it is widely acknowledged that there is a deep connection between our beliefs in indicative conditionals and the corresponding conditional beliefs.

[52] Cf. Lewis 1976. I will be using the particular formulation due to Joyce 1999.

[53] Talk of 'similarity to p-worlds' only makes sense if we assume that any two worlds $w, w' \in p$ are equally close to one another and that for any $y \in \mathcal{W}$ in which c is true, y is as close to w as it is to w'. I will henceforth make that assumption. In the terminology of Joyce 2010, p. 149 this amounts to the claim that the similarity function is *stratified* for c with partition π_P.

Now, presumably, we want $P(x \setminus c)$ to be defined even when the ratio $P(x \wedge c)/P(c)$ is not (either because $P(c)$ is undefined, or because $P(c)$ is zero), so we will need to stipulate that conditional probabilities are primitive.

Remark A.7: *Assume the $p[c]$ form a partition of c; further assume that $P(x \mid p[c])$ is defined for each $p \in \pi_P$, $c \in \mathcal{A}$. Then it follows from the probability calculus that*

$$P(x \mid c) = \sum_p P(p[c] \mid c) P(x \mid p[c]).$$

Thus, both imaging and conditionalization are weighted averages of $P(x \mid p[c])$, but the weights will differ in general.

A.5 *Imaging and expansions*

I have left implicit the dependence of $P(x \setminus c)$ on the similarity function $(p, c) \mapsto p[c]$, but it is time to make it explicit. From now on, I will write $P_\sigma(\cdot \setminus c)$ to denote the image of P on c relative to σ, where $\sigma : \mathcal{A} \times \pi_P \to \mathcal{A}$. I will also denote by $p_\sigma[c]$ the set of c-worlds that are most similar to p relative to σ.

Note that in order for the image of P on a condition c to be well-defined, relative to a similarity function σ, on a point x, all the relevant conditional probabilities of the form $P(x \mid p_\sigma[c])$ for each $p \in \pi_P$ must be well-defined. This gives us a notion of *accessibility* for similarity functions:

Definition A.8: A probability function P has *access* to a similarity function σ (relative to a set of suppositions S, and and explanandum e) just in case, for all $p \in \pi_P$, and all $c \in S$, $P(e \mid p_\sigma[c])$ is well-defined.

The reason this is relevant to our purposes is that if P is an expansion of Q (that is, $\pi_Q \subsetneq \pi_P$, and for each $q \in \pi_Q$, $P(q) = Q(q)$), P may have access to more similarity functions than Q. This because even when $Q(p)$ is well-defined, $Q(x \mid p_\sigma[c])$ may not be.

A.6 *Toy models*

I now want to revisit the first toy example of §8. Why does the postulation of a new variable—reflecting whether a Red was in state R—increase the expected counterfactual resilience of our explanandum?

Let C be your credence function before the postulation of the new variable. Let red-n stand for the proposition that a given Red was exposed to red light at

time t_n (similarly for blue, and green). Let faster stand for the proposition that a given Red moves faster than normal. The following credence assignments could represent your degrees of belief: $C(\textsf{red-1}) = .5$, $C(\textsf{faster} \mid \textsf{red-1}) = 0$, $C(\textsf{faster} \mid \textsf{blue-1}) = .1$, $C(\textsf{faster} \mid \textsf{blue-2}) = .3$, $C(\textsf{faster} \mid \textsf{blue-2}, \textsf{red-1}) = .9$.

Let e be the explanandum: that the Red is moving faster than normal at time t_2. Presumably, after your observations, $C(e)$ is quite high. Nevertheless, if we let S contain all the descriptions of possible light colors the given Red could have been exposed to at time t_1, we have that $CR_{S,C}(e)$ is not too high. (For $C(\textsf{green-1} \mathrel{\square\!\!\rightarrow} e)$ and $C(\textsf{blue-1} \mathrel{\square\!\!\rightarrow} e)$ are much lower than $C(e)$.)

Now, consider the question whether Red responds to blue light the way it does by virtue of being in state R. Call this proposition H. You have no well-defined credence over H. Nevertheless, the counterfactual resilience of e relative to C_H is high, using as a similarity function the partition generated by R—two worlds are equivalent just in case they agree on whether Red is in state R. This is because for all $s \in S$:

$$C_H(e \setminus s) - C(e) = C_H(R)C_H(e \mid R) - C(e) \approx 0.$$

To assess the value of ?H, of course, we also need to estimate the resilience of e relative to $C_{\neg H}$. But this will presumably be equal to the prior resilience of e. Thus, learning the answer to ?H can be expected to increase the resilience of e.

The second toy example from §8 can be given a similar treatment. Recall that the frequency data corresponding to your observations of second generation critters, prior to the introduction of the new hypothesis, was given by the following table:

SECOND GEN.	*red, red parents*	*red, mixed parents*	*white, white parents*
red, red parents	100% red	100% red	100% red
red, mixed parents	100% red	75% red 25% white	50% red 50% white
white, white parents	100% red	50% red 50% white	100% white

Consider a particular 2-inch red critter that you know comes from mixed parents. Suppose you want to explain e, the proposition that its pattern of distribution of offspring is as follows:

	red, red parents	red, mixed parents	white, white parents
Red offspring (%)	100	75	50

Your degree of belief in e is very high. But if with respect to S, the set of complete specification of the critter's ancestry, e is not highly resilient. For your degree of belief on e on the counterfactual supposition that the critter comes from red parents is very low.

Now consider the expansion of your credence function resulting from adding the proposition that the critter is weak-red (r_{weak}). Conditional on the critter being weak-red, your degree of belief in e will remain close to its prior degree of belief. But its degree of resilience increases. For your credence function now has access to the similarity function σ defined by the partition $\{r_{\mathsf{weak}}, \neg r_{\mathsf{weak}}\}$. And using that partition—which is made salient by the putative explanation in terms of r_{weak}—the posterior resilience of e is given by:

$$CR_{S,P_{r_{\mathsf{weak}}}}(e) = 1 - \frac{1}{|S|}\sum_{s\in S}(P_{r_{\mathsf{weak}}}(e) - P_{r_{\mathsf{weak}}}(s \mathrel{\square\!\!\rightarrow} e))^2,$$

where $P_{r_{\mathsf{weak}}}$ is the result of updating your credence function with r_{weak}.

Now, in order to compare the posterior resilience with the prior resilience, we can do a term by term comparison. The crucial observation is that for each $s \in S$,

$$P_{r_{\mathsf{weak}}}(e) - P_{r_{\mathsf{weak}}}(s \mathrel{\square\!\!\rightarrow} e)$$

is less than or equal to

$$P(e) - P(s \mathrel{\square\!\!\rightarrow} e).$$

And since (where s_{red} is the proposition that the critters parents are both red):

$$P_{r_{\mathsf{weak}}}(e) - P_{r_{\mathsf{weak}}}(s_{\mathsf{red}} \mathrel{\square\!\!\rightarrow} e) << P(e) - P(s_{\mathsf{red}} \mathrel{\square\!\!\rightarrow} e),$$

we have that

$$CR_{S,P_{r_{\mathsf{weak}}}}(e) >> CR_{S,P}(e).$$

And since

$$CR_{S,P_{\neg r_{\mathsf{weak}}}}(e) \approx CR_{S,P}(e),$$

we have that the expeceted increase in resilience of the expansion is non-null.

REFERENCES

Arntzenius, F. 1995. A Heuristic for Conceptual Change. *Philosophy of Science* 62.3, pp. 357–369.

———. 2008. Rationality and Self-Confidence. In: *Oxford Studies in Epistemology*. Ed. by T. S. Gendler and J. Hawthorne. Vol. 2. Oxford: Oxford University Press, pp. 165–178.

Baker, A. 2003. Quantitative Parsimony and Explanatory Power. *The British Journal for the Philosophy of Science* 54.2, pp. 245–259.

———. 2011. Simplicity. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Summer 2011.

Brier, G. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78.1, pp. 1–3.

Bromberger, S. 1992. *On What We Know We Don't Know*. Chicago & Stanford: The University of Chicago Press & CSLI.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, Mass.: MIT Press.

Earman, J. 1992. *Bayes or Bust?* Cambridge, Mass.: MIT Press.

Foley, R. 2009. Beliefs, Degrees of Belief, and the Lockean Thesis. In: *Degrees of Belief*. Ed. by F. Huber and C. Schmidt-Petri. Vol. 342. Synthese Library. Dodrecht: Springer Netherlands. Chap. 2, pp. 37–47.

Forster, M. R. 1999. How Do Simple Rules 'Fit to Reality' in a Complex World? *Minds and Machines* 9.4, pp. 543–564.

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

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

Gibbard, A. 2008. Rational credence and the value of truth. In: *Oxford Studies in Epistemology*. Ed. by T. S. Gendler and J. Hawthorne. Vol. 2. Oxford: Oxford University Press, pp. 143–164.

Good, I. J. 1967. On the Principle of Total Evidence. *The British Journal for the Philosophy of Science* 17.4, pp. 319–321.

Greaves, H. and D. Wallace. 2006. Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind* 115.459, pp. 607–632.

Hájek, A. n.d. Most Counterfactuals are False. Unpublished ms., Australian National University.

Horwich, P. 1982. *Probability and Evidence*. Cambridge: Cambridge University Press.

Huber, F. and C. Schmidt-Petri, eds. 2009. *Degrees of Belief*. Vol. 342. Synthese Library. Dodrecht: Springer Netherlands.

Jackson, F. 1986. What Mary Didn't Know. *Journal of Philosophy* 83.May, pp. 291–295.

Jackson, F. and P. Pettit. 1988. Functionalism and Broad Content. *Mind* 96.387, pp. 381–400.

Jeffrey, R. C. 1983. *The Logic of Decision*. University Of Chicago Press.

Joyce, J. M. 1998. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science* 65.4, pp. 575–603.

———. 1999. *The Foundations of Causal Decision Theory*. New York: Cambridge Univ Press.

———. 2009. Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In: *Degrees of Belief*. Ed. by F. Huber and C. Schmidt-Petri. Vol. 342. Synthese Library. Dordrecht: Springer Netherlands. Chap. 10, pp. 263–297.

———. 2010. Causal reasoning and backtracking. *Philosophical Studies* 147.1, pp. 139–154.

Lange, M. 2005. Laws and their stability. *Synthese* 144.3, pp. 415–432.

———. 2009. *Laws and Lawmakers*. New York: Oxford University Press.

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

———. 1976. Probabilities of conditionals and conditional probabilities. *The Philosophical Review* 85.3, pp. 297–315.

Lipton, P. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27.1, pp. 247–266.

Maher, P. 1995. Probabilities for new theories. English. *Philosophical Studies* 77 (1), pp. 103–115.

Manski, C. F. 1981. Learning and decision making when subjective probabilities have subjective domains. *The Annals of Statistics* 9.1, pp. 59–65.

Nolan, D. 1997. Quantitative Parsimony. *The British Journal for the Philosophy of Science* 48.3, pp. 329–343.

Raiffa, H. and R. Schlaifer. 1961. *Applied statistical decision theory*. Boston: Harvard University Press.

Rayo, A. 2011. A Puzzle About Ineffable Propositions. *Australasian Journal of Philosophy* 89.2, pp. 289–295.

Rooy, R. van. 2004. Utility, Informativity and Protocols. *Journal of Philosophical Logic* 33.4, pp. 389–419.

Skyrms, B. 1977. Resiliency, Propensities, and Causal Necessity. *The Journal of Philosophy* 74.11, pp. 704–713.

———. 1980. *Causal Necessity*. New Haven: Yale University Press.

———. 1990. *The Dynamics of Rational Deliberation*. Cambridge, Mass.: Harvard University Press.

Sober, E. 1998. Black Box Inference: When Should Intervening Variables Be Postulated? *The British Journal for the Philosophy of Science* 49.3, pp. 469–498.

Stalnaker, R. 1968. A theory of conditionals. *Studies in logical theory* 2, pp. 98–112.

Stalnaker, R. 2002. Epistemic Consequentialism. *Aristotelian Society Supplementary Volume* 76.1, pp. 153–168.

Strevens, M. 2004. The Causal and Unification Approaches to Explanation Unified—Causally. *Noûs* 38.1, pp. 154–176.

Sturgeon, S. 2008. Reason and the Grain of Belief. *Noûs* 42.1, pp. 139–165.

White, R. 2005. Explanation as a Guide to Induction. *Philosopher's Imprint* 5.2.

Williams, J. R. G. 2012. Counterfactual triviality: A Lewis-impossibility proof for counterfactuals. *Philosophy and Phenomenological Research* 85.3, pp. 648–670.

Woodward, J. 2010. Scientific Explanation. In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Spring 2010.

Yablo, S. 1992. Mental Causation. *The Philosophical Review* 101.2, pp. 245–280.