Learning is a Risky Business

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

Abstract

Richard Pettigrew has recently advanced a justification of the Principle of Indifference on the basis of a principle that he calls "cognitive conservatism," or "extreme epistemic conservatism." However, the credences based on the Principle of Indifference, as Pettigrew formulates it, violate another desideratum, namely, that learning from experience be possible. If it is accepted that learning from experience should be possible, this provides grounds for rejecting cognitive conservatism. Another set of criteria considered by Pettigrew, which involves a weighted mean of worst-case and best-case accuracy, affords some learning, but not the sort that one would expect. This raises the question of whether accuracy-based considerations can be adapted to justify credence functions that permit induction.

1. Introduction. Richard Pettigrew (2016*a*,*b*) has recently advanced an ingenious argument, based on considerations of epistemic accuracy, demonstrating that the attitude he calls "cognitive conservatism" (2016*a*, 46), or "extreme epistemic conservatism" (2016*b*, 166) entails that, in a state of complete ignorance, an agent's credences should satisfy the Principle of Indifference.

As Pettigrew notes, establishing the conditional "Cognitive Conservatism $\Rightarrow$ Principle of Indifference" leaves it open whether this conditional is to be used as a premise in an application of *modus ponens* or *modus tollens*. Pettigrew takes cognitive conservatism to be an appropriate epistemic attitude in the situation imagined in the proof, and hence concludes that prior credences in a state of complete ignorance should satisfy the Principle of Indifference. If, however, we have grounds for rejecting the Principle of Indifference, Pettigrew's conditional turns these into grounds for rejecting Cognitive Conservatism.

In what follows, I will argue that there are, indeed, such grounds. Using the Principle of Indifference to set priors in the sort of situation envisaged by Pettigrew brings with it the peril of foreclosing *a priori* any possibility of learning from experience. If we take it as a principle that one's priors should not render learning impossible, then the conclusion to be drawn is that Cognitive Conservatism should be rejected.

2. Pettigrew's argument, in brief. The question at issue is that of setting prior credences in the absence of any information about the way the world is. For this purposes it is helpful to imagine a Superbaby[1] who has sufficient language skills to consider the propositions of interest, is capable of deductive reasoning, is versed in the probability calculus, but has as yet no empirical information.

In the framework employed by Pettigrew, the epistemic worth of a credal state is to be judged by some measure of its closeness to the truth, or *accuracy*. Accuracy may be judged in a variety of ways, but it is essential that the measure of accuracy be a *proper scoring rule*. This means that, if an agent evaluates the *expected accuracy* of a credence function—that is, a weighted average of the accuracies that would obtain in each of an exhaustive set of mutually exclusive alternatives, where the weighting factors are the agent's own credences in the alternatives—then the agent will judge no credence function to have higher expected accuracy than the agent's own. Pettigrew favors the Brier rule, on which departure from the truth is measured by the square of a Euclidean distance on the space of all probability assignments on the set of propositions generated by the finite partition under consideration. The argument generalizes to

[1] The concept, though not the term, is due to I. J. Good (1968). Alan Hájek (ms.) reports that David Lewis coined the term. I do not know of any occasion on which Lewis used the term in print.

a broader class of accuracy functions (see Pettigrew 2016*a*, §5; 2016*b*, §12.4), the details of which are inessential for present purposes. The key features needed are the following. Take a partition $\{P_1, P_2,\ldots, P_n\}$. Consider the credence function that assigns equal credence to the members of the partition. We assume: (i) This credence function has the same accuracy score, no matter which element of the partition is true. (ii) We can improve accuracy in one eventuality only at the expense of accuracy in another eventuality.

The Principle that, according to Pettigrew, should guide the Superbaby's credence is one of conservatism. The Superbaby should minimize her risk, and adopt a credence function that maximizes accuracy in the worst case (the *Maximin* rule). It is easy to see that, given what has been said about the accuracy function, the one that maximizes worst-case accuracy is the credence function that assigns equal credence to all elements of the partition.

3. Indifference forecloses learning. Pettigrew intends the Maximin rule to be applied only at the beginning of an agent's credal life, to set the agent's prior credences. The agent then goes out into the world and incorporates empirical evidence into her credences by conditionalizing on the evidence acquired. Though her credences will shift as a result, the new, conditional credences will retain the imprint of the epistemic conservatism that begat the prior credences. This is problematic, as it conflicts with the desideratum that evidence about the past be informative about the future, and hence conditionalizing on evidence up to a certain time should shift credences about future events.

To see this, let the Baby consider propositions about the state of the world at various times. She might, for example, consider, for each time t_i in some set of times $\{t_1,\ldots,t_k\}$, a partition

$\{P_1(t_i), \ldots, P_n(t_i)\}$.[2] A maximal proposition considered by the Baby selects, for each time t_i, one P_j out of the considered partition of states of the world at t_i. The set of all such maximal propositions gives us a partition consisting of n^k possible histories of the world.

The Principle of Indifference favored by Pettigrew constrains the Baby's priors to assign equal credence to each of these n^k histories. This has the unfortunate consequence that propositions concerning the state of the world at distinct times are probabilistically independent of each other. That is, if Cr_0 is the initial credence function that assigns equal credence to each history, then, for any distinct times t, t', we have, for all i, j,

$$Cr_0(P_i(t) \mid P_j(t')) = Cr_0(P_i(t)) \tag{1}$$

In general, if H is any history specifying the state of the world at some set of times not including t, we have,

$$Cr_0(P_i(t) \mid H) = Cr_0(P_\mathrm{i}(t)). \tag{2}$$

This means that, if the Baby sets out in life with the initial credence function Cr_0, and updates her credences by conditionalizing on information received, information about the past is irrelevant to propositions about the future, and, no matter how much experience she

[2] For simplicity, we're considering a partition of the same cardinality n for each time considered, but nothing in the argument depends on this. However, for Pettigrew's argument to apply, the partition considered by the Superbaby must be finite, and so we can consider only a finite set of times and, for each time, a finite partition of ways the world could be at that time.

incorporates into her stock of knowledge, her credences about future events are affected, and remain the same as her prior credences.

For an example, consider a situation discussed by Pettigrew. Suppose that the Baby considers a partition concerning the colour of the handkerchief in my pocket: $\{Blue, Not\text{-}Blue\}$. Suppose, now, that the Baby considers, not merely the colour of the handkerchief in my pocket at a single time, but the colour on each of some set of days, say, every day for one year. The partition considered by the Baby consists of 2^{365} propositions, each of which specifies, for each day, either *Blue* or *Not-Blue* on that day. The Principle of Indifference demands that the Baby assign the same prior credence, $1/2^{365}$, to each element of the partition. Call this credence function Cr_0.

Now consider one particular day, say, December 31. Cr_0 assigns probability one-half to my handkerchief being blue on that day. Now consider the credence function that results from Cr_0 by conditionalization on the proposition that my handkerchief is blue on every day from January 1 through December 30. This conditional probability still yields credence one-half for blue on December 31. Therefore, if the Baby were to set her initial credences to Cr_0, and if she were to update her credences by conditionalizing on observations of handkerchiefs in my pocket on every day from January 1 through December 30, and if she saw that my handkerchief was blue on each of those days, she would still assign credence one-half to my handkerchief being blue on the following day, the same credence she had prior to any experience at all. That is, her credences about the future are as conservative as her priors; she is still maximizing worst-case accuracy, applied to propositions about the future.

That the prior credence function recommended by the Principle of Indifferences forecloses learning from experience is not a new point. This is an issue that Carnap (1945, 1950, 1962) faced in his attempts to set up an inductive logic, and the issue that led him to reject the prior probability function he called $\mathfrak{m}\dagger$, which assigns equal probability to every state description (see Carnap 1945, pp. 80–1; 1950, 1962, pp. 564–565).

4. Bad news for the inductivist? Though the natural reaction is to take this as bad news for the Principle of Indifference, another reaction might be to regard it as bad news for any credence function on which induction is possible, as any such function violates the Maximin condition, which enjoys one to maximize worst-case accuracy. Is the inductivist perhaps being unduly rash? If so, could a cognitive conservative convince the inductivist that she is being unduly rash?

Consider two Superbabies, Arthur and Trillian. Both consider the year-of-handkerchiefs partition from the previous section. Arthur is epistemically conservative, and adopts a credence function that maximizes the worst-case accuracy. This means assigning an equal credence to each member of the partition, a credence function that makes observation of handkerchief-colour on any day irrelevant to handkerchief-colour on any other day. Trillian, on the other hand, adopts some credence function on which conditionalizing on the proposition that the handkerchief is blue every day through December 30 raises credence that it will be blue the next day.

What might Arthur say to Trillian, in an attempt persuade her that she is being excessively rash? He may point out to her that, in the event that my handkerchief is blue every day through December 30 and red the next, the accuracy of her prior is lower than the accuracy of his. Trillian can respond that, yes, this is so, but, in the event that my handkerchief is blue every day through December 30 and blue the next, the accuracy of her prior is higher. Arthur's prior has the best worst-case accuracy, but Trillian's has better best-case accuracy, and thus we have a trade-off.

If my handkerchief is blue every day through December 30 and not blue the next, Arthur wins the accuracy competition. If, however, my handkerchief is blue every day through December 30 and blue the next, Trillian wins. Moreover, Trillian regards the latter eventuality as more probable, and so she regards the risk she is taking, of possibly losing the accuracy competition to Arthur, as worth it. As the accuracy function employed is a proper scoring rule, by her own lights—that is, judged by her own credences—a shift to Arthur's priors would involve a loss of *expected* accuracy. Arthur's conservatism-based pleas for her to change her mind will only sound convincing if she is already convinced that the scenarios in which her accuracy is worse than Arthur's are to be regarded as at least as credible as the scenarios in which her accuracy is better than Arthur's. But that is just to say: Arthur's pleas will be convincing only if she has already adopted his credences.

This points to a limitation of the use of accuracy considerations in setting prior credences. Pettigrew invokes them to set initial credences for a Superbaby who as yet has none. They cannot, however, be used as a corrective. Because the accuracy function is a proper scoring rule, an agent who has some credences that don't satisfy Pettigrew's Maximin condition will not regard a shift to credences that do satisfy the condition as an improvement in expected accuracy.

5. Balancing risk and caution. An agent who adopts an attitude of extreme epistemic conservatism maximizes worst-case accuracy, at the price of a lower best-case accuracy than is enjoyed by credence functions other than the uniform credence function. An agent might reasonably be concerned with increasing best-case accuracy. Unfortunately, as Pettigrew argues, a prescription to maximize best-case accuracy leads to dogmatism, as this prescription recommends adopting one of the extremal credence functions that assigns credence 1 to one element of the partition and 0 to all the others. An agent who adopted such a function would be taking an extreme risk, putting all of her eggs in one basket and hoping that the gamble pays off.

One can also consider a mean between these two extremes, maximizing a weighted average of worst-case and best-case accuracy. This yields a set of criteria, one for each possible value λ of the weight accorded to best-case accuracy, called the *Hurwicz$_\lambda$ criteria*, discussed in chapter 13 of Pettigrew 2016b. For an *n*-element partition, provided that the weight accorded to best-case accuracy is high enough,[3] any credence function permitted by this criterion is a weighted average of the uniform function and some extremal credence function (that is, a credence function that accords credence 1 to one element of the partition and 0 to all others). Obviously, this does not single out a unique permissible credence function; there will be as many of these as there are elements of the partition.

Adopting a Hurwicz$_\lambda$ criterion opens up the possibility of learning, of a sort.[4] Let us return to the year-of-handkerchiefs example. Suppose that a Superbaby who adopts a Hurwicz$_\lambda$ criterion suspects that I might favor blue enough to wear it every day, and adopts a credence function that, for some λ greater than $1/2^{365}$, assigns credence λ to the proposition that my handkerchief is blue every day, and distributes credence $1–\lambda$ equally among the alternative distributions of blue and non-blue handkerchiefs throughout the year. With this credence function, if the Baby sees a steady stream of blue handkerchiefs, her credence in the proposition that my handkerchief will always be blue increases, and, with it, her credence that my handkerchief will be blue tomorrow. She is capable of some learning.

[3] What counts as “high enough” is that the weight, λ, accorded to best-case accuracy be greater than or equal to $1/n$, where n is the number of elements of the partition considered.
[4] This was pointed out to the author by Pettigrew in a personal communication.

Some learning, but her ability to learn is limited. If, on any day, my handkerchief is not blue, then from that day forward her credence that, the next time she sees my handkerchief, it will be blue, is one-half. A long string of blue handkerchiefs with even one exception does not raise her credence that the next one she sees will be blue.

This generalizes. A credence permitted by a Hurwicz$_\lambda$ criterion is a weighted average of some extremal credence function and the uniform credence function. Evidence, if it has any effect on credence, adjusts the relative weighting of these two functions. If evidence is obtained that is incompatible with the element of the partition on which the extremal credence function is concentrated, its weight goes to zero, and from then on the credence function is the same as the one obtained from the Principle of Indifference.

There is an extensively studied class of prior credences over partitions like the one considered, that are such that, upon observing a long stream, of length *n*, of handkerchiefs, of which a fraction *f* are blue, the agent's credence that the next one will be blue approximates *f*, whatever that frequency may happen to be, and approximates *f* all the more closely as *n* increases.[5] Both the Minimax criterion and the Hurwicz$_\lambda$ criterion declare all such credences irrational. So, if one holds that a credence function with these properties might be permissible under certain circumstances, it is necessary to reject both types of criteria.

The question arises whether some other accuracy-based considerations could be harnessed to yield inductive credences of this sort. One avenue that could be explored is coarse-graining of the partition. If our Baby is going to receive, throughout the year, reports of the colour of the handkerchief in my pocket each day, then she must, at a minimum, consider a partition fine enough to encompass the full range of evidence-reports envisaged. Assigning equal prior

[5] See chapters 1, 2, 4, and 11 of Zabell (2005) for discussion, and some of the history.

credence to each possible evidence report forecloses learning. If, however, she assigns equal credence, not to each element of the partition, but to each possible number of blue handkerchiefs, from 0 to 365, and apportions her credence uniformly within each subset of the partition consisting of propositions that agree on the number of blue handkerchiefs, then her credences yield Laplace's rule of succession—upon observing a string of length *n* of handkerchiefs, of which *m* are blue, her credence that the next will be blue is $(m + 1)/(n + 2)$. If, somehow, it could be argued that a Superbaby ought to apply accuracy considerations, not to the full partition under consideration, but to a coarse-graining of it that lumps together scenarios in which the number of blue handkerchiefs are equal, then perhaps accuracy-based considerations could recover Laplace's rule.

6. Conclusion. Pettigrew acknowledges that some readers may not take the same attitude as he does towards extreme epistemic conservatism.

> At this point, it seems to me, we have reached normative bedrock: one cannot argue for cognitive conservatism from more basic principles. Thus, to those who reject cognitive conservatism… I can only recommend the argument of this paper as an argument for the following (subjunctive) conditional: Cognitive Conservatism $\Rightarrow$ Principle of Indifference (2016*a*¸46; see also 2016*b*, 166–67).

It seems to me that we have not reached bedrock, as there is another principle that we should hold to, namely, that an epistemic agent should be able to learn from experience. But this, as we have seen, involves *not* adopting the Principle of Indifference that Pettigrew recommends. We have argued for the conditional,

$$\text{Possibility of Learning from Experience} \Rightarrow \neg(\text{Principle of Indifference}).$$

Conjoined with the conditional that Pettigrew establishes, this gives us,

$$\text{Possibility of Learning from Experience} \Rightarrow \neg(\text{Cognitive Conservatism}).$$

If (as I think you should), you agree that one should not foreclose the possibility of learning from experience, you should be willing to take an epistemic risk, knowing full well that, in the worst case scenario, the accuracy of your priors will be lower than those of the extreme epistemic conservative.

References

Carnap, Rudolf (1945). "On Inductive Logic." *Philosophy of Science* **12**, 72–97.

——— (1950). *The Logical Foundations of Probability*. 1st ed. Chicago: The University of Chicago Press.

——— (1962). *The Logical Foundations of Probability*, 2nd ed. Chicago: The University of Chicago Press.

Good, I. J. (1968). "The White Shoe *Qua* Herring is Pink." *The British Journal for the Philosophy of Science* **19**, 156–157.

Hájek, Alan (ms.). "Staying Regular?" http://hplms.berkeley.edu/HajekStayingRegular.pdf

Pettigrew, Richard (2016*a*). "Accuracy, Risk, and the Principle of Indifference." *Philosophy and Phenomenological Research* **92**, 35–59.

——— (2016*b*). *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

Zabell, S. L (2005). *Symmetry and is Discontents.* Cambridge: Cambridge University Press.