VALUES IN MACHINE LEARNING: WHAT FOLLOWS FROM UNDERDETERMINATION?

TOM F. STERKENBURG

Abstract. It has been argued that inductive underdetermination entails that machine learning algorithms must be value-laden. This paper draws from the philosophy of induction to rather highlight the epistemic motivations and justifications that play a role in machine learning algorithm design. The analysis offered indicates that some of the arguments from underdetermination to value-ladenness are too quick, but it also supports their conclusion by indicating how the practical realization of these epistemic considerations inevitably introduces various non-epistemically value-laden judgments, too. The suggestion is that exposing value-ladenness is not inconsistent with, and even profits from, appreciation of the epistemic considerations involved.

1. Introduction

Machine learning is in many ways biased. Much contemporary work in computer science and philosophy alike is devoted to charting the various types and entry points of algorithmic bias in machine learning pipelines (d'Alessandro et al., 2017; Danks and London, 2017; Hellström et al., 2020; Mehrabi et al., 2021; Fazelpour and Danks, 2021). Several authors (Karaca, 2021; Biddle, 2022, 2023; Birhane et al., 2022; Nyrup, 2022; Sullivan, 2022, 2023) have also made a connection to the philosophy of science literature on the role of value-laden judgments in scientific inference (Douglas, 2016; Elliott and Steel, 2017; Elliott, 2022). Some have adopted arguments from this literature to reason more fundamentally that machine learning algorithms *must* be value-laden (Dotan, 2021; Johnson, 2024).

Johnson (2024, p. 28), in particular, poses the question "whether it is really possible for [machine learning] algorithms to be value-free *even in principle*." Setting aside the "[p]roblematic social patterns [...] necessarily encoded in the data on which algorithms operate," and setting aside even the "all-too-human nature of the engineers themselves," she asks "whether values are constitutive of the very operation of algorithmic decision-making, such that on *no* idealized conception could [machine learning algorithms] be value-free" (ibid.).

In addressing this question, Johnson adopts general arguments from the philosophy of science against the so-called value-free ideal. These arguments rely on the *inductive* nature of scientific inference, and the fundamental problem of the *underdetermination* of inductive conclusions by the available data; characteristics that are shared by machine learning algorithms. Johnson writes that "[t]hese arguments result in the view that both scientific and algorithmic decision procedures are deeply value-laden" (2024, p. 30).

Date: December 19, 2025. This is a (substantially revised) preliminary version. I welcome feedback.

Yet there is something unsatisfying about the blanket lesson that machine learning algorithms are the product of value-laden choices beginning to end. For instance, when one opens a textbook on machine learning, one finds various theoretical and methodological—apparently epistemic—motivations for this or that algorithm. What seems in order, in the spirit of work connecting the ethics and epistemology of artificial intelligence (Russo et al., 2024; Grote, 2025), is a more careful picture of how both epistemic and non-epistemic considerations come together in the design of machine learning algorithms.

My aim in this paper is modest. Drawing from the philosophy of induction, I seek to identify epistemic considerations in machine learning algorithm design. My strategy is to use Johnson's reasoning from inductive underdetermination, which includes an argument against the possibility of demarcating epistemic from non-epistemic values, as a foil to bring out the epistemic motivations and justifications in the theory and practice of machine learning. However, while I will suggest that Johnson's reasoning is not fully successful, my analysis will also reveal that non-epistemically value-laden judgments are inevitable in machine learning algorithm design. Thus, to be clear, I do not seek to defend a claim that machine learning algorithms are value-free, or even that there is some internal, value-neutral stage to machine learning. My suggestion is that exposing value-ladenness is not inconsistent with—and even profits from—recognition of the epistemic considerations at play.

The plan is as follows. In section 2, I rehearse Johnson's version of the argument from underdetermination, including the argument against demarcation. In section 3, I set up my analysis by clarifying and delineating the relevant notions of "value-ladenness" and "machine learning algorithm." Then, in section 4, follows my analysis of the epistemic factors at play in learning algorithm design. I introduce and refine my main observations over three levels of decreasing abstraction: from the (philosophical) theory of Bayesian inference, to the theory of supervised classification, to a classical algorithm for digit recognition. I conclude in section 5.

2. The underdetermination argument

My focus here is on Johnson's (2024) argument, as an explicit and representative articulation of the inference from inductive underdetermination to the value-ladenness of machine learning algorithms.[1] Johnson's reasoning, then, from "adopting arguments against the value-free ideal in science and extending them to the domain of machine learning" (2024, p. 29), is that underdetermination implies the need for certain canons of inductive inference (section 2.1) and that these canons introduce non-epistemic values (section 2.2).

2.1. **Problems and canons of inductive inference.**

Johnson starts by tracing the origin of the value-free ideal in the rejection of a standard of objectivity that is clearly too strong. Namely, no interesting scientific inference can be based on "just the facts" or the evidence only. The "raw data" (even granted such a thing exists) must underdetermine, essentially by definition, more general hypotheses we seek to infer. This is the problem of underdetermination of theory by evidence, which, as Johnson observes, is rooted in Hume's problem of induction (2024, pp. 30f).

[1] Another is Dotan's (2021), which I discuss briefly in section 4.2.3 below. Similar ideas are expressed in works like (Ratti, 2026). The purported philosophical lesson has already made it to machine learning textbooks: "the use of inductive inference implies that machine learning models are deeply value-laden" (Prince, 2023, p. 432).

Johnson notes two characteristics of inductive inference, or any inference that "goes beyond the information given in the premises." First, and again essentially by definition, "induction, unlike deduction, fails to guarantee truth." Second, and this was Hume's concern, induction differs from deduction in its *justification*. Whereas "the justification of deduction is *a priori* and necessary [...] the justification of induction is contingent—it depends on the world being a certain way" (2024, p. 32). These observations imply that whenever we make inductive inferences in science or beyond, we must do this by "making non-evidential assumptions" (ibid.). Johnson concludes that "any domain of inquiry in which we attempt to draw conclusions on the basis of limited data [...] therefore comes with its own set of assumptions on which it relies," and she "call[s] this broad collection of assumptions in different domains 'canons of inductive inference'" (ibid., p. 33).[2] The canons of inductive inference are "necessary means of overcoming underdetermination" (ibid.).

This raises the question which canons "scientists need to adopt in order to accomplish the aims of science" (2024, pp. 33f), and this, according to Johnson, is what the debate within the philosophy of science over the value-free ideal has centred on: "which canons are acceptable and which are impermissible" (ibid., p. 34). "A canonical answer to this question," Johnson continues, "was provided by Thomas Kuhn" (ibid.). The list of "theoretical virtues" or "epistemic values" he put forward (including accuracy, fruitfulness, consistency, breadth of scope, and simplicity) "was taken to provide at least a benchmark answer to the question of which canons scientists ought to adopt" (ibid., p. 34).[3]

At this point Johnson notes, with Douglas (2016, p. 611), that a more apt label for the value-free ideal would be the "epistemic-values-only-in-scientific-inference ideal." Namely, first, "there will always be some role for 'values' (or canons (or biases)). However, those values (or canons (or biases)), according to the ideal, will be limited to the epistemic" (Johnson, 2024, p. 35). Second, "the relevant focal point of debates surrounding the value-free ideal is scientific *inference*" (ibid.). Everyone agrees that "values can guide *some* aspects of scientific practice" (like the choice of research project), but these aspects "fall 'outside' of inductive inference itself" (ibid.). Johnson therefore explicitly limits scope to "what seems the best possible candidate for defending the value-free ideal, inference itself" (ibid.).

The relevance of the story so far to machine learning is, of course, that the dialectics are supposed to be analogous. First of all, "as inductive decision-making procedures, machine learning algorithms are subject to these same problems of induction and underdetermination" (Johnson, 2024, pp. 36f).[4] Second, this means that here too Kuhnian canons must come into play: "[i]f program engineers adhere to the value-free ideal, then they are apt to produce programs that draw conclusions from some dataset in ways that maximize accuracy, fruitfulness, consistency,

[2] Johnson follows Douglas (2016, p. 610) in this use of the term "canons of inference," originally due to Levi (1960). She uses the term interchangeably with "biases," to be understood in a "normatively neutral" manner (2024, fn. 12; following Johnson, 2020; Antony, 2016), and, later on, with "epistemic values."

[3] Kuhn introduced this list as "standard criteria" for theory choice (1977, p. 322), which "function not as rules, which determine choice, but as values, which influence it" (ibid., p. 331). Later authors have given different lists (e.g., McMullin, 1984; Longino, 1990), and have also used different terms to refer to these kind of criteria (e.g., "epistemic factors," McMullin, 1984; "cognitive values," Laudan, 1984; "constitutive values," Longino, 1990).

[4] Johnson (2024, fn. 23) makes reference here to the "No Free Lunch Theorem," which I will discuss in sections 4.2.2 and 4.2.3 below.

breadth of scope, and simplicity" (ibid., p. 37). Hence, objections to the value-free ideal in science will likewise "apply to the adoption of the value-free ideal in the production, use, and evaluation of machine learning programs" (ibid, p. 38.).

2.2. **Against epistemic values.** The next step is that even the "epistemic-values-only-in-scientific-inference ideal" cannot be upheld. Johnson here turns to arguments in philosophy of science that "strive to show that even in principle, this ideal is unattainable" (2024, p. 36). There are two "standard arguments" she reviews: the argument against the possible demarcation of epistemic and non-epistemic values (section 2.2.1) and the argument from inductive risk (section 2.2.2).[5]

2.2.1. *The argument against demarcation.* Longino (1996) argues against a neat boundary between epistemic and non-epistemic (or "cognitive" and "non-cognitive") values.[6] Johnson distinguishes two interpretations of Longino's arguments.

The "most straightforward" interpretation, which Johnson (2024, p. 39) dubs the *justification argument*, highlights the values that must drive the "meta-decision" about what canons to select. Demarcation is untenable "if your justification for choosing an epistemic virtue over a non-epistemic virtue (or vice versa) depends on social and political values" (ibid.). Johnson offers as an illustration how Longino (1996, p. 51) pits the Kuhnian virtue of external consistency (i.e., consistency with accepted theory in other domains) against the theoretical virtue of *novelty* defended in feminist philosophy of science. The novelty criterion has a socio-political basis, namely "the need for theoretical frameworks other than those that have functioned in gender oppression by making gender invisible." But on the same par, "external consistency, in a context in which theories have had that function, perpetuates this invisibility. Those satisfied with the status quo will endorse this criterion" (ibid.). Thus, Johnson writes, "in both cases socio-political values guide us [...] in accepting the canons that we do," which "renders a strict demarcation between the two lists on the grounds that one set is value-free untenable" (2024, p. 39).

Johnson's second, "more subtle" interpretation, is the *constitutive argument*, which concerns "the natures of the values themselves" (2024, p. 39). Demarcation is untenable "if the adoption of a seemingly epistemic virtue in a particular context depends constitutively on the socio-political features of the context" (ibid., p. 40). Johnson here gives the example of a sleeping drug that was approved despite the failure of clinical trials to take into account the metabolic differences between men and women, leading to women taking too high doses. The assumption that the male metabolic system is paradigmatic is a commitment to the value of simplicity, but one which "imbibe[s] the very socio-political values" on which existing male privilege is built (ibid.).

[5] As noted by Elliott (2022, p. 19, fn. 10), there is no general agreement on how precisely different arguments (or argument steps) against the value-free ideal relate. For instance, Elliott himself presents the "gap argument," or the reasoning that underdetermination leaves a gap to be filled by values, as distinct from the "error argument," the argument from inductive risk. He places the demarcation argument under the header of the gap argument; in contrast, Douglas (2016) puts the gap argument under the header of the "descriptive challenge," which she sets apart from the "boundary" (demarcation) and the "normative" (error) challenge. Others treat the error argument as a special case of the gap argument (Biddle, 2013, fn. 3; ChoGlueck, 2018). I here simply follow Johnson's presentation, based on (Douglas, 2016).

[6] Longino's argument was anticipated by Rooney (1992), who critizes the "relatively firm distinction [that] is still endorsed" between "constitutive" and "contextual" values by Longino (1990).

Returning to the context of machine learning, the demarcation argument criticizes the call on the value-free ideal for the choice of certain methods over others. The adoption of certain canons or values in such choices itself calls for justification, and "[i]t is in providing this further justification that program engineers will likely have to appeal to facts that go beyond the purely epistemic," including "considerations about the overall aim of the program and the context in which it is intended to be used, facts which themselves depend on social and political factors" (Johnson, 2024, p. 43). But by the justification interpretation of the demarcation argument, "any further justification that involves social or ethical considerations will render even those first-order decisions value-laden" (ibid.). Further, by the constitutive interpretation, "even abiding by a seemingly pure epistemic list of considerations when making design decisions might usher in socio-political values" (ibid.).

2.2.2. *The argument from inductive risk.* Douglas (2000) argues that the presence of *inductive risk*, or the chance of being wrong in inductively accepting a scientific hypothesis, necessitates value-laden judgment calls.[7] In Johnson's words, the canons of induction are "inevitably fallible," so "in all cases where we adopt canons of inference, i.e., in all cases of induction, we run the risk of getting things wrong" (2024, p. 44). As the risk of being wrong has real-world consequences, "the threshold of confidence can only be established by appeal to ethical values, thus rendering the decision to adopt any particular hypothesis value-laden" (ibid., p. 45).

This applies "equally well, if not more so, in the case of machine learning programs" (Johnson, 2024, p. 45). For instance, in the case of an image recognition program to distinguish human from non-human shapes, you would accept a lower accuracy if the program is used to automatically turn on your office light, than if it is used in a self-driving car to avoid collisions. "Algorithmic design choices about how to manage error therefore inherently involve values" (ibid., p. 46).

2.3. **A look ahead.** Below, in section 4, I will contrast my observations about the epistemic considerations in machine learning to Johnson's underdetermination argument, and indicate why I think the argument is not fully successful. More precisely, of the two arguments of section 2.2, I will there connect back to the argument Johnson also focuses on, the argument against demarcation; I will only briefly return to the inductive risk argument in the conclusion, section 5.

But first some additional clarification is in order about the underdetermination argument and what it establishes in the case at hand, the case of machine learning.

3. Values and algorithms

Specifically, we need to clarify the notions of "value-ladenness" (section 3.1) and "machine learning algorithm" (section 3.2) concerned. An example of an actual learning algorithm serves to make things more concrete (section 3.3).

3.1. **Values, choices, and reasons.** What would it mean for an algorithm to be value-laden? In the original debate about values in science it has been a recurring complaint that the notion of (non-epistemic) value is not clearly defined, and newer work still flags this as a major challenge (Biddle, 2013; Ward, 2021; Elliott and Korf, 2024). Johnson also does not analyze the notion further, and I will not myself

[7] The pedigree of this argument is (Churchman, 1948; Rudner, 1953). For modern versions, see (Brown and Stegenga, 2023; Brown, 2024).

attempt a more precise definition either; but I will adopt a useful taxonomy recently proposed by Ward (2021), who disambiguates different ways in which scientific choices can be value-laden.[8,9]

Ward takes it for granted that in the original debate, the role of values concerns scientists' *choices* (like the choice to accept a certain hypothesis). In the case of machine learning algorithms, it also makes sense to analyze value-ladenness in terms of choices. Not the choices of the algorithm itself (whatever that might mean), but the choices of human engineers in how to *design* a certain algorithm. I will take it that an assertion that an algorithm is value-laden is really an assertion about the values involved in such choices.[10]

Ward distinguishes two broad categories of how values relate to choices. To begin, values can stand in a *causal* relationship to choices. Values can act as *causal effectors* in bringing about choices. Moreover, in the other direction, values can be *affected goods*, causally impacted by certain choices. Ward argues that, at least when it comes to choices that run inductive risk (which is to say, where "potential errors have practical consequences outside of science," 2021, p. 58), claims in either direction of the existence of value-ladenness in the causal sense are trivial and so uninteresting. If we plausibly assume that any choice to pursue a certain research project is value-laden, we already have that "*every* part of science is causally downstream of values" (ibid., p. 60). Furthermore, the claim that choices that run inductive risk affect goods in the world is "basically tautological" (ibid.).[11]

When it comes to machine learning, the causal interpretation likewise trivially implies value-ladenness across the board. Save for algorithms imagined "in purely academic abstraction" to operate "wholly divorced from human endeavors," which Johnson (2024, p. 56; pp. 28f, fn. 3) rightly sets apart, machine learning algorithms are part of pipelines that start with real-world learning problems and end with real-world consequences. Just taking the one direction, choices to embark on such problems with machine learning are not purely epistemic, and these choices obviously precede and causally affect the further choices of design and use of the actual machine learning algorithm.

I will therefore focus on Ward's other category, which is in terms of *reasons*. Namely, values can provide reasons for choice. Following a distinction made in the philosophy of action, these reasons can be either *motivating* or *justifying*. Motivating reasons are simply "the reasons for which a person does something or decides to do something" (Bond, 1974, p. 335). In contrast, justifying reasons are "reasons supporting 'ought' judgments" or "reasons for or against" doing or deciding something (ibid., p. 334). The first type of reasons are tied to a person's "desires, beliefs, and emotions," whereas the second are "tied to the world beyond" (Bond, 1983,

[8] My concern, like Ward's, is with the *role* values play, rather than what values really *are* (cf. Elliott and Korf, 2024, p. 7).

[9] Ward's proposal is also not uncontroversial: for instance, from his decision-theoretic approach to values in science, Winsberg (2024) rejects psychologist accounts, including Ward's (ibid., pp. 399f), in favor of a "logicist" understanding of values. To make my main point, the presence of epistemic motivations and justifications in machine learning, I have nevertheless found it useful to adopt Ward's taxonomy below in terms of motivating and justifying reasons.

[10] This is also consistent with what Johnson writes about the role of values in machine learning: the demarcation and the inductive risk argument are applied to the "decision points left up to [machine learning engineers]" (2024, p. 43) and their "[a]lgorithmic design choices" (ibid., p. 46).

[11] Of course, as Ward is also careful to note, it is still worth studying how specific scientific choices were affected by or lead to specific values. Also see Ratti and Russo (2024).

p. 30, also quoted by Ward, 2021, p. 55). Ward gives as an example a politician who votes in favour of expanding healthcare benefits for elderly people. He owns a nursing home company himself and the expansion is bound to make him money: this is actually his motivating reason for voting in favour. Nevertheless, he may cite as a reason that something needs to be done to redress healthcare inequalities and deficiencies: this is a justifying reason.

In my discussion below, I will consider both of Ward's ways of understanding values as connected to reasons for choices: as motivating and as justifying reasons.

3.2. **Inference itself.** In posing the question of value-ladenness, Johnson seeks to isolate the actual algorithm from other stages of a machine learning pipeline, like the stage of training data selection. Moreover, her question specifically concerns the algorithm's "inductive inference itself" (2024, p. 35). However, Johnson's own account still leaves ambiguous what stage in a machine learning pipeline the inductive inference and indeed the algorithm is supposed to correspond to.

I will now follow Johnson in trying to limit scope in this way. Taking up the conception of values as connected to design choices from section 3.1, I will seek to isolate the stage of the design of the algorithm for inductive inference.

3.2.1. *The machine learning algorithm?* Biddle (2022) gives an account of epistemic risks and value-laden choices in the various stages of a typical machine learning pipeline (and in particular for recidivism-prediction systems). Biddle argues that "developers must navigate epistemic risk that reflects values at (at least) the following stages: (1) problem identification and framing, (2) data decisions and model competencies, (3) algorithm design: accuracy and explainability, (4) algorithm design: conceptions of fairness, (5) algorithm design: choices of outputs, and (6) deployment decisions about transparency and opacity" (ibid., p. 322). In Biddle's taxonomy, where is the design of the algorithm for inductive inference located?[12]

Stage (1) of problem identification/framing precede the stage of the algorithm design: this is a typical point of entry of value judgments that we, with Johnson, would want to put aside. The same would hold for decisions about the data sets used for training and for benchmarking, located at stage (2). But the stages (3) to (5), prefixed "algorithm design," sound closer to our mark.

There is, however, an ambiguity in Biddle's understanding of the relevant notion of "algorithm." He first writes that a "machine-learning [...] algorithm, in contrast to a traditional algorithm, is one that 'learns for itself' in a bottom-up manner on the basis of data" (2022, p. 322). This is a straightforward description of the notion of a *learning algorithm*: an algorithm that on the basis of training data produces a certain output, like (in a classification problem) a classifier. The output of a learning algorithm, when the training is done, is also called the machine learning *model*. By the end of the same section, however, Biddle appears to use the term "algorithm" to refer, not to the learning algorithm, but to the learned model. In discussing the training of a deep learning model, and a standard learning algorithm that iteratively updates the deep network's parameters (which determine the model), he writes that

[12] As mentioned in the introduction, there exist several efforts in the literature to chart the biases entering at various stages of a (typical) machine learning pipeline, including more detailed and principled taxonomies than Biddle's (e.g., d'Alessandro et al., 2017 use the lens of the CRISP-DM standard for model building and deployment). I follow Biddle's stages here because they are at a helpful level of coarse-grainedness for the current discussion.

each adjustment in weights corresponds to a "change in the algorithm," and that when "the training stage has ended [...] the algorithm is fixed" (ibid., p. 333).

3.2.2. *The trained model.* It is, for instance, the latter view of the algorithm as the learned model that is at stake in Biddle's stage (4), the evaluation of fairness. The conception of *algorithmic fairness* that Biddle links up to here is about formal definitions of disparate impact in terms of statistical properties of learned classifiers or predictive models.[13] For example, the notorious COMPAS recidivism-prediction system, which Biddle also focuses on (2022, sects. 4–5), was charged with failing to satisfy the criterion of equalized odds, meaning, roughly, that false positive and false negative rates were not the same for different sensitive groups.[14] In fact, Johnson herself also brings in COMPAS and algorithmic fairness as an "application" of her general argument to a concrete case (2024, sect. 3.3).[15]

When it comes to choices concerning the learned model, the obvious direction to look is the after-training stage, when the model is evaluated. In the typical machine learning procedure, one trains a model and then evaluates the error (and perhaps additional criteria like fairness) on a test set. An important choice is the decision that the test performance is good enough, and in that sense to accept the model. Since this decision will normally take into account the model's intended use, we here also already find ourselves at Biddle's final stage (6), the stage of deployment; and at a natural place for launching the argument from inductive risk.[16]

3.2.3. *The learning algorithm.* However, when we seek to isolate the stage of the inductive inference of the algorithm, this stage, the evaluation of a trained model, does not appear to be the most natural candidate. A more natural candidate is how the model is arrived at in the first place, and more specifically the core inductive inference step which, after all, gives machine learning its name: the actual learning step, executed by the learning algorithm.

The design of the learning algorithm is what I will focus on in my analysis below. My main reason is ultimately not that this constitutes the best answer to the question of what is the "inference itself," but rather that this will be helpful for my main aim of bringing out epistemic considerations in machine learning algorithm

[13] For entries to this literature, see Barocas et al. (2023); Pessach and Shmueli (2022).

[14] Strictly speaking criteria of algorithmic fairness are not just properties of a model, but also of (an independent estimate of) a ground truth. The COMPAS system was later shown to satisfy a criterion of predictive equity, meaning, roughly, that the proportion of individuals with the same risk score who recidivate is the same for different sensitive groups. Subsequent work showed that in non-trivial cases these two criteria are mutually exclusive, so that there is not only a trade-off between accuracy and fairness, but also between different fairness criteria (Biddle, 2022, sect. 3d).

[15] Initially, Johnson specifies "'machine learning programs,' 'algorithmic decision-making,' and 'algorithms'" as a "broad class of automated programs that function by [...] 'learning' from patterns manifest in the data [...] in order to build a predictive model" (2024, fn. 6), which suggests the notion of learning algorithm. But her discussion of the COMPAS case, and also her response to objections to the argument from inductive risk (emphasizing "how these programs are used for decision making", ibid., pp. 51f), suggests that the value-ladenness concerns the models.

[16] As follows: the choice of what is good enough is inevitably also an assessment of practical consequences of model errors, which must involve non-epistemic values.

design.[17,18] Indeed, at the end of my analysis I will problematize the very idea that, for the question of value-ladenness, we can neatly isolate the stage of learning algorithm design from other stages in the pipeline.

3.2.4. *Inference itself, conclusion.* In order to say a little bit more about the restriction to the learning algorithm in my main analysis in section 4 below, let me briefly reconnect to Biddle's (2022) taxonomy and a few particularly contentious aspects of algorithm design.

Biddle's stages (3)–(5) of algorithm design, recall, have to do with the model's interpretability, with the model's fairness, and with the model's outputs. As aspects of learned models, these stages clearly feature choices in evaluation and acceptance of a model.[19] Yet these aspects can also clearly already play a role in designing the learning algorithm. This is obvious for the choice of outputs: in selecting a model, the algorithm normally evaluates the outputs of potential models on the training data, and different types of outputs ask for different ways of doing this. This immediately points at a related design choice which Biddle does not discuss, but others in this context do, namely the choice of cost or loss function.[20] But also in the case of fairness, there exist various *in-processing* techniques to optimize towards certain fairness criteria during the learning process, particularly again by choice of loss function (Pessach and Shmueli, 2022, sect. 4.2; Mehrabi et al., 2021, sect. 5); so that there is here a clear sense in which a choice of fairness metric is part of the learning algorithm design.

However, we do not need to view fairness criteria as choices at the level of learning algorithm design, or indeed as canons of induction necessary to bridge underdetermination, as Johnson (2024, p. 47) suggests. An alternative perspective is that rather than assumptions needed to bridge the inductive gap, fairness criteria give a certain refined accuracy criterion or *goal* in learning. More generally, from this perspective, the choice of outputs and of loss function precedes the design of the learning algorithm: they are choices in formulating what the inductive learning problem actually *is*. The issue of underdetermination, and the need for further assumptions, only arises in the next step, of how to actually solve the inductive inference problem: how to generalize from the data to conclusions of a certain form under a certain accuracy criterion.

This is the perspective that I will adopt in my analysis in section 4 below. For my main aim, highlighting the epistemic considerations at play in algorithm design, this

[17] I might here indeed be departing from Johnson's own view. She immediately clarifies "inductive inference itself" as "the point at which we decide to accept or reject some conclusion" (Johnson, 2024, p. 35), which could be interpreted as pertaining to the model acceptance step.

[18] I should also note that people can mean different things when talking about "the inference" in the context of machine learning. In the statistics literature, the "inference" in the statistical inference commonly refers rather to assessment of an estimator's uncertainty, and so is again closer to the evaluation and acceptance step. In machine learning, the phrase "inference time" actually refers to using the final model.

[19] For instance, Sullivan (2022, 2023) discusses how non-epistemic values are relevant to the extent the opacity of a trained model poses problems for understanding and explanation.

[20] Karaca (2021) gives a careful account of values entering in the construction and evaluation of machine learning models for binary classification, and argues that "value judgments based on social values are involved in the construction of ML classification models mainly through cost-sensitive ML optimisation" (ibid., p. 18). Johnson (2024, p. 43) gives the choice of loss function as an illustration of the constitutive argument against demarcation. I will say more about the loss function in sect. 4.3.3 below.

is a helpful theoretical simplification, which allows me to make the connection to formal frameworks of induction in philosophy and generalization in machine learning theory. My main observations will survive, I think, the realization that a clean separation between problem formulation and learning algorithm design becomes hard to sustain when we look at the design of real machine learning algorithms—as I will discuss at the end of my analysis. There I will return to the following concrete example of a specific machine learning algorithm, developed for a specific real-world problem, which will serve to make the question and the analysis a little more tangible.

3.3. **An example: handwritten digit recognition.** For simplicity, I pick an example from the early days of machine learning, namely the algorithm developed by LeCun et al. (1989b,a) for the recognition of hand-written digits. The aim is for a system that can read off the correct symbol from images of single handwritten digits. The learning problem is to infer, from a training set of correctly labeled such images, a general classification model for reading handwritten digits. The authors' approach is to use a neural network, and it is indeed one of the first uses of a convolutional network for image recognition.[21]

I did *not* choose this example because it is already a clear example of a value-free machine learning application. It is not. The decision that hand-written digit recognition is a relevant problem ("of great practical value," LeCun et al., 1989b, p. 397), to be tackled with machine learning, is clearly not purely epistemic.[22] This holds even more so for any further decision to deploy such a system, like for automating zip code reading in postal processing (and replace the humans previously doing that); the choice of embarking on this project is already value-laden because of the obvious promise (or risk) of such practical applications. This is therefore certainly not an example that can be set aside as "wholly divorced from human endeavors" (Johnson, 2024, p. 56). But also important aspects of the model construction are arguably not free of value judgments. One such aspect is again the training data, consisting of "segmented numerals digitized from handwritten zipcodes that appeared on real U.S. Mail passing through the Buffalo, N.Y. post office" (LeCun et al., 1989b, p. 397). There are arguably inevitable non-epistemic judgment calls in choosing these data as sufficiently representative for the purpose at hand.[23] As such, again, a trained model based on these data is inevitably value-laden, too.

However, our concern for now is the *learning algorithm*. What is the learning algorithm here? To a first approximation, this is the automated inductive inference procedure that goes from training data of a certain form (16x16 pixel grayscale images with labels 0 to 9) to a general model mapping any such image to a label.

[21] This extends the group's earlier work in using neural networks for image recognition (Denker et al., 1989), which still relied on extensive pre-processing of images into feature vectors. Convolutional networks made a forceful reappearance in the modern deep learning boom (LeCun et al., 2010; Goodfellow et al., 2016, ch. 9).

[22] This is perhaps less clear for the problem of digit recognition as an interesting problem for machine learning research. The authors' motivation is basically that the problem is neither too hard nor too simple, writing that the "handwritten digit-recognition application was chosen because it is a relatively simple machine vision task," yet one that "deals with objects in a real two-dimensional space and the mapping from image space to category space has both considerable regularity and considerable complexity" (LeCun et al., 1989b, p. 397).

[23] The authors note failure of generalization due to "writing styles not present in the training set" (1989a, p. 547).

More precisely, this is the procedure that, on the basis of the training data, infers to a certain configuration of parameters of the network, expressing such a general rule. There are actually two different components we can discern here. On the one hand, there is the actual training or optimization algorithm, here a standard gradient descent algorithm for neural networks. On the other, there is the neural network architecture itself, which the training algorithm works on, and which determines which models are expressible by the network (and so learnable) to begin with. Indeed, I structure my analysis below around the decomposition of learning algorithms in a general learning rule and a more domain-specific component; and at the end of this analysis I return to the example of digit recognition.

4. Underdetermination and epistemic reasons

I will now proceed to use insights from the philosophy of induction to bring out the epistemic reasons involved in bridging the inductive gap in machine learning. An important tool throughout my analysis is the distinction, introduced by Sterkenburg and Grünwald (2021), between domain-general *learning rules* and the domain-specific *inductive biases* these must be equipped with, together forming the actual learning algorithms. I will highlight how there are epistemic motivations for the domain-specific inductive biases and epistemic justifications for the domain-general learning rules. However, I will also highlight how, from this perspective, one can readily see that non-epistemically value-laden reasons must come in, too. In the course of the analysis, I will also point out where Johnson's argument from underdetermination appears too quick.

My strategy is to formulate and refine my observations over three different learning settings. I start with the framework of Bayesian inference, as employed in the philosophy of science (section 4.1). I then turn to the setting of statistical classification, as studied in machine learning theory (section 4.2). Finally, I come back to the concrete example of handwritten digit recognition (section 4.3).

4.1. Bayesian learning.

I will here consider the basic subjective Bayesian learning procedure, as set out in many works in the philosophy of science (e.g., Huttegger, 2017; Sprenger and Hartmann, 2019). While it is a significant step from this basic picture to actual (Bayesian) machine learning, it has the advantage of being both relatively simple and familiar, while sufficient to already introduce the main observations.[24]

4.1.1. *The Bayesian learning rule.*

In the basic Bayesian learning procedure, one starts with a probability function p over propositions in some formal language.[25] This probability function is the *prior*; and the *learning* from a piece of evidence E consists in updating the prior p into the *posterior* p' by conditionalization or Bayes's rule,

$$p'(\cdot) := p(\cdot \mid E), \tag{1}$$

[24] Johnson also refers in various places to the philosophy of science literature on inductive inference, and in particular Bayesian learning.

[25] I simply assume here the propositional framework common in philosophy, rather than the measure-theoretic framework standard in statistics and machine learning. Nothing hinges on this.

where $p(H \mid E)$ can be calculated using Bayes's theorem,[26]

$$p(H \mid E) = \frac{p(E \mid H)p(H)}{p(E)}. \tag{2}$$

The learning procedure that forms the core of the Bayesian approach is therefore simply Bayes's rule (1). Note that this rule asks for two input components on the right-hand side: apart from a proposition E (the *data)*, it also needs a prior probability function p.

I will discuss in more detail below how the Bayesian learning rule has a domain-general justification in terms of the epistemic value of *rationality*. The prior adds to this domain-general rule a more domain-specific component, together yielding what I will call a Bayesian learning algorithm.

4.1.2. *Bayesian learning algorithms.* Such a learning algorithm is a procedure that takes input data and returns an output probability function; the particular prior is, so to speak, part of the inner mechanism of the algorithm. This is an algorithm for inductive inference, and therefore subject to the underdetermination of its outputs (probability functions) by the inputs (data).

This might not have been so if there existed fully "neutral," "objective," or "universal" priors, and a Bayesian algorithm with such a prior could be said, perhaps, to merely extract to the posterior what is in the given data. But it is generally accepted that there is no such thing: any choice of prior must express restrictive assumptions (see, e.g., Howson, 2000; Huttegger, 2017; Sterkenburg, 2018). The assumptions expressed in the prior (in tandem with the Bayesian conditionalization rule) bridge the inductive gap between the data and the inductive conclusion (the posterior function), and are therefore also called *inductive assumptions*.

4.1.3. *Inductive assumptions and epistemic motivation.* The usual picture is that the prior expresses our beliefs about the domain at hand, so that the relevant inductive assumptions are assumptions we believe hold true for the domain. To adapt a toy example from Norton (2021, sect. 1.9),[27] suppose we want to draw an inductive inference from the data that salt A has crystallographic form B. We might further have a high credence in Haüy's principle that each crystalline substance has a single characteristic crystallographic form. This domain-specific principle we can formulate in a Bayesian prior, so that by the Bayesian inductive inference (conditionalization of the prior on the data), we draw the inductive conclusion that with high posterior probability all samples of salt A have crystallographic form B.

In this picture, where we seek to formulate our prior as an honest expression of what we believe to be the factual structure of the relevant domain, the inductive assumptions clearly have an epistemic motivation (cf. Ward, 2021, p. 60).[28]

[26] Bayes's theorem just follows from the probability axioms. What characterizes Bayesian learning is that the posterior is set to the conditional probability, in accordance with Bayes's rule.

[27] According to Norton's material theory of induction, all inductive inferences are solely "powered" by local facts. This may be taking it too far into the other extreme: on Norton's account there is no role left for domain-general learning rules, which also renders his critical discussion of Bayesian learning somewhat off (Sterkenburg, 2024).

[28] The usual interpretation of the classic discussion by Jeffrey (1956) is that he countered Rudner's inductive risk argument by observing that scientists can just report their credences (cf. Johnson, 2024, pp. 50–51); an observation that resembles the one I made here (also see Hatchwell, 2024). However, as pointed out by Harvard and Winsberg (2022, sect. 4), Jeffrey wrote that this is the view we "*seem* to have been driven to" if (his actual point) it is not the business of the

4.1.4. *Inductive assumptions and epistemic justification.* The question of the justification for inductive assumptions depends on what we take to be a proper justification. In her discussion of the justifying-reasons interpretation of arguments against the value-free ideal, Ward likewise insists on the "need to provide an independently motivated account of scientific justification," including "what sorts of things are potential justifiers for any given choice" (2021, pp. 60f).

In the extreme case, we can ask for the kind of foundational justification that is at stake in Hume's skeptical argument.[29] If we accept the skeptical argument, then we must conclude that any inductive assumptions are ultimately lacking such justifying reasons. But then, nor would any non-epistemic values count as such foundational justifiers, or they would provide a solution to the problem of induction after all. More generally, it does not yet follow that in the lack of ulterior epistemic justifying reasons, non-epistemic values must enter the picture (cf. Intemann, 2005).[30]

4.1.5. *Inductive assumptions and values.* Recall the argument from underdetermination to values: we start with the claim that the inductive gap must be bridged by general canons of induction or epistemic values, and subject those in a subsequent step to an argument against the demarcation of epistemic and non-epistemic values. Can we apply this reasoning to the inductive assumptions which bridge the inductive gap in the Bayesian picture?

Not straightforwardly so, as it is not so clear, or even question-begging, that we need to think of inductive assumptions as epistemic values, ready to be subjected to the argument against demarcation.[31] However, what we can say is that there is some truth in the conclusion. Namely, in practice, when we look at the choices that need to be made in the formulation of inductive assumptions, both epistemic and non-epistemic reasons must come into play.

The obvious concern with the above Bayesian picture is that it is overly stylized, a far cry removed from actual science or machine learning. When we descend from the philosophical picture to the level of the actual practice of Bayesian statistics or machine learning, then we can observe that the formulation of a prior, even if it is motivated by capturing the structure of the relevant domain, will also be constrained and informed by various other, practical or pragmatic rather than epistemic, considerations.[32]

---

scientist to accept or reject hypotheses, and that this would be a view with "great difficulties" (1956, p. 245, emphasis mine). Jeffrey focused on the presupposition of a satisfying theory of confirmation (see fn. 33 below); on a subjectivist picture, the (practical) difficulty would be that 'having' a probability is one thing, but actually providing probabilities in constructing a model (prior) is another (cf. Harvard and Winsberg, 2022, pp. 13f): I come to this in section 4.1.5 below.

[29] Johnson (2024, p. 41) writes that her demarcation argument is "deeply related to Hume's problem of induction." Dotan (2021) also stresses the connection to Hume's argument.

[30] Johnson (2024, fn. 32) writes that "the nail in the coffin for the value-free ideal [...] would be to demonstrate that non-epistemic values alone can end the regress," and makes the suggestion that "justification has got to stop somewhere [...] surely the decision to cut off justification at any particular point will therefore be a pragmatic decision, and thus one that depends on non-epistemic values." This is a natural suggestion, but does presuppose an account of justification under which such an active decision on the part of algorithm designers is indeed inevitable.

[31] Winsberg (2024, fn. 20) objects to the interpretation of Kuhn's "criteria for choice" as values: "this is a source of confusion in the values-in-science literature. The 'epistemic values' of Kuhn's famous essay are not the values of values in science."

[32] For example, in Bayesian machine learning, considerations of computational tractability usually limit choice to certain flavours of default priors, i.e., default parametrized hypothesis

In fact, already in the philosophical literature on Bayesian learning, we directly discern the appearance of non-epistemic considerations in the formulation of the prior. For instance, central to the account of Huttegger (2017) are symmetry restrictions on inductive assumptions, which he presents as a tool to reduce otherwise insurmountable complexity (ibid., pp. 24f).[33]

This does not mean, however, that there is no longer a concern to formulate inductive assumptions which match the domain, or that this epistemic concern is itself actually non-epistemic. For example, symmetry assumptions can serve to reduce complexity, but should still accord to what we know about the domain. It is therefore more natural and helpful to say that design choices are motivated both by epistemic and non-epistemic considerations.[34]

4.1.6. *Rationality and epistemic justification.* Let me finally return to the general Bayesian learning rule and its justification. One can distinguish two main components in philosophical justifications for the Bayesian approach, both in terms of the epistemic value of *rationality* (see again, e.g., Sprenger and Hartmann, 2019).

The first component concerns the justification for rendering rational degrees of belief as probabilities (i.e., quantities that satisfy the standard axioms of probability). Such justifications include Dutch book arguments (only probabilistic beliefs shield one from sure-loss bets), axiomatic characterizations (only probabilities satisfy natural constraints on a quantitative plausibility measure), and accuracy arguments (only probabilistic beliefs minimize one's total epistemic inaccuracy).

The second component concerns justifications for the actual learning rule, the Bayesian updating procedure. Here we find "dynamic" versions of the previous arguments, but also arguments that Bayesian updating is the most conservative way of moving from a prior to a posterior distribution, and arguments based on convergence-to-the-truth or merger-of-opinion results. These arguments purport to provide general, epistemic justifications for Bayesian updating as the rational way of learning from evidence.[35]

4.1.7. *Rationality and values.* Can we apply the demarcation argument to the epistemic value of rationality which underlies the Bayesian approach?[36]

---

classes plus a default prior distribution over parameters. Steel (2015) flags this aspect in the context of Bayesian statistics. Also see (Biddle, 2013).

[33]Huttegger's project is in the tradition of Jeffrey's subjectivist "radical probabilism," but the emphasis on symmetry assumptions in inspired by Carnap's program of inductive logic. Carnap thought that there are objective, purely semantic probabilistic confirmation relations between hypotheses and evidence; however, this semantic relationship depends on the choice of logical language, which he thought is pragmatic (cf. Jeffrey, 1956, fn. 7).

[34]I think this is consistent with Winsberg's (2024) decision-theoretic approach to values in science, and in particular his entanglement thesis (which he traces back to Rudner and Jeffrey, and also Putnam) that it is practically impossible to cleanly separate credences from utilities (values). He notes, with reference to Putnam (2012), that it is still "perfectly possible to think that epistemic judgments and values are conceptually distinct" (ibid., p. 392); similarly, it is possible, and indeed also helpful for understanding design choices in machine learning, to conceptually distinguish epistemic and non-epistemic considerations. (Even if Winsberg will not agree with the psychologistic gloss on values I have chosen to adopt here, see fn. 9.)

[35]Note that at least some of these justifications in terms of rationality trade on the (more fundamental?) epistemic value of *accuracy*. Also see footnote 38.

[36]Johnson (2024, p. 33) indeed lists "Bayes' Rule (in the case of belief formation)" as an example canon of induction, which could then presumably be subjected to the demarcation argument.

Again, not straightforwardly so, in this case because an important lever of the argument is the presumption that there is always a particular socio-political context, itself imbued with non-epistemic values, in which the development of a learning algorithm takes place.[37] Arguably, however, the context of the "design" of the general Bayesian learning procedure is really the context of foundational work in philosophy and learning theory. This is the apparently epistemic project of philosophers and theoreticians to formulate principles of rational learning and develop arguments that the Bayesian approach (and in particular the Bayesian updating rule) satisfies those. Furthermore, even in a particular socio-political context, it seems that an algorithm designer could defer to these context-independent and epistemic reasons for the Bayesian learning approach.

One route towards a more refined argument for the influence of non-epistemic values would be to observe that the various arguments for the rationality of the Bayesian approach are not uncontroversial, in particular because they rely on commitments that are not purely epistemic.[38] Another would be to simply observe that in practice the choice for going the Bayesian way will also involve various non-epistemic considerations, like computational feasibility.[39]

I will not further discuss the case of Bayesian (machine) learning, and move on to the standard, non-Bayesian, approach to classification. But the upshot of the discussion is that there are epistemic motivations and (arguably) justifications in Bayesian learning, even if there must also be non-epistemic factors involved.

4.2. **Classification and empirical risk minimization.** The prototypical machine learning paradigm is supervised classification. Recall the digit recognition example: we have a set of possible instances that we seek to classify using a finite number of labels. A learning algorithm receives for input a finite training sample of instances that are already labeled (this is what makes the problem supervised), and outputs a classifier (a trained model) that labels all possible instances.

4.2.1. *The ERM rule.* The most basic learning rule for supervised classification, the *empirial risk minimization* (ERM) rule, proceeds as follows. It works with a class $\mathcal{H}$ of possible classifiers or models, the *model class*, and on receiving a training sample, it selects a model from $\mathcal{H}$ that minimizes the error on the training sample. (Standardly the error here is the mean number of misclassifications on the training sample, also called the 0/1 error.)

What is the justification for this algorithm? A theoretical basis for ERM can be found in statistical learning theory (SLT).[40] Here we first assume that the instances and labels are sampled i.i.d. from some unknown probability distribution $\mathcal{D}$. We

[37] In introducing Longino's argument, Johnson writes that "which virtue is adopted in any particular instance of scientific theorizing is a contextual matter, and crucially will be settled in virtue of the socio-political features of that context" (2024, p. 39).

[38] For instance, justifications in terms of betting are vulnerable to the complaint that they are more pragmatic than epistemic. More recent "accuracy-first" justifications are a response to this complaint (Pettigrew, 2016). Critics have argued that a choice of quantitative accuracy measure is still an unavoidable pragmatic element in such accounts (Mayo-Wilson and Wheeler, 2019).

[39] Lenhard (2022) argues that the success of Bayesian methods in statistics since the 1990s is largely a matter of new computational approaches, and has less to do with (is even problematic for) their philosophical justification in terms of rationality.

[40] See, in increasing order of formal detail, Grote et al. (2024, sect. 2); Harman and Kulkarni (2007); Sterkenburg (2025, sect. 2); von Luxburg and Schölkopf (2011) for explanations of SLT for a philosophy audience. A standard textbook is Shalev-Shwartz and Ben-David (2014).

do not assume anything about the structure of this distribution; we only assume that data come from *some* unknown distribution $\mathcal{D}$. Second, we adopt as our learning goal finding a classifier that minimizes the *true risk*, which is the probability of misclassifying an instance randomly drawn from this unknown $\mathcal{D}$. Since the distribution is unknown, so is the true risk; but it turns out we can still analyze it.

Namely, a fundamental result in SLT (indeed often called the *fundamental theorem*) says that we can derive a so-called *probably-approximately-correct* (PAC) guarantee for ERM. If the complexity or *capacity* of $\mathcal{H}$ is small enough,[41] then for *any* unknown distribution $\mathcal{D}$, we have that for a large enough training set S sampled from $\mathcal{D}$, ERM will with high probability select a classifier that has a risk approximately as low as the lowest-risk classifier in $\mathcal{H}$.[42]

4.2.2. *ERM algorithms.* Similarly to the case of the Bayesian algorithm and the prior distribution, the ERM rule requires, aside from the data, a further input component: the model class. I will use the term "ERM algorithm" for any implementation of the ERM rule with already a particular model class $\mathcal{H}$ provided. The resulting procedure takes a training sample and outputs a classifier, with the model class part of the algorithm's inner mechanism.

Again, this is an algorithm for inductive inference, which is subject to the underdetermination of its outputs (classifiers) by the inputs (training data). This might not have been so if there existed some "universal" model class, and if a "universal" ERM algorithm equipped with such a model class would still have a PAC guarantee. Namely, a guarantee of finding the approximately-best model in this universal class would actually be a guarantee of finding the approximately-*absolutely*-best model. Such a universal algorithm could be said, perhaps, to objectively extrapolate the patterns in the data to a choice of model.

But there can be no such universal ERM method, as shown by impossibility results usually referred to as "no-free-lunch theorems." In particular, for any statistical learning algorithm A (including ERM with any particular choice of $\mathcal{H}$), there will be possible true distributions $\mathcal{D}$ such that A is a bad method, meaning that with high probability its true risk is high (Shalev-Shwartz and Ben-David, 2014). This means that any choice of hypothesis class for ERM that preserves the PAC guarantee must encode restrictive assumptions, which fit some possible situations but not others. These assumptions encoded in the model class (in tandem with the ERM rule and the PAC guarantee) bridge the gap between the data and the inductive conclusion, and are therefore also called the *inductive bias* (Mitchell, 1980; Sterkenburg and Grünwald, 2021; Rendsburg, 2024).[43]

More specifically, the strength of the PAC guarantee is a direct function of the complexity or capacity of the model class (and so, more informally, of the strength of the inductive bias). This translates in a certain means-ends justification for a

[41] Capacity in machine learning refers to the size or complexity of a hypothesis class. There exist various formal notions of capacity for different types of learning problems; the most well-known, which also figures in the fundamental theorem, is the Vapnik-Chervonenkis (VC) dimension.

[42] The fundamental theorem concerns the specific case of binary classification with the 0/1 loss function (implicit in the above notion of true risk). There exist other results for more general classification and regression problems with different loss functions, and while these results tend to be more involved, they still generally give guarantees for (versions of) ERM provided the hypothesis class is in some sense of limited capacity (see Shalev-Shwartz and Ben-David, 2014).

[43] Note that the term "inductive bias" does not yet have a normative connotation, in line with the use of "bias" by Johnson (footnote 2 above; also see Kelly, 2022).

simple class of hypotheses (strong inductive bias): in order to have a good (better) learning guarantee for ERM, one needs to make strong(er) assumptions.[44]

4.2.3. *Inductive bias and values.* Dotan (2021) invokes the no-free-lunch theorems to argue for the essential role of non-epistemic values in theory choice. She notes that earlier types of argument (which would include those given by Johnson) are vulnerable to the objection that they only apply to specific "historical, practical, or political contexts" (ibid, p. 11083).[45] In contrast, "drawing from a mathematical theorem avoids some of the difficulties faced by other arguments because it is independent of human contingencies and contextual particularities" (ibid., p. 11082).

Dotan proceeds in three steps. Her first observation, based on the no-free-lunch theorems, is that "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" (ibid., p. 11090). This is the observation of underdetermination of "theory choice" (selection of a model) by the data, and the resulting need for further inductive assumptions; as given a precise expression by the no-free-lunch result sketched in section 4.2.2 above.[46]

Dotan's second step is to consider bringing in "other traditional epistemic virtues" (Dotan, 2021, p. 11090). But the observation that accuracy is not enough also entails that "[i]f we want to use epistemic virtues other than accuracy, we need to justify them without relying on accuracy" (ibid., p. 11091), and so "NFL challenges the ability to provide pure epistemic justifications for using other traditional epistemic virtues" (ibid., p. 11094). Her third and final step is that "non-epistemic values are natural candidates to supplement accuracy or other considerations" (ibid.).

The core of Dotan's argument is therefore essentially the same as Johnson's move from underdetermination to the need for further general canons or epistemic values in inductive inference. The call on no-free-lunch results does not make an essential difference: in the end, such results are merely a more precise formulation of the lesson of underdetermination. The main difference is that Dotan does not call upon the demarcation argument to proceed to the need for non-epistemic values. Her argument is rather that any candidate epistemic value would have to somehow reduce to accuracy, which is not enough to bridge the underdetermination.

This reasoning again seems overly quick. Similarly to the formulation of Bayesian priors, one can seek to formulate model classes that encode local assumptions about the domain, viz., about what classifiers one thinks are likely to be accurate for the problem at hand (cf. Sikorski and Liu, forthcoming, sect. 3). It is not clear what it would mean for such local assumptions to have to reduce to accuracy. The more

[44] See (Sterkenburg, 2025) for details. Importantly, simplicity does not act here as an independent Kuhnian epistemic value, but as a provable pre-condition for a certain accuracy guarantee.

[45] Specifically, Dotan mentions arguments from the history of science, from inductive risk, and from the impossibility of demarcation.

[46] Dotan actually makes the much stronger claim that all hypotheses have the same "expected accuracy"—in our terms, the same *true* risk. This claim follows from her discussion of the original no-free-lunch theorem for supervised learning, due to Wolpert (1996). But this result crucially relies on the assumption of a uniform distribution over possible learning situations, which is a question-begging assumption of full-blown randomness or "unlearnability" (Sterkenburg and Grünwald, 2021; also see Rushing, 2022). However, all Dotan needs is that further assumptions are required to "supplement predictive accuracy" (Dotan, 2021, p. 11090), and this follows from the general observation of underdetermination, as made precise by versions of the no-free-lunch result that do not rely on this uniformity assumption.

natural thing to say is that the corresponding inductive bias is motivated by the epistemic concern of accuracy.

Nevertheless, there is something true about the conclusion that accuracy is not enough. In practice, as we will see in the example below, there will be various non-epistemic considerations at play in the formulation of an appropriate model class. Moreover, the theoretical basis and its justification in terms of accuracy also involves and prompts various non-epistemic choices.

4.2.4. *Predictive accuracy and values.* When it comes to the epistemic value of predictive accuracy which underlies classification and the general ERM rule, the demarcation argument as presented by Johnson again seems too quick. The relevant context of the design and analysis of the SLT framework (and the ERM rule) is the apparently epistemic project of theoretical computer scientists to provide a foundation for accurate classification. In particular, the general learning objective that is center to this theoretical project, predictive accuracy, is usually just accepted as an axiomatic epistemic goal—including, as we saw, by Dotan. Nevertheless, there are again more refined routes to bringing out the influence of non-epistemic values.

One likely culprit is the assumption of an unknown distribution from which instances are sampled in an i.i.d. manner, which is crucial in the theoretical analysis and justification. This assumption is clearly not fully domain-general: there are certainly real-world learning problems where the i.i.d. assumption is not plausible, or indeed outright problematic.[47] This does not yet straightforwardly imply that, in the spirit of Johnson's constitutive demarcation argument, this restricted range of application is characterized by certain non-epistemic values, which therefore motivate the theoretical framework and method. However, we can more readily observe that this choice is not purely epistemic either: an important motivation for theoreticians' assumption of i.i.d. is surely its mathematical convenience.[48]

Following the recommendations of the theory involves further choices to be made. For instance, the theory instructs us to choose a simple model class, but this push towards simplicity must not only be checked by epistemic considerations about what model class still has a reasonable inductive bias, but will also be mediated by such factors as the amount of training data available (cf. Sterkenburg, 2026).[49] Another decision, already mentioned, is the choice of loss function to actually measure accuracy. I will return to these and other factors, as we now move on to our example of an actual learning algorithm for classification.

4.3. **Example: handwritten digit recognition.**

4.3.1. *The training algorithm.* Learning with a neural network means learning a configuration of the network's free parameters (connection weights). The configuration of parameters determines which classification function the network represents.

[47] The i.i.d. assumption in machine learning is often flagged as akin to Hume's principle of the uniformity of nature (e.g., Steel, 2009, p. 475; Li, 2023; Ratti, 2026).

[48] Theoretical computer scientists also study other general frameworks, with different assumptions. For instance, in the framework of *prediction with expert advice* (Cesa-Bianchi and Lugosi, 2006), the assumption of a data-generating probability distribution is dropped.

[49] Moreover, the classical theoretical analysis and its recommendations have to some extent been overhauled by the phenomenon in modern deep learning that over-parametrized (i.e., highly complex) model classes still generalize very well, even if we do not yet understand why on a theoretical level (Belkin, 2021). This indicates a further choice on the part of algorithm designers to either follow the classical analysis or aim for the so-called effect of "benign interpolation."

So a learning or training algorithm sets, on the basis of the training data, the values of the free parameters, thus selecting a classification function (trained model).

LeCun et al. use a "backpropagation network," which refers to the fact that the network's parameters "are trained using backpropagation" (1989a, p. 542). At the time newly introduced, back-propagation or simply backprop (Rumelhart et al., 1986; see Goodfellow et al., 2016) is now a standard auxiliary function for learning network parameters, normally (including in this instance, LeCun et al., 1989a, p. 546) part of a version of the stochastic gradient descent (SGD) algorithm. SGD is used for the optimization problem of tweaking network weights towards minimization of training error (Goodfellow et al., 2016). That is to say, it is used for the optimization problem of implementing the learning rule of ERM.

It is a leap from the nice theoretical specification of ERM to the practical implementation of SGD. The SGD algorithm only approximates a solution to an optimization problem (find a minimal-training-error function, i.e., configuration of parameters) which cannot be solved analytically. Moreover, SGD requires a loss function with well-behaved derivatives, which rules out the standard 0/1 loss function; the authors instead opt for mean squared error over the real-valued outputs of the ten outputs units corresponding to the possible labels (LeCun et al., 1989a, p. 546). Then there are various further implementation choices, like the activation functions for the hidden units (in this case, the hyperbolic tangent, ibid.), and the parameter initialization before SGD can be applied (in this case a certain uniformly random initialization motivated by the shape of the activation functions, ibid.).

Nevertheless, despite these various messy engineering considerations, we can still discern the theoretical story of sections 4.2.1 and 4.2.2 above. It is still the learning rule of ERM that is being approximated by SGD. Moreover, the authors explicitly evoke the reasoning that there is only a guarantee of good generalization if the capacity of the hypothesis class (as given by all possible weight configurations) is sufficiently constrained (LeCun et al., 1989a, p. 541; also see LeCun, 1989):

> "The basic design principle is to minimize the number of free parameters in the network as much as possible without overly reducing its computational power. This principle increases the probability of correct generalization because it results in a specialized network architecture that has a [...] reduced Vapnik-Chervonenkis dimensionality."

4.3.2. *The network architecture.* How did the authors go about minimizing the capacity of the model class? For important part: by applying local domain knowledge.

The paper opens (LeCun et al., 1989a, p. 541),

> "The ability of learning networks to generalize can be greatly enhanced by providing constraints from the task domain. This paper demonstrates how such constraints can be integrated into a back-propagation network through the architecture of the network."

In learning with a neural network, the model class is given by the network architecture. The architecture (i.e., the neurons, or activation functions, and their mutual connections) determines what classifiers can be represented by the network, and so learned at all. The authors aim at "designing a network architecture that contains a certain amount of prior knowledge about the task" (1989a, p. 541), particularly "prior knowledge about shape recognition" (1989b, p. 399). This motivates

them to introduce a convolutional neural network as an architecture specifically suited for learning from images.

Goodfellow et al. (2016, sect. 9.4) write that "[w]e can imagine a convolutional net as being similar to a fully connected net, but with an infinitely strong prior [that encodes our beliefs about what models are reasonable] over its weights." The relevant prior rules out functions that do not satisfy certain properties, specifically having to do with "local interactions" and "invarian[ce] to small translations" (ibid., p. 336). Correspondingly, LeCun et al. (1989a, sect. 3; 1989b, sect. 4) describe how local knowledge about the task domain[50] informs design choices characteristic of convolutional nets,[51] which we can imagine as representing or implementing the kind of prior Goodfellow et al. describe. That is, local knowledge about the task domain motivates the network's inductive bias.

4.3.3. *The role of non-epistemic values.* Naturally, it is not just local domain knowledge that led to the design of the final network architecture and so the inductive bias. Even if the general choice for a convolutional net was epistemically motivated, various further implementation choices had to be made, prompted and driven by practical considerations such as computational convenience and tractability. As we saw before, the same holds for the implementation choices for SGD, which includes such things as the activation functions and the parameter initialization.

Such choices are, at the very least, "epistemically unforced" (Parker, 2014, p. 26). For example, while partly epistemically motivated, the weight initialization the authors pick is not clearly the unique epistemically superior choice. The same holds for the choice of the mean squared error as substitute for the untractable mean 0/1 loss.[52] I will leave here open whether it is useful to set apart such practical or pragmatic considerations from (other) non-epistemic values, specifically the ethical and social values Johnson has in mind.[53] What we see is that at least a certain kind of non-epistemic judgment inevitably comes into play when translating epistemic considerations into actual learning algorithms.

4.3.4. *The inference itself, revisited.* The absence of clearly social or ethical values in my analysis does point at a serious limitation of this analysis, namely the narrow focus on the learning algorithm. Even if the actual learning forms the defining core of machine learning, it is still only one part of a machine learning pipeline. Indeed, as I already intimated in section 3.2 above, one may wonder whether, for the discussion of value-ladenness, this narrow focus can be maintained at all.

[50] Including "well-known advantages to performing shape recognition by detecting and combining local features," that "if a feature detector is useful on one part of the image, it is likely to be useful on other parts of the image as well," and "a certain level of invariance to distortions and translations," (LeCun et al., 1989b, p. 399).

[51] Specifically, "sparse interactions, parameter sharing, and equivariant representations," (Goodfellow et al., 2016, pp. 324f).

[52] Karaca (2021) focuses on "cost-sensitive" loss functions ("using different costs for different types of training errors," ibid., p. 12), and still holds that "cost-insensitive" functions, which would presumably include the 0/1 loss in our example, merely serve the aim to "maximize the predictive accuracy" (ibid., p. 13). However, while perhaps not clearly associated with "social values," a choice for the 0/1 loss function is still not epistemically forced (cf. footnote 38). For more on the choice of loss function, and the complicated interaction between epistemic and non-epistemic considerations in real-world statistical analysis, see the discussion by Hennig and Kutlukaya (2007).

[53] For instance, Henschen (2021) takes "conventional or pragmatic reasons" to be distinct from value-laden reasons. Brown (2024, p. 13) criticizes his distinction, writing that "pragmatic considerations are [...] no less problematic than other nonepistemic values."

I conveniently used Johnson's idea of "the inference itself" to legitimize this focus on the learning algorithm, and this helped me to bring out my points clearly. Having done this, I readily admit that we may have to throw away this ladder: it is not obvious that one can so neatly demarcate the choices involved in learning algorithm design from choices earlier in the pipeline. For instance, when we consider the choice of loss function in our example, then we see that the clear boundary I presupposed in section 3.2.4 between problem formulation (including choice of loss function) and the design of the learning algorithm is not so clear at all.[54]

In general, it is plausibly the case that choices over the entire machine learning pipeline, from general problem formulation to model acceptance, interact in complicated ways. If this is so, then it is also plausible that considerations more directly related to the social and ethical specifics of the context play a role in the choices involved in learning algorithm design.

4.3.5. *Epistemic and non-epistemic values.* Furthermore, some of the theoretical tools I used, like the distinction between general learning rule and inductive bias, also become more blurry in practice.[55] Nevertheless, even amid all this added complexity, my point stands. There are epistemic reasons for choices in machine learning algorithm design, even if there are also always non-epistemic ones. The concern for accuracy is an epistemic concern, which as explained brings in certain epistemic motivations and justifications for reaching it. This is so even when, for instance, already the necessary choice of a particular way of measuring accuracy (via a loss function) brings in non-epistemic considerations, and even if we have to trade off accuracy for non-epistemic factors like computational feasibility, or choose to trade off accuracy for more clearly social or ethical factors like fairness.

I think that the picture offered here, that there are both epistemic and non-epistemic choices at play in machine learning, is more helpful than the picture suggested by Johnson's demarcation argument, that epistemic considerations reduce to non-epistemic values and so everything in machine learning is value-laden.

5. Conclusion

I have in this paper highlighted the epistemic motivations and justifications that underly machine learning algorithm design, in particular the epistemic motivation for appropriate inductive biases and the epistemic justification for domain-general learning rules. However, this discussion also showed how the practical realization of these epistemic considerations in actual learning algorithms inevitably brings in various non-epistemically value-laden choices. In contrast to Johnson's argument from underdetermination and demarcation, which as indicated also does not appear fully successful, the point of the inevitability of non-epistemic values did not follow so much from general philosophical argument as from looking more closely at the theory and practice of machine learning.

[54] In the epistemic projection approach to values in science of Parker (2024), as well as the proposal of hypothetical value-judgments by Schurz (2024), it seems that some such boundary between learning problem formulation and the actual inference would have to be upheld. It would be interesting to try to apply these proposals to (a particular case study in) machine learning.

[55] I discussed how the basic story of statistical learning theory is still discernible in the example of the first convolutional network for imagine recognition, but this is already much more questionable for present-day deep learning algorithms (also see fn. 49). The relevant inductive bias here appears not just determined by the (heavily overparametrized) architecture, but also by the general optimization rule (Sterkenburg and Grünwald, 2021, p. 10011).

I am not claiming that there is no successful principled philosophical argument for the necessary presence of non-epistemic values in machine learning algorithm design. I have not here discussed the argument from inductive risk, which Johnson did also bring in, and which is generally taken to be the strongest such argument (Brown, 2024). But for the subsequent problem of value management, it is surely helpful again to take a close look at the actual theory and practice of machine learning, and to also appreciate the epistemic considerations at play.

References

L. M. Antony. Bias: Friend or foe? Reflections on Saulish skepticism. In M. Brownstein and J. Saul, editors, *Metaphysics and Epistemology*, volume 1 of *Implicit Bias and Philosophy*, pages 157–190. Oxford University Press, 2016.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

M. Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

J. B. Biddle. State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science Part A*, 44(1):124–133, 2013.

J. B. Biddle. On predicting recidivism: Epistemic risks, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3):321–341, 2022.

J. B. Biddle. Values in artificial intelligence systems. In G. J. Robson and J. Y. Tsou, editors, *Technology Ethics: A Philosophical Introduction and Readings*, pages 132–140. Routledge, 2023.

A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. The values encoded in machine learning research. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 173–184. ACM, 2022.

E. Bond. Reasons, wants, and values. *Canadian Journal of Philosophy*, 3(3):333–347, 1974.

E. Bond. *Reason and Value*. Cambridge University Press, 1983.

M. J. Brown. For values in science: Assessing recent arguments for the ideal of value-free science. *Synthese*, 204(4):112, 2024.

M. J. Brown and J. Stegenga. The validity of the argument from inductive risk. *Canadian Journal of Philosophy*, 53(2):187–190, 2023.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

C. ChoGlueck. The error is the gap: Synthesizing accounts for societal values in science. *Philosophy of Science*, 85(4):704–725, 2018.

C. W. Churchman. *Theory of Experimental Inference*. Macmillan, 1948.

D. Danks and A. J. London. Algorithmic bias in autonomous systems. In C. Sierra, editor, *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 4691–4697, 2017.

J. S. Denker, W. R. Gardner, H. P. Graf, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, H. S. Baird, and I. Guyon. Neural network recognizer for hand-written zip code digits. In D. S. Touretzky, editor, *Proceedings of the First Conference on Advances in Neural Information Processing Systems (NIPS 1989)*, pages 323–331. Morgan Kaufmann, 1989.

R. Dotan. Theory choice, non-epistemic values, and machine learning. *Synthese*, 198(11): 11081–11101, 2021.

H. Douglas. Inductive risk and values in science. *Philosophy of Science*, 67(4):559–579, 2000.

H. Douglas. Values in science. In P. Humphreys, editor, *The Oxford Handbook in the Philosophy of Science*, pages 609–630. Oxford University Press, 2016.

B. d'Alessandro, C. O'Neil, and T. LaGatta. Contentious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, 2017.

K. C. Elliott. *Values in Science*. Elements in the Philosophy of Science. Cambridge University Press, 2022.

K. C. Elliott and R. Korf. Values in science: what are values, anyway? *European Journal for the Philosophy of Science*, 14(4):53, 2024.

K. C. Elliott and D. Steel, editors. *Current Controversies in Values and Science*. Current Controversies in Philosophy. Routledge, 2017.

S. Fazelpour and D. Danks. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), 2021.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, 2016.

T. Grote. Machine learning in healthcare and the methodological priority of epistemology over ethics. *Inquiry*, 68(4):1218–1247, 2025.

T. Grote, K. Genin, and E. Sullivan. Reliability in machine learning. *Philosophy Compass*, 19(5):e12974, 2024.

G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*. The Jean Nicod Lectures. A Bradford Book. MIT Press, 2007.

S. Harvard and E. Winsberg. The epistemic risk in representation. *Kennedy Institute of Ethics Journal*, 32(1):1–31, 2022.

R. Hatchwell. Individual values and inductive risk: Remotivating the Bayesian alternative. *Synthese*, 204(1):37, 2024.

T. Hellström, V. Dignum, and S. Bensch. Bias in machine learning - what is it good for? In A. Saffiotti, L. Serafini, and P. Lukowicz, editors, *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with the 24th European Conference on Artificial Intelligence (ECAI2020)*, pages 3–10, 2020.

C. Hennig and M. Kutlukaya. Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal*, 5(1):19–39, 2007.

T. Henschen. How strong is the argument from inductive risk? *European Journal for Philosophy of Science*, 11(3):92, 2021.

C. Howson. *Hume's Problem: Induction and the Justification of Belief*. Oxford University Press, 2000.

S. M. Huttegger. *The Probabilistic Foundations of Rational Learning*. Cambridge University Press, 2017.

K. Intemann. Feminism, underdetermination, and values in science. *Philosophy of Science*, 72(5):1001–1012, 2005.

R. C. Jeffrey. Valuation and acceptance of scientific hypotheses. *Philosophy of Science*, 23(3):237–246, 1956.

G. M. Johnson. The structure of bias. *Mind*, 129(516):1193–1236, 2020.

G. M. Johnson. Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*, 21(1–2):27–61, 2024.

K. Karaca. Values and inductive risk in machine learning modelling: The case of binary classification models. *European Journal for Philosophy of Science*, 11(4):102, 2021.

T. Kelly. *Bias: A Philosophical Study*. Oxford University Press, 2022.

T. S. Kuhn. Objectivity, value judgment, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, pages 320–399. University of Chicago Press, 1977.

L. Laudan. *Science and Values: The Aims of Science and their Role in Scientific Debate*. Pittsburgh Series in Philosophy and History of Science. University of California Press, 1984.

Y. LeCun. Generalization and network design strategies. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, and L. Steels, editors, *Connectionism in Perspective*, pages 143–156. Elsevier, 1989.

Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989a.

Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In D. S. Touretzky, editor, *Proceedings of the Second Conference on Advances in Neural Information Processing Systems (NIPS 1989)*, pages 396–404. Morgan Kaufmann, 1989b.

Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *International Symposium on Circuits and Systems (ISCAS 2010)*, pages 253–256. IEEE, 2010.

J. Lenhard. A transformation of Bayesian statistics: Computation, prediction, and rationality. *Studies in History and Philosophy of Science*, 92:144–151, 2022.

I. Levi. Must the scientist make value judgments? *The Journal of Philosophy*, 57(11): 345–357, 1960.

D. Li. Machines learn better with better data ontology: Lessons from philosophy of induction and machine learning. *Minds & Machines*, 33(3):429–450, 2023.

H. E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Knowledge*. Princeton University Press, 1990.

H. E. Longino. Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. Hankinson and J. Nelson, editors, *Feminism, Science, and the Philosophy of Science*, volume 256 of *Synthese Library*, pages 39–58. Kluwer, 1996.

C. Mayo-Wilson and G. Wheeler. Epistemic decision theory's reckoning. PhilSci-Archive preprint 16374, 2019.

E. McMullin. The rational and the social in the history of science. In J. R. Brown, editor, *Scientific Rationality: The Sociological Turn*, volume 25 of *The University of Western Ontario Series in Philosophy of Science*, pages 127–163. Reidel, 1984.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):115,1–35, 2021.

T. M. Mitchell. The need for biases in learning generalizations. Technical Report CMB-TR-117, Department of Computer Science, Rutgers University, 1980.

J. D. Norton. *The Material Theory of Induction*, volume 1 of *BSPS open series*. University of Calgary Press, 2021.

R. Nyrup. The limits of value transparency in machine learning. *Philosophy of Science*, 89(5):1054–1064, 2022.

W. S. Parker. Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science Part A*, 46:24–30, 2014.

W. S. Parker. The epistemic projection approach to values in science. *Philosophy of Science*, 91(1):18–36, 2024.

D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys*, 55(3):51,1–44, 2022.

R. Pettigrew. *Accuracy and the Laws of Credence*. Oxford University Press, 2016.

S. J. Prince. *Understanding Deep Learning*. MIT Press, 2023.

H. Putnam. For ethics and economics without the dichotomies. In H. Putnam and V. Walsh, editors, *The End of Value-Free Economics*, pages 111–129. Routledge, 2012.

E. Ratti. Machine learning and the ethics of induction. In J. M. Durán and G. Pozzi, editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, volume 527 of *Synthese Library*, pages 361–380. Springer, 2026.

E. Ratti and F. Russo. Science and values: A two-way direction. *European Journal for Philosophy of Science*, 14(1):6, 2024.

L. S. Rendsburg. *Inductive Bias in Machine Learning*. PhD Dissertation, University of Tübingen, 2024.

P. Rooney. On values in science: Is the epistemic/non-epistemic distinction useful? In K. Okruhlik and D. L. Hull, editors, *Proceedings of the Biennial Meeting of the Philosophy of Science Association (PSA 1992)*, volume 1, pages 13–22, 1992.

R. Rudner. The scientist *qua* scientist makes value judgments. *Philosophy of Science*, 20 (1):1–6, 1953.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

B. Rushing. No free theory choice from machine learning. *Synthese*, 200(5):414, 2022.

F. Russo, E. Schliesser, and J. Wagemans. Connecting ethics and epistemology of AI. *AI & SOCIETY*, 39(4):1585–1603, 2024.

G. Schurz. Hypothetical value judgments: Reconciling value-neutrality and value-engagement in science. In L. Magnani, editor, *Proceedings of the Académie Internationale de Philosophie des Sciences*, volume 2, pages 149–178. 2024.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

M. Sikorski and D. Liu. Non-epistemic values and the automation of science. *Journal for General Philosophy of Science*, forthcoming.

J. Sprenger and S. Hartmann. *Bayesian Philosophy of Science*. Oxford University Press, 2019.

D. Steel. Testability and Ockham's razor: How formal and statistical learning theory converge in the new riddle of induction. *Journal of Philosophical Logic*, 38(5):471–489, 2009.

D. Steel. Acceptance, values, and probability. *Studies in History and Philosophy of Science Part A*, 53:81–88, 2015.

T. F. Sterkenburg. *Universal Prediction: A Philosophical Investigation*. PhD Dissertation, University of Groningen, 2018.

T. F. Sterkenburg. Review of John D. Norton's *The Material Theory of Induction*. *Philosophy of Science*, 91(4):1030–1033, 2024.

T. F. Sterkenburg. Statistical learning theory and Occam's razor: The core argument. *Minds and Machines*, 35(1):3, 2025.

T. F. Sterkenburg. Statistical learning theory and Occam's razor: Structural risk minimization. PhilSci-Archive preprint 27569, 2026.

T. F. Sterkenburg and P. D. Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3):9979–10015, 2021.

E. Sullivan. Inductive risk, understanding, and opaque machine learning models. *Philosophy of Science*, 89(5):1065–1074, 2022.

E. Sullivan. How values shape the machine learning opacity problem. In I. Lawler, K. Khalifa, and E. Shech, editors, *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Routledge Studies in the Philosophy of Mathematics and Physics, pages 306–322. Routledge, 2023.

U. von Luxburg and B. Schölkopf. Statistical learning theory: Models, concepts, and results. In D. M. Gabbay, S. Hartmann, and J. Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. Elsevier, 2011.

Z. B. Ward. On value-laden science. *Studies in History and Philosophy of Science Part A*, 85:54–62, 2021.

E. Winsberg. Managing values in science: A return to decision theory. *Kennedy Institute of Ethics Journal*, 34(4):389–418, 2024.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

Munich Center for Mathematical Philosophy (MCMP), LMU Munich
Munich Center for Machine Learning (MCML)
Email address: tom.sterkenburg@lmu.de