Episteme, 14, 1 (2017) 59–69 © Cambridge University Press
doi:10.1017/epi.2016.50

ACCURACY UNCOMPOSED: AGAINST CALIBRATIONISM

B. A. LEVINSTEIN
balevinstein@gmail.com

ABSTRACT

Pettigrew offers new axiomatic constraints on legitimate measures of inaccuracy. His axiom called 'Decomposition' stipulates that legitimate measures of inaccuracy evaluate a credence function in part based on its level of calibration at a world. I argue that if calibration is valuable, as Pettigrew claims, then this fact is an explanandum for accuracy-first epistemologists, not an explanans, for three reasons. First, the intuitive case for the importance of calibration isn't as strong as Pettigrew believes. Second, calibration is a perniciously global property that both contravenes Pettigrew's own views about the nature of credence functions themselves and undercuts the achievements and ambitions of accuracy-first epistemology. Finally, Decomposition introduces a new kind of value compatible with but separate from accuracy-proper in violation of Pettigrew's alethic monism.

INTRODUCTION

Credal states can be epistemically good for a number of reasons. They can be informative, justified by the evidence, coherent, or explanatory. They can also be accurate – i.e., they can place high confidence in truths and low confidence in falsehoods. According to *accuracy-first epistemology*, it's this last virtue and only this last virtue that matters.

Many norms of epistemic rationality, however, don't simply tell us to be accurate. They tell us, for instance, to obey probabilism, update by conditionalization, defer to chance, and follow the evidence. Accuracy-firsters must then justify these norms by demonstrating their instrumental connection to the rational pursuit of accuracy. Toward this end they treat accuracy as the epistemic analog of practical utility. By appeal to legitimate measures of accuracy and standard principles of rational choice, they argue that rational epistemic agents who care solely about accuracy will obey them. In other words, by treating accuracy as *epistemic utility*, they justify important derivative norms governing credences.

Richard Pettigrew's *Accuracy and the Laws of Credence* is a fantastic contribution to this project. Pettigrew presents the tremendous accomplishments of epistemic utility theory to date (many of which he himself is responsible for) and then goes on to make novel contributions, including new arguments for conditionalization and probabilism based on his latest account of alethic epistemic value.

Despite these accomplishments, however, there remains one major chink in the armor of accuracy-first epistemology. All arguments in favor of fundamental norms such as probabilism only work for a privileged set of measures of inaccuracy. They fail miserably, in fact, according to some very natural alternative measures. In other words, these arguments

all rely on the notion of 'proximity to the truth' being formalized in a contentious way. Accuracy-firsters then owe a philosophical account of why only their favored measures of inaccuracy are legitimate.

Joyce (1998, 2009) and Leitgeb and Pettigrew (2010) have attempted to provide such an account, but Pettigrew devotes Chapter 3 of his book to explaining why such attempts fail. In Chapter 4, he offers his own new axiomatic constraints on legitimate measures of inaccuracy. He goes on to show that any measure that meets his constraints delivers the desired results. Most importantly, if a measure is Pettigrew-legitimate, then every non-probability function is accuracy-dominated, and no probability functions are dominated.

The maths is right, but I don't think Pettigrew's axioms are. Pettigrew's most interesting axiom, which he calls DECOMPOSITION, stipulates that legitimate measures of inaccuracy evaluate a credence in part based on its level of *calibration* at a world. I argue that if calibration is valuable, as Pettigrew claims, then this fact is an *explanandum* for accuracy-first epistemologists, not an *explanans* for three reasons. First, the intuitive case for the importance of calibration isn't as strong as Pettigrew believes. Second, calibration is a perniciously global property that both contravenes Pettigrew's own views about the nature of credence functions themselves and undercuts the achievements and ambitions of accuracy-first epistemology. Finally, DECOMPOSITION introduces a new kind of value compatible with but separate from accuracy-proper in violation of Pettigrew's alethic monism. Therefore, the main gap in the accuracy-based arguments of epistemic utility theory remains. Even if the only final epistemic end is accuracy, epistemic utility theory has no convincing argument for the claim that only its favored measures are legitimate.

1. THE CHALLENGE

To understand the importance of restrictions on measures of inaccuracy, let's first examine an epistemic-utility argument for probabilism. This argument comes in many different versions, but the basic idea runs as follows. If an agent has a credence function b that doesn't obey the laws of probability, then according to any *legitimate* measure of inaccuracy, b is guaranteed to be less accurate at every world than some probability function b'. In other words, b is *accuracy-dominated*. No probability function, however, is even weakly dominated. That is, for any probability function c and any function c', c is strictly more accurate than c' at some world. Therefore, if accuracy is epistemic utility, adopting a non-probability function as one's credence function is epistemically irrational because that function is dominated by an undominated alternative.

Now, for this argument to work, we need a precise understanding of what counts as a legitimate measure of inaccuracy. The fundamental idea, of course, is that credences closer to truth-values are less inaccurate. A credence of .6 in a truth is less inaccurate than a credence of .5 in the same proposition. Likewise, an entire credence function that's uniformly closer to the truth than another is overall less inaccurate.

To formalize this notion, let W be a set of worlds, and $\mathcal{F}$ be a set of propositions over W. For $X \in \mathcal{F}$, we let $v_w(X) = 1$ (=0) if X is true (false) at w. bel($\mathcal{F}$) is the set of belief functions over $\mathcal{F}$, where a belief function assigns some number x in [0,1] to each proposition in $\mathcal{F}$. Note that probability functions are belief functions.

A measure of inaccuracy (also known as a scoring rule) is a function $\mathcal{I} : \text{bel}(\mathcal{F}) \times W \to \mathbb{R}^{+}_{\geq 0}$ that is intended to measure how 'close' a belief function is to the truth at a given world.

The most common inaccuracy measure, and the one Pettigrew himself eventually endorses, is the *Brier Score*:

$$\mathfrak{B}(c, w) = \sum_{X \in \mathcal{F}} (v_w(X) - c(X))^2$$

The Brier Score is quite natural. It measures the squared Euclidean distance between c and v_w. As de Finetti (1974) shows, if inaccuracy is measured with the Brier Score, then the dominance argument for probabilism rehearsed above goes through. That is, all and only non-probability functions are undominated.

However, the same cannot be said for other natural measures, such as the *Absolute Value Score*:

$$\mathfrak{A}(c, w) = \sum_{X \in \mathcal{F}} |v_w(X) - c(X)|$$

This score identifies the inaccuracy of c at w with the sum of the positive pointwise differences between v_w and c.

To see why $\mathfrak{A}$ won't work for our dominance argument, consider:

URN: Suppose an urn contains a Red, Green, and White ball, one of which is sure to be drawn, with *R*, *G*, and *W* denoting the relevant propositions. Alice has credence 1/3 in each of *R*, *G*, and *W*, while Bob has credence 0 in each. According to $\mathfrak{A}$, Bob is sure to have a total score of 1, regardless of which ball is in fact drawn. Alice, however, is sure to receive a score of 1/3 + 2/3 + 2/3 = 5/3. So, according to $\mathfrak{A}$, Bob's non-probabilistic credence function *accuracy dominates* Alice's probabilistic one.

It is hard to see *ex ante* why $\mathfrak{B}$ but not $\mathfrak{A}$ should count as legitimate. Both are standard measures of divergence, and both are continuous in their first argument. Most importantly, both obey the following Pareto Principle, which is a fundamental constraint on any measure of inaccuracy:

TRUTH-DIRECTEDNESS: Suppose $\mathcal{I}$ is a legitimate measure of inaccuracy. If $|b(X) - v_w(X)| \geq |c(X) - v_w(X)|$ for all $X \in \mathcal{F}$, and $|b(X) - v_w(X)| > |c(X) - v_w(X)|$ for some $X \in \mathcal{F}$, then $\mathcal{I}(c, w) < \mathcal{I}(b, w)$.

That is, if b is at least as inaccurate as c for *every* proposition and sometimes more inaccurate, then b should count as more inaccurate overall. We will frequently refer back to this principle below.

The challenge that accuracy-firsters face is to justify constraints that rule out measures that don't support the argument for probabilism (such as $\mathfrak{A}$) while permitting at least some measures that do (such as $\mathfrak{B}$). As mentioned above, Pettigrew thinks (and I agree) that all previous attempts are unsuccessful. The rest of this essay is devoted to examining the constraints Pettigrew currently advocates.

1.1. *Pettigrew's Constraints*

My main focus in this article is on Pettigrew's axiom of DECOMPOSITION. Nonetheless, to get a better picture of Pettigrew's view of accuracy, it's worth laying out all of five of his axioms in brief.[1] The axiom DIVERGENCE ADDITIVITY in particular will be of importance to later discussion.

The first constraint Pettigrew endorses in his official argument is:

ALETHIC VINDICATION: The omniscient credence function at a world is the ideal credence function to have at that world. Thus, v_w is the ideal credence function at w.

Given alethic monism, this constraint is mandatory. The best credence function to have at a world is the one that assigns credence 1 to all truths and credence 0 to all falsehoods at that world. Pettigrew's second constraint is slightly more loaded, but relatively unobjectionable to accuracy-firsters. It says simply that legitimate measures of inaccuracy should be *divergences* from the omniscient credence function. A function $\mathfrak{D}$ is a divergence if $\mathfrak{D}(c, c') \geq 0$ and $\mathfrak{D}(c, c') = 0$ just in case $c = c'$. The constraint is:

PERFECTIONISM: If $\mathcal{I}$ is a legitimate inaccuracy measure, there is a divergence $\mathfrak{D}$ such that $\mathcal{I}(c, w) = \mathcal{I}_{\mathfrak{D}}(c, w) = \mathfrak{D}(v_w, c)$. We say that $\mathfrak{D}$ generates $\mathcal{I}$.

Inaccuracy is supposed to measure how 'far' a credence function is from the truth. Together with ALETHIC VINDICATION, this axiom entails that $\mathcal{I}(c, w) \geq 0$, and $\mathcal{I}(c, w) = 0$ just in case $c = v_w$. Furthermore, because $\mathcal{I}$ is generated by $\mathfrak{D}$, we can appeal to $\mathfrak{D}$ to measure how far apart two credence functions are from one another.

The third axiom relates the divergence between entire credence *functions* (the global divergence) to the divergence between the credences they assign to *individual* propositions (the local divergences). In particular, it requires that we simply identify the global divergence with the sum of the local divergences:

DIVERGENCE ADDITIVITY: If $\mathcal{I}$ is a legitimate inaccuracy measure generated by $\mathfrak{D}$, there is a one-dimensional divergence $\mathfrak{d}$ such that $\mathfrak{D}(c, c') = \sum_{X \in \mathcal{F}} \mathfrak{d}(c(X), c'(X))$. We say that $\mathfrak{d}$ generates $\mathfrak{D}$.

Note that the three axioms listed so far jointly entail that if $\mathcal{I}$ is a legitimate measure of inaccuracy, then $\mathcal{I}(c, w) = \sum \mathfrak{d}(v_w(X), c(X))$. That is, the inaccuracy of an entire credence function simply is the sum of the inaccuracy (i.e., divergence) between an agent's credence in each proposition and the truth-value of that proposition at a world. We'll have more to say about this third axiom later.

The fourth axiom rules out major jumps in inaccuracy or divergence.

1 I'll only discuss the list of axioms on p. 65 of Pettigrew's book. He also endorses an additional axiom that ends up entailing that the Brier Score is the only correct measure of inaccuracy, but that won't come into play here.

Divergence Continuity: If $\mathfrak{I}$ is a legitimate inaccuracy measure generated by an additive divergence $\mathfrak{D}$ that is generated by $\mathfrak{d}$, then $\mathfrak{d}$ is continuous in its first and second argument.

This axiom may be controverted, but I have no objection to it. I wish only to pause to point out:

(1) $\mathfrak{A}$ and $\mathfrak{B}$ are both compatible with the four axioms listed so far.
(2) These four axioms entail Truth-Directedness.

So, these axioms are not yet strong enough to deliver the desired results, but they do entail the most obvious restriction on legitimacy, i.e., Truth-Directedness.

1.2. *Calibration*

The main axiom of interest requires a bit of preliminary discussion. Pettigrew appeals to a notion he calls *calibration inaccuracy*, which is different to the notion of inaccuracy proper.

We say that a credence function c is well-calibrated if $100x\%$ of the propositions c assigns credence x to are true. That is, c is *well calibrated* at a world w if, for every x in c's range:

$$x = \frac{|\{X \in \mathcal{F} : c(X) = x, v_w(X) = 1\}|}{|\{X \in \mathcal{F} : c(X) = x\}|}$$

Perfect calibration is, intuitively, an attractive feature of a credence function. If you're a weather-forecaster, approximately 80% of the time you assign forecast rain with 80% probability, it should actually rain. If not, something's off: The set of propositions you assign credence .8 to does not have the right *truth-frequency*.

Indeed, as Pettigrew notes, philosophers such as van Fraassen (1983) and Shimony (1988) are pure calibrationists. They claim that credences aim not at truth but at frequency of truth. That is, they claim that c is perfectly vindicated at w not when $c = v_w$, but when c is well-calibrated.[2]

If we adopt such a purely calibrationist view, we can try to measure the value of a credence function at a world by its proximity to perfect calibration instead of by its proximity to truth. This is a bit tricky, since many different credence functions are calibrated at the same world. v_w, for instance, is well calibrated. But if $\mathcal{F}$ is closed under negation, then the function $c_{.5}$ that assigns credence .5 to every single proposition in $\mathcal{F}$ is also well calibrated.

Pettigrew claims that each credence function c is naturally compared on this view to its *well-calibrated counterpart* at w, denoted c^w, defined as:

$$c^w(Z) := \frac{|\{X \in \mathcal{F} : c(X) = c(Z), v_w(X) = 1\}|}{|\{X \in \mathcal{F} : c(X) = c(Z)\}|}$$

2 Ramsey (1931) also appeals to a notion of calibration to determine which credence functions are ideal, but his is different to the one explicated here.

c^w corrects c's mis-calibration. For instance, suppose only 70% of the propositions c assigns credence .8 to are true. Then c^w assigns each of those propositions credence .7. Thus, c^w functions as a re-calibration of c at w.

Whereas c's inaccuracy proper is divergence from v_w, c's calibration inaccuracy is its divergence from c^w. That is, the *calibration inaccuracy of c at w relative to divergence* $\mathfrak{D}$ is $\mathfrak{D}(c^w, c)$.

As should be clear, we can't identify calibration inaccuracy with inaccuracy proper. After all, $c_{.5}$ defined above is well calibrated, but it's not perfectly accurate. Indeed, calibration inaccuracy isn't even truth-directed. Suppose:

- $\mathcal{F} = \{X, \neg X\}$
- $v_w(X) = 1$
- $c_{.99}(X) = .99$, and $c_{.99}(\neg X) = .01$.

Then $c_{.99}$ is not well calibrated, but $c_{.5}$ is. By TRUTH-DIRECTEDNESS, however, $c_{.99}$ is strictly more accurate than $c_{.5}$.

So, Pettigrew disagrees with those who identify vindication with calibration. However, he thinks that calibration is nonetheless desirable *ceteris paribus*. In particular, according to Pettigrew, c^w should be less inaccurate overall than c at w (assuming $c \neq c^w$). Calibration should count for something.

In addition to any miscalibration, the inaccuracy of c's well-calibrated counterpart should count toward c's inaccuracy as well. If $c^w = c_{.5}$, for instance, then even after c is recalibrated, it's still not especially accurate. That is, once we correct for one of c's flaws (its *calibration* inaccuracy), another flaw still remains (its remaining inaccuracy proper). On the other hand, if $c^w = v_w$, then after recalibration c is perfectly accurate, so there should be no additional penalty.

Pettigrew's final constraint puts these two intuitions together. c's inaccuracy, according to Pettigrew, should simply be the (weighted) sum of its calibration inaccuracy (i.e., its divergence from c^w) and c^w's inaccuracy (i.e., c^w's divergence from v_w). More explicitly, the final constraint is:

DECOMPOSITION: If $\mathcal{I}$ is a legitimate inaccuracy measure generated by a divergence $\mathfrak{D}$, then there are [positive real numbers] α, β such that:

$$\mathfrak{D}(v_w, c) = \alpha\mathfrak{D}(c^w, c) + \beta\mathfrak{D}(v_w, c^w)$$

Note that DECOMPOSITION helps explain what's wrong with the Absolute Value Score $\mathfrak{A}$. Recall in URN, a credence function $c_\circ$ that assigned credence 0 to R, G, and W $\mathfrak{A}$-dominated a credence function $c_{1/3}$ that assigned credence 1/3 to each of those propositions even though exactly one of those propositions was sure to be true. $c_{1/3}$ just is the well-calibrated counterpart of itself and of $c_\circ$ at each world. Therefore, DECOMPOSITION entails that $c_{1/3}$ is less inaccurate than $c_\circ$, a fact with which the Brier Score, but not the Absolute Value Score, agrees.

2. OBJECTIONS TO DECOMPOSITION

I'll now argue that Pettigrew in particular and accuracy-firsters in general should reject the axiom of DECOMPOSITION. Although it's indeed true that highly accurate credences tend to be nearly well-calibrated, this is a fact that should be derived and not taken for granted.[3]

2.1. *The Intuitive Case*

The first problem with the appeal to calibration is that the intuitive case for its importance is weaker than Pettigrew claims. One important idea motivating DECOMPOSITION is that for any function c, it's natural to evaluate c not just against v_w but against c^w as well. c^w, in other words, is intuitively a function against which we should measure c's success at a world.

It does seem true that professional weather-forecasters should be well-calibrated. Surely, as we already observed, of the days when a weatherperson forecasts rain with x probability, it should rain about x proportion of the time. So, initially, it seems that c^w is a clear benchmark that's at least as good overall as c at w.

However, I think this intuition has bite only for special kinds of credence functions and sets of propositions. There's little motivation for thinking c^w is a good bar for comparison for general c and $\mathcal{F}$. To see why, consider the following three cases:

(1) Suppose c_1 assigns credence .9 to the propositions that it will rain on Monday (R_m) and that it will rain on Tuesday (R_t). c_1 has no opinion about any other propositions. So, $c_1^w(R_m) = c_1^w(R_t) = 0, .5$, or 1 depending on whether it rains on one day, both days, or neither day.

(2) c_2 is defined on the large set of propositions $\mathcal{F} = \{R_1, \ldots, R_{1,000}\}$, where R_i is the proposition that it rains on day i. Suppose exactly 60% of the R_i's are true. For each i, $c_2(R_i)$ is very close to .6, but always just a little above or below. Furthermore, no two credences are exactly the same. That is, $c_2(R_i) \neq c_2(R_j)$ if $i \neq j$. Note that $c_2^w(R_i) = v_w(R_i) = 1$ or 0 for all i.

(3) Suppose $\mathcal{B}$ is a large set of propositions about weather in Bristol, 90% of which are true. $\mathcal{T}$ is an equally large set of propositions about toadstool biology, 50% of which are true. c_3 assigns credence .7 to each and every proposition in $\mathcal{F} = \mathcal{B} \cup \mathcal{T}$. Note that $c_3 = c_3^w$.

If we were trying to design a legitimate measure of inaccuracy, these cases would hardly motivate any evaluative appeal to Pettigrew's notion of calibration or well calibrated counterpart. Indeed, in each case c_i^w seems evaluatively irrelevant. In (1), $\mathcal{F}$ is too small for c_1^w to be intuitively important for the purposes of evaluation. That is, if we were to

3 Mathematically, for the accuracy-dominance argument to work, Pettigrew needs to restrict the class of legitimate measures to those that are additive, continuous, and *strictly proper*. A measure $\mathcal{I}$ is strictly proper if every probability function assigns itself strictly lower expected inaccuracy than it assigns to any other measure. DeGroot and Fienberg (1982, 1983) prove that all such measures in fact obey the axiom of DECOMPOSITION. Pettigrew argues in Chapter 3, however, that we should not rule out improper measures *ex ante*. So, his result that measures obeying his five axioms are strictly proper can be thought of as the inverse of the DeGroot-Fienberg Theorem.

evaluate what's right or wrong with c_1 from an intuitive standpoint, we may appeal to its proximity to truth, but we would hardly mention its proximity to calibration.

In (2), we might think of c_2 as nearly calibrated in some sense, but under the official definition its well-calibrated counterpart is v_w, so it ends up with high calibration inaccuracy according to the official notion. Calibration accuracy here, despite the large set of propositions, is not especially relevant save for the fact that it coincides perfectly in this case with accuracy proper.

In (3), c_3 is fairly *unreliable* over both $\mathcal{B}$ and $\mathcal{T}$. That is, when we restrict c_3's domain to its natural subject matter, c_3 is mis-calibrated. However, it all averages out over the full domain, and c_3 ends up perfectly calibrated over the full, rather gerrymandered set of propositions $\mathcal{F}$.

Intuitively, then, Pettigrew's notion of calibration seems to come out of left field when we think about a number of different kinds of cases. c^w simply isn't a natural counterpart to compare c to in general. It is an appealing comparison, admittedly, when we are thinking about weatherpersons who frequently forecast *exactly* the same probability many times about propositions with similar subject matter. Near calibration here indicates a kind of reliability. However, it's a mistake to take calibration to be such an important intuitive notion relevant to the concept of accuracy based on these sorts of cases. The intuition that calibration matters for accuracy, or even that a credence function is naturally compared to its well-calibrated counterpart, is insufficiently general to motivate DECOMPOSITION.

2.2. *Globalism*

The second problem with DECOMPOSITION is that it invokes an irreducibly *global* concept to constrain the notion of legitimate inaccuracy.

By *global*, I mean that the value of $c^w(X)$ depends on more than just $c(X)$ and $v_w(X)$. It also depends on the credence c and v_w assign to other propositions as well. For instance, if c assigns credence .5 only to X and Y, then the value of $c^w(X)$ is affected by the truth-value of Y even if X and Y have nothing to do with one another. Because Pettigrew invokes such a global notion in DECOMPOSITION, he renders the notion of a legitimate inaccuracy measure irreducibly global as well.

This nonlocality is problematic for Pettigrew in particular. In fact, it conflicts with the motivation he provides for his DIVERGENCE ADDITIVITY axiom, which requires that the overall inaccuracy of a credence function c be the sum of the divergences between $c(X)$ and $v_w(X)$ for each X:

> When we say that we represent an agent by her credence function, it can sound as if we're representing her as having a single, unified doxastic state. But that's not what's going on. *Really, we are just representing her as having an agglomeration of individual doxastic states, namely, the individual credences she assigns to the various propositions about which she has an opinion.* A credence function is simply a mathematical way of representing this agglomeration; it is a way of collecting together these individual credences into a single object.
>
> To illustrate the point, it might help to compare a credence function to a musical melody. Suppose I were to ask how far one melody lies from another. I would not simply treat each as a sequence of notes (pitches and durations) and measure the distance between each note in one and its counterpart in the other, and then sum them up. Rather, I would treat each melody as

> an integrated whole and I would ask how far the overall 'shape' of one lies from the overall 'shape' of the other. *A credence function, on the other hand, is not an integrated whole – it is simply a mathematical representation of a list of credence-proposition pairings.* Thus, we need not look to its 'shape' when we measure its distance from another credence function. (p. 49, emphasis mine)

The idea here is that credence functions are not holistic entities, but simply a way of listing out individual doxastic attitudes which are to be assessed without regard to one another. By appealing to calibration, however, Pettigrew requires precisely the opposite. How we assess what you think about one proposition depends on what you think about other propositions.

Additionally, regardless of Pettigrew's motivation for other axioms, invoking calibration in DECOMPOSITION undercuts the achievements of accuracy-first epistemology. One goal of AFE is to justify global rational constraints on credence functions that aren't themselves explicitly alethic by appeal to local, alethic evaluations. For instance, the norm of probabilism is a global constraint: whether c is a probability function or not depends on the relationship between c's attitudes toward the various propositions in $\mathcal{F}$. It also is not explicitly alethic: probabilism doesn't say anything about the pursuit of the truth or accuracy. The accuracy-first justification for probabilism aims to explain why probabilism is nonetheless a legitimate norm without appealing directly to any structural relationships between c's credences. Instead, it shows that such a relationship emerges from an agent's rational pursuit of accuracy. If the rational agent simply wants each credence individually to be accurate, she'll end up with a probability function.

However, whether an agent is calibrated depends in part on the relationship between her credences. Moreover, since calibration is not even truth-directed, it is up to the accuracy first epistemologist to explain why this global property of credence functions is related to the rational pursuit of accuracy. She should not assume it from the start.

2.3. *Veritism*

The justification for DECOMPOSITION ultimately rests on Pettigrew's view that calibration is itself valuable. As he puts it:

> The motivating intuition for the calibrationist accounts of accuracy that I would like to retain as far as possible is that credences are better the closer they are to being well calibrated. … To retain this motivating intuition, we say that, while … calibration accuracy cannot be the whole story about accuracy because it is not truth-directed, it is nonetheless part of the story – calibration accuracy is a component of accuracy. The other component of accuracy, I claim, is directly motivated by the desideratum of truth-directedness. (p. 63)

There are two *prima facie* ways to understand Pettigrew's views about the value of calibration. On the first reading, calibration accuracy, like accuracy-proper, is an intrinsically valuable feature of a credence function. It's one component of what makes a credence function good that's partly separable from, but compatible with, the goal of having credences close to truth-values.

On the second, weaker reading, high calibration accuracy is merely a property that we somehow know accurate credence functions have. It may, on this view, be a byproduct or spandrel property of valuable credence functions, in the sense that credence functions

approximating the truth by necessity approximate the frequency of the truth. Any measure that fails to appreciate this fact must be a lousy measure. In other words, the axiom of DECOMPOSITION is justified because we notice that credences we judge inaccurate are always either calibration inaccurate, or their well-calibrated counterparts are themselves inaccurate, or both. This relationship is robust even though we don't here value calibration accuracy *per se*. Proximity to truth, not proximity to frequency of truth, is the sole source of goodness.

Now, valuing proximity to truth and to frequency-of-truth are compatible. After all, scoring rules like $\mathfrak{B}$ satisfy all of Pettigrew's axioms and therefore satisfy both TRUTH-DIRECTEDNESS and DECOMPOSITION. Nonetheless, if Pettigrew endorses this first reading, he must embrace a kind of value-pluralism. However, his whole project is based on value-monism: Credences are valuable insofar and only insofar as they are near truth-values. They cannot then acquire any *additional* value because of their proximity to certain *frequencies* of truth-values. Compare: if Pettigrew were to value both accuracy and proportioning one's beliefs to the evidence, then he could not be a value monist even if he maintained that these two values were never in conflict.[4] So, the first reading is untenable for accuracy-firsters like Pettigrew.

Moreover, this first reading would belie Pettigrew's claim that all legitimate inaccuracy measures satisfy his five axioms regardless of his stance on value-monism. For, if the motivation for DECOMPOSITION appeals to the *sui generis* value of calibration, then there's no reason to think that legitimate measures of pure *inaccuracy* should satisfy it. Instead, DECOMPOSITION would be a constraint on the class of legitimate inaccuracy measures that are also plausible measures of total epistemic disutility. This reading thus undermines the purely *alethic* argument for probabilism.

However, even the second reading is problematic for accuracy-firsters. Even if Pettigrew does not think calibration accuracy is an independent and unalloyed good, it is nonetheless an explicitly anti-alethic notion that violates the fundamental constraint on any legitimate measure of inaccuracy, *viz.* TRUTH-DIRECTEDNESS. It is therefore surprising that calibration accuracy has much to do with accuracy proper. The two notions – proximity to truth-values and proximity to well-calibrated counterpart – appear to be in tension. The fact that some measures such as the Brier Score satisfy DECOMPOSITION and Pettigrew's other axioms shows that this tension is not irresolvable, but it does not show what calibration accuracy has to do with accuracy full-stop. Calibration is a derivative good from a purely veritistic point of view. The relationship between reasonable measures of inaccuracy and calibration, or any other notion that violates TRUTH-DIRECTEDNESS, should be derived and not assumed.

3. CONCLUSION

One of the great challenges of accuracy-first epistemology is to characterize the class of legitimate measures of inaccuracy. Pettigrew's latest attempt explicitly appeals to the

4 In his (2013), Pettigrew explicitly endorses the view that following one's evidence is a mere byproduct of the goal of accuracy. He goes on to argue that alternative views (wrongly) endorse a kind of epistemic value pluralism.

notion of calibration. I've argued that such an appeal is philosophically out of bounds for Pettigrew and other accuracy-first epistemologists.

First, the intuitive case for the evaluative importance of calibration isn't as strong as Pettigrew and other philosophers claim. Second, appealing to calibration renders the concept of inaccuracy perniciously global. Third, calibration inaccuracy is non-alethic: some perfectly calibrated credence functions are farther from the truth than uncalibrated ones. Although it is true that there's an important relationship between accuracy and calibration, this relationship should be derived. The importance of calibration is an *explanandum* for accuracy-firsters, not an *explanans*.

REFERENCES

de Finetti, B. 1974. *Theory of Probability*, Volume 1. New York, NY: John Wiley and Sons.

DeGroot, M. H. and Fienberg, S. E. 1982. 'Assessing Probability Assessors: Calibration and Refinement.' In S. S. Gupta and J. O. Berger (eds), *Statistical Decision Theory and Related Topics* III, Volume 1. New York, NY: Academic Press.

—— **and** —— 1983. 'The Comparison and Evaluation of Forecasters.' In *Proceedings of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics*, Volume 32, pp. 12–22. Oxford: Blackwell Publishing.

Joyce, J. M. 1998. 'A Nonpragmatic Vindication of Probabilism.' *Philosophy of Science* 65, 575–603.

—— 2009. 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief.' In F. Huber and C. Schmidt-Petri (eds), *Degrees of Belief*, Volume 342, pp. 263–97. New York, NY: Springer.

Leitgeb, H. and Pettigrew, R. 2010. 'An Objective Justification of Bayesianism. I: Measuring Inaccuracy.' *Philosophy of Science*, 77: 201–35.

Pettigrew, R. 2013. 'Accuracy and Evidence.' *Dialectica*, 67: 579–96.

Ramsey, F. P. 1931. 'Truth and Probability.' In R. Braithwaite (ed.), *Foundations of Mathematics and Other Essays*, pp. 156–98. London: Routledge & Kegan Paul.

Shimony, A. 1988. 'An Adamite Derivation of the Calculus of Probability.' In J. Fetzer (ed.), *Probability and Causality*. Dordrecht: D. Reidel.

van Fraassen, B. 1983. 'Calibration: Frequency Justification for Personal Probability.' In R. Cohen and L. Laudan (eds), *Physics, Philosophy, and Psychoanalysis*, pp. 295–319. Dordrecht: Springer.

BEN LEVINSTEIN works primarily on ethics, epistemology, and decision theory. After receiving his PhD from Rutgers in 2013, he served as a researcher on Richard Pettigrew's Epistemic Utility Theory project. Since then, he's held post-doctoral fellowships at Oxford and Rutgers.

