

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 offers a more precise account of what it would mean for a “machine learning algorithm” to be “value-laden,” and, building on this, argues that a general argument from underdetermination does not warrant this conclusion.

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 non-epistemic value 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).

Yet there is something unsatisfying about the lesson that machine learning algorithms must be the product of value-laden choices beginning to end. For one thing, it seems odd to be led to the conclusion that, simply in virtue of their being procedures for inductive learning, standard learning algorithms like stochastic gradient

Date: December 20, 2024. This is a preliminary version. I welcome feedback.

descent or Bayesian updating must already be inherently value-laden. More generally, 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., 2023; Grote, forthcoming), is a more careful picture of how both epistemic and non-epistemic factors come together in the design of machine learning algorithms. One step towards such a picture is to show why a general argument from underdetermination does *not* already settle the matter that learning algorithms must be value-laden. That is the work of this paper.

To be clear, I do not seek to defend a claim that machine learning algorithms are not, in fact, value-laden. In the course of my analysis I indicate more specific paths for exposing value-ladenness, particularly there in algorithm design where epistemic considerations must meet practical demands. My point is that these paths require more detailed engagement with the actual learning algorithms, and as such require more work than a general argument from underdetermination.

The plan is as follows. In section 2, I rehearse Johnson’s version of the argument from underdetermination, which includes an argument against the possibility of demarcating epistemic from non-epistemic values. In section 3, I set up my critique by clarifying and delineating the relevant notions of “value-ladenness” and “machine learning algorithm.” Then, in the main section 4, I show why the underdetermination argument does not suffice to establish, at least as understood in the terms of section 3, the value-ladenness of learning algorithms. I introduce and refine my main counterobservations 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 underdetermination to value-ladenness in machine learning.¹ While the manner I make precise the claim of value-ladenness in section 3 may very well depart from Johnson’s own view,² I take her reasoning as summarized here to remain representative of how the argument from underdetermination goes, and as such it serves as the backdrop to my critical analysis of section 4.

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

¹Another is Dotan’s (2021), which I discuss briefly in section 4.2.3 below. Similar ideas are expressed in works like (Ratti, forthcoming). 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).

²I will indicate possible departures either in the text or in footnotes.

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).

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).³ 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).⁴

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).⁵ Second, this means that here too Kuhnian canons must come into play: “[i]f program engineers adhere

³Johnson 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.”

⁴Kuhn 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).

⁵Johnson (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.

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, 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 maintained. Johnson here turns to arguments in the 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).

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 against demarcation*, highlights the socio-political values that must drive the “meta-decision” 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 (that is, 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.).

Returning to the context of machine learning algorithms, the demarcation argument criticizes appeal to the value-free ideal in choosing 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

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

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 a call on non-epistemic values.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. None of the arguments that Johnson draws from are uncontroversial. For instance, Steel (2010) defends, against the demarcation argument, a distinction between epistemic and non-epistemic values; and Ward (2021) reasons that at least on one interpretation of “value-ladenness” the inductive risk argument is not plausible. I will in my critique call upon some of this existing work. However, my main strategy is to precisify what the underdetermination argument looks like, and where it fails, in the specific case of machine learning algorithms.

By “the underdetermination argument” I mean Johnson’s overall argument, namely that the inductive gap implies the need for further non-evidential factors (the argument step of section 2.1) which must be non-epistemic values (the argument step of section 2.2).⁸ My strategy will be to point out how the inductive gap at least *could* be bridged by epistemic factors, and why the demarcation argument does not succeed in refuting this possibility.⁹

⁷The pedigree of this argument is (Churchman, 1948; Rudner, 1953). For modern versions, see (Havstad, 2022, sec. 5; Brown and Stegenga, 2023; Brown, 2024).

⁸As 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 follow Johnson, trusting that her presentation (based on Douglas, 2016) is not too controversial.

⁹Johnson qualifies her claim of the in-principle value-ladenness of machine learning algorithms by restricting the “range of conceivable algorithms” to ones that are in some sense for “real-world use” (2024, pp. 28f, fn. 3), but by the conclusion she appears to have walked back her claim

My focus (like Johnson’s) in the second step will thus be on the demarcation argument, rather than the inductive risk argument. Indeed, I think that, in the case of machine learning algorithms, the question of the inductive inference itself *could* be separated from the question of acceptance, blocking the inductive risk argument at this level.¹⁰ At least, this is so on my proposed way of making the question of the value-ladenness of learning algorithms more precise, to which I turn now.

3. THE VALUE-LADENNESS OF LEARNING ALGORITHMS

In order to evaluate the underdetermination argument for the value-ladenness of machine learning algorithms, we need to clarify the notions of “value-ladenness” (section 3.1) and “machine learning algorithm” (section 3.2). An example of an actual learning algorithm serves to make things more concrete (section 3.3).

3.1. Values, choices, and reasons. What does 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 delineated, 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. I will here commit the same sin: I will not attempt a more precise account of the concept of value. However, I will adopt a useful taxonomy recently proposed by Ward (2021), who disambiguates different ways in which scientific choices can be value-laden.¹¹

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, I think it is also natural 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* or *construct* 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.¹²

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

somewhat further, writing that “machine learning programs [...] are value-laden to the extent that they are connected to and dependent on matters that we care about as human beings” (ibid., p. 56). I simply focus on what follows from the general underdetermination argument she gives, and while my analysis in the main section 4 starts from an abstract theoretical perspective, ultimately my interest is also in “real-world” algorithms.

¹⁰Such a separation is of course the classical response to the inductive risk argument, originally due to Jeffrey (1956) in a Bayesian picture. Note that I am not using this strategy to try and resurrect a version of the value-free ideal for machine learning. Since I will be narrowly concerned with the learning algorithm, I think that, in the taxonomy by Brown (2024) of possible responses (and their failure) to the modern inductive risk argument, the current work is closest to what Brown calls “partial rapprochements,” or “specifying value-laden and value-free moments of scientific inquiry” (ibid., p. 22). Brown’s view is that this strategy rather addresses the “value-management question,” which I am happy with. Ultimately, the purpose of this work is to get a clearer picture of how epistemic and non-epistemic factors interact in the design of learning algorithms.

¹¹My concern, like Ward’s, is with the *role* values play, rather than what values really *are* (cf. Elliott and Korf, 2024, p. 7).

¹²This 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).

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.).¹³

When it comes to the question of the necessary value-ladenness of machine learning algorithms, I likewise think that the causal interpretation is trivial and therefore not so interesting. 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, 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 not pin down Johnson’s argument to either interpretation. I will instead in each step 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). I will follow her in seeking to narrow down the question in this way. However, I think her own account still leaves ambiguous what stage in a machine learning pipeline the inductive inference—and indeed the algorithm—is supposed to correspond to.

3.2.1. The machine learning algorithm? Biddle (2022) gives an account of epistemic risks and value judgments 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

¹³Of 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).

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). As a way towards precisifying our question, let us ask: in Biddle’s taxonomy, what stage or stages does the question of the value-freeness of the algorithm actually pertain to?¹⁴

Stage (1) of problem identification/framing clearly precedes the stage of the algorithm, and is a typical point of entry of value judgments that we here want to put aside. The same holds for decisions about the data sets used for training and for benchmarking, located at stage (2). It gets more interesting with the stages (3) to (5), which are all three prefixed “algorithm design.” Just from this label, it seems these stages must be relevant to our question, given our understanding of values as pertaining to choices in algorithm design.

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 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, clearly 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.¹⁵ 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.¹⁶ 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). She also does not clearly

¹⁴As mentioned in the introduction, there exist several efforts in the computer science and the philosophy literatures 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 clarifying the question at hand, namely what aspects are part of the actual machine learning algorithm.

¹⁵For entries to this literature, see Barocas et al. (2023); Pessach and Shmueli (2022).

¹⁶Strictly 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).

keep the two different views on what is the relevant algorithm separate, or indeed defaults to a view of the relevant algorithm as the learned model.^{17,18}

To cast our problem as the value-ladenness of learned models has the advantage of an apparent direct parallel to the original debate about values in science. Where the question there mainly concerned scientists’ choices for (acceptance of) certain hypotheses or theories, the question would now be about developers’ choices for (acceptance of) certain machine learning models. But where exactly are the choices made in the latter case? Perhaps the most obvious place 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 kind of choice is the decision that the test error 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.¹⁹

However, this choice, whether or not to accept an already trained model, does not really seem to be the place of “the inference itself” in machine learning. More naturally, “the inference itself” pertains to how the model is arrived at in the first place: the model construction.²⁰ Now we can analyze the question of the value-ladenness of a machine learning model by also taking the various development choices in its construction into account. This includes the choice for the learning algorithm. But this also includes, for instance, the choice for the training data: the data is crucial in what the final model looks like. The problem with setting up the question in this way is that we end up in a similar situation as with the causal interpretation of values: it makes the question trivial. If the value-ladenness of a learned model is indeed already a matter of the values involved in early stages like data selection, then no argument from underdetermination is required: the question is settled as soon as we (plausibly, cf. Biddle, 2022, sect. 3.b) accept such choices are inevitably value-laden.

3.2.3. *The learning algorithm.* I think that at this point the natural thing to do is to explicitly restrict attention to what is, after all, the core inductive inference step in machine learning: the actual learning step, executed by the learning algorithm. I think that when we ask about the value-ladenness of machine learning algorithms, and particularly what follows from underdetermination of “the inference itself,”

¹⁷Initially, 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.

¹⁸As a matter of fact, the infamous COMPAS system “is not a ML model—it was not created by any standard ML algorithm. It was designed by experts based on carefully designed surveys and expertise” (Rudin, 2019, p. 209; also see Rudin et al., 2020). So, as an illustration or application of an argument for the value-ladenness of machine learning algorithms, irrespective of one’s view of what “the algorithm” actually is, the COMPAS case is not the most fortunate choice.

¹⁹As 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.

²⁰Here I might be departing from Johnson’s own view. In fact, she immediately clarifies “inductive inference itself” as “the point at which we decide to accept or reject some conclusion” (Johnson, 2024, p. 35), which might be interpreted as pertaining to the model acceptance step.

the natural interpretation of this question adopts the original view of what is the relevant machine learning algorithm: not the learned model, but the actual learning algorithm.²¹ In this paper I will therefore consider the question whether it follows from underdetermination that learning algorithms must be value-laden, that is, whether it follows that non-epistemic values must enter as reasons for choices of design of learning algorithms.

3.2.4. *Inference itself, conclusion.* To further clarify the scope of this question, let me 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. One might again say that as aspects of models, these stages primarily concern choices of evaluation and acceptance of trained models.²² Yet that is too quick: these aspects can clearly already play a role in designing the learning algorithm. This is quite obvious for the choice of outputs; and the same holds for a related design choice which Biddle does not discuss, but others in this context do, namely the choice of cost or loss function.²³ But also in the case of fairness, for instance, 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, there is also something odd about viewing fairness criteria 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 nice and clean perspective that I will adopt in my critical analysis in section 4 below. A neat separation between problem formulation and inductive inference step will be important to get clear on what exactly follows from inductive underdetermination. This is so even if such a neat separation becomes hard to sustain when we look at the design of real machine learning algorithms—as I will do at the end of my analysis. There I will return to the following concrete example

²¹This is not to deny 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.

²²For 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.

²³Karaca (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.

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.²⁴

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.²⁵ 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 people 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.²⁶ As such, again, a trained model based on these data is inevitably value-laden, too.

However, our question is whether the *learning algorithm* must be value-laden. 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. 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

²⁴This 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).

²⁵This 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).

²⁶The authors note failure of generalization due to “writing styles not present in the training set” (1989a, p. 547).

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. AGAINST THE UNDERDETERMINATION ARGUMENT

Does it follow from inductive underdetermination (as discussed in section 2) that learning algorithms must be value-laden, that is (as made more precise in section 3), that non-epistemic values must enter as reasons for choices in the design of learning algorithms? In this section I argue, no.

I do so by formulating and refining a number of counterobservations 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).

An important tool throughout my analysis is a distinction, introduced by [Sterkenburg and Grünwald \(2021\)](#), between domain-general *learning rules* and the domain-specific *inductive biases* they must be equipped with, together forming the actual *learning algorithms*.

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., [Earman, 1992](#); [Sprenger and Hartmann, 2019](#)). While it is still a significant step from this basic picture to actual (Bayesian) machine learning,²⁷ the basic picture has the advantage of being both relatively simple and familiar, while sufficient to already introduce the main observations against the argument from underdetermination.²⁸

In the basic Bayesian learning procedure, one starts with a probability function p over propositions in some formal language.²⁹ 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,

$$(1) \quad p'(\cdot) := p(\cdot \mid E),$$

where $p(H \mid E)$ can be calculated using Bayes's theorem,³⁰

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

4.1.1. The Bayesian learning rule. 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 . What is the motivation for this rule? One can distinguish two main components in philosophical justifications for the Bayesian approach, both in terms of the epistemic value of *rationality*.

²⁷For machine learning textbooks from a Bayesian perspective, see [Bishop \(2006\)](#); [Murphy \(2012\)](#); [Barber \(2012\)](#).

²⁸Johnson also refers in various places to the philosophy of science literature on inductive inference, and in particular Bayesian learning.

²⁹I 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.

³⁰Bayes'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.

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).³¹ 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 justifications for Bayesian updating as the rational way of learning from evidence.³² The justifications are therefore explicitly intended to be domain-general and epistemic. At the same time, none of them are uncontroversial: objections have been raised against many if not all of these proposed justifications.

I will discuss in more detail below what this means for the question of value-ladenness. But first I will complete the picture of Bayesian learning by considering the component of the prior probability function. This is the component where the actual underdetermination comes into focus.

4.1.2. *Bayesian learning algorithms.* As mentioned, the prior probability function is an indispensable input component to the conditionalization rule (1). I use the term “Bayesian algorithm” to refer to any implementation of the Bayesian rule with already a particular prior provided. Thus a Bayesian algorithm is a procedure that just takes input data and returns an output probability function; with a particular prior, so to speak, part of the inner mechanism of the algorithm.

Such a Bayesian algorithm 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 encode restrictive assumptions (see, e.g., Howson, 2000; Huttegger, 2017; Sterkenburg, 2018). These assumptions encoded 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* (ibid.).

4.1.3. *Canons of induction or local assumptions?* As we have seen, Johnson assumes that the inductive gap must be bridged by Kuhnian values or “canons of induction.” In the case of Bayesian algorithms, this would mean that the priors must stem from or even encode such general canons.

One certainly *can* utilize or encode such general criteria in the formulation of a Bayesian prior. In a well-known paper, Salmon (1990) argued that at least some

³¹This component deals with the kind of quantities that the Bayesian algorithm manipulates, and so rather with the formulation of learning problem (cf. section 3.2.4 above). But the question of the value-ladenness of the choice for Bayesian learning rule is already hard to set apart from the overall choice for Bayesian framework. Consequently, my discussion in section 4.1.5 below of the Bayesian learning rule considers the value-ladenness of the general Bayesian approach.

³²Note that at least some of these justifications in terms of rationality trade on the (more?) fundamental epistemic value of *accuracy*. Also see footnote 39.

of Kuhn’s theoretical virtues naturally inform plausibility judgments that can go into Bayesian priors for problems of theory evaluation. But it not clear that one *must*. Another possibility is to bridge the inductive gap by context-specific or *local* assumptions about the learning problem at hand. Salmon indeed notes that “the assignment of prior probability by the Bayesian can be regarded as the best estimate of the chances of success of the hypothesis or theory on the basis of all relevant experience *in that particular scientific domain*” (ibid., p. 186, emphasis mine).

Norton’s (2003; 2021) material theory of induction even holds that all inductive inferences are solely “powered” by local facts. This may be taking it too far into the other extreme³³—but it seems at least *possible* to formulate, for some learning problems, priors that are an expression of local, problem-specific beliefs. To adapt an example from Norton (2021, sect. 1.9), 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*.

This is, of course, a very stylized example, far removed from realistic machine learning problems. But the basic observation, that there is the further possibility of bridging the inductive gap, not with general canons, but with local assumptions, stands; and the digit recognition example will provide a more realistic illustration in section 4.3 below. The observation is important, because it already blocks the move from underdetermination to the need for general canons or values, and therefore the main premise for the subsequent argument against demarcation.³⁴

4.1.4. *Value-ladenness of local assumptions.* Still, even if the need for general canons or values does not automatically follow from underdetermination, could we not directly argue that local inductive assumptions must be value-laden? Harking back to Ward’s taxonomy (section 3.1 above), we can distinguish two questions here: must non-epistemic values enter as *motivating* reasons for making such local assumptions, and must non-epistemic values enter as *justifying* reasons for such assumptions?

When it comes to motivating reasons, it is not at all clear that they must (cf. Ward, 2021, p. 60). Again, it seems at least possible to try and formulate a Bayesian

³³In particular, 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).

³⁴While Johnson invokes Kuhn’s general virtues as typical canons of induction, she also writes that “each domain comes with its own set of assumptions [canons] on which it relies” (2024, p. 33). Moreover, she notes (ibid., p. 33, fn. 13) Norton’s (2021, ch. 5) critique that the terms epistemic “virtues” or “values” misleadingly suggest that these are free-to-choose ends (rather than means for finding the truth, objectively better or worse), suggesting that she would be open to understanding epistemic values or canons in Norton’s sense, as (surrogates for) local assumptions. But then the problem remains that on this understanding of canons or epistemic “values” as local assumptions, the demarcation argument does not seem applicable.

prior as an honest assessment of what one believes to be the factual structure of the relevant domain, an assessment with a purely epistemic motivation.^{35,36}

The case of justifying reasons for local inductive assumptions is more difficult. There is here an immediate further question of what counts as a proper justification for such assumptions. 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. If we accept the skeptical argument, then we must conclude that any inductive assumptions are ultimately lacking such justifying epistemic reasons. But of course, 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 just follow that in the lack of ulterior epistemic justifying reasons, non-epistemic values must enter the picture. That the justificatory gap is bridged by values is not obviously the default option: it is something that would require additional argument (cf. Intemann, 2005).³⁷

4.1.5. *Rationality and demarcation.* Still, even if the need for general canons or values does not automatically follow from underdetermination, do such values not already come into play at an earlier stage? Namely, irrespective of the prior, it is the Bayesian approach (and so in particular the Bayesian updating rule) itself that is underwritten by certain general considerations: specifically, considerations in support of the idea that the Bayesian procedure instantiates the epistemic value of rationality. Could we not already launch the demarcation argument at this stage, against the Bayesian updating rule?³⁸

Let me start by asking directly: is it plausibly the case that the motivating reasons for the design of the Bayesian learning rule must be (or must have been) non-epistemic? Note here that 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 project of philosophers and theoreticians to formulate principles of rational learning and develop arguments that the Bayesian approach

³⁵Biddle (2013) discusses the influence of “contextual factors” (non-epistemic values) on the prior, and concludes that “on the Bayesian account, it is *impossible* to screen out all contextual factors from the epistemic appraisal of transiently underdetermined research” (p. 128). However, this must be understood as the weak claim, consistent with my observations, that the Bayesian apparatus cannot *guarantee* that “contextual factors can always be excluded” (p. 127).

³⁶One can again counter that this only holds for stylized scenarios, far removed from realistic machine learning. For example, in Bayesian machine learning, considerations of computational tractability normally limit choice to certain flavours of default priors, i.e., default parametrized hypothesis classes plus a default prior distribution over parameters. Steel (2015) flags precisely this aspect in the context of Bayesian statistics. In section 4.3.4 below, I return to the role of pragmatic factors in actual learning algorithm design.

³⁷Johnson (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.

³⁸Johnson (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.

(and in particular the Bayesian updating rule) satisfies those. It seems hard to deny that this is an emphatically epistemic project, and that these research efforts are explicitly motivated by epistemic reasons.

What about the demarcation argument? Recall that 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. 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). The presence of a value-laden context might certainly be a reasonable presumption for specific learning algorithms developed for specific real-world problems. But it does not clearly apply to the foundational project described above, where the aim is not to solve a real-world learning problem in a particular socio-political context, but to formulate a general learning approach (and learning rule), supported by domain-general epistemic considerations.

Of course, the adoption of the Bayesian approach (and so the Bayesian rule) in any particular real-world learning problem might very well be motivated by non-epistemic considerations, inherited from the particular socio-political context. But I do not see that it *must*. Even in a particular socio-political context, it is not clear why an algorithm designer *could* not defer to (and be motivated by) the existing context-independent and epistemic reasons for the general Bayesian learning approach (and learning rule), thus evading the demarcation argument.

I have so far discussed the motivating reasons for the Bayesian rule: what about justifying reasons? Recall that while there are various justificatory arguments for the rationality of the Bayesian approach (and Bayes’s rule in particular), these arguments are not uncontroversial.³⁹ To take this as an opening to argue for the presence of non-epistemic values would also not be uncontroversial, though; it would amount to taking a side in a (*the?*) central debate in Bayesian foundations.

A final “nuclear” option is to already deny that rationality is a purely epistemic value. Here the idea might be something along the lines of Johnson’s constitutive interpretation of the demarcation argument, namely that the value of rationality (and its justification) cannot be seen apart from wider (sociological, cultural, historical) circumstances under which rationality came to be valued; for instance, how rationality as “cold calculation” came to be valued in a patriarchal society (cf. Buckner, 2023, p. 78ff). This could be a promising line for the value-ladenness of the Bayesian approach, but still one that clearly needs more development.

In any case, the value of rationality might be central to Bayesianism, but does not play such a central role in (non-Bayesian) machine learning approaches. This includes the standard approach to statistical classification, to which I turn next.

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

³⁹ 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).

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 *empirical risk minimization* (ERM) rule, proceeds as follows. It works with a class \mathcal{H} of possible classifiers, also called the *hypothesis class*, and on receiving a training sample, it selects a classifier 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 motivation 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 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. Namely, 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

Thus, from this abstract learning-theoretic perspective, the ERM rule is a recipe for attaining a certain formal accuracy. This is a means-ends justification: for this specific end of minimizing true risk, ERM is the right means. Provided, moreover, that the accuracy goal of minimizing true risk counts as an epistemic goal, this learning-theoretic result counts as an epistemic justification for the ERM rule. I will unpack this epistemic basis more below, but first I will complete the picture of the ERM method with the component of the hypothesis class.

4.2.2. *ERM algorithms.* Similarly to the case of the Bayesian algorithm and the prior distribution, the ERM rule requires, apart from the data, a further input component: the hypothesis class. I will use the term “ERM algorithm” for any implementation of the ERM rule with already a particular hypothesis class \mathcal{H} provided. The resulting procedure takes a training sample and outputs a classifier, with the hypothesis 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

⁴⁰See, 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 \(2014\)](#).

⁴¹Capacity 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.

⁴²The 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](#)).

might not have been so if there existed some “universal” hypothesis class, and if a “universal” ERM algorithm equipped with such a hypothesis class would still have a PAC guarantee. Namely, a guarantee of finding the approximately-best classifier in this universal class would actually be a guarantee of finding the approximately-*absolutely*-best classifier. Such a universal algorithm could be said, perhaps, to objectively extrapolate the patterns in the data to a choice of classifier.

But there can be no such universal ERM method, as shown by impossibility results that are usually referred to as “no-free-lunch theorems.” In particular (see [Shalev-Shwartz and Ben-David, 2014](#)), 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 absolute true risk is high. 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 hypothesis 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](#); [Shalev-Shwartz and Ben-David, 2014](#); [Sterkenburg and Grünwald, 2021](#)).⁴³

More specifically, the strength of the PAC guarantee is a direct function of the complexity or capacity of the hypothesis class (and so, more informally, of the strength of the inductive bias). This translates in a certain means-ends justification for a 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. *No-free-lunch 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 classifier) by (classifiers’ accuracy on) 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.

However, Dotan actually makes a much stronger observation. She writes that all hypotheses have the same “expected accuracy”—in our terms, the same *true* risk. This observation follows from her discussion of the original no-free-lunch theorem

⁴³Note that the term “inductive bias” does not yet have a normative connotation, in line with the use of “bias” by Johnson (footnote 3 above; also see [Kelly, 2022](#)).

⁴⁴[Sterkenburg \(2025\)](#) argues in detail that if there is some sort of justification for a simplicity preference (a justification for Occam’s razor) to be had from the fundamental theorem of SLT, it is in the form of this reasoning. Importantly, simplicity does not act here as an independent Kuhnian theoretical virtue, but as a provable pre-condition for a certain guarantee of accuracy.

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

for supervised learning, due to [Wolpert \(1996\)](#). But as discussed by [Sterkenburg and Grünwald \(2021\)](#), this result crucially relies on the assumption of a uniform distribution over possible learning situations. This is an assumption that is not just lacking in motivation, but is an explicit assumption of full-blown randomness and therefore “unlearnability”—no matter the past data, the best prediction will always remain a random guess. As a premise to an argument that all hypotheses have the same expected accuracy, this assumption is clearly question-begging.⁴⁶ Fortunately, Dotan’s stronger observation does not seem necessary for her argument. The essential input for the subsequent two steps of her argument is the need for further assumptions 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 the uniformity assumption.

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 really 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 argument from demarcation to proceed to the need for non-epistemic values. Her argument is rather that any candidate epistemic virtue would have to somehow reduce to accuracy, which is not enough to bridge the underdetermination.

The same observation as before (section 4.1.3) therefore also blocks Dotan’s argument: it is not clear that underdetermination entails the need for further general canons or epistemic values. Namely, similarly to Bayesian priors, it seems at least possible to formulate hypothesis classes that encode context-dependent, local assumptions, *viz.*, about what classifiers we think are likely to be accurate for the problem at hand. The corresponding inductive bias is still motivated by the epistemic concern of accuracy.

Again, this leaves the worry that even if the inductive bias can be epistemically motivated, it cannot be epistemically justified: the Humean justificatory regress cannot be halted by epistemic reasons. But here we can repeat the other observation from before (section 4.1.4). Namely, the mere suggestion that “non-epistemic values are natural candidates to supplement accuracy” hardly settles that non-epistemic reasons must come in to halt the regress: this would require further argument.

4.2.4. Predictive accuracy and demarcation. Like in the Bayesian case, one might directly raise a demarcation worry for the epistemic values that underly the ERM rule. Since ERM is a provably good means towards a specific end, it seems that this worry must focus on this end: predictive accuracy, as made precise in SLT.

Like in the Bayesian case, however, the demarcation arguments’ presumption of a particular socio-political context, imbued with non-epistemic values, does not

⁴⁶[Rushing \(2022\)](#) also argues that the uniformity assumption in Wolpert’s no-free-lunch theorem is a problem for Dotan’s argument.

compellingly hold true. The relevant context of the design and analysis of the SLT framework (and the ERM rule) is really the explicitly epistemic project of theoretical computer scientists to provide a foundation for successful classification. In particular, the general learning objective that is center to this theoretical project, predictive accuracy, is quite clearly epistemically motivated.

That said, there are further assumptions involved in the specific way predictive accuracy is made precise in SLT. The formal accuracy goal of minimizing true risk relies, in particular, on the assumption of an unknown distribution from which instances are sampled in an i.i.d. manner, and this assumption is clearly not fully domain-general. There clearly are real-world learning problems where the i.i.d. assumption is not plausible: indeed, there surely are real-world problems where the i.i.d. assumption would be a value-laden modeling choice.

But that in itself only shows that the range of application of the framework of SLT (and the ERM rule), while general, is not unrestricted. It does not automatically follow that this restricted range is characterized by certain non-epistemic values, which therefore motivate the theoretical framework and method. We can still hold that the i.i.d. assumption is a component in the service of a precise epistemic notion of accuracy that is appropriate across a wide range of domains, and for this reason formulated and studied by theoreticians within a general framework for the analysis of learning algorithms. This is perfectly consistent with the observation that the framework is not fully general (indeed, theoretical computer scientists also study other general frameworks, with different assumptions), and hence that for various particular real-world problems, it may be inappropriate.

When it comes to justifying reasons, the epistemic value of accuracy seems even less controversial than that of rationality. Rather than standing in need of further justification, (predictive) accuracy is normally just accepted as an axiomatic epistemic goal—including, as we saw, by [Dotan \(2021\)](#). Still, again, there is the specific way accuracy is made precise in SLT, and the i.i.d. assumption is subject to the question of inductive justification.⁴⁷ There is an obvious question of the justification for the i.i.d. assumption in any particular learning problem; and it is perhaps an interesting question what justification there is for incorporating this assumption in a general framework for learning. But in both cases we can repeat the now familiar observation that the lack of rock-bottom epistemic justification does not, without further argument, prove that there must be non-epistemic reasons involved.

4.3. Example: handwritten digit recognition. The natural concern about the discussion so far is that it took an overly theoretical perspective on machine learning, and was therefore a level of abstraction too far removed from actual machine learning algorithms. Here I meet this concern by showing how my observations still apply to an actual learning algorithm, namely the digit recognition example.

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. 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).

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

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 output 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. *Theory v. practice.* One might wonder, of course, whether the authors were here in actual fact motivated by the theory they evoke, or rather by their hard-won practical experience. Indeed, it is probably safe to say that most of their design choices were driven by empirical experience (ranging from informal hunches to more systematic experimentation), or at best by a mix of practice and theory.⁴⁸ Unsurprisingly, the choices that go into practical algorithm design are not exclusively motivated by theoretical considerations.

Does this role for practical experience already mean that non-epistemic values must be involved? Not directly. The empirical fact that certain design choices or principles led to good prediction can clearly be an epistemic motivation for implementing these. When it comes to justifying reasons, one might take the position that expectations that previously successful practices continue to work are not justified,

⁴⁸For example, SGD’s superior convergence compared to a “‘true’ gradient procedure” was found by “empirical study (supported by theoretical arguments)” (LeCun et al., 1989a, p. 546).

unless underwritten by theoretical guarantees. This may be a plausible account of justification, but still one that stands in need of development and defense.

However, the main take-away from the previous section is perhaps not so much the role of practical experience, but the need for various practical or pragmatic choices in translating theoretical considerations into an implementable algorithm. This *is* an important place for certain non-epistemic judgments. I return to this below, after discussing the inductive bias.

4.3.3. *The network architecture.* We saw that general epistemic considerations motivated the authors to minimize the capacity of the model class. But how did they go about this? 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 domain⁴⁹ informs design choices characteristic of convolutional nets,⁵⁰ which we can imagine as representing or implementing the kind of prior Goodfellow et al. describe. That is, local knowledge about the task domain informs the network’s inductive bias.

4.3.4. *The role of pragmatic choices.* But of course, it is not just local domain knowledge that led to the design of the final network architecture, and therefore 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.

⁴⁹Including “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).

⁵⁰Specifically, “sparse interactions, parameter sharing, and equivariant representations,” (Goodfellow et al., 2016, pp. 324f).

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.⁵¹ The loss function is particularly interesting here, because it already shows how in practice the lines between problem formulation (including the choice of loss function) and design of learning algorithm for “the inference itself” are blurred.

I will leave here open whether it is useful to set apart such practical or pragmatic considerations from (other) non-epistemic values.⁵² What we see is that at least a certain kind of non-epistemic judgments virtually inevitably come into play when translating epistemic considerations into actual learning algorithms. But rather than the result of a general argument from underdetermination,⁵³ this observation is a starting point for a more careful analysis of how values enter algorithm design.

5. CONCLUSION

I have argued that a general argument from the underdetermination of the inductive “inference itself” does not suffice to establish that machine learning algorithms must be value-laden. Of course, this in turn does not mean that real-world learning algorithms are, in fact, value-free. But it does suggest that the more fruitful approach is a more fine-grained study of the design of such learning algorithms.

In support of such projects, I think the work in this paper constitutes a small step towards charting the possible interactions between epistemic and non-epistemic aspects of machine learning. A small step, because of the narrow focus on the *learning algorithm*: even if the actual learning forms the defining core of machine learning, it is still a small part of a practical machine learning pipeline. A small step, furthermore, because the approach I took in this paper was to start from a lofty theoretical perspective on learning methods. I think this was the right approach to confront the general argument from underdetermination, but it is a valid question to what extent this perspective is still discernable in real-world learning algorithms. Neat contrasts I drew in this paper, like that between the inductive problem formulation and the inductive learning method, and that between a general learning rule and a context-specific inductive bias, are much more blurry in many actual algorithms.

A central notion here is inductive bias. Indeed, along with the question of the inductive bias’s (non-)epistemic motivations or justifications, there is the question

⁵¹As mentioned before (section 3.2.4), the choice of loss function is a natural place to look for value-ladenness. Karaca (2021) focuses on “cost-sensitive” loss functions (“using different costs for different types of training errors,” *ibid.*, p. 12); in contrast, “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, and in that sense a pragmatic choice. In general, one can argue that while the goal of accuracy is epistemically pure, any one of the endless possible ways of quantifying inaccuracy is not (cf. footnote 39). For more on the choice of loss function, and the complicated interaction between epistemic and practical considerations in real-world statistical analysis, see the discussion by Hennig and Kutlukaya (2007).

⁵²For instance, Henschen (2021) takes “conventional or pragmatic reasons” to be distinct from value judgments. Brown (2024, p. 13) criticizes his distinction, writing that “pragmatic considerations are [...] no less problematic than other nonepistemic values.”

⁵³For instance, the need to approximate ERM by SGD is not an issue of inductive underdetermination.

what the inductive bias actually *is*. There is the question of what inductive assumptions particular learning algorithms actually operate under, which motivates work in computer science on uncovering the inductive biases in popular learning algorithms (Rendsburg, 2024). But there is also the fundamental question of what distinguishes—if a principled distinction can be made at all—inductive bias from other types of biases in machine learning. This question points towards a philosophical project of conceptual clarification of the notion of inductive bias itself.

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.
- D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.
- S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- 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.
- C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Sciences and Statistics. Springer, 2006.
- 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, 112:1–31, 2024.
- M. J. Brown and J. Stegenga. The validity of the argument from inductive risk. *Canadian Journal of Philosophy*, 53(2):187–190, 2023.
- C. J. Buckner. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence*. Oxford University Press, 2023.
- 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 Twenty-Sixth 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 1st 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*, 198:11081–11101, 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.
- J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. A Bradford Book. MIT Press, 1992.
- 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, 53:1–24, 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*, forthcoming.
- 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.
- J. C. Havstad. Sensational science, archaic hominin genetics, and amplified inductive risk. *Canadian Journal of Philosophy*, 52(3):295–320, 2022.
- 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, 92:1–23, 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*, 72:1001–1012, 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, 102:1–27, 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 2nd 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.
- 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*, 33:429–450, 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.
- K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. MIT Press, 2012.
- J. D. Norton. A material theory of induction. *Philosophy of Science*, 70(4):647–670, 2003.
- 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. Parker. Values and uncertainties in climate prediction, revisited. *Studies in History and Philosophy of Science Part A*, 46:24–30, 2014.
- 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. D. Prince. MIT Press, 2023.
- E. Ratti. Machine learning and the ethics of induction. In J. Duran and G. Pozzi, editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Synthese Library. Springer, forthcoming.
- E. Ratti and F. Russo. Science and values: A two-way direction. *European Journal for Philosophy of Science*, 14, 6:1–23, 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 one: contributed papers, pages 13–22, 1992.
- C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.
- C. Rudin, C. Wang, and B. Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1):206–215, 2020.
- 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, 414:1–21, 2022.
- F. Russo, E. Schliesser, and J. Wagemans. Connecting ethics and epistemology of AI. *AI & SOCIETY*, 2023.
- W. C. Salmon. Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage, editor, *Scientific Theories*, volume 14 of *Minnesota Studies in the Philosophy of Science*, pages 175–204. University of Minnesota Press, 1990.
- S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- 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. Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77(1):14–34, 2010.
- 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, 3:1–28, 2025.
- T. F. Sterkenburg and P. D. Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199:9979–10015, 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.
- D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996.