View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by PhilPapers

This is a slightly edited version of a paper published in K. Ahlstrom-Vij and J. Dunn, *Epistemic Consequentialism*, OUP, 2018.

# GOOD QUESTIONS

Alejandro Pérez Carballo
*University of Massachusetts, Amherst*

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. 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. The central 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.

I will situate myself within a broadly Bayesian picture of our cognitive economy. On this picture, a cognitive agent can be represented by a probability function—the agent's *credence* function. I will also think of epistemic rationality in broadly decision-theoretic terms: epistemic rationality is a matter of maximizing expected *epistemic* value, where the notion of epistemic value will be modeled using an *epistemic utility function*—an assignment of numerical values to credence functions relative to a given state of the world. An agent's epistemic value function can be seen as incorporating information about which lines of inquiry are more epistemically valuable for an agent. Judgments about the value of questions, I will suggest, can be used to motivate incorporating considerations other than accuracy into our account of epistemic value. In particular, I will argue that we should incorporate explanatory considerations into the epistemic decision-theoretic framework, and offer a proof of concept: a way of doing so that is friendly to the overall consequentialist picture.

## 1 EVALUATING QUESTIONS RELATIVE TO A DECISION PROBLEM

Think of a course of inquiry as a collection of questions. We can identify any such collection with a single question: what is the answer to each of the questions in the collection? 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.[1] 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.

An example might help illustrate this in more detail. 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 |

After computing the expected utility of each possible action, you realize that TH is the action that maximizes expected utility.[2]

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

(?H1) Did the first coin land heads?

(?H2) Did the second coin land heads?

[1] As will become clear below, I will rely on the working hypothesis—widely accepted in the linguistics literature and in work in the erotetic logic tradition (e.g. Hamblin 1958, 1973, Karttunen 1977, Belnap 1963, Groenendijk & Stokhof 1984)—that we can identify a question with the collection of its possible answers. But with some of the questions most central to inquiry—why-questions in particular—it is sometimes far from trivial to figure out what the possible answers are—a point famously emphasized in Bromberger 1962. (See also Friedman 2013 for recent, relevant discussion.) Exactly how to extend the framework I introduce below so as to evaluate such questions is a task for another day.

[2] 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. (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.)

If you have nothing to lose, you should ask one of these questions.[3] But 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 $p$, 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 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), let us 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.[4] 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 \$1.5, so that $V(\text{H1}) = \$1.5$.

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 whichever action maximizes your posterior expected utility minus the expected value of TH, both

[3] 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 & Schlaifer 1961. For a general discussion of Good's theorem, see Skyrms 1990.

[4] Throughout, I will use lowercase italics as variables ranging over propositions—including 'act' propositions in the sense of Jeffrey 1983. I will reserve lowercase small caps for names of specific propositions (e.g. 'H1', etc.).

[5] As it turns out, one could also assign value to a proposition $p$ by looking at the difference between the *posterior* expected values of the action that maximizes expected value relative to the result of conditionalizing on $p$ and the prior expected value of 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.

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 ?H1 to the weighted average of the values of its answers, so that $V(\text{?H1}) = \$.75$.[6] And in the same way, we can assign a value to ?H2—it is easy to verify that $V(\text{H2}) = \$0$, and $V(\text{T2}) = \$7.5$, so that the weighted average of the value of H2 and T2, i.e. $V(\text{?H2})$, equals $\$2.25$.[7] The upshot is that the value of ?H2 is higher than that of ?H1, so that Good's strategy recommends you ask ?H2, 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

Suppose you could will to believe. That is, suppose you could control what credence function you have. Then you could be seen as facing a decision problem: that of deciding, among different possible credence functions, which one to adopt. Like any other decision situation, this one would take place against the backdrop of a given utility function: an assignment of numerical values to each possible option—in this case, to each credence function—relative to a given state of the world.

To take a simple example, suppose you have newly minted coin which will be tossed once (tomorrow) and then destroyed. Suppose you assign .2 credence to H and .8 credence to T (nevermind why). You are now faced with the following decision situation. If the coin lands heads, then you will get $x$ dollars, where $x$ is the credence you assign to H. If the coin lands tails, then you will get $y$ dollars, where $y = (1 - x)$ is the credence you assign to T. If all you care about is money, and you are able to control what credence function to have, you should adopt the credence function that assigns zero to H. This is because the expected utility of adopting credence $x$ in T is given by

$$0.2 \times x + 0.8 \times (1 - x),$$

which is maximized at $x = 0$.

A *cognitive decision problem*, as I will understand it, is a decision problem where the options are the agent's possible credence functions. An *epistemic decision problem* is a cognitive decision problem where the utility function captures one (or more) epistemic dimension(s) of evaluation. In the example above, the utility function in question was defined as follows:

$$u(C, s_H) = \$1 \times C(\text{H})$$
$$u(C, s_T) = \$1 \times C(\text{T}),$$

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

where $s_H$ is the state of the world in which the coin lands heads and $s_T$ the one in which the coin lands tails. Note that, relative to this function, the utility of $C$ at $s_H$ (resp. $s_T$) is higher the closer $C(\textsc{h})$ (resp. $C(\textsc{t})$) is to the truth value of $s_H$ (resp. $s_T$). Arguably, then, such a utility function captures a concern for the truth, and the corresponding decision problem thus counts as an epistemic decision problem.[8]

Whether you think that a particular cognitive decision problem counts as an epistemic decision problem will then depend on whether you think the relevant utility function counts as an *epistemic* utility function. But suppose we can agree that a particular decision problem is an epistemic decision problem. Then we can use Good's strategy in order to evaluate questions relative to that decision problem. This, I submit, would count as an *epistemic* way of evaluating questions.

Granted, the discussion above was framed under the supposition that we can believe at will. And while some still try to maintain that some form of doxastic voluntarism is right, it would be a pity if the applicability of my proposal depended on their success.[9]

Fortunately, I think we can ultimately discharge the assumption in one of at least two ways. In order to use expected utility theory we need not assume that among the options ranked in terms of expected utility, it is 'up to the agent' which one to take. Furthermore, we would learn something about epistemic agents like ourselves if we looked at what epistemic changes are rational for agents who can form beliefs at will.

On the first point: 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. Nothing in the apparatus requires assuming that it is 'up to the agent' which option to take. If we think of all possible epistemic states of an agent as options, and we have a utility function defined for each such option (relative to a state of the world), we can then use expected utility theory to evaluate each of those options. The value of a question can then be understood along the following lines: the better position the question puts you in with respect to evaluating your epistemic options, the better the question.

On the second point: we talk, sometimes, of beliefs 'aiming at the truth'. But this, one might object, borders on incoherence unless beliefs can be formed at will. Talk of the aim of belief, one might continue, suggests that beliefs are the

[8] Cf. Horwich 1982, p. 127ff, as well as Maher 1993, p. 177ff. As is well known, this epistemic utility function is not *proper*, in that the expected epistemic utility of a credence function $C$ relative to $C$ itself may sometimes be something other than $C$.

[9] I'm treating doxastic voluntarism as entailing that one can decide what to believe. But this is something that proponents of doxastic voluntarism will plausibly want to deny—for discussion, see e.g. Shah 2002. All that matters for my purposes, however, is that the decision-theoretic framework can be profitably deployed in epistemology without presupposing any form of doxastic voluntarism.

result of a voluntary choice.[10] The response, of course, is to insist that talk of beliefs 'aiming at the truth' is metaphorical. Still, the complaint should put some pressure on us to spell out the metaphor a bit more clearly.

One way of doing so—not the only one[11]—is due to Allan Gibbard. The suggestion, as I understand it, is to suppose that while *we* cannot form beliefs at will, there could be agents that can. Thinking about such agents can shed light on questions about what we should believe:

> If a person is epistemically rational, we can then hypothesize, then it is *as if* she chose her beliefs with the aim of believing truths and shunning falsehoods. She doesn't literally set out to believe truths, the way she might set out to get a high score on a test by intentionally putting down the right answers. But it is as if she did: it is as if she aimed at truth and away from falsehood in her beliefs in the same way one aims at any other goal.[12]

From this, Gibbard claims, we can extract a constraint on epistemic rationality:

> A way of forming beliefs should at least satisfy this condition: if one forms beliefs that way, it will be as if one were, by one's own lights, forming beliefs voluntarily with the aim of believing truths and not falsehoods.[13]

Assuming this is a good strategy, the following should seem quite plausible. A way of evaluating questions should at least satisfy this condition: it will be as if one were evaluating the question with an eye towards solving an epistemic decision problem.

## 3 EVALUATING QUESTIONS WITH ACCURACY MEASURES

Consider another example. Again, fix a coin and suppose it will be tossed exactly three times. You have a choice among all credence functions defined over the smallest collection of propositions closed by conjunction and negation that includes each of the following:

- The first toss of the coin will land heads. (H1)
- The second toss of the coin will land heads. (H2)
- The third toss of the coin will land heads. (H3)
- The coin is 80% biased towards tails. (B)
- The coin is fair. (F)

[10] Cf. Shah & Velleman 2005, p. 498f.
[11] See e.g. Velleman 2000, p. 244ff.
[12] Gibbard 2008, p. 144f.
[13] *Ibid.*, p. 146.

To keep things simple, let's restrict our attention to credence functions such that

$$C(\textsc{b}) = C(\neg\textsc{f}),$$

so that

$$C(x) = C(\textsc{b}) \times P_{.2}(x) \; + \; C(\neg\textsc{b}) \times P_{.5}(x),$$

where $P_n$ is a probability distribution that treats the three tosses as independent random variables with $P_n(\textsc{h}1) = P_n(\textsc{h}2) = P_n(\textsc{h}3) = n$, $P_n(\textsc{b}) = 0$ if $n = .5$ and $P_n(\textsc{b}) = 1$ otherwise. In short, we are restricting our attention to the class $\mathcal{C}$ of *mixtures* of two fixed probability functions, $P_{.5}$ and $P_{.2}$, which correspond to the two different possible biases of the coin.[14]

You are then facing a cognitive decision problem—that of selecting one credence function among those in $\mathcal{C}$. Given an epistemic utility function, you can evaluate different questions from an epistemic perspective, for example:

($?\textsc{b}$) Is the coin biased?

($?\textsc{h}1$) Will the first toss land heads?

The choice is not straightforward, in part because learning the answer to each of the questions will give you information about the answer to the other one. Learning about the bias of the coin will change your credence in $\textsc{h}1$. And learning $\textsc{h}1$ will change your credence about the bias of the coin.

### 3.1 *Comparing questions from an epistemic perspective*

To fix ideas, let's stipulate that we are dealing with a set $\mathcal{W}$ of sixteen possible worlds: eight on which the coin is fair—one for each possible outcome of the three tosses—and eight on which the coin is biased.

Let us also stipulate that our utility function is given by the well-known *Brier score*:[15]

$$\beta(C, x) = -\sum_{w \in \mathcal{W}} (C(w) - \mathbb{1}_{x=w})^2,$$

where $\mathbb{1}_{x=w}$ equals 1 if $x = w$ and 0 otherwise. Given that accuracy is a plausible dimension of epistemic evaluation and that the Brier score is a reasonable measure of accuracy, I will assume that the resulting decision problem is an epistemic decision problem.

We can use this decision problem to compare $?\textsc{b}$ and $?\textsc{h}1$ as before. First, identify $?\textsc{b}$ with the set $\{\textsc{b}, \neg\textsc{b}\}$ and identify $?\textsc{h}1$ with the set $\{\textsc{h}1, \neg\textsc{h}1\}$. To determine the value of each question, recall, we first need to figure out the value of the choice that maximizes expected (epistemic) utility relative to your prior credence function $C_0$.

[14] The inclusion of $\textsc{b}$ in the domain of $P_{.5}$ and $P_{.2}$ allows us to keep things cleaner.

[15] To avoid unnecessary clutter, I write $C(w)$ instead of $C(\{w\})$ when $w$ is a possible world.

It is well-known that $\beta$ is *strictly proper*, in the sense that for all credence functions $C$, the expected $\beta$-value relative to $C$ is maximized at $C$. Thus, we know that the choice that maximizes expected utility (relative to your prior) is $C_0$ itself. For the same reason, we know that the choice that would maximize expected utility if you were to learn B (resp. ¬B), and you conditionalized on that evidence, would be your posterior credence function $C_0(\cdot \mid \text{B})$ (resp. $C_0(\cdot \mid \neg\text{B})$).

To compute the value of ?B, we then need to determine (a) the difference in expected value, relative to $C_0(\cdot \mid \text{B})$, between $C_0(\cdot \mid \text{B})$ and $C_0$, and (b) the difference in expected value, relative to $C_0(\cdot \mid \neg\text{B})$, between $C_0(\cdot \mid \neg\text{B})$ and $C_0$. By parity of reasoning, in order to compute the value of ?H1, we need to determine (c) the difference in expected value, relative to $C_0(\cdot \mid \text{H1})$, between $C_0(\cdot \mid \text{H1})$ and $C_0$, and (d) the difference in expected value, relative to $C_0(\cdot \mid \neg\text{H1})$, between $C_0(\cdot \mid \neg\text{H1})$ and $C_0$. We can then think of the value of each question as a function of your credence in B,[16] whose value can be read off Table 1.[17]

From an epistemic perspective, then, as long as your credence in B is between 0.3 and 0.6, you should prefer learning the answer to ?B over learning the answer to ?H1.

| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $V(\text{?B})$ | 0.04 | 0.07 | 0.092 | 0.105 | 0.11 | 0.105 | 0.092 | 0.07 | 0.04 |
| $V(\text{?H1})$ | 0.105 | 0.093 | 0.086 | 0.084 | 0.086 | 0.09 | 0.097 | 0.108 | 0.124 |

*Table 1:* The values assigned to each of ?B and ?H1 as a function of the prior credence in B, using the Brier score as the relevant epistemic utility function.

We obtain similar results if, instead of the Brier score, we use a different epistemic utility function, such as the *logarithmic score*:[18]

$$\lambda(C, w) = \log C(w).$$

The values of $V(\text{?B})$ and $V(\text{?H1})$ can also be seen as a function of your credence in B (see Table 2).

There is no one General Lesson to be drawn here. The point is simply to illustrate how this framework allows for a principled way of ranking questions

[16] Recall that your credence in B determines your credence in H1.

[17] A *Mathematica* notebook with the relevant computations can be viewed at (or downloaded from) http://perezcarballo.org/files/gq.nb.pdf.

[18] This not a quirk of the particular choice of numbers. Indeed, in the majority of cases, the two utility functions will agree on which question to ask. For example, for all but 10 out of 45 combinations of multiples $i$ and $j$ of 0.1 between 0.1 and 0.9, if the possible bias of the coin is $i$ and you assign credence $j$ to the proposition that the coin is biased, both utility functions will agree on which question to ask.

|  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $V(?\mathrm{B})$ | 0.325 | 0.5 | 0.611 | 0.673 | 0.693 | 0.673 | 0.611 | 0.5 | 0.325 |
| $V(?\mathrm{H1})$ | 0.691 | 0.686 | 0.677 | 0.664 | 0.647 | 0.627 | 0.602 | 0.573 | 0.539 |

*Table 2:* The values assigned to each of ?B and ?H1 as a function of the prior credence in B, using the log score as the relevant epistemic utility function.

from an epistemic perspective, assuming that accuracy is the only dimension of epistemic value. For any two questions, you can compare the expected value of learning their true answers. This is given by the weighted average of the value of each of the questions' answers, where this equals the difference in expected value, relative to the result of updating on that answer, between your posterior and your prior.

### 3.2 *Limitations of an 'accuracy-only' account of epistemic value*

Appealing to accuracy measures will not always allow us to distinguish among different questions. Suppose you only assign non-trivial credence to four atomic propositions, $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$, and $\neg p \wedge \neg q$, and suppose you take $p$ and $q$ to be independent of one another. If $C(p) = C(q)$, then the expected gain in accuracy from learning the answer to the question whether $p$ will equal that of learning the answer to the question whether $q$.

For example, suppose the coin from above is going to be tossed exactly twice. Further suppose that you are certain that the coin is fair, so that $C(\mathrm{B}) = 0$, and $C(\mathrm{H1}) = C(\mathrm{H2}) = 0.5$. If we use the Brier score as our epistemic utility function, we have:

$$V(?\mathrm{H1}) = C(\mathrm{H1}) \cdot V(\mathrm{H1}) + C(\neg\mathrm{H1}) \cdot V(\neg\mathrm{H1}),$$
$$V(?\mathrm{H2}) = C(\mathrm{H2}) \cdot V(\mathrm{H2}) + C(\neg\mathrm{H2}) \cdot V(\neg\mathrm{H2}),$$

where

$$\begin{aligned}
V(\textsc{h1}) &= -\sum_{w} C(w \mid \textsc{h1}) \cdot (\beta(C(\cdot \mid \textsc{h1}), w) - \beta(C, w)), \\
V(\neg\textsc{h1}) &= -\sum_{w} C(w \mid \neg\textsc{h1}) \cdot (\beta(C(\cdot \mid \neg\textsc{h1}), w) - \beta(C, w)), \\
V(\textsc{h2}) &= -\sum_{w} C(w \mid \textsc{h2}) \cdot \beta(C\,(\cdot \mid \textsc{h2}), w) - \beta(C, w)), \\
V(\neg\textsc{h2}) &= -\sum_{w} C(w \mid \neg\textsc{h2}) \cdot (\beta(C(\cdot \mid \neg\textsc{h2}), w) - \beta(C, w)).
\end{aligned}$$

Given the symmetry of the situation, however, we can find a permutation $\sigma$ of the set of worlds such that, for each $w$:[19]

$$\begin{aligned}
C(w \mid \textsc{h1}) &= C(\sigma(w) \mid \textsc{h2}) \\
C(w \mid \neg\textsc{h1}) &= C(\sigma(w) \mid \neg\textsc{h2}) \\
\beta(C, w) &= \beta(C, \sigma(w)) \\
\beta(C(\cdot \mid \textsc{h1}), w) &= \beta(C(\cdot \mid \textsc{h2}), \sigma(w)) \\
\beta(C(\cdot \mid \neg\textsc{h1}), w) &= \beta(C(\cdot \mid \neg\textsc{h2}), \sigma(w)),
\end{aligned}$$

And this is enough to show that $V(\textsc{h1}) = V(\textsc{h2})$ and $V(\neg\textsc{h1}) = V(\neg\textsc{h2})$, so that $V(?\textsc{h1}) = V(?\textsc{h2})$.

This may not come as a surprise. After all, it doesn't seem as if, from an epistemic perspective, there are reasons for preferring the question whether the first toss landed heads over the question whether the second toss landed heads. There are cases, however, where things are not as clear. These are cases in which, although accuracy considerations cannot distinguish between two questions, there are seemingly epistemic considerations that favor one of the two questions.

Suppose our coin from above is going to be tossed five times in a row. As before, your credence function is defined over an algebra that contains the proposition that the coin is 80% biased towards tails (B) as well as propositions that allow you to specify each of the possible outcomes of the ten coin tosses—conjunctions, disjunctions, and negations of H1, …, H5. Now consider the following case:

> KNOWN OUTCOME: This time, the domain of your credence function includes the proposition (call it M) that the coin was minted five days

[19] Since you are certain that the coin is fair, we are essentially dealing with four possible worlds, determined by the four possible sequences of outcomes. Let $\sigma(w)$ be the world that reverses the outcome of the two coin tosses in $w$, so that (e.g.) if in $w$ the first toss lands heads and the second toss lands tails, then in $\sigma(w)$ the first toss lands tails and the second one heads.

> ago. We assume, again, that your credence in F (the proposition that the coin is fair) equals your credence in ¬B. Suppose now that four of the five coin tosses land heads, and that you update by conditionalizing on your evidence. Your resulting credence function will then be $C' = C(\cdot \mid \#\mathrm{H} = 4)$. As it turns out, $C(\mathrm{M}) = C'(\mathrm{B})$.

It follows from Bayes' theorem that your posterior credence in B will be approximately 0.7, assuming you update by conditionalization.[20] And since the outcome of the coin tosses is probabilistically independent of M, this will also be your posterior credence in M. Since M and B are also probabilistically independent of one another, it follows that the value of the question whether M and the value of the question whether B will be the same—assuming, that is, that we are dealing with an accuracy measure.[21]

If we think there are no epistemic considerations other than accuracy that matter for evaluating credence functions relative to a given state of the world, we will think any comparisons we are inclined to make in cases like KNOWN OUTCOME reflect non-epistemic considerations. But we needn't think this. We *could* think there are epistemic considerations other than accuracy, and that our judgments in cases like KNOWN OUTCOME reflect such considerations. On this view, there is something more valuable, epistemically and in light of your other beliefs, about knowing the answer to whether B as opposed to knowing the answer to whether M.

Indeed, cases like KNOWN OUTCOME lend support to the view that epistemic considerations other than accuracy are reflected in our judgments about the value of questions. Knowing whether the coin is heavily biased towards heads will be crucial for determining whether you have a good explanation of the outcome of the four coin tosses. In contrast, knowing whether the coin

[20] From Bayes' theorem we know that

$$C(\mathrm{B} \mid \#\mathrm{H} = 4) = \frac{C(\mathrm{B})C(\#\mathrm{H} = 4 \mid \mathrm{B})}{C(\#\mathrm{H} = 4)}$$

By construction, it thus follows that

$$C(\mathrm{B} \mid \#\mathrm{H} = 4) = \frac{0.5 \times 5 \times 0.8^4 \times 0.2}{0.5 \times (5 \times 0.5^5 + 5 \times 0.8^4 \times 0.2)} = \frac{0.8^4 \times 0.2}{0.5^5 + 0.8^4 \times 0.2}.$$

A bit of algebra finally gives us that

$$C(\mathrm{B} \mid \#\mathrm{H} = 4) = (1 + 0.5^4 \times 0.8^{-4} \times 0.2{-}1)^{-1} \approx \frac{81}{113} \approx 0.7.$$

[21] If we measure accuracy with the Brier score, we can use the same reasoning from above, involving H1 and H2, to show that the value of ?B and the value of ?M will be the same (relative to $C'$.) But the same will be true on any measure of accuracy that satisfies EXTENSIONALITY, in the terminology of Joyce 2009, p. 273f.

was minted five days ago would have no effect, beyond the increase in expected accuracy, on the epistemic standing of your credence function.

Note that I am *assuming* that facts about the bias of the coin can explain facts about the distribution of heads in a sequence of coin tosses. On some interpretations of probability, no such explanatory relations could hold.[22] But the point I am making is a structural one. All we need is a credence function defined over an algebra of propositions such that:

- only four atomic propositions get assigned non-trivial probability;
- two logically independent propositions, among those obtained by disjunction of two of those atomic propositions, are assigned the same non-trivial probability;
- one of those two propositions, but not the other, contributes to explaining some propositions in the algebra that get assigned credence one.

As long as you think such examples can be constructed, you should agree with this: there are situations where explanatory considerations point towards one of two questions which cannot be distinguished in terms of their expected contribution to the overall accuracy of your body of beliefs.

## 4 BEYOND ACCURACY

If we are to rely on epistemic decision problems to compare questions in cases like KNOWN OUTCOME, we need to find epistemic utility functions that take into account considerations other than accuracy. Unfortunately, most work on epistemic utility theory has only looked at measures of accuracy as a source of epistemic utility functions.[23] As a result, there is a paucity of examples for us to choose from. If we want to extend our framework so as to account for cases like KNOWN OUTCOME, we will have to go beyond examples of epistemic utility functions that are found in the literature.

### 4.1 *Weighted accuracy*

The reason the question whether B is better than the question whether M, I've been saying, is this: the value of learning B outstrips the corresponding gain in accuracy. If you were to learn that B, you would be in possession of a (reasonably) good explanation of something you believe to be true. Relative to a world in which both B and M are true, it would be better to have an accurate degree of belief in B than to have an equally accurate degree of belief in M. Our epistemic

[22] This is particularly clear on the most straightforward version of finite frequentism.

[23] As we will see below, however, Joyce himself has described the form that an epistemic utility function could take if it is to incorporate considerations other than accuracy. See the discussion of 'additive scoring rules' in Joyce 2009, p. 272.

utility function, if it is to accommodate our judgment in KNOWN OUTCOME, needs to be sensitive to this difference.

Here is one way of doing so. Suppose we can measure the relative explanatory strength of each proposition $p$ and suppose we can define a function $\lambda$ that assigns a *weight* $\lambda(p)$ to each proposition $p$ that is inversely proportional to its explanatory strength.[24] We can then define an epistemic utility function that treats accuracy with respect to $p$ in a way that is proportional to $\lambda(p)$: for any credence function $C$ and world $w$, the smaller $\lambda(p)$, the more the difference between $C(p)$ and $p$'s truth-value in $w$ matters for the epistemic standing of $C$ at $w$. One such utility function is a simple modification of the Brier score, which we get to in two steps. First, let's define the *full Brier score* of $C$ at $x \in \mathcal{W}$ as follows:

$$\beta_F(C, x) = -\sum_{p \subseteq \mathcal{W}} (C(p) - \mathbb{1}_{x \in p})^2,$$

where $\mathbb{1}_{x \in p}$ equals 1 if $x \in p$ and 0 otherwise. The full Brier score is a strictly proper epistemic utility function, much like the Brier score $\beta$. Now, define the full $\lambda$-Brier score of $C$ at $x \in \mathcal{W}$ as follows:

$$\beta_F^\lambda(C, x) = -\sum_{p \subseteq \mathcal{W}} \lambda(p) \cdot (C(p) - \mathbb{1}_{x \in p})^2.$$

It is easy to see that, as long as $\lambda(p) > 0$ for all $p$, $\beta_F^\lambda$ is a strictly proper epistemic utility function.[25] For example, turn back to KNOWN OUTCOME, and let $\mathcal{W}_0$

[24] I speak of 'explanatory strength' *simpliciter*, but only to keep things simple. Nothing prevents us from building into our function $\lambda$ a particular class of explananda—say, true propositions that are in need of explanation. If we think there is an objective fact of the matter as to what are the facts in need of explanation, then we can fix $\lambda$ accordingly. If instead we think that what facts are in need of explanation depends in part on a particular agent, we will have to let $\lambda$ vary from agent to agent, so that $\lambda(p)$ measures the extent to which $p$ explains what the agent takes to be in need of explanation—where this may well be a function of the agent's credence in $p$, among other things. All that matters for our purposes is that $\lambda$ be held fixed for a given decision problem. I return to these issues in 5 below.

[25] *Proof*: The expected $\beta_F^\lambda$-score of $Q$ relative to $P$ is:

$$\sum_w P(w) \cdot -\sum_{p \subseteq \mathcal{W}} \lambda(p) \cdot (Q(p) - \mathbb{1}_{x \in p})^2.$$

Fix an enumeration $x_i$ of the members of $\mathcal{W}$ and an enumeration $p_j$ of the subsets of $\mathcal{W}$. We can now think of this sum as an $n \times 2^n$ matrix, with $n = |\mathcal{W}|$, where the cell $i, j$ is of the form $P(w_i) \times -\lambda(p_j) \cdot (Q(p_j) - 1)^2$ if $x_i \in p_j$ and of the form $P(w_i) \times -\lambda(p_j) \cdot Q(p_j)^2$ otherwise. For a fixed $j$, we can write the sum the $j$-th row as

$$-\lambda(p_j) \cdot \left( \sum_{x_i \in p_j} P(x_i) \times (Q(p_j) - 1)^2 + \sum_{x_i \notin p_j} P(x_i) \times Q(p_j)^2 \right),$$

or, equivalently,

$$-\lambda(p_j) \cdot \left( P(p_j) \cdot (1 - Q(p_j))^2 + (1 - P(p_j)) \cdot Q(p_j)^2 \right).$$

denote the relevant set of possible worlds and let $\mathcal{F}_0$ denote the collection of subsets of $\mathcal{W}_0$. For any proposition $p$ in the domain of your credence function *other* than B, let $\lambda_0(p) = 1$, and set $\lambda_0(\textsc{b}) = 1/2$. This is an assignment of weights to the relevant propositions that gives B special status. Intuitively, since we are supposing that B has more explanatory strength than any other proposition, we want to give more importance to accuracy with respect to B than to accuracy with respect to any other proposition. Since the full $\lambda$-Brier score of $P$ at $w$ is defined as *minus* the weighted average of the distance between $P(p)$ and $p$'s truth-value at $w$, in order to give more weight to the distance between $P(\textsc{b})$ and B's truth-value we need to multiply the term of the sum corresponding to the distance between $P(\textsc{b})$ and B's truth-value by a *smaller* factor.

It is worth taking a moment to check that the full $\lambda_0$-Brier score is an epistemic utility function that treats distance from the truth with respect to B differently from distance from the truth with respect to any other proposition. To fix ideas, let $w_0 \in \textsc{b}$ and suppose $P$ and $Q$ are probability functions defined over $\mathcal{F}_0$ with $\beta(P, w_0) = \beta(Q, w_0)$. Since $\beta(P, w_0) = \beta(Q, w_0)$, we know that $\beta_F(P, w_0) = \beta_F(Q, w_0)$. Now,

$$\beta_F^{\lambda_0}(P, w_0) = -\left( 1/2(P(\textsc{b}) - 1)^2 + \sum_{p \neq \textsc{b}} (P(p) - \mathbb{1}_{w_0 \in p})^2 \right).$$

Thus:

$$\beta_F^{\lambda_0}(P, w_0) = \beta_F(P, w_0) + 1/2(P(\textsc{b}) - 1)^2 = \beta_F(Q, w_0) + 1/2(P(\textsc{b}) - 1)^2,$$

and

$$\beta_F^{\lambda_0}(Q, w_0) = \beta_F(Q, w_0) + 1/2(Q(\textsc{b}) - 1)^2$$

so that $\beta_F^{\lambda_0}(P, w_0) < \beta_F^{\lambda_0}(Q, w_0)$ iff $(P(\textsc{b}) - 1)^2 < (Q(\textsc{b}) - 1)^2$ iff $P(\textsc{b}) > Q(\textsc{b})$. Thus, the epistemic utility of $P$ at $w_0$, relative to $\beta_F^{\lambda_0}$, will be less than that of $Q$ at $w_0$, again relative to $\beta_F^{\lambda_0}$, iff $P(\textsc{b}) < Q(\textsc{b})$. In other words, if two credence functions are equally accurate with respect to $w_0$, where B is true in $w_0$, $\beta_F^{\lambda_0}$ will favor $P$ over $Q$ iff $P$ assigns higher credence to B than $Q$ does.

Thus, the expected $\beta_F^{\lambda}$-score of $Q$ relative to $P$ can be written as:

$$-\sum_{p \subseteq W} \lambda(p)\,(P(p) \times (1 - Q(p))^2 + (1 - P(p))Q(p)^2)\,.$$

Now, note that the function

$$a \cdot (1 - x)^2 + (1 - a) \cdot x^2$$

takes its minimum at $x = a$. Thus, for each $p$, $P$ and $Q$,

$$\lambda(p) \cdot (P(p) \cdot (1 - Q(p))^2 + (1 - P(p)) \cdot Q(p)^2) > \lambda(p) \cdot (P(p) \cdot (1 - P(p))^2 + (1 - P(p)) \cdot P(p)^2),$$

since $\lambda(p) > 0$. As a result, if $P \neq Q$, the expected $\beta_F^{\lambda}$-score of $Q$ relative to $P$ is strictly smaller than that of $P$ relative to $P$.

### *4.2 Justifying a weight function*

Admittedly, the particular choice of our weight function $\lambda_0$ can seem somewhat *ad hoc*.[26] Even if it made sense to assign more importance to accuracy with respect to B than to accuracy with respect to any other proposition, this was only because of the details of the case at hand. Relative to the limited range of propositions we were considering, B, unlike M, has the benefit of providing a potential explanation for some of the propositions you take to be true. But things would have been different if we had been considering a different range of propositions, some of which might have been very well explained by the truth of M.

One response to this worry would be to relativize the choice of epistemic utility function to the particular collection of propositions over which the agent's credence is defined. If we are working with credence functions defined over an algebra that contains propositions that are well-explained by B, but no proposition that is explained by M, then we should adopt an epistemic utility function that gives greater weight to accuracy with respect to B than to accuracy with respect to M.

Still, this will tell us nothing about what to do if our algebra contains the proposition that the coin is very shiny—which could be well explained by M—as well as the proposition that four out of the first five tosses of the coin landed heads—which could be well-explained by B.

Granted, we could always *count* the number of propositions that could be well-explained by one vs the other. But even if that strategy is on the right track (I doubt that it is), it won't be fine-grained enough for many purposes. The explanation in terms of M of the proposition that the coin is very shiny may not be as good as the explanation in terms of B of the proposition about the distribution of heads in a given sequence of tosses. So the relative epistemic merit of accuracy with respect to a given proposition $p$ cannot be a function simply of the number of propositions in a given algebra that admit of an explanation in terms of $p$.

Better then to define epistemic utility relative to specific explanatory goals. Given a specific explanandum $e$, say that an epistemic utility function is *$e$-centered* iff it assigns greater weight to accuracy with respect to $p$ the more $p$ would contribute to the best explanation of $e$.[27] Given an $e$-centered epistemic utility function, we can use it to compare credence functions relative to any given

[26] I am setting aside the question of how to justify specific numerical assignments. After all, all that mattered to our reasoning above was that the weight assigned to B was strictly smaller than the one assigned to every other proposition.

[27] This is not, of course, the only option. One might want the function that determines how much weight to give to accuracy with respect to $p$ to be sensitive not only to how much it contributes to the *best* explanation of $e$, but also to how much it contributes to non-optimal explanations that meet some adequacy conditions.

world. This, in turn, would allow us to compare questions with respect to the goal of giving an explanation of $e$.

Now, justifying any *particular* $e$-centered epistemic utility function would require having something close to a full theory of explanation: we would need a way of determining what the best explanation of $e$ (at a given world) is and a way of comparing propositions in terms of how much they contribute to that explanation. Alas, I do not have a full theory of explanation to offer. In the next section, I aim to give something like a proof of concept: a way of incorporating explanatory considerations in a specific way in order to define what is, arguably, an $e$-centered epistemic utility function.

## 5 EXPLANATION & STABILITY

Let us start by taking on a few non-trivial theoretical commitments. A diagnosis for a good explanation of $e$, let us say, is that it makes $e$ very *stable*: given the putative explanation, $e$ couldn't have easily failed to be the case.[28] This is no doubt an overly simplistic theory of explanation,[29] but it does capture a strand of many independently attractive accounts of explanation.

### 5.1 *The stability of being well-explained*

Start by thinking of laws of nature. Laws of nature have a high degree of stability.[30] They are also some of the best candidates for explanatory bedrock. We all know the explanatory buck has to stop somewhere. We all agree that stopping at the laws of nature is as good a place as any. I say it is no coincidence that their high stability goes hand in hand with their not being in need of an explanation. It is because laws of nature are so stable—because they would have obtained (almost) no matter what—that they do not cry out for explanation.[31]

Jim Woodward, for example, has argued that a good explanation is one that subsumes the explanandum under an *invariant* generalization, where *invariance*

[28] This characterization of stability is taken, almost verbatim, from White 2005. White argues that stability, thus understood, is a virtue of explanations—at least of those explanations whose explananda cry out for explanation. His goal is to appeal to explanatory considerations in order to solve Goodman's 'new riddle' of induction.

[29] Among other things, it will not do when it comes to explanations of very low probability events. But these are vexed issues beyond the scope of the paper. See Woodward 2010 for discussion and references.

[30] Indeed, some would go so far as to use stability in order to *characterize* what laws of nature are. See, e.g. Lange 2005, 2009. For a different take on the relationship between stability and law-likeness, see Mitchell 1997 and Skyrms 1977, 1980.

[31] 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 stability. For my purposes, however, all I need is that there be an important dimension of explanatory value that is captured by the notion of stability (the same applies to the worries about low probability events mentioned in fn. 29.)

is essentially a form of stability across a specific range of cases.[32] If an explanandum is subsumed under a stable generalization, the explanandum itself will also be stable. (Note that, on this view, what makes for a good explanation of *e* is not that the explanation makes *e* stable, but rather that the explanation itself is stable. Still, we can use the extent to which an explanation makes *e* stable as a diagnosis of how good an explanation it is.)

Finally, on Michael Strevens' *kairetic* account of explanation, a sure sign of a good explanation is that it makes the explanandum stable: for, according to Strevens, a good explanation of *e* is (roughly) one that isolates the facts that 'make a difference' to the causal path that ends with *e*.[33] And what distinguishes difference-makers (you guessed) is that they are as stable as possible.[34]

Let us tentatively accept, then, that whether *e* is sufficiently stable according to *p* is a reasonable proxy for whether *p* contributes to a good explanation of *e*, so that the more stable *e* is, according to *p*, the more *p* contributes to an explanation of *e*. This allows us to specify a strategy for determining how much a particular proposition would affect the explanatory status of *e*:[35]

> EXPLANATION PROVIDES RESILIENCE (EPR): The contribution of *p* to explaining *e* is proportional to the stability of *e* according to *p*.

What is it for *e* to be stable *according to p*? Intuitively, a particular proposition *p makes e* stable if *p* entails that *e* couldn't have easily failed to obtain. More precisely, *p* makes *e* (where *e* is a true proposition) stable iff for a 'wide range' of background conditions *b*, *p* entails that, had *b* not obtained, *e* still would have obtained. Whether *p* makes *e* stable will thus depend on what counts as a 'wide range' of background conditions. For our purposes, we can assume that this is settled by context. Indeed, for our purposes we can assume that for any context there is a range of background conditions *B* such that whether *e* is stable depends on the proportion of $b \in B$ such that *e* would have obtained even if *b* had been false.

[32] Woodward 2005. The details of Woodward's account need not concern us here, but see Woodward 2001.

[33] Strevens 2008.

[34] This isn't quite right. What distinguishes difference-makers from other causal factors that played a role in *e*'s occurrence is that they are the result of abstracting away, from a given 'causal model' of *e*, all the details that aren't necessary to secure the entailment of *e*. Still, the resulting difference-makers will turn out to be those features of the causal history of *e* that suffice to guarantee the occurrence of *e* while being as stable as possible.

[35] Note that saying that *p* contributes a lot to a good explanation of *e* does not amount to saying that *p* alone is a good explanation of *e*. It could be that *p* does more than what is necessary to make *e* stable. For our purposes, however, we need not concern ourselves with the question what is *the* explanation of *e*.

### 5.2 *Explanation sensitivity in epistemic utility functions*

We can now formulate a constraint on epistemic utility functions along the following lines:[36]

> EXPLANATION SENSITIVITY: Relative to the goal of explaining $e$, accuracy with respect to $p$ matters, epistemically, to the extent that $p$ would contribute to an explanation of $e$.

Now, given EPR, EXPLANATION SENSITIVITY entails the following condition:

> STABILITY BIAS: Relative to the goal of explaining $e$, accuracy with respect to $p$ matters, epistemically, to the extent that $p$ makes $e$ stable.

Thus, we can get some traction out of EXPLANATION SENSITIVITY if we find a way of comparing the stability of an explanandum according to different propositions.

There are no doubt many ways of doing so. For concreteness, let us pick a relatively simple measure of stability. Assume first that we have a fixed set $B$ of background conditions for a given explanandum $e$. The stability of $e$ according to a proposition $p$, which we denote by $\mathbf{s}(e,p)$, is the proportion of $b \in B$ such that $p$ entails $\neg b \mathrel{\Box\!\!\rightarrow} e$. We can now define an $e$-centered epistemic utility function as follows. For each $p$, let

$$\lambda_{\mathbf{s}}(p) = \frac{1}{1 + \mathbf{s}(e,p)}.$$

For a given $p$, then, $\lambda_{\mathbf{s}}(p)$ will be a number between 1 and ½ that is inversely proportional to the extent to which $e$ is stable according to $p$. Thus, the full $\lambda_{\mathbf{s}}$-Brier score

$$\beta_F^{\lambda_{\mathbf{s}}}(C,x) = -\sum_{p \subseteq \mathcal{W}} \lambda_{\mathbf{s}}(p) \cdot (C(p) - \chi_p(x))^2.$$

is an $e$-centered epistemic utility function, for accuracy with respect to $p$ will be assigned a weight proportional to how much $p$ contributes to an explanation of $e$.

Turn back, once again, to KNOWN OUTCOME. Assume, as seems plausible, that M—the proposition that the coin was minted two days ago—contributes nothing to the stability of #H = 4—the proposition that four of the five tosses

[36] We may want to qualify this further. We may, for example, restrict this to explananda that 'cry out for explanation'. Or we may want to make this a condition only on the epistemic utility of a credence function at worlds in which the explanandum is true. For present purposes, however, I suggest we stick to the simplest formulation, especially since the examples we will consider involve explananda that *do* cry out for explanation, and of whose truth the relevant agent is certain.

landed heads. In contrast, B—the proposition that the coin was 80% biased towards heads, does increase the stability of #H = 4. After all, given the truth of B, #H = 4 would have obtained even if the initial conditions of the coin tosses had been slightly different.[37]

We can illustrate this further by means of a different example. On your desk in your office sits a paper by a colleague. You owe her comments but haven't had a chance to look at the paper. Unfortunately, student papers are just in, and you need to attend to those before your friend's. You print all of your student papers and, without looking at them, put them on a pile on your desk as you rush out. The next day, you arrive in your office and notice something strange. At the top of the pile sits your colleague's paper. And as you look through the pile, you notice that it consists entirely of copies of your colleagues' paper.

Let R be the proposition that every paper on your desk is a copy of your colleague's paper. You consider two possible explanations of R. One, which we'll call G, tells the following story: your colleague gave her paper to each of the other members of your department; all of them printed a copy at the same time; the pile of papers came out right before your students' papers were printed; since you were in a rush you didn't notice that the papers you picked up from the printer weren't those you had printed. The other, which we'll call F, tells a simpler story: your colleague, who really wanted you to look at her paper, got the custodian to let her into your office; she replaced the pile of your student papers with copies of her own papers, hoping you would just take the hint and read her paper already.

Now, as things stand you think G and F are equally likely given R. And you think R is equally likely given G as it is given F. Nonetheless, your (high) credence in R given F is more stable than is your credence in R given G. This is because, conditional on G, your ending up with a bunch of copies of your colleague's paper on your desk was just a fluke: had one of your other collegues printed her copy a few seconds later, your print job would have taken precedence and the pile would have contained a bunch of your students' papers; had you gone to the printer a few seconds later you would have noticed a print job in progress and would have checked to make sure the pile on the printer corresponded to your print job; had you hit print on your machine a few seconds earlier, you

[37] OK, this isn't quite right, for at least two reasons. For one, the bias of a coin has little to no effect on the outcome of the coin flips we are familiar with—see Jaynes 2003, ch. 10 for the surprising details. But even if we ignore that complication, the number of tosses we're dealing with is small enough that it simply isn't reasonable to suppose that the bias of the coin can have such a big effect on the sequence of outcomes. Better then to think of the case as one involving a large number of tosses, with 80% of them landing heads. Then we can appeal to the Law of Large Numbers to ensure that B really does have an effect on the stability of the claim that the proportion of heads in the sequence is 80%. But I will stick to the simpler formulation, at the cost of having to lie a little (as Paul Halmos would put it), for ease of exposition.

would have arrived at the printer before any of your other colleagues' print jobs came out. In contrast, conditional on F, G was pretty much bound to happen.

If having a good explanation of R is among our epistemic goals, I have been claiming, accuracy with respect to a proposition should matter more to the extent that it contributes to an explanation of R. So, from that perspective, accuracy with respect to F should matter more than accuracy with respect to G, at least if we suppose (as I will) that F is a better explanation of R than G is.[38]

Note that I'm *assuming* that my judgments about the stability of the explanandum given each of G and F are correct. That is, I'm assuming that, according to G, R could have easily failed to obtain; and I'm assuming that, according to F, R couldn't have easily failed to obtain. Thus, whatever set of background conditions $B$ is fixed by the context so that we are allowed to vary those conditions while still remaining within the sphere of worlds that could have easily obtained, the proportion of those $b \in B$ such that F entails $\neg b \Box\!\!\rightarrow \mathrm{R}$ is *much* bigger than the proportion of those $b \in B$ such that G entails $\neg b \Box\!\!\rightarrow \mathrm{R}$. As a result, $\mathbf{s}(\mathrm{R}, \mathrm{F}) \gg \mathbf{s}(\mathrm{R}, \mathrm{G})$, which entails that $\lambda^{\mathrm{R}}_{\mathbf{s}}(\mathrm{F}) \ll \lambda^{\mathrm{R}}_{\mathbf{s}}(\mathrm{G})$. Hence, the full $\lambda_{\mathbf{s}}$-Brier score will assign greater weight to accuracy with respect to F than to accuracy with respect to G.[39]

### 5.3 *Methodological aside*

On the proposal currently on the table, epistemic utility functions can be used to incorporate considerations other than accuracy into an evaluation of epistemic states. In particular, we can use them to take explanatory considerations into account for the purposes of comparing different epistemic states at different states of the worlds. More specifically, I've suggested we use $e$-centered epistemic utility functions in order to compare epistemic states in terms of how well they are doing relative to the goals of accuracy and of having a good explanation of $e$.

Now, there is one way of thinking about the epistemic utility framework on which it essentially provides us with a way of assessing an agent's 'epistemic decisions' by *her own* epistemic lights. On this way of interpreting the framework, the epistemic standing of an agent's epistemic state relative to a given world is relative to *that agent's* epistemic utility function. So if we think, as seems plausible,

[38] I am assuming that F and G are, and that you take them to be, logically independent. As a result, accuracy with respect to one of the two propositions is perfectly compatible (even by your own lights) with inaccuracy with respect to the other.

[39] Note that nothing in what I've said so far requires taking a stand on the debate between those who think that Inference to the Best Explanation is compatible with (subjective) Bayesianism (e.g. Lipton 2004) and those who think it is not (e.g. van Fraassen 1989). In particular, nothing in what I've said so far requires endorsing a form of 'explanationism' that goes beyond subjective Bayesianism (cf. Weisberg 2009 for discussion). All I've claimed is that, relative to the goal of explaining $e$, accuracy with respect to propositions that contribute to such an explanation should matter more than accuracy with respect to propositions that do not. But this is compatible with the claim that such explanatory considerations have no additional evidential import.

that this can only get off the ground if an agent can have reasonable access to what her own epistemic utility function is, then we will find no use for $e$-centered epistemic utility functions. After all, an $e$-centered epistemic utility function is in part determined by what *in fact* counts as a good explanation of $e$. And if an agent is mistaken about what counts as a good explanation of $e$—or if she simply lacks a view as to what a good explanation of $e$ is—her own way of epistemically evaluating an epistemic state may not correspond to an $e$-centered epistemic utility function.

To be sure, this interpretation of the epistemic utility framework is optional. We could say that an agent is epistemically rational just in case she forms her beliefs as if she were aiming to maximize the expected epistemic utility of her epistemic state—where it is up to *us*, as theorists, to determine what the epistemic utility of an epistemic state at a world is. (The contrasting claim would be: an agent is rational just in case she forms her beliefs as if she were aiming to maximize the expected epistemic utility of her epistemic state—where what epistemic utility an epistemic state has, at a given world, is determined by the agent's own views.)

Still, it would be surprising if in order to take into account considerations other than accuracy we are forced to choose a particular way of interpreting the epistemic utility framework.

Fortunately, we can recast most of what I've said so far so that it is compatible with an 'internalist' interpretation of the framework. Recall the suggestion: an $e$-centered epistemic utility function is one that gives more weight to accuracy with respect to $p$ the more $p$ contributes to explaining $e$—where we measured $p$'s contribution to explaining $e$ in terms of how much the truth of $p$ would increase the counterfactual stability of $e$. We could have instead made the alternative suggestion: an $e$-centered epistemic utility function *for an agent* is one that gives more weight to accuracy with respect to $p$ the more the agent takes $p$ to contribute to explaining $e$—where we measure the agent's estimation of $p$'s contribution to explaining $e$ in terms of how much the truth of $p$ would increased the counterfactual stability of $e$ *according to the agent's beliefs*. More specifically, for a fixed set of background conditions $\mathcal{B}$, say that the stability of $e$ according to a proposition $p$ *and a credence function* $C$, which we denote by $\mathbf{s}_C(e,p)$, is the proportion of $b \in B$ such that $C(\neg b \mathbin{\Box\!\!\rightarrow} e \mid p)$ is sufficiently high.[40] We can now define an $e$-centered epistemic utility function *for an agent with credence*

[40] If we assume, as we have so far, that $C(e) = 1$, then $\mathbf{s}_C(e,p)$ will be proportional to the following measure:

$$\rho_C(e,p) = \frac{1}{|\mathcal{B}|} \cdot \sum_{b \in \mathcal{B}} \mid C(e \mid p) - C(\neg b \mathbin{\Box\!\!\rightarrow} e \mid p) \mid,$$

which is essentially a counterfactual version of Skyrm's measure of *resilience*—see e.g. Skyrms 1977, 1980.

*function C* as follows. For each $p$, let

$$\lambda_{\mathbf{s}_C}(p) = \frac{1}{1 + \mathbf{s}_C(e, p)}.$$

For a given $p$, then, $\lambda_{\mathbf{s}_C}(p)$ will be a number between 1 and ½ that is inversely proportional to the extent to which $e$ is stable according to $p$ and $C$.

Of course, in order for this to work, we'll need to assume that the agent's credence function is defined over a much richer algebra—in particular, one that includes all the relevant counterfactuals. But this is to be expected if we are going to rely on an agent's own judgment about the explanatory worth of a proposition to be what determines her epistemic utility function.

## 6 CLOSING

The epistemic utility framework yields a natural way of evaluating questions from an epistemic perspective. And the richer our notion of epistemic utility—the more it departs from an accuracy-only perspective—the more fine-grained our comparison of questions will be. Of course, any move away from a pure accuracy-centered perspective needs to be justified on epistemic grounds. I've offered one possible way of doing so: allowing explanatory considerations to play a role in determining the epistemic merits of a credence function at a world. In doing so, I relied on a somewhat narrow account of what it takes for a proposition to contribute to an explanation, but one that I hope can give a sense of how the full story could eventually be told.

A few issues are left outstanding. First, can we avoid having to relativize epistemic utility to individual explanatory goals? One possibility worth exploring involves identifying a feature of an agent's credence in $e$ that reflects whether the agent takes $e$ to be in need of explanation.[41] This would allow us to incorporate the value of 'explanatory closure' *simpliciter* into epistemic utility functions.

A second issue is whether stability is the best tool for measuring the explanatory worth of a proposition. While the particular examples I've considered in this paper do presuppose that the extent to which $p$ increases the stability $e$ is a reasonable proxy for whether $p$ contributes to explaining $e$, the overall structure of the proposal could be preserved even if we ended up favoring a different measure of explanatory worth. As long as we can assign a numerical value to a given proposition $p$ that measures the contribution of $p$ to explaining $e$, we can use those numbers to generate weighted accuracy measures which, if I am right, will result in $e$-centered epistemic utility functions.

Finally, there is the question whether considerations other than explanatory power and accuracy ought to be incorporated into an epistemic utility function.

[41] For a proposal amenable to the present framework, see the discussion of the 'Salience Condition' in White 2005, p. 3ff.

If so, we will need to understand whether all such considerations can be agglomerated into an over-all, 'all epistemic things considered' notion of epistemic utility. That would give rise to a very powerful apparatus, one that could allow us to better understand the role of good questioning in inquiry.[42]

## REFERENCES

Belnap, Nuel D. Jr. 1963. *An Analysis of Questions: Preliminary Report*. Technical Memorandum 1287. System Development Corporation.

Bromberger, Sylvain. 1962. An Approach to Explanation. In R. J. Butler (ed.), *Analytical Philosophy*, vol. 2, 72–105. Oxford University Press. Reprinted in Bromberger 1992, pp. 18–51.

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

van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

Friedman, Jane. 2013. Question-directed Attitudes. *Philosophical Perspectives* 27(1). 145–174.

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

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

Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD dissertation, University of Amsterdam, Amsterdam.

Hamblin, Charles L. 1958. Questions. *Australasian Journal of Philosophy* 36(3). 159–168.

Hamblin, Charles L. 1973. Questions in Montague English. *Foundations of Language* 10(1). 41–53.

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

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Pres.

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

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

[42] A much earlier version of this paper—a descendant of chapter 3 of Pérez Carballo 2011—has been circulating for a little while. (That version contains a discussion of how to adapt the epistemic utility framework to compare credence functions that have different domains, which is now part of Pérez Carballo 2016.) Thanks to Rachael Briggs, Fabrizio Cariani, David Christensen, Paolo Santorio, Mark Schroeder, James Shaw, Katia Vavova, Roger White, members of the Coalition of Los Angeles Philosophers, participants in my Fall 2013 graduate seminar (joint with Chris Meacham) on Epistemic Value at UMass Amherst, as well as audiences at the Australian National University, the University of Sydney, and the University of Southern California for helpful comments and advice at different stages of this project. Thanks also to Jeff Dunn, Kristoffer Ahlstrom-Vij, and an anonymous referee for Oxford University Press for helpful comments on a previous draft of this paper. For extensive and invaluable feedback, special thanks to Sylvain Bromberger, Chris Meacham, Agustín Rayo, Bob Stalnaker, and Steve Yablo.

Karttunen, Lauri. 1977. Syntax and Semantics of Questions. *Linguistics and Philosophy* 1(1). 3–44.

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

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

Lipton, Peter. 2004. *Inference to the Best Explanation*. Second. London: Routledge.

Maher, Patrick. 1993. *Betting on Theories* (Cambridge Studies in Probability, Induction, and Decision Theory.). Cambridge: Cambridge University Pres.

Mitchell, Sandra D. 1997. Pragmatic Laws. *Philosophy of Science* 64. S468–S479.

Pérez Carballo, Alejandro. 2011. *Rationality without Representation*. PhD dissertation, Massachusetts Institute of Technology.

Pérez Carballo, Alejandro. 2016. New boundary lines. Unpublished manuscript, University of Massachusetts, Amherst.

Raiffa, Howard & Robert Schlaifer. 1961. *Applied Statistical Decision Theory*. Boston: Harvard University Press.

van Rooy, Robert. 2004. Utility, Informativity and Protocols. *Journal of Philosophical Logic* 33(4). 389–419.

Shah, Nishi. 2002. Clearing Space for Doxastic Voluntarism. *The Monist* 85(3). 436–445.

Shah, Nishi & J. David Velleman. 2005. Doxastic deliberation. *The Philosophical Review* 114(4). 497–534.

Skyrms, Brian. 1977. Resiliency, Propensities, and Causal Necessity. *Journal of Philosophy* 74(11). 704–713.

Skyrms, Brian. 1980. *Causal Necessity*. New Haven: Yale University Press.

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

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Cambridge, Mass.: Harvard University Press.

Velleman, J. David. 2000. On the Aim of Belief. In *The Possibility of Practical Reason*, 244–81. Oxford: Oxford University Press.

Weisberg, Jonathan. 2009. Locating IBE in the Bayesian framework. *Synthese* 167. 125–143.

White, Roger. 2005. Explanation as a Guide to Induction. *Philosopher's Imprint* 5(2).

Woodward, James. 2001. Law and Explanation in Biology: Invariance Is the Kind of Stability That Matters. *Philosophy of Science* 68(1). 1–20.

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Woodward, James. 2010. Scientific Explanation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2010.