Theory Choice, Non-epistemic Values, and Machine Learning

Ravit Dotan, UC Berkeley
Forthcoming in *Synthese*

Abstract:
I use a theorem from machine learning, called the "No Free Lunch" theorem (NFL) to support the claim that non-epistemic values are essential to theory choice. I argue that NFL entails that predictive accuracy is insufficient to favor a given theory over others, and that NFL challenges our ability to give a purely epistemic justification for using other traditional epistemic virtues in theory choice. In addition, I argue that the natural way to overcome NFL's challenge is to use non-epistemic values. If my argument holds, non-epistemic values are entangled in theory choice regardless of human limitations and regardless of the subject matter. Thereby, my argument overcomes objections to the main lines of argument revealing the role of values in theory choice. At the end of the paper, I argue that, contrary to common conception, the epistemic challenge arising from NFL is distinct from Hume's problem of induction and other forms of underdetermination

What is the role of values in empirical reasoning? It was never controversial that values shape the goals of scientific and other empirical inquiries, as well as the choice of the projects that are pursued. It was also never controversial that people are in fact prone to be influenced by their values in all areas of life, including in their empirical reasoning. The more interesting question is whether values have an inherent role in assessing hypotheses. For the last few decades, the consensus has been that a completely value-free assessment of hypotheses is impossible (McMullin, 1982). Generally, the reason is that available data is never enough to uniquely determine which hypothesis is true (due to epistemic puzzles such as underdetermination and induction). Therefore, we need to use other considerations when comparing hypotheses. The traditional considerations are theoretical virtues, such as simplicity, fruitfulness, or applicability to human needs. These theoretical virtues are carriers of values: we call simplicity a "virtue" because we value simplicity.

But which kinds of values are inherent to assessment of hypotheses? Often, a distinction is drawn between two kinds of virtues: epistemic and non-epistemic. Typically, epistemic virtues are theoretical characteristics that are valued because they promote epistemic goals, such as the attainment of truth, knowledge, understanding, or explanation (for this reason, the epistemic virtues are sometimes just called "epistemic values"). For example, if simpler theories are more likely to satisfy our epistemic goals, then simplicity is epistemically valuable and is an epistemic virtue. Other traditional epistemic virtues include internal consistency, consistency with other theories, empirical adequacy, and explanatory power. Non-epistemic virtues are theoretical characteristics that are valued because they promote non-epistemic goals, such as creating a more just society or making money.

The aim of this paper is to support the view that non-epistemic values are essential to assessment of hypotheses. Other arguments have been criticized for depending on particularities that are special to political and practical circumstances. My argument focuses on accuracy and draws from a theorem from machine learning, called the "No Free Lunch Theorem". While there are surely differences between human reasoners and algorithms, drawing from a mathematical theorem avoids some of the difficulties faced by other arguments because it is independent of human contingencies and contextual particularities.

The structure of this paper is as follows. I start by overviewing the main existing lines of argument in section 1. Then, in section 2, I motivate the focus on accuracy and explain which aspect of accuracy I am targeting. In section 3, I give a brief and informal overview of the basics of the No Free Lunch theorem (you can find a formal statement of the theorem in the appendix). I explain why NFL supports the claim that non-epistemic values are essential to theory choice in sections 4, 6, and 7. In brief, I argue that (i) NFL entails that predictive accuracy is insufficient to

discriminate between hypotheses on its own, (ii) NFL challenges our ability to epistemically justify the usage of other traditional epistemic virtues in theory choice, and (iii) a natural way to overcome NFL's challenges is to use non-epistemic values in theory choice. In section 5 I respond to an objection. In the last section, section 8, I explain why the epistemic challenge that NFL poses is different from Hume's problem of induction, Goodman's problem of induction, and other forms of underdetermination.

1. Three existing lines of argument

One line of argument for the claim that non-epistemic values are essential to assessment of hypotheses is based on an analysis of the history of science. Kuhn (1962) argues that scientific paradigms drastically differ in standards, language, values, modes of engaging with the world, institutions, etc. It's not the case that a scientific paradigm is replaced when another scientific paradigm is shown to satisfy epistemic goals better because paradigms differ in their epistemic goals. You can only judge that the new paradigm is better than the old one after you have already switched to the new epistemic goals, the new ways of thinking, the new language, the new mode of engagement with the world, and so on. These kinds of changes are partially driven by non-epistemic factors. In this respect, scientific revolutions are similar to political revolutions and religious conversions. One line of criticism against this argument is due to Kuhn's reliance on history. Some (e.g. Toulmin, 1970; Lauden, 1990; Bird, 2013) argue that Kuhn's historical account of science is incorrect. Others (e.g. Boghossian, 2006) argue that, even if Kuhn's historical account is right, it is not enough to support his conclusion. The problem is that Kuhn's conclusion is universal, about all paradigm shifts, but his evidence is empirical and contingent. Other lines of criticism focus on conceptual, metaphysical, and linguistic issues. For example,

some argue that Kuhn's view entails a metaphysical and/or linguistic relativism which is untenable (e.g. Davidson, 1973).

The second line of argument is based on inductive risk. Rudner (1953), Douglas (2009), Steel (2013), and others argue that decisions on how much evidence is sufficient to confirm or refute hypotheses require non-epistemic values. The basic claim is that there is no strict epistemic standard for how much evidence is enough to accept a theory, and the decision depends on the risks involved. If there are dire consequences for wrongfully accepting or rejecting the hypothesis, we will require more evidence before deciding. Whether the consequences are dire depends on what we care about and what is at stake, and therefore decisions to accept hypotheses are value-laden. Opponents of this argument put pressure on what "acceptance" means in a scientific context. Broadly, the objection is that while inductive risk is relevant in practical contexts, it is irrelevant when the practical consequences are far removed (e.g. Lacey 1999, 2017; Levi 1962).

The third line of argument blurs the distinction between epistemic and non-epistemic virtues. Longino (1990, 1996, 2002, 2014), Okruhlik (1994), and others argue that traditional epistemic values are sometimes just manifestations of non-epistemic values. Focusing on examples from social science and biology, authors promoting this line of argument argue that traditional epistemic values can be politically loaded. Think about consistency with other theories, for example. A theory may be biased, but still consistent with other theories because the other theories share the same bias. For example, for a long time it was thought that the egg is passive in the process of fertilization. One thing that was going for this view is that it was consistent with sociological theories and views about gender roles. But this consistency was due to the fact that both biological and sociological theories were influenced by social views on the

passivity of women (The Biology and Gender Study Group, 1988). One may wonder how well blurring the line between epistemic and non-epistemic virtues generalizes. The objection is that traditional epistemic values can be neutral in contexts that are less politically loaded than social science and biology, such as quantum mechanics or astronomy.

In sum, these three lines of argument emphasize certain historical, practical, or political contexts. Therefore, they are vulnerable to two objections. First, if non-epistemic values happen to influence theory choice only in specific cases, perhaps this only shows that people are sometimes imperfect; it doesn't seem to show that non-epistemic values are essential to reasoning itself. Second, if the specific cases involve subject matters with obvious practical or political implications, then one might object that non-epistemic values are irrelevant for subject matters that are theoretical and politically neutral. Instead of thinking of ways in which proponents of these arguments can respond, in this paper I support the claim that non-epistemic values are essential to assessments of hypotheses by constructing a new argument. I start by focusing on accuracy.

2. Focus on accuracy and average expected error

Accuracy is one of the most influential theoretical virtues today, and arguably even the most influential. In particular, accuracy is highly regarded in feminist epistemology and philosophy of science. Longino (1996), Rolin (2017), and others highlight its usage and importance for feminist as well as traditional scientists. In formal epistemology, members of the "accuracy first" school, such as Pettigrew (2016), argue for the priority of accuracy over other theoretical virtues. Moreover, many of the other virtues seem to inextricably involve accuracy. Think about explainability, fruitfulness, and applicability for human needs, for example. Can we

say that a theory explains, is fruitful, or helpful for human needs without addressing its accuracy at all, e.g. even if all of its predictions are false? This is not plausible, at least prima facie. The centrality of accuracy, both in common perception and in its involvement with the other virtues, makes it a worthy focal point in discussing theory choice.

Spelling out what accuracy means is notoriously tricky. I will use average expected error as an approximation of "more likely to be accurate". As I will argue later, average expected error is flexible in that it can capture a wide range of accuracy measures. The result is a broad interpretation of accuracy.

What is average expected error? The error of a hypothesis is the distance between its predictions and the results of observations. To illustrate, suppose I have a hypothesis that my friend Ankita has the flu. If she really does have the flu, we might want to say that I am maximally correct. For the sake of the illustration, say that this means that my error is 0. I.e.:

$$e(\text{"Ankita has the flu"}, \mathit{Ankita\ has\ the\ flu}) = 0.$$

However, until Ankita sees a doctor I won't know whether she really has the flu so this error wouldn't be helpful in assessing my hypothesis in advance. It would be more helpful to calculate the expected error of my hypothesis. Here's one way to go. If Ankita is has the flu, my error is 0 as discussed. If Ankita doesn't have the flu, my hypothesis is false. We can say that I am maximally mistaken and my error is 1. I.e.:

$$e(\text{"Ankita has the flu"}, \mathit{Ankita\ doesn't\ have\ the\ flu}) = 1.$$

We get the *expected* error by averaging over my errors on what we take to be the possible outcomes. For example:

EE("Ankita has the flu")

$$= \frac{e(\text{"Ankita has the flu"},\ Ankita\ has\ the\ flu) + e(\text{"Ankita has the flu"},\ Ankita\ doen't\ have\ the\ flu)}{2}$$

$$= \frac{0+1}{2} = 1/2$$

Average expected error allows us to measure accuracy with respect to multiple predictions. Suppose my policy is to always predict my friends have the flu when they ask me. My average expected error would be the average of the expected errors:

$$\frac{\sum_i\ EE(\text{"friend i has the flu"})}{\text{number of friends}}$$

Average expected error is a good approximation of "more likely to be accurate". Intuitively, if some hypothesis h_1 has a lower average expected error than some other hypothesis h_2, we would say that h1 is more likely to be accurate than h_2. This interpretation of accuracy is very broad. In the flu example, there were only two kinds of errors, 0 or 1, and we used the absolute difference to compare between the prediction and the true value. However, for the purposes of NFL, we can use any accuracy measure that is a function of predicted value and the real value alone.

The No Free Lunch theorem (NFL) compares between algorithms based on average expected error. In the next section, I present the theorem using simple cases and drop simplifying assumptions as I go. For a more formal statement of the theorem, see the appendix.

3. The No Free Lunch theorem – the basics

Machine learning algorithms look for patterns in data and use them to make predictions. Unlike hand-coded algorithms, machine learning algorithms use data to extract the values of crucial parameters themselves. Two of the most well-known machine learning tasks are

regression and classification (both are kinds of what is called "supervised learning"). Regression algorithms predict numerical values given some input. For example, regression algorithms can predict the price at which a house will sell, based on information about previous sales. Classification algorithms specify to which of K categories an input belongs. For example, classification algorithms can be used to recognize different people in photos. In both cases, and in all machine learning algorithms, some of the key parameters the algorithm requires to perform its task are hand-coded in advance, but some are learned by the algorithms themselves from past data. [1]

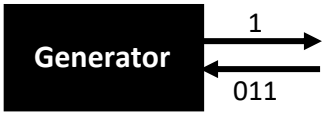
A family of impossibility results by the name of "No Free Lunch" theorems reveals limitations on what algorithms can do. In particular, Wolpert's (1996) No Free Lunch Theorem roughly says that no learning algorithm universally performs better than any other.[2] To informally explain what that means, imagine the following thought experiment: run some algorithm on some dataset. For example, run a certain regression algorithm on some housing prices dataset. Choose some error measure and use it to assess the algorithm's predictions on this dataset, on a second dataset, and on all possible datasets. What NFL says is that no matter which error measure you chose, the average expected error on all possible datasets is the same for all regression algorithms. This includes the most sophisticated learning algorithms, algorithms that make random guesses, and all other algorithmic ways to make predictions based on information.

For a more concrete illustration, suppose you want to diagnose whether a patient has a certain kind of flu based on three symptoms: fever, cough, and runny nose. [3] You construct an algorithm that takes a binary sequence of three digits as input: the first digit signifies whether the

[1] For a more technical, yet accessible, introduction to machine learning algorithms see Russell & Norvig (2010).
[2] To be more precise, Wolpert's (1996) theorem applies to supervised learning algorithms.
[3] This thought experiment is loosely based on the adversary argument from Culberson (1998) and the OR/XOR example from Wilson & Martinez (1997).

patient has a fever, the second digit signifies whether they have a cough, and the third digit signifies whether they have a runny nose. “1” means that the patient has the symptom, and “0” means that they don’t. For example, “011” means that the patient doesn’t have a fever, does have a cough, and does have a runny nose. The algorithm’s output is whether the patient has the flu. “1” means that they have it, “0” means that they don’t. The task of the algorithm is classification, as the goal is to sort patients into two categories – having the flu and not having it.

The story about the flu doesn’t matter much for the algorithm. What matters is that the algorithm gets a three-digit binary sequence, and outputs either 1 or 0. The dataset could represent anything – symptoms and disease diagnosis, studying habits and whether a student will pass a test, financial profile and whether a client will pay back a loan, and so on. In the general case, we can imagine an interaction with a generator. We produce three-digit binary sequences, e.g. 011, and send them to the generator. For every new sequence, the generator sends back either 0 or 1. We can presuppose that for every sequence that was already evaluated, the generator will send back the same digit that was sent in the past. To illustrate:



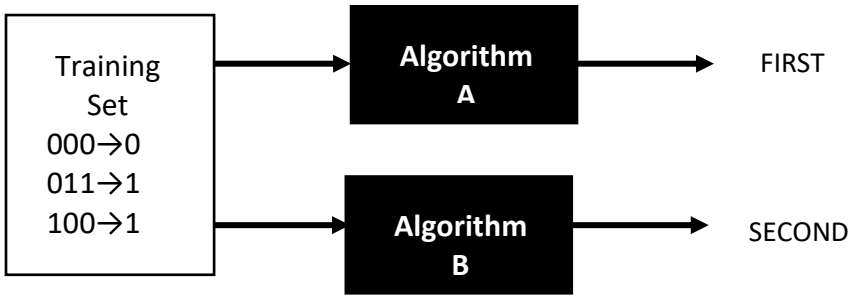
We can also presuppose that the generator works in accordance with a function, and our algorithm’s job is to guess what that function is. Further, we can presuppose that the generator doesn’t change. That is, the function won’t change. In some cases, we may have some information about which function the generator is using. However, in the general case, we make no further assumptions about the generator and the patterns it produces. That is the case the NFL targets – making no assumptions about the problem we are trying to solve, except that the generator stays the same.

The generator plays the role that nature plays when we think about experimentation in science. We interact with it, and it produces some patterns. In this rendering, the stipulation that the generator always uses the same function means assuming that the regularities in nature, or the laws of nature, are not changing. The stipulation that we make no assumptions about how the generator works beyond consistency with its past (in the sense that if 001 outputted 1 in the past, it will do so every time it is inputted) means that we make no assumptions about the content of the regularities we are trying to discover beyond consistency with their past.

Before we sit down to write our algorithm, we interact with the generator. We send some inputs to it and write down the responses we get. The resulting repository of past responses is called a training set. Suppose the training set is:

Training Set: 000→0, 011→1, 100→1

Now, suppose you constructed two different algorithms, A and B, to diagnose the flu. You run both algorithms on the training set and they produce two different hypotheses, FIRST and SECOND:



Algorithm A → FIRST – the output is identical to the first digit of the input.

Algorithm B → SECOND – the output is identical to the second digit of the input.

For example, FIRST predicts that the output for 011 is 0, and SECOND predicts that the output for it is 1. In flu diagnostics' terminology, FIRST says that having a fever is a necessary and

sufficient condition for having this type of flu, and SECOND says that having a cough is a necessary and sufficient condition for having it.

How do we determine which algorithm is better? NFL focuses on minimizing average expected error. 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.

In our example, comparing the hypotheses' errors on inputs that appear in the training set is not helpful because both hypotheses do equally well on it: they both get 2/3 of the training set right: FIRST is right about 000 and 100 and wrong about 011. SECOND is right about 000 and 011 and wrong about 100. But, in any case, the error on the training set is not reliable. The problem is that the hypotheses were constructed to fit that training set in particular. They fit not only the data but also any noise that might be influencing the sample we happen to have. In the example above we assumed that there is no noise, but we can't assume so in the general case. In general, high accuracy on the training set might indicate that the hypothesis is, in a sense, too ad hoc, and will not generalize well. Thus, NFL focuses on hypotheses' accuracy on inputs that are not included in the training set (see Wolpert, 1996, especially pp. 1345-1348, for further discussion).

The error on inputs that aren't in the training set is called the off-training-set error (OTS):

Off-Training-Set error (OTS): the average expected error on inputs that are not included in the training set.

For example, this is how to calculate A's off-training-set error:

1. Choose one possible input that is not in the training set.

For example, let's choose 010.

2. Choose an error measure

   NFL holds no matter which error measure we choose , so it doesn't matter which one we use for the purpose of illustrating NFL.

   Let's decide that the error is the difference between the prediction and the real output. For example, if the algorithm guesses 0 and the real output is 1, the error is |0-1|=1.

3. Calculate the algorithm's expected error on this input.

   For example, let's calculate A's expected error on 010. There are two possible outputs, 0 and 1, and A's guess is 0. If the output is really 0, A's error is 0. However, if the output will be 1, A's error is 1. Overall, A's expected error on this input is the average of the potential errors. Since we're not making any assumptions about the problem we are trying to solve, we're not going to use any weights:

$$ee(A, 011?) = \frac{\sum_{y \in Y} e(A, y)}{|Y|} = \frac{e(A, 0) + e(A, 1)}{2} = \frac{0 + 1}{2} = 1/2$$

   Where: Y is the set of all possible outputs, i.e. 0 and 1; |Y| is the number of possible outputs, i.e. 2.

4. Calculate the algorithm's expected error for all possible inputs not in the training set.

   In this case, there are 5 possible inputs not in the training set: 010, 001, 101, 110, 111. By the same reasoning in the previous step, A's expected error on all possible inputs is the same: 1/2.

5. The Off-Training-Set error is the average of the expected errors of all possible inputs not in the training set.

   For example, A's OTSE is 1/2:

$$OTSE(A) = \frac{\sum_{x \in X} ee(A, x)}{|X|} = \frac{ee(A, 010?) + \cdots + ee(A, 111?)}{5} = \frac{5 \cdot 1/2}{5} = 1/2$$

Where: X is the set of all possible inputs not in the training set.

The same reasoning applies to all algorithms. For example, B's OTS error is also 1/2. To see this, consider the same possible input as before, 010. B's prediction is that the output for it will be 1. If the true output is 1, B is correct and its error on this input is 0. However, if the actual output is 0, B is wrong and its error is 1. Therefore, B's expected error on this input is (1+0)/2=0.5. The same is true for all possible inputs, and so B's OTS error is also 0.5.

The result of the above discussion is that if we make no assumptions about the generator in the example except for consistency with the past, then all algorithms have the same average expected error in trying to discover the regularities it produces.

Time to drop some simplifying assumptions. First, in the example, the error of a hypothesis on a given input was the difference between the prediction and the actual output. However, we could use any error measure. Second, in the example, the inputs and outputs were binary, there was no noise, and the function was deterministic. NFL doesn't require making these assumptions. There are no restrictions on the kinds of inputs and outputs, the function may be non-deterministic, and the training set may include noise. Third, the algorithms I compared, A and B, perform equally well on the training set. However, NFL applies regardless of performance on the training set. Even algorithms that perform poorly on the training set have the same average expected error as all others. The general result is that, if we make no assumptions

about the generator, all algorithms have the same average expected error. That is the point of NFL. [4]

For a formal derivation of NFL, see Wolpert (1996). However, 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 assumptions 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.

Notice that NFL doesn't require that we assume that the world is structureless. Rather, it presupposes the existence of some structure – that the same input will result in the same output. In this sense, that NFL is less skeptical than Hume's problem of induction, which questions the justification of assuming that the future is like the past. I'll further address similarities and differences between NFL, problems of induction, and underdetermination in section 8. Before, I spell out the relevance of NFL to theory choice in sections 4 and 6, and 7, and discuss an objection in section 5.

[4] Since NFL allows to use any error measure that is only a function of the relevant values and prominent distance measures are also functions of the same values, we can manipulate NFL's results to bear on popular error measures. For example, suppose we use square Euclidian distance as our error measure for NFL: $|Y_F(\mathrm{x}) - Y_H(\mathrm{x})|^2$, where $Y_H(\mathrm{x})$ is the algorithm's prediction for input x and $Y_F(\mathrm{x})$ is the true output. Then, according to NFL, all algorithms have the same average expected error: $\sum_{x\in \mathrm{X}}|Y_F(\mathrm{x}) - Y_H(\mathrm{x})|^2 \,/|X|$ (where X is the set of all relevant inputs). However, since |X| is just the number of items in X, the quantitiy $\sum_{x\in \mathrm{X}}|Y_F(\mathrm{x}) - Y_H(\mathrm{x})|^2$ is also the same for all algorithms. But $\sum_{x\in \mathrm{X}}|Y_F(\mathrm{x}) - Y_H(\mathrm{x})|^2$ is the Brier inaccuracy measure. Therefore, we get that the predictions of all algorithms are equally inaccuate relative to the Brier inaccuracy measure.

4. NFL and theory choice

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. For example, in the case above the comparison is between the two hypotheses produced by algorithms A and B: FIRST and SECOND. 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 comparisons of hypotheses rather than comparisons 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? It is perhaps for this reason that Wolpert himself argues that NFL has wide-ranging implications for science. In the context of science, he thinks of different algorithms as analogous to different scientists who are "producing accurate theories from data" (2012, p. 5).

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 if 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 average expected error. If average excepted error is a measure of how likely a hypothesis 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.[5]

What could we use to supplement predictive accuracy? We could use accuracy on the data we already have, i.e. the training data. However, as discussed above, a good fit with the training data may be misleading because it may be a result of a good fit with the noise.

Another option is to use other traditional epistemic virtues, e.g. simplicity, coherence with other theories, explainability, or understanding. NFL poses a challenge to this approach. What makes these virtues valuable? Why use them for theory choice? Some justify the usage of these virtues by arguing that they help us discover the truth. For example, Swinburne argues that simple hypotheses are more likely to produce true predictions, all else being equal:

> I seek…to show that—other things being equal—the simplest hypothesis proposed as an explanation of phenomena is more likely to be the true one than is any other available hypothesis, that its predictions are more likely to be true than those of any other available hypothesis, and that it is an ultimate a priori epistemic principle that simplicity is evidence for truth (Swinburne 1997, p. 1).

Similarly, Lipton (2004) argues that the epistemic value of explanation is that "the explanation that would, if true, provide the deepest understanding is the explanation that is likeliest to be true" (p. 61). For this reason, he thinks that "the exciting promise of Inference to the Best explanation" is "showing how explanatory considerations are our guide to truth" (p. 62). Lipton argues that this view of the inference to the best explanation helps make sense of the fact that scientists use aesthetic considerations such as theoretical elegance, simplicity, and unification in making scientific inferences. It makes sense that they use them because these considerations are marks of good explanations, which are guiding us to truth (p. 66).

[5] See Dotan (forthcoming) for more discussion of the implication of the No Free Lunch theorem on using accuracy in theory choice.

Swinburne and Lipton use the language of "more likely to be true". However, if true predictions are accurate predictions and more likely to be true is more likely to be accurate, NFL contradicts their arguments. NFL shows that all hypotheses, no matter their properties, are equally likely to be accurate if we are unwilling to make any assumptions about the regularity about which we are hypothesizing. In particular, simple and explanatory hypotheses are as likely to be accurate as complex and non-explanatory hypotheses.

The challenge that NFL poses to theory choice is that, if we aren't making further assumptions about the problems we are trying to solve, accuracy considerations alone are insufficient for theory choice. If we want to use epistemic virtues other than accuracy, we need to justify them without relying on accuracy. One option is to explain what "more likely to be true" means not in terms of accuracy. However, disconnecting between truth and accuracy is difficult. Prima facie, a hypothesis that produces inaccurate predictions is false. Another option is to justify the usage of epistemic virtues using their aesthetic features. Perhaps having a deep explanation or an elegant theory are valuable on their own, regardless of truth. As Lipton points out, scientists do use these aesthetic features for theory choice in practice. However, if we use theoretical virtues due to their aesthetic value, then theory choice relies on non-epistemic considerations. Thus, NFL's challenge to theory choice is giving an epistemic justification to appealing to theoretical virtues without appealing to accuracy, or at least predictive accuracy.

I argue later in the paper that NFL poses a challenge to theory choice even if we are happy to make assumptions about the regularity we are trying to discover. The reason is that NFL applies to whatever methods we use to decide which assumptions to make. I discuss this more general case in sections 6 and 7. However, before that, I consider an objection to NFL.

5. Objection: What about validation error?

NFL focuses on measuring errors on inputs that do not appear in the training set. As discussed, the reason is that the algorithm finds a function that fits the entire training set, including any potential noise. When the training set error is too low, it might indicate that the hypothesis will not generalize well, i.e. won't produce good predictions for new data. A standard solution in machine learning is to evaluate hypotheses not on their training set error, but rather on their validation error.

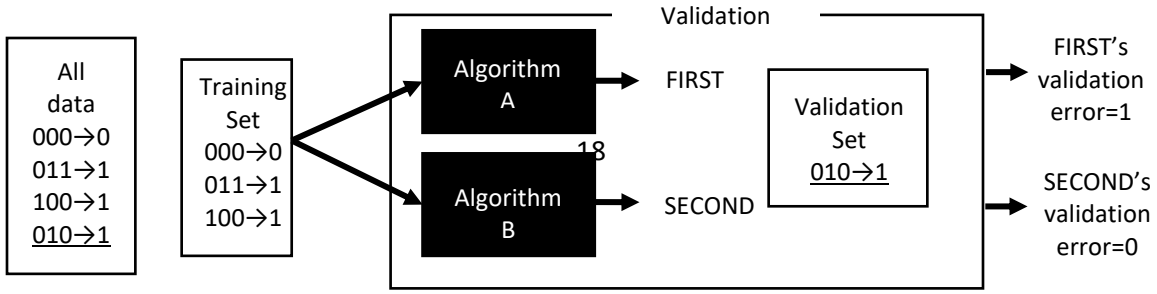
The general idea is to split the data we have into two and only use a part of it to train the algorithm. The rest, which is called a "validation set" is used to evaluate the algorithm. For example, suppose we start with all the data we had before, plus one additional data point: 010→1. We divide the data set into two: a training set which is identical to the one we had before, and a validation set consisting of the new data point.

Entire data set: 000→0, 011→1, 100→1, 010→1

Training Set: 000→0, 011→1, 100→1 (same as before)

Validation Set: 010→1

When we train our algorithms, A and B, on this training set, they will come up with the same hypotheses, FIRST and SECOND. However, now we can compare them using the validation set, the data we haven't used to train A and B. The error on the new data, the validation set, is called "validation error". We calculate the validation error by comparing the prediction of the hypothesis and the information in the validation set. We can see that FIRST's validation error is 1 since it predicts 010→0, but the validation set is 010→1. SECOND's validation error is 0 since it predicts 010→1. Schematically:



Should we choose between our hypotheses based on minimizing validation error? The problem is that there are many other ways to do validation. To give just a few examples, we could single out a different data point as the validation set. Alternatively, we could use multiple data points as a validation set, instead of just one like in the example above. Moreover, we could run the validation process multiple times, each time splitting the data differently, and average over the results for each hypothesis. Arlot and Celisse (2010) survey various validation algorithms in use today, focusing on their strengths and weaknesses. They argue that different validation algorithms work well in different circumstances. Given that different algorithms are suited to different situations, it is at least not typical that various validation algorithms will give the same recommendation.

We need to choose between validation algorithms somehow. This brings us back to the same problem with which we started – how should we choose between algorithms/the hypotheses they produce? For the same reasons as before, comparing their OTS error is a good option. Validation algorithms are still algorithms giving predictions based on a training set, just like A and B. The only difference is that they use the training set in a more sophisticated way. However, just like all other algorithms, they are susceptible to overfitting. Illustrating this point, using a series of experiments, Schaffer (1993a; 1993b) has shown that using validation techniques sometimes leads to worse performance on new data than not using validation at all. However, if we are comparing between validation algorithms based on their OTS error, we are

back in NFL territory. All algorithms, including validation algorithms, have the same average expected error.

This result may be surprising at first glance. Certain machine learning algorithms are successful in practice, much more than others. In particular, validation techniques are widely used, usually with great success. However, these observations do not conflict with NFL. NFL only determines that all algorithms perform equally well on average. Some algorithms can still be better than others on individual problems. Thus, many take the moral from NFL to be that we need to know more about the problems we encounter in reality. Using this information, we could explore which algorithms work well for which types of problems (e.g. Gomez and Rojas 2016; Fernadez-Delgado et al., 2014), or which assumptions we can safely make about our problems (e.g. Igel & Toussaint, 2004; Lattimore and Hutter, 2011). However, as I discuss in the next section, the methods we use to discover the assumptions are also subject to NFL.

6. Implications for theory choice with assumptions

The discussion so far made no assumptions about the generator, i.e. no assumptions on the problem we are trying to solve. NFL uses average expected error with equal weights because the point is that all hypotheses have the same performance if we average over *all* possible problems. However, some algorithms do better than others on a given set of problems. You might want to say that we could compare between algorithms or hypotheses using epistemic considerations alone, if we restrict our attention to their performance on the kinds of problems that we encounter in the actual world, which is a subset of all possible problems. The way forward would then be finding the right assumptions to make to restrict our attention to the right kinds of problems.

The question is how to decide which assumptions to make, i.e. which problems to prioritize. We could construct another algorithm to analyze data to come up with more basic hypotheses on what our world is like. However, NFL would apply to this algorithm as well, and therefore this strategy just pushes the bump under the rug.

Therefore, while we must choose a subset of all possible problems to escape NFL, we cannot choose our assumptions based on accuracy considerations alone. We could supplement accuracy with other traditional epistemic virtues, such as simplicity or explainability. However, as discussed above, NFL challenges the ability to provide a pure epistemic justification for using these virtues. It is not the case that the assumptions that are simplest or most explanatory will guide us to truth in the sense of producing more accurate predictions.

7. Putting things together

We learn from NFL that the standard of accuracy is insufficient to discriminate between hypotheses. If we make no assumptions about the regularities we are trying to discover, all hypotheses are equally likely to be accurate. If we want to make assumptions, accuracy is not enough to choose assumptions. NFL invites us to explore which considerations we do want to use for theory choice, and why. Whichever consideration we use, e.g. simplicity or explanatory power, the justification cannot be based on accuracy alone. NFL applies to all hypotheses, simple or complex, explanatory or not.

NFL supports the claim that non-epistemic values are necessary for theory choice in three ways. First, accuracy, possibly the most influential epistemic virtue, is insufficient to discriminate between hypotheses. Second, NFL challenges the ability to provide pure epistemic justifications for using other traditional epistemic virtues for theory choice. Third, non-epistemic

values are natural candidates to supplement accuracy or other considerations. NFL only applies when we compare the accuracy of theories on all possible problems. However, some hypotheses are more likely to be accurate than others once we restrict our attention to some problems. Since we can't restrict the set of problems we consider based on accuracy, it makes sense that we restrict ourselves to measuring accuracy on the problems we care about. That is a value-based decision.

The argument from NFL avoids the vulnerabilities of other arguments revealing the role of non-epistemic values in theory choice. Unlike Kuhn's argument, the argument from NFL is not based on the contingent history of science and is not committed to relativism. Unlike the argument from inductive risk, the argument from NFL isn't sensitive to contexts. The point is that all hypotheses have the same average expected error, no matter whether accuracy measurements are used for practical or theoretical purposes. Unlike arguments blurring the distinction between epistemic and non-epistemic virtues, the argument from NFL is not focused on politically loaded subject matters. Learning algorithms are an idealization of inductive reasoning, and NFL is an impossibility theorem that applies to all applications of inductive reasoning. Nothing is specific to any particular field.

Some think that non-epistemic influences on assessment of theories are inherently bad. For example, Lacey argues that the cost of admitting non-epistemic values as an essential component of hypothesis assessment is losing "all prospects of gaining significant knowledge" (1999, p. 216), and exposing ourselves to the dangers of wishful thinking or to the "back and forth play of biases, with only power to settle the matter". Of course, even if it is bad for non-epistemic values to influence science, it doesn't mean that they don't. In my view, to use Steel and Elliott's (2017) metaphor, values in science are like knives in cooking. They can be

dangerous if used irresponsibly, but we are very limited if we don't use them at all. Various views have been developed to explain how we can handle non-epistemic values responsibly. For example, Longino (1990, 2002), Rolin (2017), and others argue for versions of a "value-management" ideal of science. According to them, objectivity does not stem from the judgments of individuals, but rather from community practices. To be objective, scientific communities must be open and responsive to the right kind of critical discourse. NFL gives us another reason to develop conceptions of objectivity and of science which, like this one, manage the influence of non-epistemic values.

8. Comparing NFL with problems of induction and underdetermination

In closing, I'd like to discuss the uniqueness of NFL's epistemic challenge. You might wonder whether NFL reduces to the problem of induction or underdetermination. However, first, even if NFL is reducible to a familiar epistemic puzzle, it's still helpful in noticing that theory choice involves non-epistemic values while overcoming challenges other arguments face. Second, as I argue next, NFL is different from Hume's problem of induction and from other versions of underdetermination.

8.1 NFL and Hume's problem of induction

NFL is not discussed much from a philosophical perspective. When it is, it is often assumed to be a manifestation of Hume's problem of induction (e.g. Domingos, 2012; Giraud-Carrier & Provost, 2005; Korb, 2004; Schaffer, 1994; Wolpert, 1996, 2012). Yet the two issues are distinct.

Hume's problem of induction is about how to justify inductive inferences. For example, how to justify moving from the following premise to the following conclusion (Henderson, 2020):

I. All observed instances of bread (of a particular appearance) have been nourishing.

II. The next instance of bread (of that appearance) will be nourishing.

Hume is interested in the justification of moving from premise (I) to conclusion (II). First, no deductive argument can be used to do so. The reason is that in deductive arguments the falsity of the conclusion is inconsistent with the premises, but the negation of (II) is consistent with (I). Second, no non-deductive arguments can be used to move from (I) to (II), because it would rely on circular reasoning. Moving from (I) to (II) requires presupposing what is sometimes called the "principle of the uniformity of nature", according to which the unobserved is similar to the observed. To justify the principle of the uniformity of nature we would need another non-deductive argument, which would itself presuppose the principle of the uniformity of nature. Therefore, according to Hume, no argument can be given to justify inductive inferences.

NFL is similar to Hume's problem of induction in that both point to difficulties in moving from past observations to predictions. However, NFL is different from Hume's problem because NFL doesn't question the principle of the uniformity of nature and doesn't look for a justification for a general form of inference.

First, NFL presupposes the uniformity of nature. In the terminology of this paper, presupposing that nature is uniform is presupposing that the generator, which represents nature, stays the same. For example, all patients with the same symptoms have the same diagnosis. If the dataset was about the nourishment of bread, the assumption would have been that all instances of bread of the same appearance are equally nourishing. The issue that NFL is concerned with is

what we can infer from instances of bread of a certain appearance on instances of bread of a *different* appearance. Because of this difference, solutions to Hume's problem of induction that are looking to justify the uniformity of nature are irrelevant to NFL.

Second, NFL doesn't question a general form of inference. Some solutions to Hume's problem of induction reject the need to justify induction at all. For example, Strawson (1952) argues that inductive inferences are foundational. Asking whether induction is valid is senseless, like asking whether the legal system is legal. In his alternative account of induction, inductive support simply consists in observing enough positive instances of the inductive claim. In other words, inductive standards are baked into the meanings to terms such as "inductively justified". There is no need to give any further support or justification for inductive arguments. However, even assuming that solutions like Strawson's deflate Hume's problem of induction, they do not deflate the argument from NFL. Admitting that inductive reasoning is foundational or built into the meaning of terms like "justification" doesn't point to any epistemic reasons to choose between inductive hypotheses, which is the point NFL presses on. NFL is simply not about the justification of inductive inferences in general.

Other approaches to Hume's problem of induction may be conducive to the NFL discussion. What I have in mind here are solutions that account for induction non-epistemically. Take Hume's own solution for example. Hume argues that we accept the principle of uniformity of nature not because of any reasoning, but because of some psychological mechanisms. For example, when the sight of fire has generally been accompanied by a feeling of heat, our instinctual mental mechanisms will lead us to expect heat when we see fire in the future. Applying this to NFL, some instinctual mental mechanisms may lead us to have certain expectations about the fluwhich we may use in designing our algorithm. NFL supports the claim

that these expectations don't have epistemic value (in the sense that using them won't lead to lower average expected error), but they may still constitute good non-epistemic reasons to choose between hypotheses. Hume's solution is only one example of how we can use non-epistemic considerations to choose between inductive hypotheses. The right kind of non-epistemic considerations to use, if there are any, may or may not piggy-back on solutions to Hume's problem of induction.

8.2 NFL and other forms of underdetermination

The case I focused on in the discussion above was the comparison between FIRST and SECOND, two hypotheses that do equally well on existing data. The difficulties in choosing between them are difficulties of underdetermination, as the available data is insufficient to determine which hypothesis is true. Therefore, you might think that NFL reduces to some form of underdetermination, even if not to Hume's problem of induction. However, I argue that NFL is different from familiar cases of underdetermination in that it extends the class of underdetermined theories.

First, you might think, like Lauc (2018), that NFL is a rediscovery of Goodman's new riddle of induction. Goodman's new riddle of induction is concerned with how artificial predicates like "grue" can give rise to underdetermination. The point is that the fact that all emeralds observed before time t have been green is insufficient to determine whether emeralds observed after t will also be green. The reason is that our observations are consistent with various hypotheses that make different predictions, such as "all emeralds are green" and "all emeralds observed before time t are green, but the rest are blue". Another form of underdetermination is given by van Fraassen (1980). van Fraassen is concerned with underdetermination between

theories that make the same predictions. The point is that theories can make the same predictions but still differ, for example by making different untestable empirical assumptions. Such theories are underdetermined in principle.

NFL's variety of underdetermination extends the class of underdetermined theories. Goodman, van Fraassen, and others only consider the set of hypotheses that are equally supported by observations. However, NFL also includes hypotheses that do less well on the existing data. NFL determines that *all* algorithms have the same average expected error, and that includes algorithms that produce hypotheses that are incompatible with the training set.

For example, consider algorithm C, which produces the hypothesis Least Common[6]:

Algorithm C → Least Common (LC): the output is the digit that is least common in the input. For example, 011 corresponds to 0 because the digit 0 appears less than the digit 1 in the input, and 000 corresponds to 1. In the flu example, LC says that having at most one of the symptoms is a necessary and sufficient condition for having this type of flu.

LC gets only 1/3 of the samples in the training set right (recall that our training set consists of 000→0, 011→1, 100→1). This is worse than FIRST and SECOND, which get 2/3 of the training samples right. Yet, LC has the same off-training-set error as FIRST and SECOND. Consider LC's error on 010. LC predicts that the output will be 1. If the output turns out to be 1 LC is correct and has 0 error, but if the true output is 0 LC's error is 1. On average, LC's error on 010 (and on any other possible input) is 1/2. The same is true for all possible inputs not in the training set and therefore C's OTS error is 1/2, just like FIRST and SECOND.

Should we favor FIRST and SECOND over LC? NFL highlights that accuracy on past and future data can come apart. If we care only about accuracy with respect to future data, then

[6] Based on the OR/XOR example from Wilson & Martinez (1997).

there are no epistemic reasons to favor FIRST and SECOND over LC, as they all satisfy the epistemic goal of predictive accuracy equally well. However, if we care about accuracy with respect to the data we already have then we do have reasons to favor FIRST and SECOND. The question we come back to again is – with respect to what do we want our theories to be accurate? The answer to this question depends on non-epistemic values.

9. Conclusion

NFL supports the claim that non-epistemic values are needed for theory choice in three ways. First, NFL shows that accuracy, which is a central epistemic virtue, is insufficient for discriminating between hypotheses (while the theorem is strictly speaking about predictive accuracy, the discussion about it includes a critique of accuracy on the existing data). Second, NFL challenges our ability to give a purely epistemic justification for using other virtues for theory choice, as illustrated on simplicity and explanatory power. Third, a natural way to overcome NFL's challenge is to use our values to restrict the set of problems we are trying to solve. Unlike other arguments in the vicinity, the argument from NFL is independent of human and contextual contingencies. In addition, I have shown that NFL is distinct from Hume's problem of induction and other forms of underdetermination

Acknowledgements:
For extensive feedback on this paper, I would like to thank Lara Buchak and Shamik Dasgupta. For comments on earlier drafts, I would like to thank Greyson Abid, Michael Arsenault, Nick French, Alvin Goldman, Tyler Haddow, Daniel Harman, Dan Hicks, John MacFarlane, Sven Neth, Emily Perry, Daniel Warren, and two anonymous referees for Synthese. For extensive conversations, I thank Gil Rosenthal. I am also grateful for comments and discussion from the conferences where versions of this paper were presented, including the 2020 Eastern APA, the 2019 Congress on Logic, Methodology, and philosophy of Science and Technology, the 2019 Canadian Society for the History and Philosophy of Science conference, the 2019 Society of

Exact Philosophy conference, the 2019 Values in Medicine, Science, and Technology conference, and the 2018 Philosophy of Science Association conference.

Appendix: The No Free Lunch theorem(s)

"No Free Lunch" is the name of a family of theorems. Differences between No Free Lunch theorems include differences between the kinds of algorithms they consider. For example, initially No Free Lunch theorems were proven for optimization algorithms (Wolpert and Macready, 1992). Wolpert (1996, 2001, 2012) proves No Free Lunch theorems for supervised learning algorithms, and this is what I have focused on in this paper. Schaffer (1994) gives an elegant formulation of Wolpert's main No Free Lunch Theorem for classification learning algorithms, based on a preprint of Wolpert (1996). In this appendix, I state Schaffer's formulation to illustrate what NFL theorems say more formally (Schaffer calls it the "Law of Conservation of Generalization of Performance"). See Montanez (2017, chapter 2) for a review of various No Free Lunch results, and see Schaffer (1994) and Wolpert (1996) for a proof of the theorem which I will state here.

We start with defining cases in a classification problem. Each case in a classification problem, A_i, is a vector of attributes. For simplicity, we assume that each component in the vector is a finite number. $\{A_1, \ldots, A_m\}$ is the set of all possible attribute vectors, where m is finite. C is a class probability vector, which defines the relationship between attribute vectors and classes. Each component of C, C_i, is the probability that a case with attribute A_i belongs to class 1. We assume that data is generated in the same way in training and testing a learner: attribute vectors are sampled with replacement according to an arbitrary distribution D and a class is assigned to them using C. We also assume that the training set contains n samples. A learning situation S is a triple (D, C, n).

The Generalization Accuracy of a learner (GA_L) is the expected prediction performance of a learner on cases with attribute vectors not represented in the training set. For example, the generalization accuracy of a random guesser in a two-class problem is 1/2 for every D and C. We

use the generalization accuracy of a random guesser as a baseline and define Generalization Performance of a learner (GPL) the difference between its generalization accuracy and the generalization accuracy of a random guesser:

$$GP_L = GA_L - GA_{random\ guesser}$$

Generalization performance greater than zero means better than chance performance. $GP_L(S)$ is the generalization performance of learner L in learning situation S.

Using the notation above we can write Schaffer's Law of Conservation of Generalization Performance:

$$\sum_S GP_L(S) = 0 \quad \text{, for every D, n}$$

In words, this law says that any positive performance by a learner in a certain learning situation must be exactly balanced by negative performance in other learning situations.

If we allow for the possibility of noise, then the law is properly written with an integral instead of a summation:

$$\int_S \ GP_L(S)ds = 0 \quad \text{, for every D,n}$$

In this case, the components of C are taken from the real interval [0,1] and the integral runs over the space $[0,1]^m$ of class probability vector. Without noise, the components of C are taken from $\{0,1\}$ and the summation runs over 2^m possible class probability vectors.

From the conservation law, it follows that all learners have the same average generalization performance if we average over all possible learning situations (or, as I put it, that all algorithms have the same expected error if we make no assumptions about the problem we are trying to solve). Here's why.

For any learner:

$$\sum_{S} GP_L(S) = \sum_{S} \left(GA_L(\mathrm{S}) - GA_{random\ guesser}(S)\right) = 0$$

Add $\sum_S GA_{random\ guesser}(S)$ to both sides and get:

$$\sum_S GA_L(S) = \sum_S GA_{random\ guesser}(S)$$

Divide by the number of learning situations and get the formulation that was used in this paper – that the average generalization performance of any learner L is equal, and in particular equal to that of the random guesser:

$$\frac{\sum_S GA_L(S)}{\#s} = \frac{\sum_S GA_{random\ guesser}(S)}{\#S}$$

References

Arlot, Sylvain, & Celisse, Alain. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. [2018 version available at arXiv:0907.4728 [math.ST]]

The Biology and Gender Study Group. (1988) The importance of feminist critique for contemporary cell biology. *Hypatia* 3.1, 61-76.

Bird, Alexander. (2012). The structure of scientific revolutions and its significance: An essay review of the fiftieth anniversary edition. *The British Journal for the Philosophy of Science*, *63*(4), 859–883. 10.1093/bjps/axs031

Boghossian, Paul A. (2006). *Fear of Knowledge: Against Relativism and Constructivism*. Oxford: Clarendon Press. 10.15713/ins.mmj.3

Culberson, Joseph. (1998). On the futility of blind search: An algorithmic view of "no free lunch." *Evolutionary Computation*, *6*(2), 109–127. https://doi.org/10.1162/evco.1998.6.2.109

Davidson, Donald. (1973). On the very idea of conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, *47*, 5–20. 10.1075/pc.3.1.12bus

Domingos, Pedro. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Dotan, Ravit. (forthcoming). What can we learn about accuracy from machine learning? *Philosophy of Science*

Douglas, Heather. (2009). *Science, policy, and the value free ideal*. Pittsburgh: University of Pittsburgh Press.

Elliott, Kevin, & Steel, Daniel. (Eds.). (2017). *Current Controversies in Values and Science*. Taylor & Francis.

Fernández-Delgado, Manuel, Cernadas, Eva, Barro, Senén, et al. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*, 3133–3181. 10.1016/j.csda.2008.10.033

Giraud-Carrier, Christopher, & Provost, Foster. (2005). Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper. *Proceedings of the ICML-2005 Workshop on Meta-Learning.*

Gómez, David, & Rojas, Alfonso. (2015). An Empirical Overview of the No Free Lunch Theorem and Its Effect on Real-World Machine Learning Classification. *Neural Computation*, *28*.

Henderson, Leah, "The Problem of Induction", *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/induction-problem/>.

Igel, Christian, & Toussaint, Marc. (2005). A no-free-lunch theorem for non-uniform distributions of target functions. *Journal of Mathematical Modelling and Algorithms*, *3*(4), 313–322.

Korb, Kevin B. (2004). Introduction: Machine learning as philosophy of science. *Minds and Machines*, *14*(4), 433–440. 10.1023/B:MIND.0000045986.90956.7f

Kuhn, Thomas S. (1962). The structure of scientific revolutions. Chicago; London: The University of Chicago Press.

Lacey, Hugh. (2017). Distinguishing between cognitive and social values. In Kevin Elliott & Daniel Steel (Eds.), *Current Controversies in Values and Science*. New York, NY: Routledge.

Lacey, Hugh. (1999). *Is science value free? Values and scientific understanding*. *Science Teacher* (Vol. 53). London and New York: Routledge.
Lattimore, Tor, & Hutter, Marcus. (2011). No free lunch versus Occam's razor in supervised learning. [ArXiv preprint available at arXiv:1111.3846]
Lauc, Davor. (2018). How Gruesome are the No-free-lunch Theorems for Machine Learning? *Croatian Journal of Philosophy*, *18*(54), 479–485.
Lauden, Larry. (1990). *Science and Relativism: Some Key Controversies in the Philosophy of Science*. Chicago, IL: The University of Chicago Press.
Levi, Isaac. (1962). On the seriousness of mistakes. *Philosophy of Science*, *29*(1), 47–65.
Lipton, Peter. (2004). *Inference to the best explanation* (2nd ed.). New York: Routledge.
Longino, Helen. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In Lynn Hankinson Nelson & Jack Nelson (Eds.), *Feminism, Science, and the Philosophy of Science* (pp. 39–58). Kluwer Academic Publishers.
Longino, Helen. (1990). *Science as social knowledge: values and objectivity in scientific inquiry*. Princeton University Press.
Longino, Helen. (2002). *The fate of knowledge*. Princeton University Press.
Longino, Helen. (2014). Values, heuristics, and politics of knowledge. In Martin Carrier (Ed.), *The Challange of the Social and the Pressure of the Practice: Science and Values Revisited*. University of Pittsburgh Press.
McMullin, Ernan. (1982). Values in science. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 3–28.
Montañez, George D. (2017). *Why machine learning works*. Carnegie Mellon.
Okruhlik, Kathleen. (1994). Gender and the biological sciences. *Canadian Journal of Philosophy*, *24*(sup1), 21–42.
Pettigrew, Richard. (2016). *Accuracy and the laws of credence*. Oxford University Press.
Rolin, Kristina. (2017). Can social diversity be best incorporated into science by adopting the social value management ideal? In D. Steel & Kevin C. Elliott (Eds.), *Current controversies in values and science* (pp. 113–129). Routledge.
Rudner, Richard. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, *20*(1), 1–6.
Russell, Stuart, & Norvig, Peter. (2010). *Artificial intelligence: A modern approach* (3rd ed.). New Jersey: Pearson Education Inc.
Schaffer, Cullen. (1994). A conservation Law for Generalization Performance. *Machine Learning: Proceedings of the Eleventh International Conference*.
Schaffer, Cullen. (1993a). Overfitting avoidance as bias. *Machine Learning*, *10*(2), 153–178.
Schaffer, Cullen. (1993b). Selecting a classification method by cross validation. *Machine Learning*, *13*(1), 135–143.
Steel, Daniel. (2013). Acceptance, values, and inductive risk. *Philosophy of Science*, *80*(5), 818–828. 10.1086/673936
Strawson, Peter Frederick. (1952). *Introduction to logical theory*. London: Methuen.
Swinburne, Richard. (1997). *Simplicity as Evidence for Truth*. Milwaukee: Marquette University Press.
Toulmin, Stephen. (1970). Does the distinction between normal and revolutionary science hold water? In Imre Lakatos & Alan Musgrave (Eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press.
van Fraassen, Bas C. (1980). *The scientific image*. New York: Oxford University Press.

Wilson, D. R., & Martinez, T. R. (1997). Bias and the probability of generalization. *Proceedings Intelligent Information Systems. IIS'97*, 108–114. 10.1109/IIS.1997.645199

Wolpert, David H. *On overfitting avoidance as bias*. Technical Report SFI TR 92-03-5001. Santa Fe, NM: The Santa Fe Institute, 1993.

Wolpert, David H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1391–1420. 10.1162/neco.1996.8.7.1391

Wolpert, David H. (2012). What the no free lunch theorems really mean ; how to improve search algorithms, 1–13.