Synthese (2021) 199:9979–10015
https://doi.org/10.1007/s11229-021-03233-1

ORIGINAL RESEARCH

The no-free-lunch theorems of supervised learning

Tom F. Sterkenburg[1] · **Peter D. Grünwald**[2,3]

Received: 17 December 2020 / Accepted: 22 May 2021 / Published online: 4 June 2021

Abstract
The no-free-lunch theorems promote a skeptical conclusion that all possible machine learning algorithms equally lack justification. But how could this leave room for a learning theory, that shows that some algorithms are better than others? Drawing parallels to the philosophy of induction, we point out that the no-free-lunch results presuppose a conception of learning algorithms as purely data-driven. On this conception, every algorithm must have an inherent inductive bias, that wants justification. We argue that many standard learning algorithms should rather be understood as model-dependent: in each application they also require for input a model, representing a bias. Generic algorithms themselves, they can be given a model-relative justification.

Keywords No-free-lunch theorems · Problem of induction · Machine learning

1 Introduction

The no-free-lunch (NFL) theorems of supervised learning (Wolpert 1992a, 1996b; Schaffer 1994) are an influential collection of impossibility results in machine learning. Computer scientists have ranked these results "among the most important theorems in statistical learning" (von Luxburg and Schölkopf 2011, p. 695), while some philosophers have read them as "a radicalized version of Hume's induction skepticism" (Schurz 2017, p. 825, p. 830).

For helpful comments we would like to thank Gordon Belot, Kathleen Creel, Sam Fletcher, Bas van Fraassen, Timo Freiesleben, Konstantin Genin, Daniel Herrmann, Wouter Koolen, Jürgen Landes, Jonathan Livengood, Daniel Malinsky, Conor Mayo-Wilson, and Eric-Jan Wagenmakers.

✉ Tom F. Sterkenburg
tom.sterkenburg@lmu.de

Peter D. Grünwald
pdg@cwi.nl

1 Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany

2 Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

3 Mathematical Institute, Leiden University, Leiden, The Netherlands

In a nutshell, the results say—or rather, are usually *interpreted* as saying—that we cannot formally justify our machine learning algorithms. That is, we cannot formally ground our conviction that some learning algorithms are more sensible than others: that we have reason to think some algorithms perform better in attaining the epistemic goals that we designed them to attain. In Wolperts original interpretation, "all learning algorithms are equivalent" (Wolpert 1995a, p. 129; 2002, p. 35), so that, for instance, a standard learning method like cross-validation has as much justification as *anti*-cross-validation (Zhu and Rohwer 1996; Wolpert 1996b, p. 1359f; 2021, p. 6f).

Yet for many such standard learning algorithms we *do* seem to have a justification. The field of machine learning theory is concerned with deriving mathematical learning guarantees, that purport to show that standard procedures, like minimizing empirical error on the training set, are better than other possible procedures, like *maximizing* empirical error (Shalev-Shwartz and Ben-David 2014). This raises a puzzle. How can there exist a learning theory at all, if the lesson of the NFL theorems is that learning algorithms can have no formal justification?

While this tension has been noted from the start (Wolpert 1996b, p. 1347), existing explanations of the consistency of the NFL theorems with learning theory (e.g., Wolpert 1996b, p. 1368ff, Bousquet et al. 2004, p. 202ff, von Luxburg and Scholköpf 2011, p. 692ff) are partial at best. In this paper, we investigate in detail the implications of the NFL results for the justification of machine learning algorithms. The main tool in our analysis is a distinction between two conceptions of learning algorithms, a distinction that has a parallel in the philosophical literature promoting a *local* view of inductive inference. This is the distinction between a conception of learning algorithms as purely data-driven or *data-only*, as instantiating functions that only take data, and a conception of learning algorithms as *model-dependent*, as instantiating functions that, aside from input data, also ask for an input *model*.

We argue that the NFL theorems rely on the former, data-driven conception of learning algorithms; but that many standard learning methods, including empirical risk minimization and cross-validation, should not be viewed as such. By their specification, such algorithms take two inputs: data, and an explicitly formulated model or hypothesis class, which constitutes a choice of bias. What we can reasonably demand from such model-dependent algorithms is that they perform as well as possible *relative* to any chosen model. Consequently, learning-theoretic guarantees are *relative to* the instantiated models the algorithm can take, and it is in this form that there *is* justification for standard learning algorithms. It is in this sense that learning theory allows one to say that empirical risk minimization is preferable to risk maximization, and that cross-validation is preferable to anti-cross-validation.

This is all consistent with the valid lesson of the NFL results, namely that every data-only learning procedure must possess some inductive bias. Our point is that this lesson should not be taken as a stick to wield against any possible learning algorithm. On the contrary: in model-dependent learning algorithms, this lesson is accounted for from the start.

The plan of the paper is as follows. First, in Sect. 2, we introduce the original Wolpert-Schaffer results. Still granting here the data-only conception of learning algorithms, we dispute the results' interpretation that *all algorithms are equivalent*. We discuss how this interpretation relies on an unmotivated assumption of a uniform

distribution over possible learning situations, that can in fact be seen as an explicit assumption that learning is impossible. We advance the alternative statement that *there is no universal data-only learning algorithm*. As instantiations of this statement, the NFL results illustrate and support the central insight in machine learning that every mechanical learning procedure, understood as a mapping from possible data to conclusions, must possess an inductive bias.

Next, in Sect. 3, we develop the model-dependent conception of learning methods, and show how this conception makes room for a justification for standard learning methods that is consistent with the NFL results. We start by pointing out that discussions surrounding the NFL theorems share a questionable presupposition with Hume's original argument for inductive skepticism: the idea that the performance of our inductive methods must be grounded in a general postulate of the induction-friendliness of the world. We discuss philosophical work that denies the cogency of such a principle, and that advances a local view of induction. This leads us to a local view of learning algorithms: the model-dependent perspective, and the accompanying possibility of a model-relative learning-theoretic justification. We discuss this in more detail for Bayesian machine learning, empirical risk minimization, and cross-validation, making explicit why learning theory allows us to say, for instance, that cross-validation is more sensible than anti-cross-validation. We conclude in Sect. 4.

Finally, we provide two appendices that complement the main argument. In "Appendix A" we investigate the formal consistency of the original NFL results with learning theory, and in "Appendix B" we list some important nuances to our discussion about model-dependent learning algorithms.

2 All learning algorithms are equivalent?

The first mentions in print of the "no-free-lunch theorems" of supervised learning are in Wolpert (1995a; 1996b, also see 1995b),[1] although an earlier version of the results already appeared in Wolpert (1992a, b). Around the same time, Schaffer (1994) presented a version of these results, with reference to Wolpert, as a "conservation law for generalization performance."

We start this section with presenting some basic versions of the Wolpert-Schaffer results, within a problem setting of prediction (Sect. 2.1), and within the original setting of classification (Sect. 2.2). Next, we discuss Wolpert's interpretation of his results that "all learning algorithms are equivalent, on average." We discuss the results' concern with all possible learning algorithms vis-à-vis the traditional philosophical concern with "inductive method," and note its restriction to data-only algorithms (Sect. 2.3). We then critically analyse Wolpert's equivalence claim and the underlying assumption of a uniform distribution over possible learning situations (Sect. 2.4). Finally, we advance the alternative NFL statement that there is no universal data-only learning algorithm (Sect. 2.5).

[1] Wolpert (1996b, p. 1343) attributes the term "no-free-lunch theorems" to the computer scientist D. Haussler. Wolpert and others also derived NFL theorems for mathematical optimization (Wolpert and Macready 1997; Ho and Pepyne 2002), which we do not discuss in this paper.

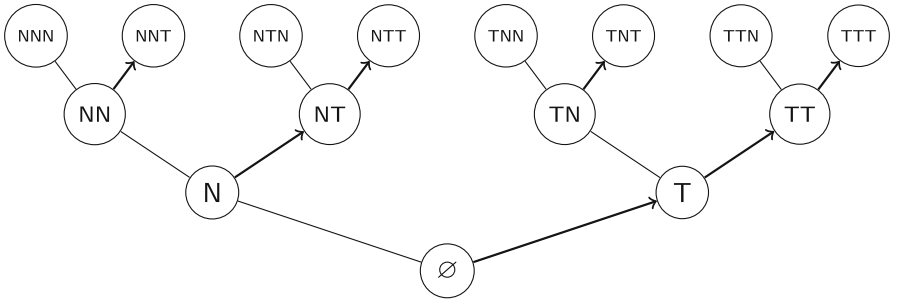

Fig. 1 NFL for prediction. For any possible learning method (say, the method that always chooses T, here represented by the arrows), there is one learning situation (path through the tree) with error 0 (follow the arrows), one with error 1 (never follow the arrow), and three situations each with error 1/3 and 2/3. Assigning each learning situation the same probability 1/8, the algorithm's expected error is 1/2

2.1 Prediction

Imagine that every day we are given a bowl of oatmeal for breakfast. Every morning on waking up, before we have our breakfast, we seek to predict whether it will be tasty (T) or not (N), based only on when it was the days before. A *learning algorithm* in this simple learning framework makes a guess whether the oats we are served today will be tasty, based on the data of the previous days. For a sequence of three days (see Fig. 1), there are in this scenario 2^3 logically possible histories or *learning situations* (of the form TTT, TNT, NTT, …), and already 2^7 possible learning algorithms (functions from {∅, T,N, TT, NT, TN, NN} to {T,N}). Let an algorithm's *error* be the ratio, among all predictions, of those predictions that are *incorrect* (e.g., a prediction of T and then obtaining N). Then a no-free-lunch statement in this scenario is that *for each possible level of error, every learning algorithm suffers this error in equally many possible learning situations*. Namely, one can verify that every single algorithm predicts perfectly (has error 0) in exactly one possible learning situation, predicts maximally badly (error 1) in exactly one other possible situation, suffers error 1/3 in three possible learning situations, and error 2/3 in the remaining three.[2]

Note that in thus counting learning situations and comparing these counts, we treat all possible learning situations on a par. Another way of doing this is to assume a *uniform probability distribution* on all possible learning situations, that is, a distribution that assigns the same probability to each of the finitely many possible learning situations. Then the above NFL result can be restated as the observation that, under the uniform distribution on learning situations, *every learning algorithm has the same expected error* of exactly 1/2. That is, every learning algorithm can be expected to do no better (or worse) than random guessing.[3]

[2] A similar example is given by Forster (1999, p. 551f).

[3] This statement generalizes to learning algorithms that issue probabilistic predictions, and the error of a single prediction $p \in (0, 1)$ is given by p in case of outcome N, and $1 - p$ otherwise. Then the more cautious a learning algorithm (the closer its predictions are to 1/2), the smaller the number of histories on which it attains either very low or high error, but this evens out in such a way that every learning algorithm still has an expected error of 1/2 (see Schurz 2021, p. 7ff).

2.2 Classification

The original Wolpert-Schaffer results were derived in a problem setting more standard in machine learning theory, the setting of *classification*. We first discuss the simplified setting of *non-stochastic* classification (Sect. 2.2.1), before we turn to the more general setting of *stochastic* classification (Sect. 2.2.2).

2.2.1 Non-stochastic classification

Imagine we want to learn to successfully classify whether a bowl of oats will be tasty or not, based on three different features we can determine before trying it: its temperature, its color, and its smell. Formally, supposing that these attributes are binary (either hot or cold, either bright or dull, either reeking or not), every *instance* of a bowl of oats can be represented by a length-three attribute vector of binary (write 0 or 1) components. This gives a total of eight (2^3) different possible instances, collected in the *domain set* $\mathcal{X} = \{0, 1\}^3$. A *classifier* is a function $f : \mathcal{X} \to \mathcal{Y}$ from the possible instances to their *labels* (tasty or not), collected in the *label set* $\mathcal{Y} = \{\mathsf{T},\mathsf{N}\}$. Supposing that the true labels are indeed fully determined by the attributes, the possible learning situations—possible *true* labelings of all instances of oats—are given by the possible classifiers. A *learning algorithm* A maps a sample $S = (x_1, y_1), \ldots, (x_n, y_n)$ of *training data*, pairs of instances and true labels, to a particular classifier f.

We are now interested in a learning algorithm's *generalization error* $L_{\overline{S}}(A(S))$: given some training sample S, how accurate is the classifier $f = A(S)$ selected by A on the instances that lie outside of S? Suppose the training data includes six of the total number of eight different possible instances of oats, determining the true tastiness labels for these six instances (see Table 2). There are four possible ways of classifying the two unseen instances, or four remaining possible learning situations f^*. Each possible learning algorithm selects a particular classifier in response to the training data, which classifies the two unseen instances in one of the four possible ways. That means that each possible learning algorithm (selected classifier f) has the same generalization error (ratio of incorrectly classified unseen instances over all unseen instances: either 0, 0.5, or 1) in the same number (one, two, one) of still possible learning situations f^*.[4]

Alternatively, we can put things again in terms of a uniform distribution $\mathcal{U}$ over all possible learning situations. So for this specific sample S of instances and labels, we have that uniformly averaged over the four remaining possible learning situations, the error of each learning algorithm is equal to 1/2. More generally, we can consider the same sample S_X stripped of its labels, and move the averaging to the front, so to speak, to cover how the possible f^* (now *all* possible f^*) assign labels to S_X, and how the algorithm fares for the resulting $S = S_X \times f^*(S_X)$ of instances and labels. But since for any four learning situations that label S_X in an identical way, an algorithm's average generalization error is 1/2, it remains 1/2 when averaged this way over *all* learning situations; and this reasoning goes through for any non-exhaustive $S_X \subsetneq \mathcal{X}$. Thus we

[4] Implicit here (and elsewhere in our presentation) is the use of a particular function to measure error or *loss*, the (standard) 0/1 loss function. The NFL theorems as stated do not necessarily go through for other loss functions (Wolpert 1996c).

	temp.	color	smell	tastiness, according to									
				$\hat{f}$	f_1^*	f_2^*	f_3^*	f_4^*	f_5^*	f_6^*	f_7^*	...	f_{256}^*
	0	0	0	N	N	N	N	N	N	N	N	...	T
	0	0	1	N	N	N	N	N	N	N	N	...	T
training	0	1	0	N	N	N	N	N	N	N	N	...	T
sample S	1	0	0	N	N	N	N	N	N	N	N	...	T
	0	1	1	N	N	N	N	N	N	N	N	...	T
	1	0	1	N	N	N	N	N	T	T	T	...	T
unseen	1	1	0	N	N	N	T	T	N	N	T	...	T
instances	1	1	1	T	N	T	N	T	N	T	N	...	T

Fig. 2 NFL for nonstochastic classification. For any learning algorithm A, any non-exhaustive training sample S (here of size six) and any possible labeling of S (say, all N, leading A to output classifier $\hat{f}$), there is the same number (here, four) of remaining possible learning situations (here, the classifiers f_1^* to f_4^*) that each label the (here, two) remaining instances differently. (Table adapted from Giraud-Carrier and Provost 2005.)

arrive at the statement that for any non-exhaustive training sample S_X of instances every learning algorithm A has expected generalization error $\mathbf{E}_{f^* \sim \mathcal{U}}\left[L_{\overline{S}}(A(S))\right] = 1/2$.[5]

2.2.2 Stochastic classification

An additional refinement in the standard framework for classification (see Shalev-Shwartz and Ben-David 2014) is that the true connection between instances and labels can itself be stochastic. Moreover, we assume some unknown probability distribution for the drawing of instances. Thus a learning situation is given by a distribution $\mathcal{D}$ over pairs of instances and labels.[6]

We now also measure generalization error in expectation over drawing an instance from $\mathcal{D}$: we shall call this the *risk*. But we have a choice here: do we take the expectation over all over $\mathcal{X}$, so including instances that were already in the training set, or do we discard the latter? Wolpert's "off-training-set" (OTS) risk, write $L_{\mathcal{D}\setminus S}(A(S))$, explicitly discounts already seen instances. He actually departs here from most of learning theory, where the error is standardly evaluated over all instances. We shall follow Wolpert in calling the latter quantity "i.i.d." (IID) risk, write $L_{\mathcal{D}}(A(S))$. Formally, for given sample $S = (x_1, y_1), \ldots, (x_n, y_n)$, $L_{\mathcal{D}}(A(S))$ is the probability, under $\mathcal{D}$, that an independently sampled example (X, Y) has $f(X) \neq Y$, where $f = A(S)$ is the classifier output by algorithm A on input S. This can also be written as

$$L_{\mathcal{D}}(A(S)) = \mathbf{E}_{(X,Y)\sim\mathcal{D}}[|Y - A(S)(X)|], \tag{1}$$

[5] Similar illustrations of the NFL theorems are given by Duda et al. (2001, p. 454ff); Carrier and Provost (2005, p. 9f); Barnard (2011, p. 1900f); Ortner and Leitgeb (2011, p. 720ff); Lattimore and Hutter (2013, p. 224ff). A precursor to this variant is the "theorem of the ugly duckling" due to Watanabe (1969, p. 376ff).

[6] The problem setting presupposed by Wolpert (his "extended Bayesian formalism," also see footnote 18) is more general still, in that both the classifier and the learning algorithm are stochastic: a classifier, like the true connection between instances and labels, is a $\mathcal{Y}$-conditional distribution over $\mathcal{X}$, and a learning algorithm is a distribution over classifiers. This additional generality does not affect the NFL statement in this section (cf. Rao et al. 1995, p. 473; 478f).

that is, the expected 0/1-error. In contrast, $L_{\mathcal{D}\setminus S}(A(S))$ is the probability that $f(X) \neq Y$, with $f = A(S)$, conditional on $(X, Y) \notin \{(x_1, y_1), \ldots, (x_n, y_n)\}$. This can also be written as

$$L_{\mathcal{D}\setminus S}(A(S)) = \mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[|Y - A(S)(X)| \;\middle|\; X, Y \notin \{(x_1, y_1), \ldots, (x_n, y_n)\}\right]. \quad (2)$$

A central claim in Wolpert's works is that OTS risk is a more natural measure of *generalization* performance than IID risk (1996b, p. 1345ff; 2002, p. 25ff). Note that it is certainly more similar to the generalization error in the previous nonstochastic case (where the labels of already seen instances are conclusively learned). But this does not make it clearly better in the stochastic case, where there is still an estimation problem even for already seen instances. We discuss the relation between the two notions (and the relevance of their difference in the context of the consistency of the NFL results with positive results in learning theory) in more detail in "Appendix A.1", and in the following always make clear what risk we mean.

Abstracting away from the oatmeal classification example, suppose instances are given by some finite-length set of features that can take a finite number of values, so that there is a (possibly huge yet) finite number m of possible instances. Given some training set S of n labeled instances, consider again any single unseen instance x. For each learning algorithm (selected classifier, assigning label y to x), there is a possible learning situation $\mathcal{D}$ in which the classifier's risk on this particular x is 0 (namely, a $\mathcal{D}$ that assigns probability 1 to label y, conditional on instance x). Likewise, there is a possible learning situation $\mathcal{D}$ in which the classifier's risk on this particular x is 1. Indeed, for each value in the unit interval there is a possible learning situation in which the classifier has *that* risk on x, as well as a counterpoint situation where the classifier has *one minus* that risk on x. The intuition that these risks all even out finds again a precise expression under the assumption of an (in this case, *continuous)* uniform distribution $\mathcal{U}$ over all learning situations—in this case, a uniform distribution[7] *over distributions*. Thus for any given set of training data, for any learning algorithm, the selected classifier's $\mathcal{U}$-expected risk on any single unseen instance is 0.5. This concerns a specific unseen instance, given some specific set of training data. But, crucially, we can again move the expectations to the front, to range over the whole process of drawing training data and measuring risk.[8] In this way we reach the statement of the NFL theorem, or the conservation law of generalization performance: every learning algorithm A, for any sample size $n < m$, has the same expected OTS risk $\mathbf{E}_{\mathcal{D}\sim\mathcal{U}, S\sim\mathcal{D}^n}\left[L_{\mathcal{D}\setminus S}(A(S))\right] = 1/2$.[9]

[7] In general, "uniform distributions" over general spaces are ill-defined or (if the spaces are noncompact) do not even exist: in "Appendix A.2" we show how in the current setting, an unambiguous definition is possible.

[8] See the exhaustive reconstruction by Rao et al. (1995) for details. An illustration similar to ours is given by Luxburg and Schölkopf (2011, p. 693f).

[9] The original statement by Wolpert and also Schaffer is still slightly different from the statement we give here. They actually take apart the marginal distribution $\mathcal{D}(X)$ that generates the instances and the conditional distribution $\mathcal{D}(Y \mid X)$ that labels the instances, allow the former to be any distribution, and let the uniform distribution $\mathcal{U}$ only range over the latter. We state this precisely in "Appendix A.2".

2.3 All learning algorithms …

We presented some versions of the Wolpert-Schaffer results, leading up to what is essentially the original form. But already the first example in the framework of prediction brings out an important characteristic of the NFL theorems: their concern, for the given learning problem, with *all possible learning algorithms*, understood as mappings from data to conclusions.[10]

There is, to begin with, no special regard for particular subclasses of learning algorithms, say those that we would intuitively call "inductive" (or indeed *learning* algorithms). In the prediction setting, for example, an "inductive" function that extrapolates the past data NN to the prediction N is no less a learning algorithm than an "anti-inductive" function that extrapolates data NN to prediction T, or indeed than the "learning-resistant" constant function that outputs T no matter what. As such, the NFL theorems can be seen to simply bypass the main companion problem to that of *justifying* induction: the problem of *specifying* or *describing* what actually constitutes inductive method or methods (see Lipton 2004, p. 7ff).[11]

That said, when it comes to the assessment of the results' implications, it seems there is only a small subset of all logically possible algorithms that we are really interested in. These are the *algorithms that are actually used*. There is a limited number of standard algorithms developed and analyzed in machine learning, generic algorithms that are employed in a wide variety of different domains. Naturally enough, the motivating discussions in Wolpert's writings focus on the ramifications of his results for the justification for *these* algorithms. We will discuss the justificatory implications of the NFL in detail in Sect. 3 below.

While the "all possible" in the NFL results' characteristic concern with *all possible learning algorithms* can be seen as a useful generality in the results' scope, there is also an important sense in which this scope is limited. This has to do with the restriction to "learning algorithms," understood as well-defined mappings from data to conclusions. The NFL results apply to formal learning rules that fully specify what conclusion follows which observed data. They clearly do not apply to a non-algorithmic conception of inductive method(s) that involves irreducibly informal factors (like, perhaps, everyday human and even scientific reasoning). But they do not even apply to a conception of learning methods as taking for input other (context-dependent) elements: the NFL

[10] We speak of "algorithms," following the custom in discussions surrounding the NFL results, even if it is perhaps better to speak of (for instance) learning *functions*. In reality, the same "algorithm" (map from data to conclusions) can be implemented—or indeed approximated—by many different (say, more or less computationally efficient) algorithms.

[11] In particular, we can understand the infamous riddle of induction due to Goodman (1954) as an expression of this problem of description. Suppose, on a minimal understanding of induction as "extrapolating the pattern from the past into the future," that we somehow were assured that inductive inference *is* justified: then, Goodman writes, we still would not know how to actually *do* induction. There are always multiple patterns we can find in the past, hence always multiple (and inconsistent) ways we can extrapolate these. Rather than following the route of attempting to specify which of the many possible extrapolations constitutes a proper inductive inference (the route Goodman himself took with his notion of *projectability*), we can remain agnostic and refrain from excluding *any* formal extrapolation rule: and indeed the NFL theorems apply to all of them.

results apply to a conception of learning algorithms as purely data-driven or *data-only*. We will also return to and expand on this point in Sect. 3 below.

2.4 … are equivalent?

The interpretation that Wolpert attached to his formal results, and that we went along with in our presentation, is that "for any two learning algorithms A and B … there are just as many situations (appropriately weighted) in which algorithm A is superior to algorithm B as vice versa" (Wolpert 1996b, p. 1360), or that "all algorithms are equivalent, on average" (Wolpert 1995a, p. 129; 2002, p. 35). The obvious worry about the significance of the NFL theorems concerns the qualifiers "appropriately weighted" and "on average" in these statements: that is, the presupposition of a uniform distribution on learning situations. This is indeed what the immediate responses in the literature focused on.

Perhaps the main criticism is that a uniform distribution is really a *worst-case* assumption for the purpose of learning. The "rational reconstruction" by Rao et al. (1995) shows that Schaffer's conservation law of generalization performance is equivalent to the (trivial) statement that for any unseen example, both possible classifications result in a generalization error of 0.5, *if* we measure the latter by uniformly averaging over both possible true classes. On a more conceptual level, this procedure of uniformly averaging corresponds to assuming that however many examples we have seen, we cannot have *learned* anything: the best guess for the label of any new example will always still be fifty-fifty. Thus these authors conclude that "the uniform concept distribution … in which every possible classification of unseen cases is equally likely … is the definition of a uniformly random universe, in which learning is impossible" (ibid, 475).[12] Obviously the NFL theorems cannot be said to hold much significance if we understand them as the observation that every learning algorithm is equivalent *in a universe where learning is impossible*.[13]

It has been suggested that this particular criticism can be countered by the observation that a uniform distribution is not a necessary condition for NFL theorems to go through (e.g., Giraud-Carrier and Provost 2005, p. 10). Rao et al. (1995, p. 475ff) show that generalization performance is conserved under a wider class of distributions; and indeed Wolpert (1996b, p. 1361f) also already gives "extensions for nonuniform aver-

[12] In "Appendix A", where we discuss the formal consistency of the original Wolpert-Schaffer results with learning theory, we bring out the same point in yet another way.

[13] Also see the discussion by Schurz (2017) of the NFL theorems in the setting of prediction, with the corresponding "state-uniform" prior that assigns equal probability to every same-length sequence of outcomes. Schurz connects this to Carnap's discussion of the corresponding confirmation function $c\dagger$, "tantamount to the principle never to let our past experiences influence our expectations of the future … in striking contradiction to the basic principle of all inductive reasoning" (Carnap 1950, p. 565), and also points out that the corresponding uniform measure on the space of all infinite sequences assigns probability 1 to infinite sequences (1) having a limiting relative frequency 1/2 and (2) being incomputable. Schurz concludes that "proponents of a state-uniform prior distribution are strongly biased: they are a priori certain that the world is irregular so that induction cannot have any chance" (2017, p. 834). The point that certain uniform distributions lead to "unlearnability" goes back at least to Boole (1854): if we assume that each ball in a bag has an equal probability of being black or white, he writes, then "past experience [of drawing with replacement] does not in this case affect future experience" (ibid., p. 372).

aging." But as long as the results do not extend to *all* distributions (and they do not: there is a certain symmetry that must be retained, Rao et al. 1995, p. 477), the worry remains that the NFL results are simply an expression of the induction-hostileness of the presupposed weighing distribution.

Wolpert was aware of this perspective on his results.[14] In (1992a), he himself refers to the assumption of a "maximum-entropy universe"; the way he puts his point there is that "[s]ince such a universe cannot be ruled out on an a priori basis, it is theoretically impossible to come to any conclusions about how to generalize using *only* a priori reasoning." But the statement that it is a priori *possible* that there are (in expectation) no distinctions between learning algorithms is weaker than the categorical statement that there *are* (in expectation) no a priori distinctions between learning algorithms, the claim of the later paper (1996b).

In this paper (ibid., 1362ff), Wolpert actually argues that the uniform distribution does have a preferred status. He starts by allowing that if we change the weighing of learning situations, then there could arise "a priori distinctions" between learning algorithms. However, he continues, "a priori" such a change of weighing could just as well favor algorithm A as B: "[a]ccordingly, claims that 'in the real world [the distribution over learning situations] is not uniform, so the NFL results do not apply to my favorite learning algorithm' are misguided at best" (ibid., 1363). Indeed, he points at results in the same paper regarding averages over *prior distributions* over learning situations, with the interpretation that there are as many *priors* for which A is superior to B as the other way around. From this perspective, "uniform distributions over targets are not an atypical, pathological case … [r]ather they and their associated results are the average case (!)" (ibid.).

This jump to a higher level is clearly inconclusive: we can restate the same worry at *that* level.[15] Most remarkable, however, is Wolpert's dialectical move of turning the table on the critic: "the burden is on the user of a particular learning algorithm. Unless they can somehow show that [the true prior] is one of the ones for which their algorithm does better than random … they cannot claim to have any formal justification for their learning algorithm" (ibid.).

Curiously, responses in the computer science literature critical of the significance of Wolpert's results have essentially *followed* him here. Rao et al., after discussing how NFL theorems must depend on a symmetrical prior, conjecture that "our world has strong regularities, rather than being nearly random. However, only time and further testing of physical theories can refine our understanding of the nature of our universe [and] might lead to a reasonable estimate of [the true prior] in our world" (1995, p. 477). Giraud-Carrier and Provost emphatically set forth as an implicit yet generally accepted "weak assumption of machine learning" that "the process that presents us with learning

[14] His discussion of the intuition behind his results in (1996a, p. 133f; 2002, p. 38ff) is in fact very similar to the analysis of Rao et al. in tracing the results back to the assumption of fully random labels of unseen instances.

[15] Elsewhere (1995a, footnote 3; 2002, footnote 4), Wolpert presents the situation rather in terms of the critic of the uniform distribution attempting to "jump a level" in questioning the uniform distribution on priors, "arguing that some [prior distributions over learning situations] are 'more likely' than others"—but to no avail, "the math responds the same way as it did to the [lower-level] objection." This is a reiteration of the dialectical move we criticize next.

problems … induces a non-uniform probability distribution [over learning situations]" (2005, p. 11).[16] But this Wolpert would not disagree with: he writes himself that a nonuniform distribution "is why some algorithms tend to perform better than others *in the real world*" (Wolpert 1996b, p. 1361, emphasis ours).[17] The point is to give a "formal justification" for believing in any such distribution. Indeed, if we seek to criticize the assumption of a uniform distribution in Wolpert's claim that all algorithms are a priori equivalent *by postulating a different, nonuniform, distribution*, then we better provide a justification for postulating *that* distribution. The result is that we find ourselves in a corner, because it is not clear where to look for such a justification. What we should have done, of course, is to insist that Wolpert justify *his* assumption.

In fact, a more fundamental reply is to demand a reason for postulating *any* prior distribution over learning situations. Doing so is a formal requirement in Wolpert's "extended Bayesian formalism" (unlike in the conventional classification framework); but that merely shows that the framework is constraining in a way that we may find inpalatable.[18,19] Indeed, it is not at all clear what it is supposed to *mean* to assign probabilities to possible learning situations. An epistemic interpretation, as some (ideally rational) agent's degrees of belief, is perhaps the easiest to make sense of, but immediately throws us back to the justification for any specific choice of prior distribution: in particular, the idea of a uniform distribution as an objective-logical "indifference prior" has long been abandoned by philosophers and statisticians alike as a viable option (see, e.g., van Fraassen 1989, p. 293ff; Zabell 2016). This is, in any case, not what Wolpert appears to have in mind: the suggestion is rather that we should think of these probabilities as objective-*physical*, as chances.[20] But in the absence of a fuller

[16] The accompanying *strong* assumption of machine learning is that this distribution is actually "explicitly or implicitly known, at least to a useful approximation" (Giraud-Carrier and Provost 2005, p. 11).

[17] In the introduction of his paper, Wolpert writes that "[i]t cannot be emphasized enough that no claim is being made … that all algorithms are equivalent *in practice*, in the real world … The sole concern of this paper is what can(not) be formally verified about the utility of various learning algorithms if one makes no assumptions concerning targets" (1996b, p. 1344). Also see Wolpert (1992a, p. 61).

[18] In Wolpert's EBF, one defines a probability distribution P that ranges over "target functions" f as well as "hypotheses" h, where the latter stand for the learning algorithm's possible guesses for the true target f. So $P(f)$ and $P(f \mid d)$ represent the "true" or "objective" priors and posteriors over targets, and $P(h \mid d)$ the learning algorithm. The need to thus specify a prior over targets f deviates strongly from the conventional learning theory framework, where it is only assumed that there is some unknown distribution governing instances and labels.

[19] The name of Wolpert's framework derives from its aim to generalize the conventional Bayesian framework where "there is no direct analogue to $P(h|d)$ … Viewed another way, [it] has $P(h|d)$ pre-fixed, to be the 'Bayes-optimal' $P(f|d)$" (Wolpert 1995a, p. 122). Thus, under the Bayesian interpretation of probability as degree of belief, "you automatically know $P(f)$ exactly." But "a 'truth' f and a guess h are different objects … a formal statement connecting $P(f|d)$ and $P(h|d)$ corresponds to an extra assumption not demanded by the mathematics." (1996a, p. 83f, notation aligned with the previous). This is, to put it mildly, an idiosyncratic rendering of the Bayesian approach. Rather than conflating, absurdly, an epistemic and an ontic interpretation of the prior $P(f)$, a Bayesian would stay clear of the latter—that is neither demanded by mathematics, but presupposed by Wolpert.

[20] See especially Wolpert (1996a, p. 84): "if we had sufficient knowledge of the laws of physics (in particular, the boundary conditions of the universe) and of the (resultant) laws of human psychology, and if we were sufficiently competent to perform the appropriate quantum mechanical calculatations, then we might say that we could calculate [the distribution of learning situations] exactly." Again, the *objective* interpretation of distributions over learning situations, complementary to the *learning algorithm's* distribution over "hypotheses," is central to Wolpert's framework.

account of the nature of these chances we do not see much reason for going along with the idea that the universe is governed by some objective distribution generating learning situations—let alone that this distribution should be uniform.

In sum, it would be granting Wolpert too much to accept that it is on us to show, contra his equivalence claim, that some algorithms are generally better than others. (We do not even need to think that "generally" is a qualifier that can be made meaningful here.) The burden is rather on Wolpert to justify the presuppositions that back his claim, in particular the assumption of a uniform distribution on learning situations, and this he has not done.

2.5 There is no universal data-only learning algorithm

We can, however, formulate a weaker variation of the NFL results, a statement that is implied by the original but that does away with the uniformity assumption. In stating it, we also make explicit the observation from Sect. 2.3 that we are still talking of *data-only* algorithms, functions from data to conclusions:

> For any data-only learning algorithm, there exists a learning situation in which this algorithm does not perform well, while in this same situation another data-only learning algorithm does perform well. (∗)

In other words, there is no single data-only learning algorithm that performs well *whenever some* data-only algorithm performs well: there is no *universal* data-only learning algorithm.

Note right away that the truth of any instantiation of this statement depends on the learning problem in question, including the possible methods and the adopted notion of good performance. It is not too hard to come up with (artificial) learning problems for which the statement is false (e.g., a problem that is formulated such that the possible learning situations explicitly accommodate a particular learning method).[21] The statement is relevant insofar it holds for problems within most standard learning frameworks and natural measures of good performance.

For instance, we retrieve this statement from the original Wolpert-Schaffer result if we drop the uniformity assumption and make "good performance" precise as (say) "having expected risk strictly smaller than 1/2." Namely, for every learning algorithm A_1, for any sample size n, there exists a learning situation $\mathcal{D}$ such that the algorithm has expected OTS risk $\mathbf{E}_{S\sim\mathcal{D}^n}\left[L_{\mathcal{D}\setminus S}(A_1(S))\right]$ at least 1/2, while *another* algorithm A_2 has expected OTS risk below 1/2 (indeed *zero*, for choice of $\mathcal{D}$ that labels instances deterministically via some f^*, and A_2 that always outputs this f^*).

A variant for IID risk is the NFL theorem in the standard textbook by Shalev-Shwartz and Ben-David (2014, p. 61ff). Here the notion of performance is that in $\mathcal{D}$-expectation

[21] In fact, Wolpert's own framework provides another example, if we see the uniformity assumption as *part of the formulated learning problem*. This renders the problem trivial because there is only one possible "truth" or learning situation (the "no-learning" truth where the correct classifications are uniformly random) so that the statement must be false (in this case, because all algorithms are equally good, in terms of expected risk, in this one situation, hence "universal").

over samples of size n no more than half the total number of possible instances, the algorithm's IID risk is smaller than $1/4$. Correspondingly, their NFL theorem states that for every learning algorithm A_1 there is a learning situation $\mathcal{D}$ such that its expected IID risk $\mathbf{E}_{S\sim\mathcal{D}^n}[L_{\mathcal{D}}(A_1(S))]$ is higher than $1/4$, while that of another algorithm A_2 is lower than $1/4$ (indeed again zero). The authors write that the "theorem states that for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner" (ibid., 61).

Another example is given by the NFL theorems collected by Belot (2021) for problems of prediction. (He calls these results "of the *absolute* variety," as opposed to "measure-relative," which would include the original Wolpert-Schaffer results.) The learning situations in these problems are (probability measures over) *infinite* sequences of binary outcomes, and he considers different types of effectively computable learning functions (namely, "extrapolators" that are, as in our example in Sect. 2.1, functions from past outcome sequences to next outcomes, and "forecasters" that output probabilities of next outcomes) and for both of these types various notions of good performance. In each case he derives two types of results, that are both instantiations of the general NFL statement that there is no universal algorithm: that for each learning algorithm A_1 there is a second algorithm A_2 that performs well in those situations in which A_1 does, *and* in other situations still ("better-but-no-best"),[22] and that for each A_1 there is an A_2 such that the situations in which they perform well are *disjoint* ("evil-twin").

These examples also illustrate that statement (*) retains much of the spirit of the original Wolpert-Schaffer statement. In particular, it is a clear expression of the central insight in machine learning (Mitchell 1980, 1997; Dietterich 1989; Russell 1991; Shalev-Shwartz and Ben-David 2014) that no purely data-driven learning algorithm—no formal inductive function from data to conclusions—can be successful in all circumstances. That is, every such data-only algorithm must possess some *inductive bias* that determines in which restricted class of situations it performs well, and hence in which situations it does not. What statement (*) still adds to this is that such a learning algorithm's inevitable inductive bias excludes it from learning successfully in some situations that are not *unlearnable*: situations in which *some other algorithm* would perform well. But it does not go as far as the original Wolpert-Schaffer statement that all (data-driven) algorithms are *equivalent* in their performance, depending as this does on the additional and unmotivated assumption of a uniform prior distribution.

3 Generic algorithms and local models

In this section, we investigate the significance of the NFL-statement (*) for the justification for machine learning algorithms. The route we take is to first relate the NFL results to Hume's skeptical argument about induction (Sect. 3.1). We note that both Hume's original argument and discussions of the original Wolpert-Schaffer results presuppose that justifying inductive methods requires justifying a general postulate of the induction-friendliness of the world. Subsequently, we discuss philosophical work

[22] Note that such results, while instances of NFL statement (*), go against the Wolpert-Schaffer statement that all algorithms are equivalent. There are in this learning framework strictly better and better performing algorithms—just no *best*.

that denies this presupposition, and that promotes a *local* perspective on induction (Sect. 3.2). We argue that a local conception of induction, applied to machine learning, points at a more natural conception of learning algorithms: rather than one-place functions *on data only*, many standard learning algorithms are better conceived of as two-place functions that for their operation also require some *model* (Sect. 3.3). Learning-theoretic guarantees do justify the use of such algorithms, in a local, *model-relative* manner (Sect. 3.4).

3.1 The road to skepticism

The NFL theorems, both the original Wolpert-Schaffer results and instantiations of the statement (*) of Sect. 2.5, are mathematical theorems. They say something about the impossibility of mathematically *proving* that some learning algorithms, conceived of as purely data-driven, perform better than others. As such, they can be seen as versions of the first, *deductive*, horn of the fork that constitutes Hume's orginal argument against a justification for induction. This first horn concerns the impossibility of inferring good performance of inductive inference using only deductive, *a priori* reasoning: since it implies no logical contradiction that induction does not perform well, we can never deductively derive, from *a priori* premises only, that it does.[23] Similarly, the NFL results show for any learning algorithm that it implies no contradiction that this algorithm does not perform well (does not perform at least as well as other algorithms), by showing that there are *a priori* possible situations in which it does not.[24]

This does not yet constitute a skeptical argument that we can offer *no* rational grounds for thinking that one algorithm performs better than another. Likewise, the first horn of Hume's fork did not yet establish a skeptical conclusion about the grounds for inductive inference. Arguably, the novelty and force of Hume's argument lay in the second horn of his fork: the assertion that neither can we offer, on pain of circularity, good *non*deductive or *empirical* grounds for thinking that inductive inference must perform well. Only the two horns taken together lead to the skeptical conclusion that we can offer *no* rational, epistemic ground for using inductive inference: that we cannot justify induction.

[23] Recall that "performs well" in NFL statements could mean, for instance, "has in expectation a sufficiently high probability of a correct conclusion." In the analogous reconstruction of Hume's argument the deductive horn would thus amount to more than the "boring" observation that induction is *fallible* (Okasha 2001, p. 309): it is not just that inductive inference is not itself deductively valid and could lead to false conclusions, it does not even imply a contradiction that it is not *likely* to give correct conclusions (see, e.g., Skyrms 2000, p. 30ff). The purpose of our very compressed presentation of Hume's argument here is mainly to draw analogies to the NFL results and their implications; and we pass over some issues regarding the proper reconstruction of Hume's original argument that are not uncontentious (including whether the argument was intended to extend to probabilistic induction, and indeed the possible differences in conceiving of the two horns as "deductive" v. "inductive" or as "a priori" v. "empirical"). See Lange (2011); Henderson (2020) for entries to the literature on Hume's (historical) argument.

[24] This is clearly what the NFL results of type (*) do. The original Wolpert-Schaffer results fit this statement less well, at least in the usual interpretation, because of the (non-tautological) assumption of a uniform distribution. But it fits an earlier interpretation by Wolpert, mentioned in Sect. 2.4: it is logically *possible* that learning situations are generated by a uniform distribution, hence that no algorithm performs better than any other.

Perhaps the Wolpert-Schaffer results were not intended to support a skeptical conclusion, and we should read conclusions of the sort that "methods for induction to unseen cases cannot be justified rigorously" (Schaffer 1994, p. 264) or that "one can not formally justify [standard learning algorithms]" (Wolpert 2002, p. 38) as merely indicating the limits of mathematically founding the performance of learning algorithms. However, something more than that is suggested in the original discussion surrounding these results, by the nods to Hume (Wolpert 1996b, p. 1341; Schaffer 1994, p. 264), but also by the outlines of a move very reminiscient of Hume's. This is the idea, discussed before in Sect. 2.4, that the only way remaining to found the good performance of our learning algorithms is to postulate that "the world" (or "nature," or the "universe") has a certain structure that guarantees this. Hume's original argument in fact *starts* with the premise that inductive reasoning proceeds upon the principle that "nature is uniform." It is this principle that is subjected to the two horns; in particular, that we cannot justify it *inductively* or *empirically*. Namely, any attempt to derive the uniformity of nature from past such observed uniformity would require the very principle at stake and thus be viciously circular.

Hume's argument and most of its later reconstructions simply concerned "inductive inference" or "inductive method," exemplified by something like enumerative induction but beyond that largely left unspecified (prompting a distinct problem of description, recall Sect. 2.3). The NFL theorems concern all possible purely data-driven learning algorithms. Still, the skeptical threat of the NFL results lies in their application to "our standard algorithms," the generic learning algorithms that we actually use (recall again Sect. 2.3). So both Hume's argument and discussion surrounding the NFL results envisage some restricted collection of generic inductive methods. And in both cases we see that the performance of these inductive methods is paired to a particular structure the world may or may not have. If the world has the matching structure, then our inductive methods perform well; if not, they do not.[25] Consequently, the dialectics turns on the justification for such an assumption on the world: in Hume's argument from the start, in the case of the Wolpert-Schaffer results in the ensuing discussion. The NFL statement (*) is similarly susceptible to this move: if we do want to uphold the existence of well-performing generic (*universal*) learning algorithms, then it seems we must postulate that the world has a structure that facilitates such algorithms' performance. But in all cases, it appears impossible to justify, without question-begging, such an assumption on the world, whence we are driven towards a skeptical conclusion.

3.2 Localizing induction

An idea that has been advanced in the philosophical literature is that we may avoid being driven there by *denying that inductive inference relies on universal uniformity principles* (Okasha 2005b). This idea builds on arguments that it is hopeless to try

[25] In Hume's argument, this structure is given by the principle of the uniformity of nature; in the case of the NFL results, by a *non*-uniform distribution. Note the opposed denotations of the term "uniform" here: a uniform distribution intuitively signifies complete randomness and thus lack of structure and regularities, so that it corresponds to an assumption of *non*-uniformity of nature in the sense of Hume.

and give a precise account of a principle of the "uniformity of nature" (Salmon 1953; Sober 1988, p. 55ff).[26]

Sober (1988, p. 58ff; also see Okasha 2005b, p. 245ff) argues that in presupposing that induction relies on a single principle of uniformity, Hume actually commits a *quantifier-shift fallacy*. It is not the case, as Hume has it, that there is a certain assumption (the uniformity of nature) that every inductive inference requires; it is rather the case that *every inductive inference requires a certain assumption*. That is, rather than all relying on a single universal uniformity principle, every induction relies on a specific and *local* empirical assumption.

Arguments against universal uniformity principles usually run together with arguments that it is hopeless to try and give a precise account of "inductive method" (Putnam 1981, 1987; Rosenkrantz 1982; van Fraassen 1989, 2000; Norton 2003, 2010). Okasha (2001) indeed develops an argument analogous to Sober's where he diagnoses the fault in Hume's reasoning to be the presupposition that inductive inference is given by universally applicable rules. He, like Norton (Norton 2003, p. 666; 2014), argues that the denial of this presupposition actually blocks the skeptical argument.

These ideas offer a *local* perspective on inductive inference.[27] In order to assess the value of this perspective for machine learning algorithms and their justification, we make two observations.

First, even if we grant that Hume's original argument no longer goes through when we deny the existence of universal uniformity principles or inductive rules, it does not follow that we are safe from a skeptical argument. As Sober (1988, p. 66ff) himself emphasises, there are still always assumptions involved in an inductive inference, that themselves stand in need of justification. Even if we are safe from Hume's argument that any nondeductive justification of induction must be *circular*, it appears we will now be facing an endless *regress*, where each empirical assumption can only be justified by another induction with its own empirical assumptions.

Yet Okasha (2005b) is more optimistic: "The *form* which the inductive sceptic's argument takes on the ∀∃ picture—pushing the demand for justification further and further back—seems somehow less problematic than in the ∃∀ case," where "the whole practice of reasoning inductively seems to be premised on an enormous, untestable assumption about the way the world is" (ibid., pp. 252, 251). We do not think that this settles the matter, but it does clearly bring out a crucial advantage of a local perspective on induction. Namely, this perspective is much closer to what the problem of justifi-

[26] Arguably it is already the lesson of Goodman's new riddle that Hume's uniformity of nature principle is empty (Okasha 2001, p. 309; Lange 2011, p. 58f; also recall footnote 11), even if this is obscured by Goodman seeking to patch Hume's supposition by a restriction to "projectable" predicates (Rosenkrantz 1982; Okasha 2001, p. 320).

[27] The general idea of urging a local perspective on induction can be discerned in various philosophical currents, including the pragmatist tradition (see Levi 1967; Bogdan 1976). Nor is the questioning of a principle of the uniformity of nature (Peirce 1878, 1902) or formal schemes of induction (see McCaskey 2021) remotely new. We focus here on the relatively recent writings by Okasha, because they specifically address the Humean problem of (global) justification, and because they bring out nicely the points that are important to our argument.

cation looks like in *actual enquiry*.[28] Plausibly, in an actual enquiry, each inference takes place within a constellation of context-specific or local empirical assumptions.[29] The motivation for such an inference will focus on one or more of these assumptions, and not on a universal uniformity principle. Furthermore, the question of justification does not only target these assumptions: even *given* these assumptions, there can still be room for different inferences, in which case there is still the question of the justification for the inference of choice, or the method used for the inference. We will argue below that both aspects are important to the question of the justification for machine learning algorithms.

Second, it might seem that a local conception of induction, inasmuch as it is coupled to the position that inductive inference cannot be encoded into general *rules*, actually does not sit very well with the enterprise of machine learning. After all, and arguably in contrast to day-to-day human or even scientific reasoning, machine learning is characterized by the design and use of learning algorithms: fully mechanical, generic procedures for inductive inference.

The rejection of general inductive rules in a local perspective must be qualified, though. For instance, Okasha (2001; also see 2005a), in the course of arguing against the idea of general rules for inductive inference, does endorse Bayesian conditionalization as the rational procedure for learning from experience.[30] There appears to be a tension there (cf. Henderson 2020): is updating by conditionalization not a rule? Okasha, however, makes a distinction: "a rule of inductive inference is supposed to tell you what beliefs you should have, given your data, and the rule of conditionalization does not do that … the state of opinion you end up in depends on the state you were in previously; whereas if you apply an inductive rule to your data, the state of opinion you end up in depends on the instructions contained in the rule" (2001, p. 316). The output of Bayesian conditionalization does not depend on the input data *only*: it also depends on "the state you were in previously," ultimately, a prior probability assignment. The rejection of general rules for inductive inference here thus concerns purely *data-driven* rules.

This idea is, of course, very much supported by the statement of the NFL theorems we advocated: there is no universal *purely data-driven* learning algorithm.[31] Moreover, this is perfectly consistent with allowing for general rules for induction that also require other inputs, plausibly inputs that encode local assumptions, like (in the case of the

[28] This is also a main selling point of Norton's "material theory of induction." But it is likewise far from clear that Norton can uphold his promise of escaping inductive skepticism, in particular, of escaping the endless justificatory regress (Kelly 2010).

[29] For a dissenting view, based on examples from the history of science of inductions within "theoretically impoverished contexts," see Lange (2002, 2004).

[30] So do Rosenkrantz (1982) and van Fraassen (1989). Norton (2010, 2014) does categorically argue against any formal scheme for induction (including the Bayesian scheme), which places his theory at odds with the perspective we develop here.

[31] Van Fraassen's (1989, p. 132ff; 2000, p. 256ff) rejection of "the ideal of induction" ("a rule" that is "rationally compelling," "objective," and "ampliative") relies for an important part on results that go back to a proof of Putnam's (1963) that is in effect an instantiation of NFL statement (*). Putnam shows by a diagonalization argument that for each prediction algorithm, there exist infinite data sequences on which this algorithm does not perform well, sequences that are in fact themselves effectively computable so that another algorithm predicts them *perfectly* (also see Sterkenburg 2019).

Bayesian method) a prior probability distribution. In sum, the lesson we take from a local conception of induction is not to reject rules for induction: the lesson is to fine-grain the notion of inductive rule, to conceive of it as a procedure that can also take for input local assumptions. Applying this perspective to machine learning algorithms, we will also be able to qualify the sweeping skeptical conclusion that the NFL theorems seemed to lead us to.

3.3 Model-dependent learning algorithms

We think it highly implausible that the use of machine learning algorithms relies, explicitly or implicitly, on a general "assumption of machine learning" about the learning-friendliness of the world, let alone a belief in some all-governing non-uniform prior distribution on possible learning situations. The assumptions that accompany the use of a learning algorithm in any particular context are normally themselves of a context-dependent, *local* nature. But how to square the role of local assumptions with the use of generic mechanical learning procedures?

The observations of the previous section point us at an answer. Many standard learning algorithms are not purely data-driven, but must also take for input a *model*. Such *model-dependent* algorithms instantiate, not a one-place function that maps data to conclusions, but a *two*-place function that maps data *and a model* to conclusions. Crucially, such algorithms can be given a *model-relative* justification.

In the following, we illustrate model-dependent learning algorithms using three standard machine learning examples: Bayesian machine learning (Sect. 3.3.1), empirical risk minimization (Sect. 3.3.2) and cross-validation (Sect. 3.3.3). These methods all have in common (as do most if not all standard model-dependent learning algorithms that we know of) that they select a hypothesis or combination of hypotheses with good predictive performance, measured in terms of the loss function of interest (empirical risk minimization, cross-validation) or a related measure such as the likelihood (Bayes). We discuss how these methods receive a model-relative justification in the form of learning-theoretic guarantees, and thereby bring out why such claims as "the NFL theorems indicate that cross-validation has no more inherent justification than anti-cross-validation" are misleading.[32] We conclude our examples with a discussion of the consistency of the NFL results with learning guarantees (Sect. 3.3.4).

Finally, we have delegated to "Appendix B" some nuances that distract from the argument's main thrust.

3.3.1 Bayesian learning

The Bayesian scheme, central to many philosophical accounts of rational learning, also constitutes an important approach in machine learning (Duda et al. 2001; Bishop 2006). What characterizes Bayesian learning is that an algorithm must be provided with a prior distribution over some domain of probability distributions, and this choice

[32] We here say very little about potentially useful distinctions between different *types* of justification and the exact nature of the accompanying inductive assumptions, but this is a natural avenue for further investigation (Corfield 2010).

of prior constitutes a choice of model. The role of the prior as a variable input factor lends such an algorithm a considerable genericity: the algorithm *itself* does not come with a particular model, but is flexibly supplied with a specific model in each specific application. This is also what provides room for a *model-relative* learning-theoretic justification: a demonstration that *relative to* the choice of prior distribution, a Bayesian algorithm performs well.

We now discuss this in some detail for Bayesian machine learning in the framework of classification, the realm of the original Wolpert-Schaffer results. Here, the prior Π is usually taken over a set of *conditional* probability distributions of the form $P(Y \mid X)$ with $Y \in \mathcal{Y}$, $X \in \mathcal{X}$ the possible labels and instances, respectively. (Recall the oatmeal example of Sect. 2.2, where $\mathcal{Y} = \{\mathsf{T},\mathsf{N}\}$ and $\mathcal{X} = \{0,1\}^3$.) The distributions are extended to n outcomes by assuming that the data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in the training set are sampled independently. The set of probability distributions in the prior support (that is, those with prior density or mass greater than 0) demarcate a model $\mathcal{M}$, a set of (conditional) probability distributions. A prototypical example is the logistic regression model (Hastie et al. 2009, p. 119ff), in which the $X_i = (X_{i,(1)}, \ldots, X_{i,(k)})$ are vector-valued as in our example, and the $P(Y \mid X)$ are given by linear functions $\sum_{j=1}^{k} \beta_{(j)} X_{(j)}$, rescaled by a fixed nonlinear function so as to become probabilities that sum to one.

There exist several variations of the Bayesian stance, which differ in how the prior is interpreted. For the purpose of our discussion, most relevant is the distinction between a *subjective* and a *pragmatic* stance. Under the former, the prior quite literally encodes one's beliefs (which can be elicited by, for example, testing willingness to bet on certain outcomes). That is, the relevant inductive assumption can be equated with one's beliefs. Alternatively, under a pragmatic interpretation, to which most practitioners subscribe, one still assumes the *model* (set of all distributions in the support of the prior) to be correct, but one can choose the prior Π for other, more pragmatic reasons. These could be considerations of (computational) convenience, of optimizing worst-case behaviour (this leads to "noninformative" or "flat" priors), or a mix of prior knowledge with worst-case and computational considerations. For example, a standard pragmatic approach for the logistic regression model is to take a Gaussian prior centered at 0 on the $\beta_{(j)}$'s.

Regardless of the prior's origin, it serves as an input to the Bayesian algorithm. Together with the data, i.e., the training sample, one uses Bayes's rule to update the prior to a posterior. The posterior over the distributions is then used to output a classifier $\hat{f}_{Bayes}$, defined as the function from $\mathcal{X}$ to $\mathcal{Y}$ that has the largest probability of being correct according to the Bayesian posterior predictive distribution (Bishop 2006). In contrast to the notion of learning algorithm in Sect. 2, where an algorithm only takes data, the Bayesian algorithm requires additional input: the user's inductive assumptions, codified explicitly as prior and induced model. One cannot avoid stating these explicitly—without specifying a prior and hence a model, the outcome of the Bayesian algorithm is simply undefined.

When it comes to the question of justification, the distinction between the two Bayesian stances is also relevant. Under the subjective stance, the Bayesian algorithm is simply *optimal*: among all algorithms, it leads to the best possible classifier (with smallest risk) under one's own inductive assumptions as encoded by the prior. In other

words, if the prior truly reflects one's beliefs, then one must also believe that the Bayesian procedure, with this prior, is justified. *If* one is willing to take the subjective stance, then any arguing that the Bayesian algorithm has no more inherent justification than any other algorithm, let alone "anti-Bayesian learning" (where one selects the classifier with the *highest* risk under the posterior), is futile.[33]

Under a pragmatic view of Bayesian inference, the prior weights cannot be directly related to one's beliefs, and the Bayesian algorithm cannot be said to be optimal in the previous sense. Nevertheless, under the pragmatic view one can still show that the Bayesian procedure has a certain model-relative optimality, even if the specific choice of prior over the same model now becomes important. We already mentioned how choices of noninformative priors can optimize worst-case behavior, by which we meant that $\hat{f}_{\text{BAYES}}$ has the smallest possible generalization error in the worst case under all $P^* \in \mathcal{M}$.

Furthermore, there exists a plethora of results (e.g., Ghosal et al. 2000, 2008) showing that, under very weak conditions on the model $\mathcal{M}$, one can select priors such that for all $P^* \in \mathcal{M}$, the posterior concentrates around P^* at a certain rate. In our context, this implies that the expected generalization error of $\hat{f}_{\text{BAYES}}$ converges to the generalization error one could obtain if one knew the "true" (leading to the best possible predictions) P^*. Moreover, one can give nonasymptotic bounds on the difference in generalization errors (Grünwald and Mehta 2020). These results provide a clear model-relative justification for the pragmatic Bayesian procedure: *if* one has reason to believe that the model is correct, then (with the right choice of prior over this model) one also has reason to believe that the algorithm performs well.

For the sake of brevity we do not go in more detail into the justification of Bayesian methods. Instead, we proceed with a more in-depth discussion of two methods that have received more attention in the context of the NFL results: empirical risk minimization and cross-validation.

3.3.2 Empirical risk minimization

This is probably the most standard machine learning method. Like Bayesian learning, empirical risk minimization (ERM) is a model-dependent method. The crucial difference with Bayesian learning is that the "model" is now not a set of probability distributions, but rather a user-specified set of classifiers $\mathcal{F}$, usually called a *hypothesis class*. In practice, it could be the set of all neural networks with a given number of nodes and connectivity matrix, represented by their weights; or the set of all decision trees of a given size. The generalization performance of ERM can be analyzed via the standard machinery of learning theory (Shalev-Shwartz and Ben-David 2014). Here, as in Sect. 2.2, one assumes that the data $S = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

[33] As a referee remarked, for a "hard-core" (to use the referee's terminology) subjective Bayesian there is a much more fundamental justification than optimality: that only Bayesian learning is *rational* (made precise in terms of *coherence*, or quantifying uncertainty in terms of degrees of belief that satisfy the probability axioms). For instance, Bishop (2006, p. 21) writes, "The use of probability to represent uncertainty … is inevitable if we are to respect common sense while making rational coherent inferences." From this point of view, performing well is strictly speaking not even an issue: once a prior is formulated, the only coherent and therefore reasonable way to combine it with the data to arrive at a prediction or classification is to follow the Bayesian algorithm.

are sampled independently from an unknown distribution $\mathcal{D}$. No further assumptions about $\mathcal{D}$ are made: instead all inductive assumptions go into $\mathcal{F}$. In Bayesian learning, the choice of model $\mathcal{M}$ can be seen as the inductive assumption that "there is a $P \in \mathcal{M}$ such that acting as if the data is a random sample from P leads to the best possible predictions." In learning theory, adopting a class $\mathcal{F}$ can be seen as the inductive assumption that "there is an $f \in \mathcal{F}$ that has classification risk $\ll 1/2$, small enough to be useful." Here the classification risk is IID risk, or the probability that $f(X) \neq Y$ under $\mathcal{D}$.

The ERM method A_{ERM} takes as input both a training sample S and a hypothesis class $\mathcal{F}$ as above. It proceeds by picking the classifier $\hat{f}_{\text{ERM}} = A_{\text{ERM}}(S, \mathcal{F})$ in $\mathcal{F}$ that made, among all elements in $\mathcal{F}$, the minimum number of errors on S, with some arbitrary rule for breaking ties. Assume for simplicity that $\mathcal{F}$ is finite, so that there exists an f^* in $\mathcal{F}$ that minimizes the risk. A variation of a standard result in learning theory says that ERM works well, in the following sense: the difference between the expected risk of $\hat{f}_{erm}$ and the best obtainable risk within the model, namely that of f^*, is bounded by $\sqrt{|\log \mathcal{F}|/(2n)}$. (See "Appendix A.3" for a derivation.) This result holds no matter what $\mathcal{D}$ is. Since the dependence on the size of $\mathcal{F}$ is logarithmic, the guarantee remains non-void even for exponentially large, and in this sense fairly complex $\mathcal{F}$. In fact, it can be extended to many infinite $\mathcal{F}$ as well: the $\log |\mathcal{F}|$ term is then replaced in the bound by an abstract (but computable) complexity notion such as the Rademacher, Vapnik-Chervonenkis or "PAC-Bayesian" complexity of $\mathcal{F}$ (Grünwald and Mehta 2020). Interestingly, as the latter paper explains in detail, such results are proven using essentially the same techniques as those used for proving non-asymptotic convergence of pragmatic Bayesian learning.

What about anti-ERM (or empirical risk *maximization*), that picks the $\hat{f}_{\text{A- ERM}} \in \mathcal{F}$ with *largest* error on the training set? We can precisely reverse the math behind the convergence of ERM to show that anti-ERM will converge to the *worst* element of $\mathcal{F}$, the element that maximizes risk. The difference between the expected risk of $\hat{f}_{\text{A- ERM}}$ and the worst obtainable risk is again at most $\sqrt{|\log \mathcal{F}|/(2n)}$ if $\mathcal{F}$ is finite, and an analogous result holds again with $\log |\mathcal{F}|$ replaced by Rademacher or VC dimension for infinite $\mathcal{F}$. Saying "ERM has no inherently better justification than anti-ERM" would thus amount to saying: "A method which (given a not too small sample) leads to the best possible predictions that can be obtained based on my hypothesis class, has no more inherent justification than a method which (given a not too small sample) leads to the worst possible predictions that can be obtained based on my hypothesis class." To us, this seems an aberration. [34]

Our point is certainly not that ERM is perfect: if $\mathcal{F}$ becomes "too complex" then ERM may suffer from severe overfitting and will not work in practice.[35] But if anyone advises us to use such a class in combination with ERM, we can simply reply that handling it goes beyond the power of ERM—other methods more suitable for that case exist, such as structural risk minimization (Vapnik 1998; Shalev-Shwartz and

[34] A complication is that the original NFL claim "ERM is no more justified than anti-ERM" is based on measuring classification quality in terms of OTS error, whereas the learning-theoretic claims that ERM is better are based on IID error. In "Appendix A.1" we lift this complication by explaining that in practice, both error measures often essentially coincide.

[35] See "Appendix B.1" for more details about shortcomings of ERM as well as cross-validation.

Ben-David 2014), or forms of minimum description length learning (Grünwald and Roos 2020), or ERM combined with cross-validation as below.

We thus have a well-defined condition (small enough complexity of our $\mathcal{F}$) under which ERM is provably preferable to anti-ERM. No such conditions have ever been formulated under which anti-ERM performs better than ERM (with the same model!), and it is highly implausible that something of the sort could be done.

3.3.3 Cross-validation

This method can be viewed as a meta-algorithm to select *between* different learning algorithms.[36] For ease of presentation, we concentrate on a simplification of cross-validation: *two-fold forward-validation*. This takes as input a data set of a given size $n > 1$, and a finite set of learning algorithms $A_1, A_2, \ldots, A_m$. Forward-validation runs all these algorithms on the first half S_1 of the original training set.[37] Letting $\hat{f}_k = A_k(S_1)$ denote the classifier learned by algorithm A_k, it then selects as final classifier the classifier $\hat{f}_{\hat{k}_{\text{fv}}}$ where $\hat{k}_{\text{fv}}$ is the k such that $\hat{f}_k$ has the smallest error on the second half S_2 of the training set, which is thereby used as a *validation* set. Thus, the final classifier always coincides with one of the m initial classifiers. For full two-fold cross-validation, one repeats the procedure with the two data sets interchanged, and for M-fold cross-validation we split the data in M parts with a validation set of size n/M. Everything we say below for two-fold forward-validation also holds *mutatis mutandis* for full M-fold cross-validation, but the phrasing of results becomes more cumbersome, so we stick to the two-fold forward case for simplicity.

Now, let $\mathcal{E}_k^{(n)}$ be the expected IID risk of algorithm k after having run on the first half of the data: $\mathcal{E}_k^{(n)} = \mathbf{E}_{S\sim\mathcal{D}^n}\left[L_{\mathcal{D}}\left(\hat{f}_k\right)\right]$. Let $\mathcal{E}_{\text{fv}}^{(n)}$ be the expected IID risk of two-fold forward-validation as defined above: $\mathcal{E}_{\text{fv}}^{(n)} = \mathbf{E}_{S\sim\mathcal{D}^n}\left[L_{\mathcal{D}}\left(\hat{f}_{\hat{k}_{\text{fv}}}\right)\right]$. One can now show (see "Appendix A.3") that the expected IID risk of forward-validation satisfies

$$\mathcal{E}_{\text{fv}}^{(n)} \leq \min_{k\in\{1,\ldots,m\}} \mathcal{E}_k^{(n)} + \sqrt{\frac{\log m}{n}}. \tag{3}$$

Thus, the expected IID risk of forward-validation converges, as n grows, to the expected risk of the learning algorithm that, among all algorithms under consideration, is best in the sense that it outputs the lowest-risk classifier in expectation over the training set S_1. This holds for all m and n, so if n is large, we can also take m very large; in particular, due to the logarithmic dependence on m, at sample size n we can choose between a number of learning algorithms m that is orders of magnitudes larger than n and still have a meaningful bound.[38]

[36] Cross-validation is often seen as a highly non-Bayesian method, but there are in fact close connections, as first pointed out by Dawid (1984); also see Fong and Holmes (2020).

[37] If n is odd, we take S_1 to contain the first $(n-1)/2$ data points and S_2 the remaining ones.

[38] For M-fold cross-validation, the constant in front of the square root changes from 1 to another positive value, but remains easily computable as long as M does not depend on n. For leave-one-out cross validation, $M = n - 1$ and the mathematical analysis is tricky and a subject of ongoing research, so we will stick here to the fixed M case.

Forward- and cross-validation can be fruitfully applied both to model-dependent algorithms and to algorithms that may be better viewed as data-only. A prototypical example of the latter is nearest-neighbor classification. Here $\mathcal{X}$ is a space equipped with a metric (e.g., Euclidean space with the Euclidean metric). The k-nearest-neighbor method based on a training set with n' instances plus labels $(x_1, y_1), \ldots, (x_{n'}, y_{n'})$ outputs the classifier which, for any value of x, picks the k data points $\{i_1, i_2, \ldots, i_k\} \subset \{1, \ldots, n'\}$ for which x_i is closest to x, and outputs the majority vote for the corresponding $y_{i_1}, \ldots, y_{i_k}$. Nearest-neighbor with $k = 1$ always has zero error on the training set, so typically overfits dramatically. However, one can use cross- or forward-validation to choose a value of k. The number m_n of k's that make sense at sample size n is at most n, so the generalization bounds above are meaningful, and we have the guarantee that the expected risk based on using $\hat{k}_{\text{fv}}$-nearest-neighbour is close to the error achieved with the unknown optimal $k \in m_n$ that achieves the best expected risk $\mathbf{E}_{S\sim\mathcal{D}^n}\left[L_{\mathcal{D}}(\hat{f}_k)\right]$.

When applying forward- and cross-validation to model-dependent learning algorithms, one typically takes the same learning algorithm (say ERM) for $A_1, \ldots, A_m$, turned into one-place algorithms by combining each A_k with a different hypothesis class $\mathcal{F}_k$. For example, A_k could represent ERM applied to $\mathcal{F}_k$, the set of decision trees of depth k. The class of all decision trees of arbitrary depth is too large for ERM to work well (yield nontrivial generalization guarantees), but in combination with forward- or cross-validation one can use the above result to get meaningful generalization guarantees again.

How about anti-cross-validation? We can invoke precisely the same analysis as for ERM. Our inductive bias is now explicitly specified at a meta-level, by specifying the algorithms A_k. If m, the number of algorithms taken into account, is fixed or grows subexponentially with n, cross-validation can be expected to converge to the best of them based on a finite and quantifiable sample size. In contrast, under the same conditions, anti-cross-validation will converge to the *worst* of them. Analogously to the ERM case, there is a clear condition (m subexponential as function of n) under which cross-validation is (much) better than anti-cross-validation relative to the given algorithms that encode our inductive bias. And again, we cannot imagine a condition that would allow one to prove an interesting guarantee in support of anti-cross-validation.

3.3.4 The consistency with no-free-lunch

To conclude our examples, we note that the model-dependent perspective still encapsulates the valid lesson from NFL results: the lesson that every algorithm, when operating *on data only*, must incorporate an inductive bias. A Bayesian algorithm, *when provided with a model and a prior distribution on this model*, will possess a certain bias; similarly, ERM, *when provided with a hypothesis class*, and cross-validation, *when provided with a set of hypothesis classes*, possess a certain bias. The models here represent an inductive bias, and NFL results show that any such model must indeed be biased in the sense that it must be restrictive. Any algorithm plus instantiated model performs well in some situations: those situations which the inductive bias, in some sense, is well-aligned with. But the algorithm plus this model does not perform well

in other situations, situations even in which the very *same* algorithm, with a *different* instantiated model, would perform well.

To further illustrate the consistency of negative NFL results and positive learning-theoretic results, recall the NFL version of Shalev-Shwartz and Ben-David (2014) that we described in Sect. 2.5. It states that every data-only algorithm (like ERM with any instantiated $\mathcal{F}$) does not perform well in situations $\mathcal{D}$ in which another data-only algorithm does perform well. They prove this by exhibiting a second algorithm that has an $\mathcal{D}$-expected risk at least 1/4 less than the first algorithm; specifically, the second algorithm is ERM with a class $\mathcal{F}'$ that is well (indeed perfectly) aligned with $\mathcal{D}$. Note that if the first algorithm is ERM with some $\mathcal{F}$, then this second $\mathcal{F}'$ *must* be a different class, for any significant difference in expected risk (depending on the sample size). This follows from the learning-theoretic guarantee that the expected risk of ERM cannot be much worse than that of the *best* hypothesis in $\mathcal{F}$, and therefore than that of any algorithm that uses (must select a classifier from) model $\mathcal{F}$. Again, ERM with a particular $\mathcal{F}$ may be much worse than a different data-only algorithm if $\mathcal{F}$ is not a good model. But ERM cannot perform much worse than any algorithm *with the same model*; and *if* we have reason to believe that our model is good, then we have reason to believe that ERM with this model performs well, too.[39]

3.4 The justification for learning algorithms

Learning theory thus provides us with *model-relative* justification for many standard methods. For a generic model-dependent method, such a model-relative justification is all we can ask for. For such a generic method, it simply does not make sense to speculate about empirical assumptions that would render the method *in itself* successful and in that sense justify it. This observation stands in sharp contrast to the reduction of the justification for standard learning methods to some postulate about the right structure of the world. We think that this observation within the domain of machine learning also lends further plausibility to local accounts of induction in philosophy.

One could object, however, that no method is *perfectly* generic, and some assumptions or biases are always inherent to it. To put this point differently, we have used the word "inductive bias" in a relatively narrow way, as only pertaining to the choice of hypothesis class. But one could object that, for instance, the method of ERM (anti-ERM), irrespective of the hypothesis class, embodies a substantive assumption that the evidence so far is not (is) misleading.[40] We agree these can also be called biases, or perhaps rather meta-biases (as they concern extrapolating classifiers' success rather than the data directly); but they are fundamentally linked to assumptions that are already introduced in the formulation of the relevant learning problem, in this

[39] Our discussion here does not yet fully resolve the consistency of the *original* NFL results with learning theory. Namely, according to these results, under a uniform prior over learning situations, any two data-only algorithms (including ERM and anti-ERM *with the same hypothesis class*) are equally good. In fact, the results also imply a variant of (*) where we drop the problematic uniformity assumption, namely that for any two algorithms (again, including ERM and anti-ERM with the same $\mathcal{F}$) there exists a $\mathcal{D}$ where the second does better than the first. In "Appendix A" we explain how these results can be consistent with the positive guarantees from learning theory.

[40] We thank one of the reviewers for pressing us on this point.

case the general problem of stochastic classification.[41] In particular, the use of ERM relies (and learning-theoretic guarantees for ERM rely) on the problem assumption of stochastic classification that data is sampled i.i.d. (this can be extended to a stationarity assumption but not much beyond). For this learning problem, and in particular due to the i.i.d. assumption, the "uniformity meta-bias" of ERM is provably good, and the "anti-uniformity meta-bias" of anti-ERM is provably not. In general, in the same way that any NFL statement concerns a certain learning problem (recall Sect. 2.5), any learning guarantee concerns a certain learning problem. Thus our claim is more precisely that many standard learning methods, also *relative to* the learning problem they were designed for, have a model-relative justification.

Finally, recall from our discussion in Sect. 3.2 that it is far from clear that a local conception of induction brings us closer to an absolute, *global* justification of inductive inferences. Similarly, a model-relative justification still leaves the justification for the model in any particular application of a learning algorithm, and indeed the further assumptions encoded in the very formulation of the learning problem. A global justification for the conclusions of a machine learning algorithm must also include the justification for all these assumptions. The obvious threat is an endless justificatory regress, where the motivation for these local assumptions leads us to an earlier inference that itself relies on inductive assumptions that want justification. Note, though, that this regress will soon, if not immediately, lead us to assumptions that we have not actually arrived at by machine learning methods. We will soon have left the domain of machine learning, and face the problem of induction in its full generality. Rather, therefore, than understanding the NFL theorems as somehow deepening Humean skepticism, the more sober conclusion is that the question of the global justification for the conclusions of machine learning algorithms reduces to the original problem of induction.

4 Conclusion

The NFL theorems are commonly understood to show that every learning algorithm must possess a certain bias, and must ultimately lack justification because any such bias must. We have argued that for many standard learning algorithms, this is turning things on their head. NFL results do show that any *data-only* algorithm must have an inherent bias. Presented such an algorithm, we could expose its bias, and question the justification for this bias and thereby for the algorithm. However, many standard learning algorithms are better conceived of as *model-dependent*. The need for a choice of bias is accommodated by such an algorithm from the start: on each application, one must equip it with a particular model, that represents the bias. The algorithm *itself* is generic in that it does not itself come with a bias: on each application, one must provide it with one. What is more, such algorithms can have a *model-relative* justification: relative to any given model, such an algorithm performs well. Learning-theoretic

[41] Casting a particular real-world problem as a particular formal learning problem (which includes, e.g., choice of possible instances and labels) is, of course, itself a modeling step, that can be said to introduce certain biases. See von Luxburg and Schölkopf (2011, p. 683ff) for a short discussion of the different places bias can enter.

guarantees show that in that sense some standard learning algorithms *are* sensible, *are* justified—and other possible algorithms are not. This is perfectly consistent with the valid lesson of NFL results that any data-only learning method, including a model-dependent algorithm *plus* a particular choice of model, must possess a bias.

In the course of our argument, we drew some parallels to the broader philosophy of induction. Most importantly, we discussed the role of a general postulate on the induction-friendliness of the world, and the local view of induction that challenges the cogency of such a postulate. We think of our emphasis on the model-dependence of many standard learning algorithms as an instance of the local view of induction. It is important to note, however, that the local view does not yet suffice to escape Hume's skeptical argument, and neither does the model-relative conception in the context of machine learning algorithms. Namely, an *absolute* justification for the conclusions of inductive inferences still requires a justification for the preceding choice of local assumptions or model.

For that reason, the local view of induction also does not suffice to fully explain the success of our inductive inferences. Analogously, Wolpert (1996b, p. 1364) points at "a rather profound (if somewhat philosophical) paradox," that is not yet resolved by the model-dependent perspective on learning algorithms: "How is it that we perform inference so well in practice, given the NFL theorems and the limited scope of our prior knowledge?" That this is not merely a "somewhat philosophical" issue is demonstrated, for instance, by the recent debate surrounding the "paradox of deep learning" (Zhang et al. 2017; Neyshabur et al. 2017; Arpit et al. 2017; Kawaguchi et al. 2019), which revolves around the perceived lack of a good explanation for the empirical success of deep neural networks. The case of deep learning is particularly interesting, as a clean separation of method and model is here much more contentious, and the remaining question of justification does not clearly center on the motivation for a well-articulated choice of model.

Funding Open Access funding enabled and organized by Projekt DEAL. Tom Sterkenburg was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Projektnummern 437206810, *Die Epistemologie der Statistischen Lerntheorie*; 432308570, *Grundlagen, Anwendungen & Theorie der Induktiven Logik*. Peter Grünwald was supported by the Dutch Research Council (NWO) via research programme 617.001.651, *Safe Bayesian Learning*.

Appendix A. The original NFL results and learning theory

The central question in this paper concerned the consistency of the negative NFL results with positive results from learning theory. In Sect. 3.3.4, we explained in formal detail how in the specific case of ERM, an NFL statement is consistent with a learning-

theoretic guarantee for this algorithm. However, we there discussed the NFL variant due to Shalev-Shwartz and Ben-David (2014), a version of statement (*), and not the *original* Wolpert-Schaffer result. The original case still leaves something to explain.

Consider a finite hypothesis class $\mathcal{F}$, and two different learning algorithms: ERM, turned into a data-only algorithm by equipping it with hypothesis class $\mathcal{F}$ (write $A_{\text{ERM}}(\mathcal{F})$) and *anti*-ERM with the same hypothesis class (selecting, for any training sample S, an $f \in \mathcal{F}$ with the *worst* empirical error on S; write $A_{\text{A- ERM}}(\mathcal{F})$). The Wolpert-Schaffer results tell us that, under a uniform prior over learning situations, for any sample size n, both $A_{\text{ERM}}(\mathcal{F})$ and $A_{\text{A- ERM}}(\mathcal{F})$ have the same expected OTS risk of 1/2. Yet the learning theoretic guarantee mentioned in the main text and proved below says that the expected IID risk of $A_{\text{ERM}}(\mathcal{F})$ is not more than $\min_{f^* \in \mathcal{F}} L_{\mathcal{D}}(f^*) + \sqrt{|\log \mathcal{F}|/(2n)}$, whereas the expected IID risk of $A_{\text{A- ERM}}(\mathcal{F})$ is not less than $\max_{f_* \in \mathcal{F}} L_{\mathcal{D}}(f_*) - \sqrt{|\log \mathcal{F}|/(2n)}$.

This presents us with something of a paradox: it appears that $A_{\text{ERM}}(\mathcal{F})$ and $A_{\text{A- ERM}}(\mathcal{F})$ behave equally well under the uniform prior on learning situations (NFL result), whereas one behaves much better than the other (learning theory) under arbitrary priors, including the uniform. Similar paradoxes have been noted before, and it has been suggested that the reason the former negative result and the latter positive result can exist together is that the latter and not the former relies on IID error (Wolpert 1996a, p. 1368f). But this contradicts Roos et al. (2006), who show that in many realistic learning situations the IID and OTS error are very close to each other.

Here, we show how to resolve the paradox and reconcile both results in situations in which both types of errors are essentially equally large. In A.1 we discuss OTS and IID risk and in A.2 we explain how the paradox is resolved. In A.3 we provide a short proof of the relevant learning-theoretic bounds.

A.1. OTS and IID risk

First recall that both the NFL and the learning theory setting assume that data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. $\sim \mathcal{D}$ where $\mathcal{D}$ is some distribution on instances in $\mathcal{X}$ and labels in $\mathcal{Y} = \{0, 1\}$. For given $\mathcal{D}$, let $\mathcal{D}_X$ denote the marginal distribution on $\mathcal{X}$ and $\mathcal{D}_{Y|X}$ the conditional distribution for Y given X (that is, for each $x \in \mathcal{X}$, $\mathcal{D}(Y \mid X = x)$ is a distribution on $\mathcal{Y}$). With this notation we have that $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim \mathcal{D}_X$ (we simply write $X^n \sim \mathcal{D}_X$, leaving independence implicit and abbreviating $(X_1, \ldots, X_n)$ as X^n). Similarly, given $X_1, \ldots, X_n$, we have that $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \mathcal{D}_{Y|X} \mid X_i$ (we simply write $Y^n \sim \mathcal{D}_{Y|X} \mid X^n$).

Based on an analysis of $\mathcal{D}_X$, Roos et al. (2006) show that in practically realistic settings, IID and OTS risk are often (though not always) extremely close, and analyses pertaining to the one transfer to the other. To get a first idea of why this might be so, assume that the instance space $\mathcal{X}$ is continuous and $\mathcal{D}_X$ has a probability density on $\mathcal{X}$. Then, for $i \neq j$, $\mathcal{D}_X(X_i = X_j) = 0$: the probability that we see the same feature vector twice is 0, and more generally, for any finite sample S, the probability that an independently sampled test instance $X \sim D_X$ is contained in S is also 0. Since both IID and OTS risk involve an expectation over such an independent X, this implies that for

continuous $\mathcal{X}$, we must have that IID and OTS risk coincide: $L_{\mathcal{D}\setminus S}(A(S)) = L_{\mathcal{D}}(A(S))$, almost-surely under $\mathcal{D}$.

Now the original NFL results require $\mathcal{X}$ to be finite or countable, so the above reasoning does not necessarily hold. But if $\mathcal{X}$ is finite but not too small, and equipped with a uniform distribution $\mathcal{D}_X$, then the situation is quite similar to the continuous setting (the probability that an X contained in S is also contained in a test set is *almost* zero) and we can still show that, at sample sizes of interest, the difference between both quantities is negligible. For example, suppose that features are m-dimensional binary vectors, $\mathcal{X} = \{0, 1\}^m$, $\mathcal{D}_X = \mathcal{U}_X$ is uniform, and the sample size is n. Then lemma 1 of Roos et al. (2006) gives that, under every distribution $\mathcal{D}_{Y|X}$,

$$\left|\mathbf{E}_{X^n\sim\mathcal{U}_X, Y^n\sim\mathcal{D}_{Y|X}|X^n}\left[L_{\mathcal{D}\setminus S}(A(S)) - L_{\mathcal{D}}(A(S))\right]\right| \leq n2^{-m}. \tag{4}$$

So, for example, if $m \geq 40$ (40 covariates being much less than what is common in modern machine learning) and sample size n is less than 10^6 (as is the case in many realistic scenarios), then the difference between the expected behaviour, over training samples, of the two risk measures is less than approximately 10^{-6}, which is completely negligible for practical purposes. It is true that in practice, $\mathcal{D}_X$ will usually not be uniform. But Roos et al. (2006) show that even for highly nonuniform $\mathcal{D}_X$, both error measures are often very close—the closeness can even be estimated from the obtained sample S.

So in those (realistic) situations where IID and OTS essentially coincide, how do we account for the co-existence of the NFL results and positive learning guarantees?

A.2. Consistency of the Wolpert–Schaffer results with learning theory

We first restate the NFL result in a fully precise manner. Assume that $\mathcal{X}$ is finite or countable, and recall that $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$. Then for every distribution $\mathcal{D}_X$ on $\mathcal{X}$,

$$\mathbf{E}_{\mathcal{D}_{Y|X}\sim\mathcal{U}}\,\mathbf{E}_{X^n\sim\mathcal{D}_X, Y^n\sim\mathcal{D}_{Y|X}|X^n}\left[L_{\mathcal{D}\setminus S}(A(S))\right] = 1/2, \tag{5}$$

where $\mathcal{U}$ is the uniform distribution on conditional distributions $\mathcal{D}_{Y|X}$. To be clear, this is the distribution such that conditional on each $x \in \mathcal{X}$, $p_{|x} := \mathcal{D}(Y = 1 \mid X = x)$ has a uniform distribution on the unit interval [0, 1]. To see that this is a natural definition of 'uniform' in this context, note that $p_{|x}$ indicates the mean of Y given x according to $\mathcal{D}$, and $\mathcal{U}$ is thus also uniform on the mean.

In contrast to (5), we derive below that for every distribution $\mathcal{D}_X$, and for every distribution $\mathcal{U}'$ on $\mathcal{D}_{Y|X}$ (including the uniform NFL distribution $\mathcal{U}$ defined above),

$$\mathbf{E}_{\mathcal{D}_{Y|X}\sim\mathcal{U}'}\,\mathbf{E}_{X^n\sim\mathcal{D}_X,Y^n\sim\mathcal{D}_{Y|X}|X^n}\left[L_{\mathcal{D}}(A_{\mathrm{ERM}}(\mathcal{F},S))-\min_{f^*\in\mathcal{F}}L_{\mathcal{D}}(f^*)\right]\leq\sqrt{\frac{\log|\mathcal{F}|}{2n}} \tag{6}$$

$$\mathbf{E}_{\mathcal{D}_{Y|X}\sim\mathcal{U}'}\,\mathbf{E}_{X^n\sim\mathcal{D}_X,Y^n\sim\mathcal{D}_{Y|X}|X^n}\left[\max_{f^\circ\in\mathcal{F}}L_{\mathcal{D}}(f^\circ)-L_{\mathcal{D}}(A_{\text{A- ERM}}(\mathcal{F},S)\right]\leq\sqrt{\frac{\log|\mathcal{F}|}{2n}}. \tag{7}$$

Since both the NFL and learning theory results hold for every distribution on finite or countable $\mathcal{X}$, we get an instance of our paradox if we (a) take a distribution $\mathcal{D}_X$ for which IID and OTS error differ by a negligible amount, say δ very close to 0, at the given sample size n, and (b) a combination of $\mathcal{F}$ and n for which $\sqrt{(\log|\mathcal{F}|)/2n}$ is very close, say ϵ, to 0, so that the bounds (6) and (7) are non-void. We henceforth call any combination $(\mathcal{D}_X,|\mathcal{F}|,n)$ for which both (a) and (b) are the case an (ϵ,δ)-*paradoxical learning situation*, with the understanding that the closer ϵ and δ to 0, the more paradoxical.

To be in an (ϵ,δ)-paradoxical situation, we see that it is sufficient, for any finite $\mathcal{F}$, to take n sufficiently large, and, from (4), to take $\mathcal{X}$ finite and $\mathcal{D}_X$ uniform, with m and n so that $\delta=n2^{-m}$ is negligibly small. We thus already know that such situations exist, and the result of Roos et al. (2006) implies that they exist for much more general $\mathcal{D}_X$ as well. Note that only the size of $\mathcal{F}$, not the definitions of its constituent f's, is relevant to determine whether the situation is paradoxical.

How can we reconcile (5), (6) and (7) in paradoxical learning situations? The key observation is that there would only be a real contradiction if the optimal classifier $f^*\in\mathcal{F}$ in our class and the worst classifier $f^\circ\in\mathcal{F}$ in our class differed substantially in terms of their risk $L_{\mathcal{D}}$. Thus, rather than being contradictory, taken together, (5), (6) and (7) simply imply that, in (ϵ,δ)-paradoxical situations, under the NFL prior, *no matter what $\mathcal{F}$ of given size we chose*, we expect both the optimal classifier $f^*\in\mathcal{F}$ and the worst classifier $f^\circ\in\mathcal{F}$ to have both IID and OTS error within about $\epsilon(1+\delta)$ of $1/2$. Thus, in the paradoxical cases in which ϵ,δ are very small, both f^* and f° behave essentially no better or worse than random guessing. As a consequence, in paradoxical situations, under the NFL prior, we expect $\mathcal{F}$, no matter how we chose it, to be essentially useless: it contains no useful (classification error substantially smaller than $1/2$) f, hence "there is nothing to be learned from $\mathcal{F}$." This reinforces the point made in Sect. 2.4 that the NFL prior is a prior under which learning is impossible: at least in paradoxical situations, it makes it next to impossible that any $\mathcal{F}$ we might choose provides us with anything to learn.

Next we move to a variation of the paradox: consider the instance of NFL statement (*) that is implied by the Wolpert-Schaffer result, saying that for any two data-only algorithms, there is a learning situation $\mathcal{D}$ such that the first algorithm has expected OTS risk at least 1/2, while the second has expected OTS risk strictly below 1/2. Again, this may appear paradoxical if we apply this result to $A_{\mathrm{ERM}}(\mathcal{F})$ and $A_{\text{A- ERM}}(\mathcal{F})$, but now the paradox may seem more serious because it does not refer to the NFL prior. But again, the statement can be reconciled with the learning-theoretic guarantees (6) and (7). Rather than contradicting (*), in conjunction with (*) and (4), they imply that

(1) there do exist (ϵ, δ)-paradoxical learning situations with very small ϵ and δ in which both the IID and OTS error of $A_{\text{A- ERM}}(\mathcal{F})$ is smaller than $1/2$ while both IID and OTS error of $A_{\text{ERM}}(\mathcal{F})$ are larger than $1/2$ ("anti-ERM is better than random guessing, ERM is worse"), ...

(2) ... yet in such situations $\mathcal{F}$ must be essentially useless in the sense that both IID and OTS errors of its best element f^* and its worst element f° are within about ϵ of $1/2$: the *only* (ϵ, δ)-paradoxical learning situations in which anti-ERM is better than random guessing and outperforms ERM must concern hypothesis classes $\mathcal{F}$ that only contain f that are themselves at most ϵ-better than random guessing (and then anti-ERM is itself also at most ϵ better than random guessing).

Viewed in this way, the paradoxical situations become much less paradoxical.

A.3. Derivation of expected risk bounds

We now give a compact derivation of the following result: *for all distributions $\mathcal{D}_X$ on $\mathcal{X}$, all conditional distributions $\mathcal{D}_{Y|X}$ on $\mathcal{Y}$ given X, with $\mathcal{D}$ denoting the corresponding joint distribution on $\mathcal{X} \times \mathcal{Y}$ and $S = ((X_1, Y_1), \ldots, (X_n, Y_n))$, we have:*

$$\mathbf{E}_{S\sim\mathcal{D}}\left[L_{\mathcal{D}}(A_{\text{ERM}}(\mathcal{F}, S)) - \min_{f^*\in\mathcal{F}} L_{\mathcal{D}}(f^*)\right]$$
$$= \mathbf{E}_{X^n\sim\mathcal{D}_X, Y^n\sim\mathcal{D}_{Y|X}|X^n}\left[L_{\mathcal{D}}(A_{\text{ERM}}(\mathcal{F}, S)) - \min_{f^*\in\mathcal{F}} L_{\mathcal{D}}(f^*)\right] \leq \sqrt{\frac{\log|\mathcal{F}|}{2n}}. \quad (8)$$

Since (8) holds for all conditional distributions $\mathcal{D}_{Y|X}$, it must also hold in expectation over every prior distribution $\mathcal{U}'$ on $\mathcal{D}_{Y|X}$, so that (6) follows. Then (7) follows by repeating all the steps of the proof below with obvious modifications. The result for two-fold forward-validation in the main text follows from (8) as follows. First, let $\mathcal{F} = \{\hat{f}_1, \ldots, \hat{f}_m\}$ be the m classifiers that were output by the m algorithms $A_1, \ldots, A_m$ on S_1, the first half of the sample. Now, use the result above with S_2 in the role of S, conditional on S_1. Then $\mathcal{F}$ is fixed, and n gets replaced by $n/2$, and the result follows by further using that the expectation (over S_1) of a minimum is no larger than the minimum of the expectation.

Proof Denote for each fixed classifier $f \in \mathcal{F}$, the loss it makes on the ith outcome by $\ell_f(X_i, Y_i)$ and one minus this loss as $\ell'_f(X_i, Y_i)$. Then $\ell'_f(X_1, Y_1), \ell'_f(X_2, Y_2), \ldots$ is i.i.d. Bernoulli. Let (X, Y) be another i.i.d. copy of X_1, Y_1 (think of it as a "test" example). By Hoeffding's inequality (Shalev-Shwartz and Ben-David 2014, p. 56) we have, for all $\eta > 0$, that $\mathbf{E}\left[\exp(\eta \sum_{i=1}^n (\ell'_f(X_i, Y_i) - \mathbf{E}[\ell'_f(X, Y)]))\right] \leq \exp(\eta^2 n/8)$ so that also, for any learning algorithm A that outputs, upon seeing sample $S = (X_1, Y_1), \ldots, (X_n, Y_n)$, a classifier $A(S) \in \mathcal{F}$, we have

$$\begin{aligned}
&\mathbf{E}\left[\exp(\eta\sum_{i=1}^{n}(\ell'_{A(S)}(X_i,Y_i)-\mathbf{E}[\ell'_{A(S)}(X,Y)]))\right]\\
&\quad\leq\mathbf{E}\left[\exp(\max_{f\in\mathcal{F}}\eta\cdot\sum_{i=1}^{n}(\ell'_f(X_i,Y_i)-\mathbf{E}[\ell'_f(X,Y)]))\right]\\
&\quad=\max_{f\in\mathcal{F}}\mathbf{E}\left[\exp(\eta\sum_{i=1}^{n}(\ell'_f(X_i,Y_i)-\mathbf{E}[\ell'_f(X,Y)]))\right]\\
&\quad\leq\sum_{f\in\mathcal{F}}\mathbf{E}\left[\exp(\eta\sum_{i=1}^{n}(\ell'_f(X_i,Y_i)-\mathbf{E}[\ell'_f(X,Y)]))\right]\leq\exp(\log|\mathcal{F}|+\eta^2n/8).
\end{aligned}$$

Jensen's inequality, division by ηn, using that the result holds for all $\eta > 0$, and differentiation, now give that:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\ell'_{A(S)}(X_i,Y_i)-\mathbf{E}[\ell'_{A(S)}(X,Y)])\right]\leq\min_{\eta>0}\left\{\frac{\eta}{8}+\frac{\log|\mathcal{F}|}{\eta n}\right\}=\sqrt{\frac{\log|\mathcal{F}|}{2n}}.$$

If we take $A(S) = A_{\mathrm{ERM}}(\mathcal{F}, S)$ to be an instance of ERM applied to $\mathcal{F}$, and replace $\ell = 1 - \ell'$, this gives

$$\begin{aligned}
&\mathbf{E}_{S\sim\mathcal{D}}[\mathbf{E}_{(X,Y)\sim\mathcal{D}}[\ell_{A_{\mathrm{ERM}}(\mathcal{F},S)}(X,Y)]-\mathbf{E}_{(X,Y)\sim\mathcal{D}}[\ell_{f^*}(X,Y)]]\leq\\
&\mathbf{E}_{S\sim\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_{A_{\mathrm{ERM}}(\mathcal{F},S)}(X_i,Y_i)\right]-\mathbf{E}_{(X,Y)\sim\mathcal{D}}[\ell_{f^*}(X,Y)]+\sqrt{\frac{\log|\mathcal{F}|}{2n}}\\
&\quad=\mathbf{E}_{S\sim\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_{A_{\mathrm{ERM}}(\mathcal{F},S)}(X_i,Y_i)-\frac{1}{n}\sum_{i=1}^{n}\ell_{f^*}(X_i,Y_i)]+\sqrt{\frac{\log|\mathcal{F}|}{2n}}\right]\leq\sqrt{\frac{\log|\mathcal{F}|}{2n}},
\end{aligned}$$

where we used that $\ell_{f*}(X_i, Y_i)$ is an i.i.d. random variable, and the fact that ERM's risk on the training data can by definition not be larger than that of f^*. This shows (8). □

Appendix B. Model-dependent algorithms: some details and nuances

Here we discuss some limitations of standard model-relative learning algorithms (B.1), and the existence of learning algorithms that are not clearly model-dependent (B.2).

B.1. Even good learning algorithms have limitations

Our point in Sect. 3.3 is emphatically not that (pragmatic) Bayes, ERM or cross-validation are perfect methods.

In the case of ERM, we already mentioned that it is not suitable for $\mathcal{F}$ that are very complex or large relative to the given training set size n. More generally, in this

regime it can sometimes behave in surprisingly bad ways, with ERM's risk temporarily increasing in the small n-regime (Loog et al. 2019). Additionally, even if the goal is to learn a classifier with small classification $(0/1)$-error, ERM often delivers better results if applied on the training set with a different, *proxy* loss function that has nicer mathematical properties. This includes logistic and hinge loss, which avoid the discontinuites of the $0/1$-loss.

As to cross-validation, it can perform poorly when the bound stated above does not hold, i.e., if the number of constituent algorithms grows exponentially in n. This happens, for example, in variable selection problems. For these, *penalized* ERM approaches such as the Lasso are more suitable (Hastie et al. 2009). Moreover, if we adopt different loss functions, such as the squared error loss, typical in regression, or the logarithmic loss in density estimation, the picture changes completely. Whether or not cross-validation is the right approach can then depend to some extent on the goal of the inference (Yang 2005; van Erven et al. 2012): does one want to learn the best squared-error predictor (cross-validation, behaving like AIC, is often suitable), or does one want to find out which components of the vector X are correlated with Y (cross-validation is not suitable)?

There are also inherent differences in the type of assumption codified into a probabilistic model $\mathcal{M}$, as in Bayesian machine learning, or a class $\mathcal{F}$ as in ERM, which further point to necessary conditions for the algorithms to work well. For example, if the employed probabilistic model is wrong-but-useful (that is to say, if it contains a distribution that leads to good predictions for the prediction task of interest), the inductive assumption underlying pragmatic Bayesian inference does not hold and in some cases Bayesian inference does fail dramatically in practice, both in regression and classification (Grünwald and van Ommen 2017; Grünwald and Langford 2007). Specifically, in such cases it can happen that no matter how many data are observed, the Bayesian posterior never concentrates around the best-predicting distribution in the model. In contrast, the inductive assumption underlying ERM is of a much more agnostic type, in which hypotheses are not assumed to be "true" but just "useful" (have small classification error), so by construction there can be no problems with "wrong-but-useful" models.

Indeed (viz. the references above), one of us has published extensively on the limitations and suitable domain of application of methods such as pragmatic Bayes, ERM and cross-validation. But acknowledging that these methods have limitations is very different from saying they lack any inherent justification. In fact, studying their strengths and limitations is done with, and shows the usefulness of, machine learning theory. An NFL claim that "all algorithms are equally (un)justified in principle" seems to preclude such study.

B.2. Algorithms that are not clearly model-dependent

Some popular learning algorithms that appear to be data-only can be understood as being model-dependent with a particular model already filled in. A prototypical example are support vector machines (SVM's, Vapnik 1998) with a particular choice of *kernel*. While commonly viewed as "machines" that take merely data as input, they

can be recast as a model-dependent algorithm: the model is a hypothesis class $\mathcal{F}$ of linear combinations of basis functions $\phi_1(X), \phi_2(X), \ldots$ of the instances, with the ϕ_j implicitly given by the kernel. The learning algorithm is penalized ERM with the hinge loss, a "proxy" for the 0/1-loss: it picks the $f \in \mathcal{F}$ that minimizes the hinge loss on the training sample with a penalty for the L_1-norm of the parameters (Hastie et al. 2009).

For other data-only learning algorithms, however, such a decomposition is more tenuous. For example, it is not immediately clear whether one can recast the nearest-neighbor algorithm as such. (Though see von Luxburg and Schölkopf 2011, p. 666 for a discussion of k-nearest neighbor as working with a certain function class determined by the parameter k.) We stress that this does not contradict our main point: we merely state that *many* (that is to say, *not all!*) often-used learning methods are inherently model-dependent, and for those, claims that they work well should be understood relative to the given models.

Finally, for some instances of ERM, though clearly a model-dependent procedure, an additional complication arises. What we have in mind here is deep learning, where $\mathcal{F}$ is the set of all neural networks with a given structure, parameterized by their weights. The standard learning algorithm in this setting is stochastic gradient descent (SGD), which is usually iterated until the error on the training set is zero. This makes SGD an instance of ERM, but it is a very special one. Neural nets typically allow for a multitude (billions) of different local minima in weight space, all with empirical error zero, but SGD directs learning towards very particular minima. For example, these minima tend to be "broad" (small perturbations of weights do not cause a noticeable change in predictions) or equivalently, the weights can be grossly discretized and hence the description of the network shortened without sacrificing accuracy (Dziugaite and Roy 2017). Moreover, the found weight vector tends to be small under a particular, nonstandard norm on vector spaces (Bartlett et al. 2017; Neyshabur et al. 2015).

Thus, in the use of ERM-by-SGD with a neural network model there is a complicated interaction between the learning algorithm and the model: the same model trained with different instances of ERM that end up in very different minima might lead to very different generalization behaviour. Moreover, the models typically have so many weights that they can represent basically any continuous function from input to output, so they are much too large or complex to represent our inductive bias. Rather, it seems that in those instances of learning problems on which deep learning works so well, SGD has a tendency to find a solution in a particularly "simple" (small norm, broad minima, compressible) subset of weight space. There is therefore an interplay between the learning algorithm (SGD), the "true" inductive bias (namely, those problems in which deep learning works well) and "effective" model complexity (for such problems, SGD only explores a tiny fraction of weight space making the model much less complex).

In sum, in deep learning, there is no crisp separation between inductive bias and learning algorithm. We stress again that this does not invalidate our main point: for algorithms like cross-validation, there *is* a clear separation, and quality assessments should be done in a model-relative way.

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., & Lacoste-Julien, S. (2017). A closer look at memorization in deep networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning (ICML 2017). Proceedings of Machine Learning Research* (Vol. 70, pp. 233–242).

Barnard, E. (2011). Determination and the no-free-lunch paradox. *Neural Computation*, *23*(7), 1899–1909.

Bartlett, P., Foster, D. J., & Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (eds.) *Advances in neural information processing systems 30: Annual conference on neural information processing systems (NIPS 2017)* (pp. 6240–6249).

Belot, G. (2021). Absolutely no free lunches! *Theoretical Computer Science*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning. Information Sciences and Statistics*. Berlin: Springer.

Bogdan, R. J. (Ed.). (1976). *Local Induction*. Synthese Library (Vol. 93). Dordrecht, The Netherlands: D. Reidel.

Boole, G. (1854). *An investigation of the laws of thought*. London: Macmillan.

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, & G. Rätsch (eds.), *Advanced lectures on machine learning, ML summer schools 2003 , volume 3176 of lecture notes in artificial intelligence* (pp. 169–207). Springer.

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

Corfield, D. (2010). Varieties of justification in machine learning. *Minds and Machines*, *20*(2), 291–301.

Dawid, A. P. (1984). Present position and potential developments: Some personal views. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Dietterich, T. G. (1989). Limitations on inductive learning. In A. M. Segre (ed.), *Proceedings of the sixth international workshop on machine learning (ML 1989)*, San Mateo, CA, USA (pp. 124–128). Morgan Kaufmann.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Dziugaite, G. K. & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In: *Proceedings of the 33rd conference on uncertainty in artificial intelligence (UAI)*.

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496.

Forster, M. R. (1999). How do simple rules 'fit to reality' in a complex world? *Minds and Machines*, *9*, 543–564.

Gabbay, D. M., Hartmann, S., & Woods, J. (Eds.). (2011). *Inductive Logic. Handbook of the History of Logic (Vol. 10)*. Amsterdam: Elsevier North Holland.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, *28*(2), 500–531.

Ghosal, S., Lember, J., & van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, *2*, 63–89.

Giraud-Carrier, C., & Provost, F. (2005). Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper? In *Proceedings of the workshop on meta-learning, 22nd international machine learning conference (ICML 2005)* (pp. 9–16).

Goodman, N. (1954). *Fact, fiction, and forecast*. London: The Athlone Press.

Grünwald, P. D., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*(2–3), 119–149.

Grünwald, P. D., & Mehta, N. A. (2020). Fast rates for general unbounded loss functions: From ERM to generalized Bayes. *Journal of Machine Learning Research*, *21*, 1–80.

Grünwald, P. D., & van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, *12*(4), 1069–1103.

Grünwald, P. D., & Roos, T. (2020). Minimum description length revisited. *International Journal of Mathematics for Industry*, *11*(1), 1930001.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction. Springer series in statistics* (2nd ed.). New York, NY: Springer.

Henderson, L. (2020). The problem of induction. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Ho, Y.-C., & Pepyne, D. L. (2002). Simple explanation of the no-free-lunch theorem and its implications. *Journal of Optimization Theory and Applications*, *115*(3), 549–570.
Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2019). Generalization in deep learning. Forthcoming as a book chapter in *Mathematics of Deep Learning*. Cambridge University Press. https://arxiv.org/abs/1710.05468.
Kelly, T. (2010). Norton, Hume, and induction without rules. *Philosophy of Science*, *77*(5), 754–764.
Lange, M. (2002). Okasha on inductive skepticism. *The Philosophical Quarterly*, *52*(207), 226–232.
Lange, M. (2004). Would "direct" realism resolve the classical problem of induction? *Noûs, 38*(2), 197–232.
Lange, M. (2011). Hume and the problem of induction. In Gabbay et al. (2011) (pp. 43–91).
Lattimore, T. & Hutter, M. (2013). No free lunch versus Occam's razor in supervised learning. In D. L. Dowe (eds.), *Proceedings of the Solomonoff memorial conference, volume 7070 of lecture notes in artificial intelligence* (pp. 223–235). Springer.
Levi, I. (1967). *Gambling with truth: An essay on induction and the aims of science*. New York, NY: Knopf.
Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.
Loog, M., Viering, T., & Mey, A. (2019). Minimizers of the empirical risk and risk monotonicity. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems (NeurIPS 2019)* (pp. 7478–7487).
McCaskey, J. P. (2021). Reviving material theories of induction. *Studies in History and Philosophy of Science Part A, 83*, 1–7.
Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Technical report CMB-TR-117, Department of Computer Science, Rutgers University.
Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Neyshabur, B., Tomioka, R., & Srebro, N. (2015). Norm-based capacity control in neural networks. In P. D. Grünwald, E. Hazan, & S. Kale (eds.), *Proceedings of The 28th conference on learning theory (COLT 2015), volume 40 of JMLR workshop and conference proceedings* (pp. 1376–1401).
Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (eds.), *Proceedings of the 30th international conference on neural information processing systems (NIPS 2017)* (pp. 5949–5958).
Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647–670.
Norton, J. D. (2010). There are no universal rules for induction. *Philosophy of Science*, *77*(5), 765–777.
Norton, J. D. (2014). A material dissolution of the problem of induction. *Synthese*, *191*(4), 671–690.
Okasha, S. (2001). What did Hume really show about induction? *The Philosophical Quarterly*, *51*(204), 307–327.
Okasha, S. (2005a). Bayesianism and the traditional problem of induction. *Croatian Journal of Philosophy*, *5*(14), 181–194.
Okasha, S. (2005b). Does Hume's argument against induction rest on a quantifier-shift fallacy? *Proceedings of the Aristotelian Society*, *105*(1), 237–255.
Ortner, R., & Leitgeb, H. (2011). Mechanizing induction. In Gabbay et al. (pp. 719–772).
Peirce, C. S. (1878). The order of nature. *Popular Science Monthly*, *8*, 203–217.
Peirce, C. S. (1902). Uniformity. In J. M. Baldwin (Ed.), *Dictionary of philosophy and psychology* (Vol. 2, pp. 727–731). New York, NY: Macmillan.
Putnam, H. (1963). 'Degree of confirmation' and inductive logic. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 761–783). LaSalle, IL: Open Court.
Putnam, H. (1981). *Reason, truth, and history*. Cambridge: Cambridge University Press.
Putnam, H. (1987). *The many faces of realism*. LaSalle, IL: Open Court.
Rao, R. B., Gordon, D., & Spears, W. (1995). For every generalization action, is there really an equal and opposite reaction? Analysis of the conservation law for generalization performance. In A. Prieditis & S. Russell (eds.), *Proceedings of the 12th international conference on machine learning (ICML 1995), San Francisco, CA* (pp. 471–479) Morgan Kaufmann.
Roos, T., Grünwald, P. D., Myllymäki, P., & Tirri, H. (2006). Generalization to unseen cases. In Y. Weiss, B. Schölkopf, & J. C. Platt (eds.), *Proceedings of the 18th international conference on neural information processing systems, NIPS 2005* (pp. 1129–1136). MIT Press.
Rosenkrantz, R. D. (1982). Does the philosophy of induction rest on a mistake? *Journal of Philosophy*, *79*(2), 78–97.
Russell, S. (1991). Inductive learning by machines. *Philosophical Studies*, *64*(1), 37–64.

Salmon, W. C. (1953). The uniformity of nature. *Philosophy and Phenomenological Research*, *14*(1), 39–48.
Schaffer, C. (1994). A conservation law for generalization performance. In W. W. Cohen & H. Hirsch (eds.), *Proceedings of the 11th international conference on machine learning (ICML 1994)* (pp. 259–265) San Francisco, CA: Morgan Kaufmann.
Schurz, G. (2017). No free lunch theorem, inductive skepticism, and the optimality of meta-induction. *Philosophy of Science*, *84*(4), 825–839.
Schurz, G. (2021). The no free lunch theorem: Bad news for (White's account of) the problem of induction. *Episteme, 18(1)*, 31–45.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press.
Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic*, 4th edn. Wadsworth.
Sober, E. (1988). *Reconstructing the past: Parsimony, evolution, and inference. A Bradford book*. Cambridge, MA: The MIT Press.
Sterkenburg, T. F. (2019). Putnam's diagonal argument and the impossibility of a universal learning machine. *Erkenntnis*, *84*(3), 633–656.
van Erven, T., Grünwald, P. D., & de Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 361–417. **With discussion, pp. 399–417**.
van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Clarendon Press.
van Fraassen, B. C. (2000). The false hopes of traditional epistemology. *Philosophy and Phenomenological Research*, *60*(2), 253–280.
Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: Wiley.
von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Gabbay et al. (2011) (pp. 651–706).
Watanabe, S. (1969). *Knowing and guessing: A quantitative study of inference and information*. New York, NY: Wiley.
Wolpert, D. H. (1992a). On the connection between in-sample testing and generalization error. *Complex Systems*, *6*, 47–94.
Wolpert, D. H. (1992b). *On overfitting avoidance as bias*. Technical report 92-03-5001, The Santa Fe Institute.
Wolpert, D. H. (1995a). The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In D. H. Wolpert (ed.), *The mathematics of generalization: Proceedings of the SFI/CNLS workshop on formal approaches to supervised learning, volume 20 of Sante Fe Studies in the sciences of complexity* (pp. 117–214). Boca Raton, FL: CRC Press.
Wolpert, D. H. (1995b). *Off-training set error and a priori distinctions between learning algorithms*. Technical report 95-01-003, The Santa Fe Institute.
Wolpert, D. H. (1996a). Reconciling Bayesian and non-Bayesian analysis. In G. R. Heidbreder (eds.), *Maximum entropy and Bayesian methods: Proceedings of the thirteenth international workshop volume 62 of fundamental theories of physics* (pp. 79–86). Dordrecht: Kluwer.
Wolpert, D. H. (1996b). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341–1390.
Wolpert, D. H. (1996c). The existence of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1391–1420.
Wolpert, D. H. (2002). The supervised learning no-free-lunch theorems. In R. Roy, M. Köppen, S. Ovaska, T. Furuhashi, & F. Hoffmann (Eds.), *Soft computing and industry: Recent applications* (pp. 25–42). London: Springer.
Wolpert, D. H. (2021). What is important about the no free lunch theorems? In P. Pardalos, V. Rasskazova, & M. N. Vrahatis (Eds.), *Black box optimization, machine learning and no-free lunch theorems*. Springer.
Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.
Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937–950.
Zabell, S. L. (2016). Symmetry arguments in probability. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy* (pp. 315–340). Oxford: Oxford University Press.
Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of the 5th international conference on learning representations (ICLR)*.

Zhu, H., & Rohwer, R. (1996). No free lunch for cross-validation. *Neural Computation*, *8*(7), 1421–1426.