No free theory choice from machine learning

Bruce Rushing

Received: 15 April 2022 / Accepted: 7 September 2022 / Published online: 2 October 2022
© The Author(s) 2022

Abstract
Ravit Dotan argues that a No Free Lunch theorem (NFL) from machine learning shows epistemic values are insufficient for deciding the truth of scientific hypotheses. She argues that NFL shows that the best case accuracy of scientific hypotheses is no more than chance. Since accuracy underpins every epistemic value, non-epistemic values are needed to assess the truth of scientific hypotheses. However, NFL cannot be coherently applied to the problem of theory choice. The NFL theorem Dotan’s argument relies upon is a member of a family of theorems in search, optimization, and machine learning. They all claim to show that if no assumptions are made about a search or optimization problem or learning situation, then the best case performance of an algorithm is that of random search or random guessing. A closer inspection shows that these theorems all rely upon assigning uniform probabilities over problems or learning situations, which is just the Principle of Indifference. A counterexample can be crafted that shows that NFL cannot be coherently applied across different descriptions of the same learning situation. To avoid this counterexample, Dotan needs to privilege some description of the learning situation faced by scientists. However, this means that NFL cannot be applied since an important assumption about the problem is being made. So Dotan faces a dilemma: either NFL leads to incoherent best-case partial beliefs or it is inapplicable to the problem of theory choice. This negative result has implications for the larger debate over theory choice.

Keywords Theory choice · Epistemic values · No free lunch theorem · Machine Learning · General philosophy of science

1 Introduction

A number of authors have argued that non-epistemic values should play a role in choosing between scientific theories. A non-epistemic value is an instrumental goal...
such as ending world hunger or proselytizing a favored religion; an epistemic value is a value that helps produce true beliefs such as accuracy or empirical adequacy. The claim of these authors is not that instrumental goals influence which science is produced, e.g. what questions are asked and how intellectual capital is spent. Instead, they claim that deciding the truth of scientific theories requires more than the epistemic merits of those theories. Arguments for this thesis come from the history of science (Kuhn & Hacking, 2012), from inductive risk (Rudner, 1953; Douglas, 2009; Steel, 2013), and from the claim there is no difference between epistemic and non-epistemic values (Longino, 1990, 1996, 2001; Okruhlik, 1994).

Ravit Dotan offers a novel argument. She claims that a result from computer science called the No Free Lunch theorem (NFL) limits the best accuracy of theoretical hypotheses to no more than chance, which shows that “accuracy considerations alone are insufficient for theory choice” (Dotan, 2021, 11091). Roughly, NFL states that if no assumptions about a problem are made, at best all algorithms perform the same as a random guess on the solution to said problem. Furthermore, she argues that because accuracy underpins every other epistemic value, NFL shows that epistemic values are insufficient for theory choice. Consequently, “non-epistemic values are essential to assessment of hypotheses” (Dotan, 2021, 11082).

Dotan is incorrect that NFL limits the best accuracy of theoretical hypotheses to no more than chance. NFL cannot say anything coherent about the best accuracy of hypotheses due to it assuming a uniform prior over hypotheses. This assumption is the Principle of Indifference or Principle of Insufficient reason from classical probability theory. It is justified on grounds that adopting the uniform prior is equivalent to making no assumptions about the problem at hand. But by re-describing the question, it can be shown that NFL provides contradictory probability estimates for a random guess. Were her arguments sufficient to limit accuracy, Dotan would have one committed to incoherent probabilities, i.e. irrationality. The only solution is to assert a canonical description of the problem. But having such a description contradicts the claim that no assumptions are made about the problem at hand. This leads to a dilemma faced by Dotan: either her account is committed to incoherence or she has no justification for applying NFL to the question of theory choice.

There are several implications of my conclusion with respect to the larger debate over theory choice. I identify three lessons to be learned from the failure of applying NFL to limit the best accuracy of theoretical hypotheses. Those lessons involve understanding theorems, a common confusion over what NFL says, and the neglected importance of other impossibility results for the debate over theory choice.

Here is how my argument proceeds. First, I review Dotan’s argument in more detail, including her interpretation of a relevant NFL theorem in machine learning. Second, I discuss the family of theorems referred to as NFL and its assumption of a uniform prior over target functions. Third, I argue that this uniform prior is just the Principle of Indifference, and I show that NFL’s assumption of it leads to a Bertrand’s paradox. Fourth, I consider a series of objections. Fifth, I end by summarizing and discussing larger lessons to be learned from these arguments.
2 Reviewing Dotan’s argument

Dotan argues that NFL shows that accuracy considerations alone are not sufficient for deciding on the truth of hypotheses. She writes that the “No Free Lunch theorem roughly says that no learning algorithm universally performs better than any other” (Dotan, 2021, 11085–11086). As she makes clear, this is the same as saying that all learning algorithms perform just as well as random guessing. And because either she understands hypotheses as learning algorithms or the outputs of said algorithms, she believes that this result has the implication that accuracy is not sufficient for deciding the truth of hypotheses. In this section, I review her argument. First, she either argues that hypotheses can be understood as algorithms or the outputs of algorithms. Second, she argues that NFL shows that on any measure of performance, all algorithms perform equally across all problems. Third, these two facts together militate that all hypotheses are equally accurate, i.e. they perform the same as a random guess.

Either Dotan considers hypotheses to be a type of algorithm or the product of algorithms. An algorithm is a finite, mechanical procedure that produces a unique output on its defined inputs. Cooking recipes, repair manuals, and computer programs are all examples of algorithms. Dotan argues that though NFL is about algorithms, it is also about hypotheses:

NFL is formulated as a theorem about algorithms. However, it seems to be about something else. Algorithms can be compared in many ways: using their efficiency, the year in which they were created, the number of times the letter “A” appears in them, and so on. NFL doesn’t evaluate algorithms based on these or other characteristics of algorithms. Rather, it compares the products of the algorithms—the sets of predictions, classifications, hypotheses, etc. that they produce [...]. Therefore, loosely speaking, the point of NFL is that all hypotheses have the same average expected error. I say I use the word “hypotheses” loosely because I don’t mean to be committing to any particular view on what hypotheses are, nor do I mean to say that NFL is about comparison of hypotheses, rather than comparison of theories, sets of predictions, classifications, and so on. What I do mean to do is to draw attention to the fact that NFL pertains to the question of theory choice: Which hypothesis (or theory, or a set of predictions, etc.) is better? (Dotan, 2021, 11090)

Her argument here is a bit unclear. One can view NFL as applied to hypotheses in two ways. First, the algorithms themselves are viewed as hypotheses. And hypotheses make predictions. For example, my hypothesis that Jones is married predicts that he should wear a wedding ring. This has the nice upshot of being a straightforward application of NFL. If hypotheses are evaluated on their predictions, then the claim is that NFL says across any measure of performance of prediction accuracy, no hypothesis is universally better than any other. Second, she could be claiming that the outputs of algorithms are hypotheses. For example, my hypothesis that Jones wears a wedding ring is a prediction of whatever algorithm led me to believe that he was married. It is less clear how NFL applies here. The view would be that since all algorithms perform the same across any predictive problem, the outputs of those algorithms (hypotheses)
are equally likely to be correct. So all hypotheses would have the same probability of being correct.

Regardless of each interpretation, the argument is that NFL shows all hypotheses to be universally identically accurate. Dotan’s argument applies a particular version of NFL from machine learning. Machine learning is a sub-discipline of artificial intelligence (AI) where programs have to learn to do tasks from data without prior knowledge manually programmed in like symbolic AI. The NFL theorem Dotan relies upon applies to a sub-discipline in machine learning called supervised learning. Supervised learning is a type of training regime for algorithms where humans (or other AIs) have manually assigned each example from data to a ground truth label. The task of an algorithm is to learn a function that maps data input with its correct label. Examples of this task include regression, such as predicting income based on socioeconomic variables, and classification, such as predicting the type of animal from a photograph.

When discussing the performance of an algorithm in supervised learning, researchers and engineers reference the Off-training set (OTS) error. This is because algorithms are typically trained and tested on different subsets of data. The goal is for algorithms to perform well in the real-world, but if trained on the full data available to them, they can just simply memorize the labels of data samples. Consequently, a subset of data are withheld from training to test an algorithm’s true knowledge. Performance is then measured on that set. It is the same idea that some teachers use in school where problem sets on tests are different from those done in class or homework.

In her argument, Dotan considers the OTS error to be average expected error. Let \( X \) be the data examples and \( Y \) the ground truth labels. Then expected error \( \epsilon \) is a function from algorithms \( a \in A \), examples \( x \in X \), ground truth labels \( y \in Y \), and a cost function on algorithms, \( c : A \times X \times Y \rightarrow \mathbb{R} \), to the reals. What \( c \) is exactly does not matter.\(^1\) If \( x \) is data sample and \( a \) an algorithm, then the expected error is:

\[
\epsilon(c, a, x) = \frac{\sum_{y \in Y} c(a, \langle x, y \rangle)}{|Y|}
\]  

(1)

This says that the expected error for an algorithm on a specific data example \( x \) is the unweighted average of the algorithm’s costs on each possible ground truth label. Importantly, each ground truth label is assumed to be equally likely. Otherwise the expected error would be the mathematical expectation where the probabilities of each ground truth label differ. Dotan argues one does this because “we’re not making any assumptions about the problem we are trying to solve” (Dotan, 2021, 11088). I will return to this shortly. Once in place, the average OTS expected error can be found by another unweighted average over the expected error for every test set sample \( x \) in the test set \( X \):

\[
OTSE(a, c) = \frac{\sum_{x \in X} \epsilon(c, a, x)}{|X|}
\]  

(2)

\(^1\) In Dotan’s example, she considers an error measure function from an algorithm \( a \) and a pair \( \langle x, y \rangle \) to the set \( \{1, 0\} \). The error is then defined as just \( C(a, \langle x, y \rangle) = |y - \hat{y}| \) where \( \hat{y} = a(x) \), i.e. \( a \)’s prediction on \( x \).
It can be shown that the average performance of any algorithm is uniquely dictated by the expected error of that algorithm on a sample. Since the expected error has every label equally likely for each training sample, the best an algorithm can do is a random guess on the label for said sample. So the $OTSE$ for every algorithm will be the same, i.e. the same as a random guess.

As discussed, the aforementioned summary of this NFL theorem in supervised learning illustrates that when computing the expected error or cost of an algorithm, an unweighted average is applied. Dotan recognizes this and argues that it is warranted because one should not make assumptions about nature’s regularities:

More specifically, NFL focuses on average expected error which gives all input/output pairs the same weight. This move is meant to reflect the fact that we make no assumptions about the generator, i.e. that we make no assumptions about the regularities we are trying to discover (Dotan, 2021, 11087).

The idea is that one should weight ground truth labels equally for each possible data sample because one must not assume anything about the way the world works apart from consistency, i.e. that there are regularities. For example, if one has a classification algorithm over dog and cat photos, one should assume that each possible picture has an equal chance at being a dog or a cat. Dotan thinks that the same applies to the truth of hypotheses about the laws of nature.

According to Dotan, the upshot of this discussion is that all hypotheses are equally likely to be accurate. Either she considers hypotheses to be algorithms or the output of algorithms. And she considers the particular version of NFL from supervised learning to be warranted about the best performance of prediction algorithms. Consequently, on accuracy considerations alone, one should assign equal credence to all hypotheses:

If we think of NFL in this way, as comparing between hypotheses or sets of predictions, it has implications for theory choice. NFL entails that we don’t make any assumptions about the regularities we are trying to discover except for consistency with the past, all hypotheses have the same averaged expected error. If average excepted [sic] error is a measure of how likely a hypotheses is to be accurate, then all hypotheses are equally likely to be accurate. Therefore, predictive accuracy is not a standard that can be used to discriminate between hypotheses, if we are making no assumptions about the problem we are trying to solve (Dotan, 2021, 11090).

The upshot is that accuracy considerations cannot decide the truth of a hypothesis. Nor can other epistemic values. This, Dotan argues, is because accuracy is the most basic of all epistemic values. It either underpins or supports all the others. If accuracy is not sufficient neither are other epistemic values. Consequently, non-epistemic values must be used for deciding the truth of hypotheses.
3 No free lunch theorem in detail

So far, I have discussed Dotan’s use of a NFL theorem from machine learning. She argues the theorem states that if no assumptions are made about the underlying regularities in nature, then all hypotheses have the same average expected error. The theorem she appeals to had parts discussed first in Wolpert (1992) and Schaffer (1994), with the full result proved in Wolpert (1996). Eventually, NFL was extended to search and optimization in Wolpert and Macready (1997). The latter is the most widely known and cited of the NFL results. Wolpert has argued that NFL “can be viewed as a formalization and elaboration of concerns about the legitimacy of inductive inference, concerns that date back to David Hume (if not earlier)” (Wolpert, Wolpert (2012), 1). Philosophers have examined the relationship between NFL and inductive inference (Schurz, 2017) and have argued extensively against Wolpert’s interpretation (Sterkenburg & Grünwald, 2021). Because Dotan’s argument relies upon NFL, it would be useful to review the two main NFL results in search and optimization and in supervised learning in detail along with the philosophical discussion. This will illustrate an important assumption of these theorems: that the problems they apply to are uniformly distributed.

In each NFL result, Wolpert argues that the best tool for assessing the performance of algorithms is an “extended Bayesian formalism” (Wolpert, 1996, 1348) or “Bayesian” approach (Wolpert and Macready, 1997, 67). The idea is to apply probability theory to the problems one is trying to solve, the measures of solving that problem, the algorithms measured, and the data the algorithms operate on. In the context of supervised learning, the problems being solved involve a set of exemplars (vectors of features), a set of ground truth labels, and a task of assigning exemplars labels such as image classification or regression. Wolpert describes these as consisting of input and output sets, where the task is to learn a conditional probability distribution over labels given exemplars called the target function. Algorithms are interpreted as learned conditional probability distributions called hypothesis functions. Both conditional probability distributions are treated as random variables that probabilities may be formed over. For search and optimization, those problems are the search or optimization problems, such as the traveling salesman problem, the algorithm is used to solve. Both problems and the algorithms being used to solve them are treated as functions mapping between points in search or optimization space. Probabilities are formed over problems and the algorithms used to solve those problems.

Wolpert states that the probability of success for an algorithm on a performance metric depends on the product of two parts. In the case of supervised learning, it depends on the posterior probabilities of the hypothesis function given the data and the target function given the data (equation (3.3) in Wolpert (1992), 25). Since the posterior of the target function given the data is proportional to the prior probability of the target function, the success of an algorithm on a problem is a function of said prior probability. With search and optimization, it depends on the conditional probability that the desired answer is found given the algorithm and the problem at hand and the prior

---

2 She uses the rendition given in Schaffer (1994).
3 This means that one is forming probabilities over conditional probability distributions.
probability of the problem (equation (1) in Wolpert and Macready (1997), 70). Wolpert calls both the inner products and remarks that they determine the performance of an algorithm on any performance metric. Intuitively, one can think of them as measuring how closely “aligned” an algorithm is to the problem it is being used to solve. The second of the two inner products in both cases, the posterior probability of the target function given the data and the prior probability on the problem, plays a critical role in the derivation of each NFL theorem.

In the case of search and optimization, Wolpert proves several NFL theorems. Two theorems are relevant to this discussion because they are good examples of the importance of the uniform prior on the second part of the inner product. The first is the original NFL theorem for time independent search. Let each search problem be a cost function $f: X \rightarrow Y$ where $X$ is a finite search space and $Y$ is a finite set of cost values. Let $m = \{0, 1, \ldots \}$ be a time ordering and $\mathcal{D} = \bigcup_{m \geq 0} (X \times Y)^m$ be the set of points and their costs visited at $m$ for all $m$ time steps. These can be viewed as functions where each pair $d_m(i) = (d^x_m(i), d^y_m(i))$ consists of a visited point $d^x_m(i) \in X$ and its associated cost $d^y_m(i) = f(d^x_m(i))$. The associated costs for all visited points can be enumerated as a list $d^y_m = (d^y_m(0), d^y_m(1), \ldots, d^y_m(m))$. An algorithm is a function $a : \mathcal{D} \rightarrow \{x | x \notin d^x \}$ that goes from pairs of points in search space and their associated cost to a new point in search space. The question one might ask is how algorithms, $a$, perform on the associated costs given some cost function $f$? This is measured by $\Pr(d^y_m | f, m, a)$, the first part of the inner product for search. Overall expected performance of that algorithm then depends on $\Pr(f)$, the second part of the inner product. From here it becomes clear that if $\Pr(f)$ is uniform, the average performance of any pair of algorithms $a_1$ and $a_2$ is given by Wolpert’s theorem one (Wolpert and Macready, 1997, 69):

$$
\sum_f \Pr(d^y_m | f, m, a_1) = \sum_f \Pr(d^y_m | f, m, a_2)
$$

Since performance measures depend on the costs of the search algorithm, $d^y_m$, this means that regardless of performance measure, the average performance of all algorithms across all search problems is the same, i.e. no better than random search. The intuition here behind the proof is that for any problem that an algorithm performs above average on, there is exactly one other problem where the algorithm performs equally below average. And since all problems are equally weighted by the uniform prior, these values will cancel out. This canceling behavior shows up in other NFL theorems.

Why one should assume a uniform prior on search problems? Wolpert argues one may not but proves another theorem that non-uniform priors do not lead to free lunches. In more recent work, he proves that “if any given search algorithm performs better than another over a given set of $P(f)$’s, then it must perform corresponding worse on all other $P(f)$’s” (Wolpert, 2012, 4). The trick of course is an unstated assumption that comes out in the proof: he assumes that the prior over prior probabilities for search problems is uniform (Wolpert, 2012, 10). One then gets the same canceling behavior and the desired result.
Both examples from search and optimization demonstrate the importance of the uniform prior probability distribution for the prior in inner product. The same applies in the NFL theorems for supervised learning.

In the previous section, an example of NFL for supervised learning was provided. This is one instance of a flotilla of theorems that Wolpert provides (Wolpert, 1996, 1354–1358). Schurz lumps these into two classes: a strong version of NFL and a weak version of NFL (Schurz, 2017, 830–831). The strong version asserts that for any cost value of a learning algorithm’s performance, the probability that the algorithm is in a world that leads to that cost value is the same for all learning algorithms. The weak version is that the expected prediction success of any algorithm is the same as a random guesser. Each version makes different assumptions about the cost function measuring predictive success. But both assume that a uniform probability distribution over the target functions. To see how this connects to inner product description, it would be useful to present Wolpert’s formalism in more detail.

Wolpert’s approach relies upon the geometrical concept of the unit simplex. The $n$-dimensional unit simplex is a generalization of triangles and is defined as $S_n = \{ (t_1, \ldots, t_n) \in \mathbb{R}^n | \sum_{i=1}^{n} t_i = 1$ and $t_i \geq 0$ for $i = 1, \ldots, n \}$. For example, $S_1$ would be the point $(1)$ in the line segment, $S_2$ would be the line segment going from $(0, 1)$ to $(1, 0)$ in $\mathbb{R}^2$, $S_3$ would be the triangle embedded in $\mathbb{R}^3$ defined between the points $(1, 0, 0), (0, 1, 0)$, and $(0, 0, 1)$, and so on.

Wolpert uses mappings between the input space of supervised learning algorithms and $n$-dimensional simplexes to define conditional probability distributions going from said inputs to outputs. These define the learning situations an algorithm might encounter. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite (or possibly countably infinite) set of input elements and output elements. Define the random variables $g : \mathcal{X} \rightarrow S_{|\mathcal{Y}|}$ and $h : \mathcal{X} \rightarrow S_{|\mathcal{Y}|}$ to be the target function and hypothesis function respectively going from $\mathcal{X}$ to the $|\mathcal{Y}|$-dimensional unit simplex $S_{|\mathcal{Y}|}$. For example, if the input space consists of thirty two four character long binary strings, e.g. 0010, 1010, and so on, and the output space is just the set $\{0, 1\}$, then both $g$ and $h$ will map to $S_2$, i.e. the line segment from $(0, 1)$ to $(1, 0)$. Since each point in the unit simplex must sum to one and all indices of a point are greater than or equal to zero, they can be treated as probabilities for the joint distribution of $\mathcal{Y}$. Thus, if one treats $y$ and $x$ as random variables whose ranges are $\mathcal{Y}$ and $\mathcal{X}$ respectively, $g$ and $h$ both define conditional probability distributions $\Pr(y|x)$.

Since each target and hypothesis function is a Euclidean vector, the alignment of those functions can be given by their inner product, i.e. by the dot product.

It is easy to show that the probability of the target function given the data is proportional to the prior probability of the target function. If one assumes a uniform distribution over this prior probability, one can read off a vector for the probability of the target function given the data that is in the opposite direction and magnitude of any target function that hypotheses are aligned with. The result is the same canceling behavior seen in the search and optimization NFL results. This was seen in Dotan’s example where the labels for each algorithm were weighted equally—meaning that for any algorithm that guesses the right label for a given target function, there will be another target function where it guesses wrongly and the two will cancel out.

Summarizing, the key assumption that allows the zoo of NFL theorems to be proven is the assumption of uniform probability of the second term in the inner product for-
mulation of search and supervised learning. In both cases, this allows one to weight each search problem or target function equally, allowing functions where an algorithm performs well to be canceled out exactly by one where it performs poorly. The importance of the uniform distribution is recognized by Wolpert. He argues that this assumption merely reflects the absence of assumptions over the search problem being solved or the target function being predicted (Wolpert and Macready, 1997, 67). Dotan adopts this line in her argument and explains its importance when she writes that

We can intuitively see why all algorithms have the same average expected error. When calculating the expected error for each input in step 3, we count all possible outputs as equally likely. The reason is that we make no assumptions about how the generator operates except for consistency with the past. I.e., we make no assumption about the outputs that are likely to be produced in the problem we are trying to solve. Because we make no assumptions about possible outputs, for the purposes of OTS error it doesn’t matter which predictions the algorithm makes. The contribution to the average expected error will always be the same—getting it right in one case and getting it wrong in all others (Dotan, 2021, 11089).

Mirroring Wolpert’s own arguments about NFL applied to non-uniform probability distributions, Dotan argues a need for further justification of any assumption about the target problem would be required and in the absence of that further justification, “this strategy just pushes the bump under the rug” (Dotan, 2021, 11093). So the importance of the uniform probability assumption is recognized by all proponents of NFL.

It has also been recognized by critics. As mentioned, Wolpert considers the NFL to be an instance of the problem of induction. Philosophers and computer scientists have responded to it in two directions. The first is to consider Wolpert’s theorem as assuming that the environment is maximally induction unfriendly (Rao et al., 1995) or the dogmatic assertion of an “induction hostile” prior (Schurz, 2017, 833). In the former, Rao et al argue that Wolpert’s uniform distribution amounts to the assertion that no learning is possible. Similarly, Schurz shows that in the case of random binary sequences, the uniform distribution assumption amounts to assigning probability one to the limiting frequency of one in those binary sequences being one-half. In other words, adopting the uniform distribution assumption is to be certain that the generator of those sequences are unbiased coin flips. Of course, this just is the claim that there are no patterns one could inductively learn from such a world. This means that no algorithm such as the proven best meta-inductors will do better than random guessing, where a meta-inductor is an algorithm that aggregates the predictions of other algorithms. As Schurz says, the “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” (Schurz, 2017, 834).

As mentioned, Wolpert has defended the uniform prior assumption due to a NFL result applied to prior distributions. But as I noted earlier, this result assumes a uniform prior over priors. Sterkenberg and Grünwald note that this just backs up the justification of a uniform distribution one level (Sterkenberg and Grünwald, 2021, 10). They remark that Wolpert and his interlocutors then settle into a debate over what distributions assumed in the second part of the inner product are justified. This debate
is a replay of the problem of the priors. Sterkenberg and Grünwald suggest that the real answer is to abandon any justification for prior probabilities. If Wolpert’s probabilities are viewed as subjective degrees of belief, then the right answer is to bring whatever non-dogmatic prior one has to the table, which echoes the abandonment of “a uniform distribution as an objective-logical “indifference prior”” by philosophers and statisticians (Sterkenburg and Grünwald, 2021, 11). They note that Wolpert cannot do that easily as he is committed to a framework of algorithms being purely data driven. The suggestion then is to abandon that framework for one where algorithms embody certain modeling assumptions. Justification of algorithms is then resolved to be one over the relative strengths of different modeling assumptions.

Sterkenberg and Grünwald’s mention of an “indifference prior” should alert one to the NFL’s uniform distribution assumption being just the Principle of Indifference over search problems and learning situations. This has ramifications for Dotan’s argument that NFL shows non-epistemic values are critical for theory choice.

4 The principle of indifference and a counterexample

The Principle of Indifference (PI) (also known as the Principle of Insufficient Reason) has played an important role in the history of probability theory. It holds that ignorance about a set of exhaustive and exclusive outcomes is best expressed by a uniform prior over those outcomes. This is the same reasoning that Dotan gives for her interpretation of NFL. The absence of assumptions about the learning situations faced by an algorithm is best expressed as a uniform distribution of weights to those learning situations. As a rule for forming priors, PI is a candidate solution to the problem of the priors: what constraints, apart from the probability axioms, should there be on prior partial beliefs? Considered as a solution to that problem, it is generally regarded as a failure (see Zabell, 2005 for an extended discussion). Why it fails has ramifications for Dotan’s argument that NFL limits the best case accuracy of scientific theories. A counterexample can be constructed that shows applying NFL results in assigning incoherent best case partial beliefs to theories.

A coin is to be flipped four times resulting in one of sixteen possible sequences. Absent any knowledge about the bias of the coin, what is the probability of each sequence? PI recommends that each sequence be assigned the same probability: one-sixteenth. The idea is since one lacks further information and each sequence is equally possible, they are equally probable. The claim of equal possibility refers to the set of sequences being exhaustive and exclusive. They are exhaustive in that they specify every possible trial involving the coin being flipped four times. And they are exclusive in that each sequence is logically incompatible with the others. This is what van Fraassen refers to as “the Principle of Uniform Distribution” (Van Fraassen, 1989, 299). It is one form of PI and the most famous one.

As I documented earlier, each version of NFL requires this assumption. The reasoning by Dotan tracks closely the reasoning for PI. She argues that if one does not make any assumptions about the target function an algorithm predicts, then one should assign equal weights to each target function. A lack of assumptions just is an absence of knowledge or information. It is the same justification for PI. So not only is the
assumption of a uniform distribution over outcomes common between NFL and PI, the justifications for that assumption are identical.

PI, however, has a serious problem. The problem was first hinted at by Boole (Boole, 1854, 369–375) but received its namesake examples from Bertrand (1907). Called Bertrand’s paradox, it shows that PI leads to incoherent probability estimates. The issue is that the relevant possible outcomes can be described in different but logically equivalent ways. These different descriptions lead to differing probabilities assigned to the same proposition.

This paradox applies to NFL. From it, one can show that NFL as Dotan interprets it leads to incoherent best case probabilities assigned to hypotheses. To see how it works, I make use of the following modification of van Fraassen’s cube example (Van Fraassen, 1989, 302–304).

Consider the case of the data scientist Reggie who works at a cube factory. Reggie has to answer customer queries about the cubes produced at the factory. However, except for the fact that she knows cubes to have a maximum side length of two centimeters, she knows nothing else about the cubes being produced. Reggie employs her background as a data scientist to develop algorithms that provide her answers for customer queries.

One day a customer comes to Reggie and asks her given some observable properties of a cube produced by her factory, what is the probability that said cube has a side length of less than or equal to one centimeter. That is if \( c \) is the vector of observed properties for cube \( c \) and \( s(c) \) is the side length of the cube, the customer wants to know \( \Pr(s(c) \leq 1 \text{ cm}) \).

Reggie has had a stellar graduate education at Stantech and begins reasoning through the problem using probability theory. She notes that the query is over the membership of a cube in a proposition, where a proposition is some subset of a sample space of possible outcomes. Probabilities are always formed with respect to said sample spaces and an algebra. Here the sample space or space of possible outcomes is just the set of possible cubes \( C \) produced by the factory. The algebra is a set of subsets on \( C \) that includes \( C \) and is closed under finite union (or countable union if a sigma algebra) and complement. The query then is whether a cube is in a specific subset of possible cubes found in the algebra. That subset is just the proposition containing the cubes whose side length are less than or equal to one cm. The cube having the observed properties \( c \) is another proposition in that algebra. So the query is estimating the probability that given the cube \( c \) is in the proposition of cubes with observed properties \( c \), what is the probability that \( c \) is in the proposition containing all cubes whose side length is less than or equal to one cm?

Reggie translates this problem to the supervised learning context. This requires her to specify an input space and an output space. She considers the input space as being the set of exemplars given by observations \( C \) for each cube \( c \in C \) and the output space as the set \( Y = \{1, 0\} \), where 1 says the observed cube’s side length is less than or equal to one cm and 0 says otherwise. The target function to be learned is \( f : C \rightarrow \mathcal{S}_{|Y|} \) and can be thought of as specifying the conditional probability distribution \( \Pr(y|c) \).

---

4 More specifically, Boole showed that PI cannot be consistently applied across descriptions of problems involving discrete random variables and Bertrand showed the same difficulties to apply to continuous random variables.
where $y$ and $c$ are taken to be random variables going from the set of possible cubes $\mathcal{C}$ to $\mathcal{Y}$ and $\mathcal{C}$ respectively. Thinking back on her training in supervised learning, she decides to set up the problem so that the task of an algorithm $a$ is to learn that target function and then offer its best guess to the query. If the learned conditional probability distribution is given by $h : \mathcal{C} \to S_{|\mathcal{Y}|}$, then the algorithm outputs for any input $c$:

$$a(c) = \arg\max_y h(c) = \arg\max_y \Pr(y|c)$$ (4)

That is the algorithm $a$ outputs the label with the highest probability given by $h$. Algorithms are trained on some proper subset of all cubes possibly produced by the factory to learn $f$. They are then tested on cubes not in their training set drawn randomly from the total possible number of cubes.

Now Reggie is a clever data scientist and recalls from her graduate education a paper she read that the best case probability assignment she can offer in this problem is given by NFL. This would save her considerable time in programming and compute so she reasons out analytically what that best case assignment might be. She sees two ways to think about answering the query. In both ways, the best case probability according to NFL is 0.5. First, the best case accuracy of an algorithm on the OTS can be thought as giving the most accurate probabilities desired in the query. In this case, the query is understood as one minus the expected error of the algorithm on OTS due to the algorithm offering its best guess of a cube’s label after learning $f$. To find the best estimate of an algorithm and thus the best estimate of $\Pr(s(c) \leq 1 \text{cm}^2|c)$, one takes the average OTS expected error of said algorithm. The average OTS expected error of an algorithm will be the average expected error across these test sets. Since the number of labels is two, the OTS expected error will be one-half by NFL. Since the query is one minus the expected error, $\Pr(s(c) \leq 1 \text{cm}^2|c) = 1 - 0.5 = 0.5$. Second, one can view the $h$ learned by the algorithm as estimating desired probability directly. Then the best case accuracy would be given by the alignment of $h$ with the true target function $f$. Since by NFL, all target functions $f$ are equally likely, the most aligned $h$ is the one that assigns probability 0.5 to any cube, i.e. $\Pr(s(c) \leq 1 \text{cm}^2|c) = 0.5$. Reggie writes this down and sends out the report for the customer with her best guess probability.

The next day, Reggie receives another query from a customer. This customer wants to know that given some observable properties of a cube produced in her factory, what is the probability that the cube’s side area is less than or equal to one centimeter, the probability the cube’s side area is greater than one centimeter but less than or equal to two centimeters, the probability the cube’s side area is greater than two centimeters but less than or equal to three centimeters, and the probability the cube’s side area is greater than three centimeters but less than or equal to four centimeters? That is her customer’s query wants the following probabilities, where $ar(c) = s(c)^2$ is the side area of the cube: $\Pr(ar(c) \leq 1 \text{cm}^2|c)$, $\Pr(1 \text{cm}^2 < ar(c) \leq 2 \text{cm}^2|c)$, $\Pr(2 \text{cm}^2 < ar(c) \leq 3 \text{cm}^2|c)$, $\Pr(3 \text{cm}^2 < ar(c) \leq 4 \text{cm}^2|c)$.

She again goes back to her schooling and reasons that this query partitions the sample space by cube side area. Each item in the requested query corresponds to whether the cube $c$ is in the respective element of that partition. Thinking about the
problem in terms of supervised learning, she now has an input space of $C$ and an output space given by the set $\mathcal{Y}^\prime = \{11, 10, 01, 00\}$. She reasons that 11 says that the cube $c$ is in the proposition $1\,cm^2 < ar(c) \leq 2\,cm^2$, 10 says that the cube $c$ is in the proposition $2\,cm^2 < ar(c) \leq 3\,cm^2$, and 00 says that the cube $c$ is in the proposition $3\,cm^2 < ar(c) \leq 4\,cm^2$. The target function is now given by $f'' : \mathcal{C} \rightarrow S_{|\mathcal{Y}|}$ and an algorithm’s task is to learn $f''$ and estimate the most likely label from learned distribution $h''$. She applies the trick she used from the day before to leverage NFL to estimate her best guess probabilities. First, she checks her proposed algorithms for their best case accuracy. She correctly infers that the expected error will be three-quarters according to NFL, since the number of labels is four and the best one can do with a random guess is one-quarter. Substituting the best case accuracy of the algorithm across OTS for one’s credence, the probabilities for the first item in the query will be $\Pr(ar(c) \leq 1\,cm^2|c) = 1 − 0.75 = 0.25$. She finds the same result for the other items in the query. The same applies in the case of taking $h''$ directly: the most accurate probability distribution assuming all $f''$ are equally likely is the $h''(c) = 0.25$ for any $c$.

Excited about the revolutionary time savings given by her new methodology, Reggie returns to work the following day and finds yet another query from a customer. This customer wants to know given some observed properties of a cube, what is the probability that the volume is greater than one centimeter but less than two centimeters, probability that the volume of the cube is less than or equal to one centimeter, the probability that the volume is greater than two centimeters but less than three centimeters, and so on up to eight centimeters in cube volume. Her customer’s query is to compute eight probabilities, where $v(c) = s(c)^3$ is the volume of the cube: $\Pr(v(c) \leq 1\,cm^3|c)$, $\Pr(1\,cm^3 < v(c) \leq 2\,cm^3)$, and so on.

Reggie knows her trick will work here as well. She reasons that this customer is asking about the membership of a cube in some element of an eight element partition of her sample space. In the supervised learning context, Reggie views her input set as $C$ and output set as $\mathcal{Y}'' = \{111, 110, 101, 011, 001, 010, 100, 000\}$. 111 corresponds to the proposition $v(c) \leq 1\,cm^3$, 110 corresponds to the proposition $1\,cm^3 < v(c) \leq 2\,cm^3$, and so on. The target function to be learned is now $f''' : \mathcal{C} \rightarrow S_{|\mathcal{Y}|}$ and the learned distribution is given by $h'''$. Reggie uses the same strategy and reasons her expected error according to NFL to be seven-eighths and with it the best case accuracy for each item in her query to be one-eighth. And so the credences she reports to her customer for the first item in that query is $\Pr(v(c) \leq 1\,cm^3|c) = 1 − 0.875 = 0.125$ and likewise for the other items. She finds the same answer when reasoning from $h'''$ directly to give the queried probability. There the most accurate distribution assuming all $f'''$ are equally likely is $h'''(c) = 0.125$.

While checking her math, Reggie becomes horrified at her estimates. She realizes that she has given incoherent estimates for the same proposition. In the first query, she estimated $\Pr(s(c) \leq 1\,cm|c) = 0.5$ but in her second query she estimated $\Pr(ar(c) \leq 1\,cm^2|c) = 0.25$ and in her third query $\Pr(v(c) \leq 1\,cm^3|c) = 0.125$. But she reasons, these are probabilities over the same proposition. Since side area is just side length squared and volume is just side length cubed, the propositions are translatable into one another and pick out the same set in the sample space. Yet she has assigned different probabilities to that set. So her best estimates for each of these queries are incoherent.
This story illustrates an important problem with Dotan’s argument. Dotan argues that NFL provides estimates that are the best one can do with epistemic considerations alone. In this example, the best one can do is to have three different partial beliefs to the same query. But the result is that one assigns partial beliefs that are not even functions, let alone probabilities. One is incoherent—one is not even rational. And this applies either if Dotan believes the hypotheses to be the algorithms themselves or the outputs of the algorithms. If it is the algorithms, then one has different best estimates for those algorithms’ performance. If it is the outputs of algorithms, then one assigns different partial beliefs to the same output hypotheses (due to 1, 11, and 111 all corresponding to the same proposition in each target function).

Dotan might argue that one does not need to assign different partial beliefs to the same proposition. Just pick one of the functions and use that to characterize the proposition \( s(c) \leq 1 \text{ cm} \). This would mean that Reggie should have assigned the same credence to \( \Pr(s(c) \leq 1 \text{ cm}|c) \), \( \Pr(\text{ar}(c) \leq 1 \text{ cm}^2|c) \), and \( \Pr(v(c) \leq 1 \text{ cm}^3|c) \). But here lies the bite of Bertrand’s paradox: each function is logically translatable into the others. One has merely re-described the problem by moving to \( f' \) and \( f'' \) over \( f \). The upshot is that applying a uniform distribution over one function entails applying a non-uniform distribution over the others. For example, assigning \( \Pr(s(c) \leq 1 \text{ cm}|c) = 0.5 \) means that the probabilities given for Reggie’s other queries would not be uniform. This means that applying NFL naively to just one function ensures violating it on the logically equivalent others.

Still, Dotan might insist that \( f \) or \( f' \) or \( f'' \) are somehow privileged over the other descriptions. She might say one is the correct description of the problem and the others are not. But this violates Dotan’s interpretation of NFL at a deeper level. She claims that NFL expresses our best probabilities making no assumptions about a problem. But stating that \( f \) or \( f' \) or \( f'' \) are the privileged description is making an assumption about the problem. One is saying that the problem is best described this way and not the others on grounds outside of logic. van Fraassen best describes the incoherence of asserting a privileged description:

But that response asserts that in the absence of further information we have no way to determine the initial probabilities. In other words, this response rejects the Principle of Indifference altogether. After all, if we were told as part of the problem which parameter should receive a uniform distribution, no such Principle would be needed (van Fraassen, 1989, 305).

Another way to characterize the problem for Dotan’s interpretation is through Wolpert’s framework. Wolpert says that the possible target functions map from exemplars (the observed properties of cubes) to the simplex defined by the number of labels. So the target function to be learned is either \( f : \text{C} \rightarrow \text{S}_2 \), \( f' : \text{C} \rightarrow \text{S}_4 \), and if \( f'' : \text{C} \rightarrow \text{S}_8 \), where \( \text{C} \) is the set of cube property vectors trained or tested on. The different variations on each of these functions are what NFL’s uniform distribution applies to. But NFL does not tell one which target functions are correct. It does not tell one the right simplex. But knowing which simplex is correct is to already make an important assumption about the problem at hand: it corresponds to knowing important structural features of the problem relevant for solving it. And if one does not know
which simplex, on logical grounds applying NFL would result in incoherent partial beliefs.

Summarizing, the problem confronting Dotan’s interpretation is the following dilemma. Either no assumptions are made about the problem at hand or some are. If no assumptions are made about the problem at hand, then any learning situation can be described in ways that are logically equivalent but which application of NFL leads to incoherent partial beliefs and irrationality. If some assumptions are made, then there is no justification for NFL’s assumption of uniform prior probabilities on candidate target functions. So NFL does not need to apply and non-epistemic values need not be imported into theory choice.

5 Objections

So far, I have argued that Dotan’s application of NFL either leads to incoherent partial beliefs or is unwarranted. This is due to NFL relying upon a uniform probability distribution over target functions. In this section I document four possible responses. The first is to review whether Bertrand style paradoxes can be solved via the principle of maximum entropy and symmetry arguments. This would justify PI as a representation of ignorance and with it, NFL theorems. The second is to rely upon NFL theorems that do not use uniform priors. Since these NFL theorems do not rely upon a uniform prior, Dotan can avoid the bad consequences that Bertrand style paradoxes would generate. The third is to argue that privileging descriptions of the problem justifies non-epistemic values in theory choice. And the fourth is to appeal to absolute no free lunch theorems, which show that there is no optimum inductive method.

5.1 Saving the uniform prior

Since Bertrand’s paradox was discovered, there have been a number of attempts to defuse the paradox and save PI. If an attempt was successful, it would at least partially defuse worries over applying the uniform prior in NFL. One might then be somewhat justified equating a uniform prior probability with a lack of assumptions.

The leading candidate is an appeal to geometric symmetries to specify a unique measure for inducing probabilities. Poincaré first proposed this solution (Poincaré, 1912) and Jaynes has given its most robust defense (Jaynes, 1973). The idea is that in cases like the cube example, there are translational invariances on distance or length. For example, the cube side-length measure \( m(a, b) = b - a \) is uniquely identified by how one could move the cube left or right. Shifting the cube by \( x \) centimeters does not change the value of \( m \), i.e. \( m(a, b) = m'(a + x, b + x) \). Taking such a measure that happens to be scale invariant too\(^5\) and defining one’s probabilities off such a measure, one then ends up with the same unique probability for side length, area, and volume using PI.

While Van Fraassen concedes that in Bertrand’s paradox Poincaré and Jaynes’s solution works, he shows that no such geometric symmetries are available for the

\(^5\) Van Fraassen argues that it must be the logarithm, i.e. \( M(a, b) = \log b - \log a \) (Van Fraassen, 1989, 309-310).
similar von Mises’s water and wine problem (Van Fraassen, 1989, 313–319). But that might be too generous. There are two further problems with the proposed solution. First, an appeal to geometric symmetries is either a restriction of the problem or not the correct problem to begin with. As Shackel argues in the case of the original version of Bertrand’s paradox, Jaynes’s solution requires either a restriction to finite translations or the use of empirical experiment to fix the symmetries (Shackel, 2007, 171–172). Second, that latter case proves to be important since different methods of experiment can result in the same symmetry giving different measures and thus different probabilities (Drory, 2015). So an appeal to geometric symmetries is unlikely to save PI nor are more general forms of PI such as maximum entropy able to avoid the paradox (Shackel & Rowbottom, 2020).

5.2 Non-uniform priors for NFL

There are several NFL theorems that do not rely upon uniform priors. The theorems in question are the various sharpening of NFL results first proposed by Wolpert (Wolpert & Macready, 1997) and later given mathematically exact definitions by Igel and Toussaint (2005). The primary theorem of interest is Igel and Toussaint’s theorem five (Igel and Toussaint, 2005, 320). This theorem works on equivalence classes of target functions based on histograms. A histogram is a function that maps labels to counts of input examples such that the sum of every label equals the total number of input examples. For example, suppose one has 100 cubes from the cube factory thought experiment with the possible labels \{0, 1\}. One histogram might be the function \(h_0 : \{0, 1\} \mapsto \mathbb{N}\) that returns \(h_0(0) = 35\) and \(h_0(1) = 65\). These histogram functions effectively “bin” the data according to their labels. One can then sort target functions into equivalence classes based on their histograms. Call each equivalence class \(B_{h_i}\) for each histogram \(i\). For example, if \(f_1\) and \(f_2\) share \(h_0\) as their histogram, they map the same number of input examples to the labels \{0, 1\} even though they may not map the same inputs to outputs.

Igel and Toussaint show that if one assumes the probability of any two functions in their respective \(B_{h_i}\) is the same for all histograms \(h\), then the performance of any algorithm is the same as all the others (Igel and Toussaint, 2005, 320). The intuition for the proof is that within each equivalence class \(B_{h_i}\), an algorithm that performs well on one function will perform equally poorly on another in that class. Note that this does not require one’s prior probability to be uniform. One might assign different probabilities to two target functions who are not members of the same equivalence class. What it does require is that one’s prior is uniform over every target function in the same \(B_{h_i}\). So the sharpened NFL result holds for priors that are not universally uniform but locally uniform.

This might help Dotan’s argument because no uniform prior is needed over all target functions. But one still gets the result that any algorithm performs no better than a random guess. So non-epistemic values might still play a role in theory choice.

The problem with this solution is that it still has a requirement of uniformity with respect to equivalence classes by histograms. One can still re-parameterize the learning situation within each equivalence class, which will lead to incoherent probabilities. For
example, the 100 cubes from earlier could be sorted into labels according to area, which will give a different average expected error and hence a different probability estimate per this sharpened NFL. The histogram equivalence classes may not be the same but the application of a uniform probability distribution over their members ensures that different best guess estimates may be had. And with them incoherent probabilities. If Dotan argues that some set of target functions better describes the problem, then the justification of the uniform prior within equivalence classes is unwarranted due to one’s assumptions.

Summing up, while NFL has been extended to some non-uniform priors, those priors still preserve uniformity locally. This means that the sharpened NFL cannot be used by Dotan to argue that epistemic values are insufficient for theory choice.

5.3 Privileging descriptions as non-epistemic values in theory choice

The second objection to my argument would have Dotan fall on one horn of the dilemma. She might claim that privileging certain descriptions of the problem amounts to invoking non-epistemic values in theory choice. For example, one might desire there to be more cubes with side lengths less than or equal to one centimeter when considering area and volume. So one might choose the description of \( f \) where the uniform prior is chosen based on side length and not area or volume.

This objection fails. It is not clear why epistemic values would be insufficient for privileging one description over another. Arguing that non-epistemic values are needed for such description is just begging the question. Why should I appeal to my desires and values for describing a problem? Would I not choose accuracy or any other epistemic value to make that decision? After all, I want my description to be accurate and represent the world as it is—not as how I want it to be. Dotan might reply NFL shows this to be insufficient. But this is backwards. NFL only shows that result if one has a privileged description. In which case, Dotan’s argument is that assuming non-epistemic values being instrumental in theory choice shows that accuracy considerations are insufficient. That is, she shows a tautology: if non-epistemic values are crucial for theory choice, then they are crucial for theory choice.

5.4 Absolutely no free lunches

NFL theorems come in two varieties. The sort of NFL theorems discussed so far are those that are relative to measures, such as probability measures. The other variety of NFL are those that assume no measure, so-called absolute no free lunch theorems. These are impossibility results about the existence of a universally optimum inductive algorithm. By universal I mean applies across all learning situations, and by optimum I mean performs better than any other algorithm on a given performance metric. One of the first such theorems was proven by Putnam against Carnap’s program of an inductive logic, which showed via diagonalization arguments that there is no measure function that is both computable and able to infer any computable regularity (see Putnam, 1963 and Sterkenburg, 2019 for discussion). Similar results have been proved to hold more generally (see Belot, 2020 for a full list).
These results are directly relevant for they describe a limit to epistemic methods. To illustrate, consider a modification of Putnam’s theorem against Carnap. Let the set of possible learning situations be the set of all possible binary sequences. And let the set of theoretical hypotheses be the Turing machines that learn to predict each entry of a binary sequence from the prior entries. Then it can be shown that for any machine, there exist binary sequences that said machine cannot learn. Intuitively, one can think of there being binary sequences that are the product of a computable anti-predictor that outclasses and confounds the learning machine. Furthermore, it can be shown that for any said machine, there exists another machine that can learn said confounding binary sequences as well as learn successfully the patterns the prior machine ingested. But then the previous result applies just as well to that machine. The upshot is that there is no machine that is universal and optimal. This is a hard limit to any inductive method.

The operative question for the debate over theory choice is then does one happen to live in a world that will confound any of the set of theories one might use to predict said world? If so, then the decision to select theories will involve hard trade-offs (Belot, 2020, 160) and would possibly allow for non-epistemic values to weigh into theory choice.

It should be noted that the question of which world one is in is a question that allows an answer involving a probability measure. This is just for one to ask what probability one assigns that one’s world will frustrate one’s most accurate theories? In that case, one is back to a framing of the problem in the sense that the measure relative NFL theorems attempt to address. But as I have argued, those theorems either lead to incoherent partial beliefs or are not warranted. So the argument still needs to be made by the proponent of non-epistemic values in theory choice about why one would believe that one lives in a world where epistemic values are insufficient for theory choice.

Two comments are worth making here. First, I think I do not live in a world hostile to my best inductive methods. I cannot rule out that those methods may break down and find that I live in such a world in the future. But I do not have good reason to seriously entertain that skeptical hypothesis. Second, the discussion of the type of world I live in brings the discussion over non-epistemic values in theory choice much closer to traditional discussions over inductive skepticism. Am I in a world hostile to my best inductive methods, i.e. should I be skeptical of my best methods? This question is in the vicinity of inductive skepticism.

6 Conclusion

I have argued in this paper that Dotan’s argument for non-epistemic values in theory choice proves too much. Because her argument ultimately rests on the Principle of Indifference, it is vulnerable to Bertrand-style paradoxes. This would commit her to incoherent partial beliefs if she relies upon NFL as a guide to the forming probabilities over hypotheses. If she argues that Bertrand-style paradoxes can be avoided by privileging a description of the learning situation or problem at hand, then her justification for NFL’s uniform prior is unwarranted. A privileged description is one where one
is making important assumptions about the learning situation or problem at hand. So one should not assign a uniform prior to the algorithms or hypotheses. She cannot avoid these problems by appeal to symmetry arguments. Nor can she appeal to NFL results that do not rely upon a uniform prior because those results use priors that are still “locally” uniform. She cannot just simply assume non-epistemic values are what privilege certain descriptions because this just begs the question. And while absolute no free lunch theorems might be relevant to theory choice, they are not sufficient alone for showing non-epistemic values are necessary for theory choice. Consequently, NFL is not sufficient for showing non-epistemic values play a crucial role in the deciding the truth of theories.

Here are some broader implications from my discussion of Dotan’s argument. There are three lessons to be learned from the failure of applying NFL to theory choice.

First, a theorem’s assumptions are very important, and it can be difficult to know when they apply and do not apply. A uniform prior over hypotheses might be warranted, for example, in the case of sequences of fair coin tosses. So NFL would apply here and any algorithm would only perform as well as random guessing on each coin toss. One might think the uniform prior to hold because one knows that the process that generates the sequences is independent and identically distributed according to some bias parameter. But it is a much stronger assumption when deciding over scientific theories. What corresponds exactly in the case of theories with coin flipping process? Does it have the same properties? Consequently, the prior over worlds needs to be carefully considered, lest it lead to paradox. And when the prior is unjustified, so is NFL.

Second, a common confusion over NFL is that it applies when no assumptions about a problem are made. This is partially due to Wolpert, who in his paper on supervised learning NFL says “the sole concern of this paper is what can(not) be formally inferred about the utility of various learning algorithms if one makes no assumptions concerning targets” (Wolpert, 1996, 1344). A uniform prior, however, is not the absence of assumptions but a very big assumption. This assumption when unjustified can lead to problems, as Bertrand’s paradox illustrates. Consequently, understanding what NFL actually says would help in deciding where in the debate over theory choice it might be applicable.

Third, a discussion of NFL does lead to an important observation that seems relevant to the debate over theory choice: the lack of a universally optimal inductive method. This absence pushes the argument over theory choice to whether we exist in a world where epistemic values are insufficient for finding an inductive method that works well in said world. The proponent for non-epistemic values in theory choice needs an argument for why one might think one is in such a world. My belief is that it is doubtful humanity is so unfortunate.

Acknowledgements I would like to thank and acknowledge Daniel Herrmann, Brian Skyrms, Kyle Stanford, and two anonymous reviewers for thoughtful feedback on this paper.
Declarations

financial or non-financial interest  The authors have no relevant financial or non-financial interests to disclose.

Open Access  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Belot, G. (2020). Absolutely no free lunches! Theoretical Computer Science, 845, 159–180.
Bertrand, J. L. F. (1907). Calcul des probabilités. Gauthier-Villars.
Boole, G. (1854). An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities. Dover Constable.
Dotan, R. (2021). Theory choice, non-epistemic values, and machine learning. Synthese, 198(11), 11081–11101.
Douglas, H. (2009). Science, policy, and the value-free ideal. University of Pittsburgh Press.
Drory, A. (2015). Failure and uses of Jaynes’ Principle of Transformation Groups. Foundations of Physics, 45(4), 439–460. https://doi.org/10.1007/s10701-015-9876-7.
Igel, C., & Toussaint, M. (2005). A no-free-lunch theorem for nonuniform distributions of target functions. Journal of Mathematical Modelling and Algorithms, 3(4), 313–322.
Jaynes, E. T. (1973). The well-posed problem. Foundations of Physics, 3(4), 477–492.
Kuhn, T. S., & Hacking, I. (2012). The Structure of Scientific Revolutions: 50th (Anniversary). University of Chicago Press.
Longino, H. E. (1990). Science as social knowledge: Values and objectivity in scientific inquiry. Princeton University Press.
Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. H. Nelson & J. Nelson (Eds.), Feminism, science, and the philosophy of science (pp. 39–58). Kluwer Academic Publishers.
Longino, H. E. (2001). The fate of knowledge. Princeton University Press.
Okruhlik, K. (1994). Gender and the biological sciences. Canadian Journal of Philosophy, 24(sup1), 21–42. https://doi.org/10.1080/00455091.1994.10717393.
Poincaré, H. (1912). Calcul des probabilités (Vol. 1). Gauthier-Villars.
Putnam, H. (1963). Degree of confirmation? And inductive logic. In P. Arthur (Ed.), The philosophy of Rudolf Carnap (pp. 761–783). Schilpp. 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 Machine learning proceedings (pp. 471–479). Elsevier.
Rudner, R. (1953). The scientist Q Ua scientist makes value judgments. Philosophy of Science, 20(1), 1–6. https://doi.org/10.1086/287231.
Schaffer, C. (1994). A conservation law for generalization performance. Machine learning proceedings (pp. 259–265). Elsevier.
Schurz, G. (2017). No free lunch theorem, inductive skepticism, and the optimality of meta-induction. Philosophy of Science, 84(5), 825–839.
Shackel, N. (2007). Bertrand’s paradox and the principle of indifference. Philosophy of Science, 74(2), 150–175. https://doi.org/10.1086/519028.
Shackel, N., & Rowbottom, D. P. (2020). Bertrand’s paradox and the maximum entropy principle. Philosophy and Phenomenological Research, 101(3), 505–523. https://doi.org/10.1111/php.12596.
Steel, D. (2013). Acceptance, values, and inductive risk. *Philosophy of Science, 80*(5), 818–828. https://doi.org/10.1086/673936.
Sterkenburg, T. F. (2019). Putnam’s diagonal argument and the impossibility of a universal learning machine. *Erkenntnis, 84*(3), 633–656.
Sterkenburg, T. F., & Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese.* https://doi.org/10.1007/s11229-021-03233-1.
Fraassen, V., & Bas, C. (1989). *Laws and symmetry.* Clarendon Press.
Wolpert, D. H. (1992). On the connection between in-sample testing and generalization error. *Complex Systems, 6*(1), 47.
Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390.
Wolpert, D. H. (2012). What the no free lunch theorems really mean; how to improve search algorithms. *Santa Fe Institute, 7,* 1–13.
Wolpert, D. H., & Macready William, G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.
Zabell, S. L. (2005). *Symmetry and its discontents: Essays on the history of inductive probability.* Cambridge University Press.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.