Episteme, 18, 1 (2021) 31–45
doi:10.1017/epi.2018.54

THE NO FREE LUNCH THEOREM: BAD NEWS FOR (WHITE'S ACCOUNT OF) THE PROBLEM OF INDUCTION

GERHARD SCHURZ
schurz@phil.hhu.de

ABSTRACT

White (2015) proposes an a priori justification of the reliability of inductive prediction methods based on his thesis of induction-friendliness. It asserts that there are by far more induction-friendly event sequences than induction-unfriendly event sequences. In this paper I contrast White's thesis with the famous no free lunch (NFL) theorem. I explain two versions of this theorem, the strong NFL theorem applying to binary and the weak NFL theorem applying to real-valued predictions. I show that both versions refute the thesis of induction-friendliness. In the conclusion I argue that an a priori justification of the reliability of induction based on a uniform probability distribution over possible event sequences is impossible. In the outlook I consider two alternative approaches: (i) justification externalism and (ii) optimality justifications.

1. INTRODUCTION: WHITE ON THE PROBLEM OF INDUCTION

The thesis of induction-friendliness asserts that there are by far more induction-friendly than induction-unfriendly event sequences. White (2015) proposed an a priori justification of the reliability of inductive prediction methods based on this thesis. In this paper I contrast White's account with the famous no free lunch (NFL) theorem. I will explain two versions of this theorem, the strong NFL theorem applying to binary and the weak NFL theorem applying to real-valued predictions. I will show that both versions refute the thesis of induction-friendliness and, thus, White's account of the problem of induction. In conclusion an a priori justification of the reliability of induction based on a uniform probability distribution over possible event sequences turns out to be impossible. In the outlook I will consider two alternative approaches: (i) justification externalism and (ii) optimality justifications.

White's (2015) paper consists of two parts. In the first part White clarifies what the problem of induction consists in – this part is devoted to the "problem of the problem of induction", the title of the paper. White argues that the problem of induction – the question of how we can justify induction – becomes epistemologically significant if one assumes the following

Reliability principle:
If a person S considers whether the methods or rules she followed in concluding P are reliable and she is not justified in believing that they are reliable, then she is not justified in believing P.

White gives convincing arguments in favor of the reliability principle, to which I have nothing substantial to add. In what follows I will simply assume this principle is correct. The second part of White's paper is devoted to the "problem of induction" which, according to the reliability principle, consists in justifying the reliability of inductive inferences. My critical discussion focuses on this second part of White's paper. White proposes that a justification of the reliability of induction can be given based on the following claim, which I call the

Thesis of induction-friendliness (White's thesis): There are by far more induction-friendly event sequences (or 'states of the world') than induction-unfriendly event sequences.

In the next section I will explicate White's thesis in the required detail and introduce some necessary terminology. In section 3, I contrast the thesis with a famous result from machine learning – the no free lunch (NFL) theorem – which asserts the exact opposite. In section 4, I draw the conclusion that an a priori justification of the reliability of induction based on a uniform probability distribution over all possible event sequences is impossible. In the outlook I point towards two alternative approaches: (i) justification externalism and (ii) optimality justifications.

2. WHITE'S THESIS EXPLICATED: PREDICTION GAMES AND SUCCESS MEASURES

Like White I focus here on *predictive* (inductive) inferences of *binary* events, coded by 0 and 1. Given are possibly infinite event sequences (e) $=_{def}$ $(e_1,e_2,\ldots,e_n,\ldots)$ with each e_n being either 0 or 1 ($e_n \in \{0,1\}$) and at each point in time or 'round' n ($n \geq 0$), the task is to deliver a prediction of the next event, abbreviated as $pred_{n+1}$. I assume that the compared prediction methods are *non-clairvoyant*, i.e., the prediction $pred_{n+1}$ is only a function of the (observed) past events $(e_1,\ldots,e_n)$, but not a function of future events. Let (e_{1-n}) abbreviate $(e_1,\ldots,e_n)$, $pred_{n+1}(M)$ denote the prediction of a given method M at time n, and ω be the set of natural numbers. Then each prediction method M is defined by a function f_M mapping (e_{1-n}) into a prediction $pred_{n+1}$ for each $n \in \omega$, that is, $pred_{n+1}(M) = f_M((e_{1-n}))$. A pair $((e),\{M_1,\ldots,M_q\})$ consisting of an event sequence and a finite set of prediction methods is also called a *prediction game* (Schurz 2008).

We have to distinguish two sorts of prediction games with corresponding measures of predictive success:

- Games with *binary predictions*: Here at each time a definite yes-no prediction of the next event must be delivered, i.e., $pred_{n+1} \in \{0,1\}$ for all $n \in w$. For binary predictions the straightforward measure of the predictive success of method M at a given time n is the truth-frequency of correct prediction until time n: $suc_n(M)$ = (number of correct predictions)/n.
- Games with *real-valued predictions* (including binary predictions as a special case): Here the prediction may be any real-valued number between zero and one (rounded up to a finite accuracy), i.e., $pred_{n+1} \in [0,1]$ (the closed interval of reals between 0

and 1). The natural scoring function for real-valued predictions is the absolute distance between prediction and event: $score(pred_i,e_i) = 1-|pred_i-e_i|$ ("$|pred_i-e_i|$" is called the natural loss function). The predictive success rate of method M at time n, abbreviated as $suc_n(M)$, equals the sum of scores of method M until time n, divided by n, $suc_n(M) = \Sigma_{1\leq i\leq n} score_i(M)/n$. The above truth-frequency measure is a special case of this success measure for binary predictions because the score of a binary prediction $pred_i$ is 1 if $pred_i = e_i$ and 0 otherwise.

Proper real-valued predictions of binary events are needed for several purposes, e.g., for *probabilistic* forecasters computing an estimated subjective probability (P) of the next event and predicting this probability, i.e., $pred_{n+1} = P(e_{n+1} \mid (e_{1-n}))$. In this paper, real-valued predictions are introduced to account for White's assumption that a prediction method may *refrain* from making a prediction, for example if the so-far observed event sequence is so 'unorderly' that no inductively projectable pattern can be discerned (White 2015: 285). Thus, in White's setting the *value* space of possible prediction, Val_{pred}, includes the three elements 1, 0 and n (no prediction delivered), $Val_{pred} = \{1,0,n\}$. The question is how to measure the predictive success of a non-prediction. Surely it has to be measured somehow, since otherwise a method could 'boost' its success simply by refraining from delivering a prediction whenever the predictive task is not extremely easy. Curiously, in his reply to White, Cariani (2015: 295) designs a method that tries to prevent its own failure by refraining from making predictions whenever its success rate drops below a critical level. Cariani adds that "of course [he is] not suggesting that such a method defeats scepticism about induction". In other words, a successful prediction method must not only have a high success rate when making a prediction, but must also have a high rate of *applicability*. The most straightforward way to account for this is to assign to non-predictions the same score as that of a *random guess*, which is 0.5 in binary games (this method is also applied, for example, by Martignon and Hoffrage 1999). This means that a non-prediction can equally be identified with a real-valued prediction of 0.5, because a prediction of 0.5 is also guaranteed to have a score of 0.5.[1]

Note that with real-valued predictions we can also model more fine-grained effects of non-predictions. For example, an incorrect non-prediction may lead to a higher loss when the event is 1 than when it is 0; we can model this case by identifying a non-prediction with the prediction of a value $r < 0.5$.

The difference between binary and real-valued predictions is important for the problem of induction. In the case of binary predictions one can construct, for every method M, a 'demonic' event-sequence (e) that 'conspires' against M and produces, for every time n, exactly the opposite event of what M predicts (this observation goes back to Putnam 1965 and is mentioned by Cariani 2015: 29). In contrast, if real-valued predictions are allowed, a success rate of 0.5 can be guaranteed by always predicting 0.5. In what follows, I call this method "averaging", abbreviated as Av (i.e., for all $n \in w$, $pred_{n+1}(Av) = 0.5$). We shall see in the next section that it is for this reason that a *strong* NFL theorem can be proved for games with binary predictions, while for games with real-valued predictions

1 Long after I submitted my paper I discovered a related critique of White's account by Barnett and Li (2018). The overlap of their paper (which refers to Schurz 2017) with this paper is small. They prove that the prediction rule "fool me once" – a version of Cariani's prediction method – has the same expected score as random guessing (however, they use a different scoring rule).

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press

only a *weak* NFL theorem holds. Unfortunately this does not bring much hope for White's account because, as we will see, the weak NFL is strong enough to refute the thesis of induction-friendliness.

With that said, White's thesis can be identified with the following pair of assertions:

Thesis of induction-friendliness, explication 1: There are significantly more [less] binary sequences (of a given length) for which inductive prediction methods have a high [low] success rate than there are sequences for which counter-inductive or non-inductive prediction methods have a high [low] success rate.

I understand the notion of high [low] success relative to a given success-threshold $t > 0.5$. Thus, $suc_n(M)$ is regarded as 'high' if it is at least as great as t and is regarded as 'low' if it is smaller than $1 - t$. Note that the *unbracketed* and the *bracketed* version of the thesis of induction-friendliness are *two distinct* assertions and that the thesis is obviously meant to entail both.

Explication 1 of White's thesis hides a subtle difficulty, namely what an *inductive* prediction method is. The standard understanding is a prediction method in the explained sense of a 'method' (mapping each observed event history into a prediction), based on the inductive projection of some *specific* pattern (this is also Cariani's understanding in 2015: 204). In contrast, several passages of White's support the conclusion that by "inductive method", he means the entire *family* of inductive methods, based on the inductive projection of *some* pattern. For example, White writes that the inductive method cannot only be applied to sequences 1111..., which by projecting the pattern '1-iterated' to the future leads to predictions of 1s, but also to more complicated sequences such as 11001100..., which by projecting the more complex pattern '1100-iterated' leads to iterated predictions of the form 1100.

Prima facie, the family of all induction methods is not even well-defined, because the class of 'all possible' patterns is not recursively enumerable. Even if this problem is solved (by a suitable restriction of the notion of 'possible patterns'), the problem that the predictions of this family of methods are not well-defined remains, for two reasons: there exist event sequences (i) for which two different inductive methods lead to opposite predictions, and (ii) for which a counter-inductive method leads to precisely the same prediction as an inductive method. In order to illustrate these facts, I introduce a couple of simple prediction methods.

- Majority-induction "M-I" predicts the event that so far has been in the majority and predicts 1 initially and in the case of ties. Formally, $pred_{n+1}(M\text{-}I) = 1$ if $n = 0$ or $freq_n(1) \geq 0.5$, else $pred_{n+1}(M\text{-}I) = 0$ (where "$freq_n(e)$" denotes the relative frequency of event e until time n).
- Majority counter-induction "M-CI" predicts the opposite of M-I, i.e., $pred_{n+1}(M\text{-}CI) = 0$ if $n = 0$ or $freq_n(1) \geq 0.5$, else $pred_{n+1}(M\text{-}CI) = 1$.

More generally, each inductive method has a counter-inductive dual which predicts the opposite of the method. In what follows e* denotes the opposite of a binary event e, thus $1^* = 0$ and $0^* = 1$. Note that M-I and M-CI are binary prediction methods (i.e., they always predict either 1 or 0).

First, consider the event sequence

$$(e)_1 = 1111111\ldots = 1 - \text{iterated}.$$

M-I predicts 1111... and has perfect success in application to $(e)_1$, while M-CI predicts 0000... and has perfect failure: $suc_n(\text{M-I}) = 1$ and $suc_n(\text{M-CI}) = 0$ for every $n \in w$. Not only White, but also many other philosophers have doubted that a counter-inductive method can be perfectly successful,[2] but this impression is illusionary: all *oscillatory* sequences are friendly to counter-induction. Take, for example, the alternating sequence

$$(e)_2 = 01010101 = 01 - \text{iterated}.$$

It is easy to see that M-CI predicts 0101... and is perfectly successful for this sequence, while M-I predicts 1010... and fails completely: $suc_n(\text{M-CI}) = 1$ and $suc_n(\text{M-I}) = 0$ for every $n \in w$.

White would point out that there is a more refined method of induction which achieves perfect success if applied to the sequence $(e)_2$, by recognizing the inductive regularity "01-forever" and predicting this sequence with equally perfect success as M-CI. Let us call this method "2-block-majority-induction", abbreviated as "2b-M-I". With these simple observations we have proved the above claim: (i) we have two inductive methods, M-I and 2b-M-I, which in application to the sequence $(e)_2$ lead to opposite predictions, and (ii) we have the inductive method 2b-M-I and the counter-inductive method M-CI, which in application to $(e)_2$ yield the same perfect success.

A precise definition of the 2-block majority method (with an initial guess that is true for the sequence $(e)_2$) is the following: The method 2b-M-I divides the sequence of events in consecutive blocks of two and tracks the frequencies of all four possible 2-block pattern ordered as follows: 01, 10, 00,11. The method delivers a 2-block prediction ($pred_{n+1}$, $pred_{n+2}$) at all even times n as follows: If at least one 2-block pattern has a so-far frequency of at least 0.5, the method 2b-M-I predicts the 2-block pattern that has maximal frequency (its initial 2-block prediction is "01" and in the case of ties it choses the first best pattern in the given ordering). Otherwise, it refrains from making a prediction for the two next times, which means that it delivers the 2-block prediction (0.5, 0.5). Based on its 2-block prediction ($pred_{n+1}$, $pred_{n+2}$) the method predicts $pred_{n+1}$ at time n and $pred_{n+2}$ at time n + 1 (independently from the event e_{n+1}).

Like M-I, 2b-M-I applies not only to strict but also to weak inductions. For example, the 12-membered sequence 010111010001 contains the two underlined 2-blocks that do not follow the regular pattern "01-iterated"; so $f_{12}(01) = 2/3$ holds for this sequence, whence 2b-M-I predicts $pred_{13} = 0$ and $pred_{14} = 1$. Note that, unlike M-I, 2b-M-I is no longer a binary prediction method; if all four frequencies are smaller than 0.5, it predicts (by convention) 0.5. However, we can also define a binary variant of 2b-MI, call it 2b-M-I_{bin}, which uses M-I as its fall-back method whenever 2b-M-I predicts 0.5.

The dialectics between induction and counter-induction can be replicated at the level of 2-block methods. Define 2-block majority-counter-induction, 2b-M-CI, in the explained way, predicting the exact opposite of 2b-M-I, where the opposite of a sequence of binary

2 Cf., e.g., van Cleve (1984: 561), or the entry "counterinduction" in *The Oxford Dictionary of Philosophy* (Blackburn 2016).

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press

events is defined as the sequence of opposites, e.g., (101)* = (010) (etc.). Consider the 'doubly oscillating' sequence

$$(e)_3 = 10011001 = 1001 - \text{forever}.$$

2b-M-I applied to $(e)_3$ predicts 01100110... and fails completely, while 2b-M-CI applied to $(e)_3$ predicts 10011001... and succeeds perfectly.

Again it can be pointed out that the 4-block variant of majority-induction would detect the 4-block regularity and achieve the same success as the 2-block majority counter-induction. But obviously, the same dialectics may be repeated based on the 8-block oscillating sequence 10010110-forever (etc.).

Be aware that m-block (counter-)inductive majority methods do *by no means* exhaust the space of all inductive and corresponding counter-inductive prediction methods of arbitrary complexity; this space is neither computationally tractable nor even well-defined. At least, the class of m-block (counter-)inductive methods can be reasonably defined, in precise analogy to 2-block methods (note that M-I is a 1-block prediction method).[3] These methods are sufficient for explaining the philosophical problems connected with White's thesis.

There are many more kinds of prediction methods. For example, there are the past-independent methods, whose predictions are independent from the event patterns observed in the past, but possibly dependent on the given point in time. A subcase of these are the *constant* methods that always predict the same constant (binary or real-valued) number. More importantly, there are (as explained) binary versus real-valued methods. Recall that all m-block methods for $m \geq 2$ are real-valued because they predict 0.5 (i.e., refuse to predict) in certain cases. More interesting cases of real-valued predictions are, for example,

- Cautious majority induction methods: they predict the majority-event only if the majority is sufficiently high, or formally: $pred_{n+1} = e$ if $freq_n(e) \geq t > 0.5$ where t is a suitably chosen threshold and e is 0 or 1. The same modification is possible for all m-block majority methods.
- Average induction, which predicts the so-far observed frequency of 1s, $pred_{n+1}(1) = freq_n(1)$.
- Logico-probabilistic induction methods, for example, Carnap's (1950) c* system which predicts $pred_{n+1}(c^*) = (freq_n(1) + 1)/(n + 2)$ in the binary case (and many more).

What all this shows is that in order to make sense, White's thesis has to be applied to particular inductive prediction methods and not to 'induction in general', because different induction methods may deliver opposite predictions (the same goes for counter-inductive

3 Here is the full definition: An m-block majority-inductive method, mb-M-I, does the following: (i) It assumes a given ordering among all possible m-block patterns $P_1,\ldots,P_q$, with $q = 2^m$, and initially predicts the first pattern in this ordering, (ii) at any time $n = k \cdot m$ (for integer-valued $k \geq 1$), it predicts as its m-block prediction ($pred_{n+1},\ldots,pred_{n+m}$) the first pattern in this ordering whose so-far observed frequency is ≥ 0.5 and is not smaller than the so-far observed frequency of any other m-block pattern; if there is no such pattern, its m-block prediction is "0.5 m-times", and (iii) at any given time n, it predicts that event whose number in the predicted m-block equals "n+1 moulo m" (this is the remainder after dividing n+1 through m).

and past-independent prediction methods). Thus the straightforward explication of White's thesis is the following:

Thesis of induction-friendliness, explication 2: For every inductive prediction method (of a given family of methods) there are significantly more [less] binary sequences (of a given length) for which it has a high [low] success rate than there are sequences for which the corresponding counter-inductive method or some past-independent method has a high [low] success rate.

As explained, the underlying "family" must be clearly characterized; every one of the families described above, or any union of them, would be suitable.

Based on the previous discussion, an alternative explication of White's thesis suggests itself which refers to an entire family of inductive methods and asserts induction-friendliness to the set of event sequences for which *at least one* of the methods of this family is successful. But since the methods of the family produce opposite predictions for various sequences and points in time, it is not possible to define a prediction method that is successful whenever *some* method of this family is successful. I will return to this question at the end of the next section. Now, I turn to explication 2 of White's thesis and confront it with the NFL theorem(s).

3. THE STRONG AND WEAK NO FREE LUNCH (NFL) THEOREM

The NFL theorem is a deepening of Hume's inductive skepticism developed in machine learning, a branch of computer science. NFL theorems have been formulated in different versions,[4] a most general formulation is found in Wolpert (1996). Wolpert's NFL theorem comes in a strong and a weak version; it is formulated in a highly abstract mathematical language, hardly understandable for non-mathematicians. In this section I introduce a version of the NFL theorems for prediction games and try to explain their proofs in a simple way.

The strong NFL theorem applies to binary and the weak NFL theorem to real-valued prediction methods. Both NFL theorems assume binary events (for graded events, restricted NFL theorems are possible). Consider the space of all possible binary event sequences of a given fixed length N, abbreviated as SEQ_N; the case of countably infinite N (N=ω) is admitted. (Formally $SEQ_N = \{0,1\}^N$ and SEQ_N contains 2^N sequences.) Both NFL theorems presuppose a prior probability distribution P over SEQ_N that assigns the same probability to every sequence; in what follows I call such a distribution a *state-uniform* probability distribution.

I first explain the strong NFL theorem:

Theorem 1. ***Strong NFL theorem*:** Given a state-uniform P-distribution over SEQ_N, the probability that a prediction method M has a certain success rate $\frac{k}{n}$ after n rounds ($1 \leq n \leq N$) is *the same* for all *binary* prediction methods over binary events (whether 'inductive', 'counter-inductive', 'past-independent' or whatever) and is given by the binomial formula: $P(suc_n(M) = \frac{k}{n}) = \binom{n}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{n-k} = \binom{n}{k}/2^n$.

4 Cf. Wolpert (1992, 1996), Schaffer (1994), Giraud-Carrier and Provost (2005).

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press

Note that $\binom{n}{k}$, "n-over-k", is the number of possibilities to select k distinct elements out of n elements and is given as $\frac{n!}{k!\cdot(n-k)!}$, with $n! =_{def} 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$. Theorem 1 asserts, in other words, that the state-uniform probability of method M predicting k out of n times correctly equals the probability of throwing k out of n heads with a regular coin. Here is a simple proof:

Proof of theorem 1: Let M be any binary prediction method. We evaluate M's success rate for sequences of length n. There are $\binom{n}{k}$ possible sequences of M-scores – denoted as $(s) =_{def} (score_1(M), \ldots, score_n(M))$ – with k 1s in them. Since M's prediction function f_M is deterministic (i.e., f_M maps each possible past history (e_{1-i}) into a prediction 0 or 1) it follows that each score sequence (s) can be realized by *exactly one* event sequence (e), which is recursively constructed from (s) as follows: $e_{i+1} = f_M((e_{1-i}))$ if $score_{i+1}(M) = 1$, else $e_{i+1} = (f_M((e_{1-i})))^*$. Thus there are as many event-sequences that give M a success rate of $\frac{k}{n}$ after n rounds as there are score sequences, namely $\binom{n}{k}$. Since every event sequence has the same probability, namely $1/2^n$, the result follows.
Q.E.D.

Wolpert (1996: 1349) proves a strong NFL theorem that is more general in two respects: It applies (i) not only to deterministic but to probabilistic prediction methods, which predict 1s or 0s with certain probabilities, and (ii) not only to binary-valued but to discrete-valued prediction games whose loss function is 'homogeneous', which amounts to the requirement of a *zero-one* loss: loss(pred,e) = 1 if pred = e, else loss(pred,e) = 0. For philosophical purposes these generalizations are not needed and the above version of the strong NFL theorem is sufficient.

The strong NFL theorem implies, among other things, that for each pair of prediction methods, the number – or in the infinite case the probability – of event sequences in which the first method outperforms the second is precisely equal to the number (or probability) of event sequences in which the second method outperforms the first. More generally, the *success profile* – which consists of the number of sequences for which various success rates are achieved – is the same for all methods. It is obvious that this result destroys any hope for explication 2 of White's thesis.

So far, however, we restricted our attention to binary methods, while White's framework allows for methods that refuse to make predictions. As explained in section 2, this possibility is integrated in our account by allowing for real-valued predictions (which have several other advantages). It is easy to recognize that the strong NFL theorem does not hold for real-valued predictions. For example, for sequences of length 10 the success profile of every binary prediction method is computed by theorem 1 as follows (see Table 1):

Table 1 Number of event sequences of length 10 for which every binary prediction method achieves a certain success rate.

Success rate $\frac{k}{n}$:	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
No. of sequ. $\binom{n}{k}$:	1	10	45	120	210	252	210	120	45	10	1

In contrast, the constant method Av (which always predicts 0.5) achieves a success rate of 0.5 for all 2^{10}=1024 binary sequences; it never has a success higher or lower than average. Therefore its success profile for sequences of length 10 has the following numbers in the line "No. of sequ." of Table 1: 0,0,0,0,0,1024,0,0,0,0,0. This fact can be generalized: More *cautious* prediction methods that tend to predict values close to 0.5 will not have extreme success rates (either 'high' or 'low') as often as risky methods. However, it turns out that according to a state-uniform distribution these two effects precisely compensate each other, so that the state-uniform *expectation value* of a real-valued prediction method is always the same – and this is the content of the weak NFL theorem:[5]

Theorem 2. ***Weak NFL theorem*****:** Given a state-uniform P-distribution over SEQ_N, the expectation value of the success rate after n rounds ($1 \leq n \leq N$) is *the same* for all *real-valued* prediction methods over binary events (whether 'inductive', 'counter-inductive', 'past-independent' or whatever), namely one half: $Exp_P(suc_n(M)) = 0.5$.

The easiest way to understand the proof of theorem 2 is by means of the following 'infamous' but well-known theorem in probability theory:

Theorem 3. ***Induction-hostile state-uniformity*** (Carnap 1950: 564–6; Howson and Urbach 1996: 64–6):
Assume the probability (density) distribution P is uniform over the space SEQ_N. Then $P(e_{n+1}=1|(e_{1-n})) = 1/2$ for every possible next event e_{n+1} ($n+1 \leq N \leq \omega$) and sequence of past events (e_{1-n}). Thus, P satisfies the properties of a random distribution over {0,1}.

The proof of theorem 3 is incredibly simple: At any time n there are as many continuations of the past event sequence (e_{1-n}) that continue with 1 as continuations that continue with 0, and since every continuation has the same (state-uniform) probability, it follows that $P(1|(e_{1-n})) = P(0|(e_{1-n})) = 1/2$.

Theorem 3 provides us with an easy proof of the weak NFL theorem:

Proof of theorem 2: If a real-valued prediction method M predicts a value r in [0,1] for time $n+1$, then its score is r if the event e_{n+1} is 1 and $(1-r)$ if the event is 0. According to theorem 3, at any time n the state-uniform probability that $e_{n+1} = 1$ equals 1/2, independent of the past. Therefore whatever a real-valued method M predicts, the state-uniform expectation value of its score is 1/2, because for every real value r in [0,1], $0.5 \cdot r + 0.5 \cdot (1-r) = 0.5$. Since expectation values are additive, these expected scores add up; dividing their sum through the number of rounds gives 0.5 for the expectation value of $suc_n(M)$ for every $n \in w$.
Q.E.D.

The weak NFL theorem disproves explication 2 of White's thesis for real-valued predictions. To see this, consider first the extremely incautious (binary) method M-I in

5 Wolpert (1996) mentions the weak no free lunch theorem only in one paragraph on p. 1354; for our purpose this version of his theorem is even more important.

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press

comparison to the most cautious method Av. M-I has a high success rate (at least as high as some threshold $t > 0.5$) for more event sequences than Av, but theorem 2 assures us that in compensation (since the expected success is the same) M-I must also have a low success rate (at least as low as $1 - t$) for more sequences than Av.

The same goes for methods whose 'cautiousness' is intermediate between M-I and Av. As an example, consider the inductive weak majority method, abbreviated as WM-I, which is like M-I except that it predicts 0.5 initially and in the case when the so-far observed event-average is 0.5. Analogously, we define the weak counter-inductive majority method, WM-CI. The success profile of the two methods is remarkably different: WM-I can have very high success rates but cannot have very low sequences. Reversely, WM-CI can have very low success rates, but cannot have very high success rates. The reason for this initially surprising effect is that WM-I is very good in predicting sequences with extreme event frequencies and WM-CI in predicting highly oscillatory sequences, but the number of frequency ties (in which case the method predicts 0.5) is much higher in highly oscillatory sequences than in sequences with extreme frequencies. In compensation, the number of sequences in which WM-CI does just a little better than average is higher than the corresponding number of sequences for WM-I, and reversely for number of sequences in which WM-I's success is just a little below 0.5. WM-CI's success profile is the precise mirror image of that of WM-I, reflected at the middle point 0.5 (see Table 2).

More generally, whenever one real-valued prediction method M_1 has high success values more often than another method M_2, it must also have low success values more often, because the state-uniform expected success of both methods is the same. The consequence for explication 2 of the thesis of induction-friendliness is this: Whenever an inductive method M satisfies the first (unbracketed) part of the thesis, in comparison to some other non-inductive method M', then it violates the second (bracketed) part of the thesis.

Also, the weak NFL theorem tells us that according to a state-uniform P-distribution, one cannot say that more refined induction methods are 'better' than simple induction methods. For example, one would expect that the longer the patterns that a m-block method checks for repetitiveness in the past, the more 'successful' this method is, but this expectation is wrong because the expected success rates of refined induction methods are precisely the same as those of simple induction methods.

Let us finally consider the alternative explication of White's thesis mentioned at the end of section 2, which refers to an entire family of inductive methods, for example the family of m-block majority induction methods. As explained, since the methods of this family produce opposite predictions for various sequences and points in time, one cannot have a prediction method that is successful whenever *some* method of this family is successful. In order to turn an entire family of inductive methods into a 'most general' inductive method, one needs rules of *preference* between inductive methods in case they deliver contradicting predictions. This is of course possible, for example by the following definition: The generalized m-bounded majority induction method Gen_m-M-I (for a fixed m that is significantly smaller than N) does at each time $n \in w$ the following: (i) It scans through all $k \leq m$ and k-block patterns and searches for the smallest $k \leq m$ and k-block pattern of size k that has a maximal so-far frequency which is at least as great as 0.5; (ii) if such a pattern is found, it predicts according to the k-block majority induction method; (iii) otherwise it predicts 0.5.

One would expect the so defined generalized inductive method to be superior to all particular m-block induction methods, but that is an illusion: since it is a prediction method

Table 2 Number of event sequences of length 10 for which the two real-valued prediction methods WM-I and WM-CI achieve certain success rates ("SucRate"); average success "Av" on the left. (Computer simulation performed by Paul Thorn.)

SucRate:	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1.0	**Av:**
WM-I:	–	–	–	–	–	32	64	104	136	140	128	132	80	104	28	54	4	16	1	1	–	**0.5**
WM-CI:	–	1	1	16	4	54	28	104	80	132	128	140	136	104	64	32	–	–	–	–	–	**0.5**

(in the explained sense), the weak NFL theorem applies to it. Thus, Gen_m-M-I's state-uniform expectation value is equal to that of any other method, be it as 'dumb' as you wish. In conclusion, there is no hope to validate White's thesis by moving on to generalized inductive methods which select for each sequence and time a particular inductive method out of a well-defined family of inductive methods.

Even if there is no method which is successful over all sequences for which at least one inductive method is successful, one might nevertheless be interested in the truth value of White's thesis for the existential quantification over methods. So let us consider the following explication:

Thesis of induction-friendliness, explication 3: There are significantly more [less] binary sequences (of a given length) for which *some* m-block inductive prediction method (for $m \leq N$) has a high [low] success rate than there are sequences for which the corresponding counter-inductive m-block method or some past-independent m-block method has a high [low] success rate.

Conjecture: Explication 3 of the induction-friendliness thesis is false.

Unfortunately, I could not find a proof of this conjecture, not even for its restriction to binary prediction methods (recall that each m-block method has a binary completion mb-M-I_{bin}, which predicts according to M-I whenever mb-M-I predicts 0.5). The difficulty can be illustrated as follows. Consider the binary 1-block and 2-block inductive methods and let highseq(M) be the number of sequences of a fixed length N for which a given method M has a success rate of at least t (for a given threshold $t > 0.5$); moreover let highseq(M_1 or M_2) be the number of sequences for which at least one of the two methods M_1 or M_2 has a success rate of at least t. Then the strong NFL theorem entails that

$$\text{highseq(M-I}_{bin}) = \text{highseq(M-CI}_{bin})$$
$$\text{highseq(2b-M-I}_{bin}) = \text{highseq(2b-M-CI}_{bin}),$$

but one cannot conclude from this that highseq(M-I_{bin} or 2b-M-I_{bin}) = highseq(M-CI_{bin} or 2b-M-CI_{bin}-), because the sequences for which the two methods have high success rates may *intersect* and I did not find a way to prove that these intersections have equal cardinality.[6]

Thus I leave the question of White's thesis in its third explication as an open problem. Note, however, that even if it were true and our conjecture false, it would not be of much help to the problem of justifying the reliability of induction, because no inductive method (be it as 'general' as you wish) can be successful over all sequences for which 'some' inductive method is successful.

4. CONCLUSION AND OUTLOOK

The deeper reason for why a state-uniform P-distribution is so devastating for our hopes to justify induction is revealed by theorem 3: it is the most induction-hostile distribution

6 If highseq(M-1_{bin} and 2b-M-I_{bin}) < highseq(M-CI_{bin} and 2b-M-CI_{bin}) would hold, then highseq(M-I_{bin} or 2b-M-I_{bin}) > highseq(M-CI_{bin} or 2b-M-CI_{bin}) would follow.

one can imagine. A proponent of this distribution believes with probability 1 that the event sequence to be predicted is an identically and independently distributed (IID) random sequence. This implies for infinite sequences that the proponent is a priori certain that the event sequence has (a) a limiting frequency of 0.5, and (b) is non-computable. Condition (a) follows from theorem 3 and the (strong) law of large numbers and condition (b) from the fact that there are uncountably many sequences, but only countably many computable ones. However, the sequences for which a non-clairvoyant prediction method can be better than random guessing in the long run are precisely those that *do not* fall into the intersection of classes (a) and (b). Summarizing, an epistemic agent with a state-uniform prior distribution is a priori certain that the world is completely irregular so that 'intelligent' prediction methods cannot have a chance to be better than 'unintelligent' ones.

The law of large number reflects a deep property of the binomial distribution that is crucial to understanding the induction-hostile nature of a state-uniform P-distribution. The binomial distribution with probability $p = 0.5$, $\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{k}{n} \cdot 0.5^n$, has its maximum over the point $k = n/2$, where the frequency equals the probability (or comes as close as possible to it). More importantly, this maximum increases, in comparison to the neighboring frequencies, the larger n gets, and it becomes an infinitely sharp peak for $n \to \infty$. Since the irregularity (or entropy) of a non-computable sequence is the higher the closer its probability comes to 0.5, this means that among all sequences there are much more irregular than regular ones, and the share of the irregular sequences approaches 100% when their length n grows to infinity. Having this in mind, it is no longer surprising that based on a state-uniform P-distribution the expected chances of induction to be better than random guessing are zero.

A state-uniform distribution over infinite sequences assigns a probability of 0 to every frequency limit that is different from 0.5. Thus it is extremely *biased* in regard to the distribution over possible limiting frequencies. It is well-known that if one assumes a prior distribution that is not state-uniform but *frequency-uniform*, i.e., attaches the same probability to all possible frequency limits $p \in [0,1]$ of binary sequences, then one validates Laplace's rule of induction, $P(e_{n+1} = 1 \mid \text{freq}_n(1) = \frac{k}{n}) = \frac{k+1}{n+2}$. In computer science, the idea underlying frequency-uniform distributions has been generalized by Solomonoff (1964: §4.1), who proves that a distribution P is uniform over the possible frequencies iff the prior probability of sequences, P(s), decreases exponentially with their algorithmic or 'Kolmogorov' *complexity*, K(s): $P(s) \sim 2^{-K(s)}$. This means in more informal words that according to a frequency-uniform distribution, regular sequences are overwhelmingly more probable than irregular sequences.

These considerations tell us that the requirement of a 'uniform probability distribution' is *seriously ambiguous*. Its effects depend crucially on the given *partition* of the relevant possibility space: possible sequences versus possible frequencies. Which prior distributions are more 'natural', state-uniform ones or frequency-uniform ones? In my eyes, there is no objective answer to that question, because all prior distributions are subjective and biased in some respect. An objective justification of the reliability of induction – one that does not depend on whether one assumes a certain prior distribution and not another – is not in sight.

In conclusion it seems impossible to give a non-circular justification of the reliability of induction. So far, Hume's skeptical conclusion appears to be correct; but that is not the

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press

end of the story. As an outlook I mention two major escape routes from Humean skepticism that one still can take, given that (i) this conclusion is accepted and (ii) the notion of epistemic justification is 'goal-externalist' in the sense of being oriented towards the goal of truth-conduciveness:

- One possible escape route is *justification-externalism*. This position abandons White's reliability principle explained in section 1. According to this position it is sufficient for being justified in believing in the conclusion of an inductive inference that the inference is de facto reliable, without the believing person or anybody else having to be able to justify it (cf. van Cleve 1984, 2003). This shift in the meaning of "justification" relieves us of the burden of searching for a justification of induction. However, as White's excellent arguments in the first part of his paper make clear, the epistemic price of this move is very high, since not only induction but also counter-induction or God-guided clairvoyance could possess an externalist justification, at least in principle.
- The second possibility are *optimality justifications*, as introduced in Schurz (2008, 2009, 2019). This position maintains White's reliability principle and an (access-)internalistic understanding of the notion of justification. Optimality justifications do not attempt to 'prove' that induction is reliable, but rather, that it is *optimal*, i.e., that it is the best that we can do in order to achieve predictive success. Reichenbach (1949: sec. 91) was the first philosopher who suggested an optimality account. His account failed because results in formal learning theory show that no prediction method can be universally optimal at the level of *object-induction*, that is, of induction applied at the level of events (Kelly 1996: 263). In contrast, the account proposed in Schurz (2008, 2009, 2019) is based on *meta-induction*, i.e., induction applied on the meta-level of competing prediction methods. Based on results in machine learning it is shown that there is a real-valued strategy, called attractivity-weighted meta-induction (AMI), that is universally optimal among all prediction methods accessible to the epistemic agent. AMI predicts a weighted average of the predictions of these methods, using weights being a (delicately chosen) function of their so far achieved success rates. AMI is provably predictively optimal in the long run ($n \to \infty$) among all accessible methods in strictly all possible worlds, even in worlds whose event and success frequencies don't converge but are oscillating forever, and even if clairvoyant methods are admitted. According to the author, this result provides us with a weak a priori justification of meta-induction that does not contradict the weak NFL theorem.[7] Moreover, this a priori justification of meta-induction may provide us with an a posteriori justification of object-induction in our real world, insofar – and to the extent that – object-induction has turned out to be, so far, the most successful prediction strategy. This argument is no longer circular, because it is based on a non-circular justification of meta-induction.

7 A contradiction with the NFL theorem is avoided, because the state-uniform prior distribution that is assumed by the NFL theorem assigns a *probability of zero* to all infinite event sequences for which strategy AMI enjoys "free lunches" (cf. Schurz 2017).

REFERENCES

Barnett, Z. and Li, H. 2018. 'Fool Me Once: Can Indifference Vindicate Induction?' *Episteme*, 15(2): 202–8.

Blackburn, S. (ed.) 2016. *The Oxford Dictionary of Philosophy*, 3rd edition. Oxford: Oxford University Press.

Cariani, F. 2015. 'Some Questions about the Problem of the Problem of Induction.' *Episteme*, 12(2): 291–6.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Giraud-Carrier, C. and Provost, F. 2005. 'Toward a Justification of Meta-learning: Is the No Free Lunch Theorem a Show-stopper?' In *Workshop on Meta-learning* (*Proceedings of the ICML* 2005), pp. 12–19. Burlington, MA: Morgan Kaufmann.

Howson, C. and Urbach, P. 1996. *Scientific Reasoning: The Bayesian Approach*, 2nd edition. Chicago, IL: Open Court.

Kelly, K.T. 1996. *The Logic of Reliable Inquiry*. New York, NY: Oxford University Press.

Martignon, L. and Hoffrage, U. 1999. 'Why Does One-Reason Decision Making Work?' In G. Gigerenzer, P.M. Todd and the ABC Research Group (eds), *Simple Heuristics That Make Us Smart*, pp. 119–40. Oxford: Oxford University Press.

Putnam, H. 1965. 'Trial and Error Predicates and a Solution to a Problem of Mostowski.' *Journal of Symbolic Logic*, 30: 49–57.

Reichenbach, H. 1949. *The Theory of Probability*. Berkeley, CA: University of California Press.

Schaffer, C. 1994. 'A Conservation Law for Generalization Performance.' In W.W. Cohen and H. Hirsh (eds), *Machine Learning (Proceedings of ICML 1994)*, pp. 259–65. Burlington, MA: Morgan Kaufmann.

Schurz, G. 2008. 'The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem.' *Philosophy of Science*, 75: 278–305.

—— 2009. 'Meta-Induction and Social Epistemology.' *Episteme*, 6: 200–20.

—— 2017. 'No Free Lunch Theorem, Inductive Skepticism, and the Optimality of Meta-Induction.' *Philosophy of Science*, 84: 825–39.

—— 2019. *Hume's Problem Solved: The Optimality of Meta-Induction*. Cambridge, MA: MIT Press.

Solomonoff, R.J. 1964. 'A Formal Theory of Inductive Inference.' *Information and Control*, 7: 1–22 (part I), 224–54 (part II).

van Cleve, J. 1984. 'Reliability, Justification, and Induction.' *Midwest Studies in Philosophy*, 9(1): 555–67.

—— 2003. 'Is Knowledge Easy – or Impossible? Externalism as the Only Alternative to Skepticism.' In S. Luper (ed.), *The Skeptics: Contemporary Essays*, pp. 45–59. Aldershot: Ashgate.

White, R. 2015. 'The Problem of the Problem of Induction.' *Episteme*, 12(2): 275–90.

Wolpert, D.H. 1992. 'On the Connection between In-Sample Testing and Generalization Error.' *Complex Systems*, 6: 47–94.

—— 1996. 'The Lack of A Priori Distinctions between Learning Algorithms.' *Neural Computation*, 8(7): 1341–90.

GERHARD SCHURZ is Director of the Düsseldorf Center for Logic and Philosophy of Science at the Heinrich Heine University in Düsseldorf. His research interests are located in the Philosophy of Science, Epistemology, Logic and Cognitive Science. One of his major research projects concerns the problem of induction and in particular the optimality justification of meta-induction.

https://doi.org/10.1017/epi.2018.54 Published online by Cambridge University Press