The Immortal Science of ML:
Machine Learning & the Theory-Free Ideal

Mel Andrews

July 31, 2025

Abstract

This paper contends with the notion that the methods of machine learning (ML) are unique among the tools of science in enabling a form of theory-free inductive inference. I contest these assertions of epistemic distinctness, attributing the prevalence of these views to an untenable conception of scientific objectivity: what I term a theory-free ideal, in homage to its normative counterpart. ML, as a formal method of induction, must rely on conceptual or theoretical resources to get inference off the ground. By means of two case studies, I argue that the theory-free ideal has a deleterious effect on the epistemic standing of ML-involving science.

1 Introduction

In the decade elapsed since the deep learning revolution, machine-learning (ML) techniques have established a firm foothold throughout the sciences. Successes of such methods have ranged from real-time particle-event sortation at the Large Hadron Collider (Duarte et al., 2018) to DeepMind's Nobel prize-winning accomplishments with its AlphaFold 2.0 system for protein-structure prediction (Jumper et al., 2021). According to certain spokespeople from science and engineering communities, however, these achievements are trivial in comparison to the potential such methods hold for scientific advancement.

Various commentators have asserted with some frequency that ML will instigate profound—even "revolutionary"—changes to the nature of science and the knowledge it produces (Anderson, 2008; Boge, 2022; Hey et al., 2009; Mayer-Schönberger & Cukier, 2013; Society & Institute., 2019; Spinney, 2022; Srećković et al., 2022). The people behind such claims see ML methods as holding the potential to retire or else displace the role of theorizing in science (Anderson, 2008; Mayer-Schönberger & Cukier, 2013; Spinney, 2022; Srećković et al., 2022). Desai et al. (2022) refer to this conception of an ML-enabled scientific paradigm as "the epistemically revolutionary new frontier raised by data science: the so-called 'theory-free' paradigm in scientific methodology" (p.469). Some of these statements regarding the scientific usage of ML echo proclamations that were once made of domain-generic statistical analyses: Levins and Lewontin (1985) write of the motivation behind principle component analysis (PCA) and regression techniques that researcher's "assumption is that they are approaching the data in a theory-free manner and that data will 'speak to them' through the correlation analysis" (Levins & Lewontin, 1985, 156). We see this sentiment echoed today in the assertion that big data analytic tools promise to allow the raw data to "speak for themselves" (Anderson, 2008, 1).

This theory-freedom is intended as a negation of theory-mediation, theory-drivenness, theory-involvement, and theory-ladeness. It is also, we will see from an examination of the source literature, a denial that methods rest essentially on domain-knowledge or prior conceptualization of the target phenomena, or that they should be understood as representing features of target systems in any inference-licensing respect. "Theory" is hence to be understood in a broad and colloquial sense, as incorporating domain knowledge or conceptualization of target phenomena. Subscribers to the theory-free ideal seek to purge science of what they see as epistemically compromising arbitrariness and subjectivity. This subjective element is brought on board when human critical thinking or conceptualization of target phenomena play an essential role in shaping an empirical research program.

Taken at face value, the claim that ML could enable a form of theory-free inductive inference—or even inductive inference that differs qualitatively in its degree of theoretical support—does not withstand scrutiny. That inductive inference rests on pregiven conceptual infrastructure is, I take it, an effective premise of all discussions of the subject since Hume. Neither does inductive inference move along a sliding scale between theoretically and empirically driven: induction is, definitionally, a form of inference that requires both theoretical and empirical support. These are neither quantifiable nor scalar.

Ample scholarship in epistemology dating back centuries has characterized induction. The present paper is not a thesis in epistemology, nor is it intended as a revision to philosophical conceptions of induction, of which I take Hume's characterization to be adequate. My aim in this paper is, instead, to analyze claims regarding the use of ML in science, its epistemic status, and its transformative potential. My engagement with these questions seeks to improve upon prior philosophical treatment in distinguishing between normative and descriptive agendas. Claims of the epistemic distinctness of ML, I contend, latch onto real novelty in some instances of ML deployed toward scientific ends: potential for misuse and lack of methodological standards. Instead of identifying this as the epistemic problem it represents, however, claims of epistemic distinctness and theory-freedom function to reify the (potential) misuse of ML-based tools into an account of how these tools normally function, how they necessarily function, or even how they normatively *ought* to function.

In the course of this paper, I will characterize the misdirected conception of scientific objectivity implied by these claims. I refer to the meta-narrative that endorses this notion of objectivity as a *theory-free ideal*, paralleling the *value-free ideal*, which is its normative counterpart. This meta-narrative, I will show, is detrimental to the epistemic soundness of science conducted with ML.

My argument proceeds as follows. Claims of the epistemic novelty of ML-involving methods concerning their reliance on theory have been alleged by both scientists and philosophers. I furnish exemplars of these in 2.2. Successful inductive inference, however, requires background assumptions; this is one of the defining characteristics of inductive inference, according to known formal and epistemological accounts, as I show in section 2.4. This presents us with a dilemma: either revise our understanding of induction, or conclude that ML only enables inductive inference in virtue of theoretical input. I urge that the effects of such beliefs on scientific practice with ML should be the most salient consideration; we should accept the second horn for pragmatic reasons, which I endeavor to motivate via case study. In Section 2.5.1, I show that when ML is incorporated into scientific pipelines to epistemic success, it is in virtue of working explicitly with theoretical resources. These come in at the junctures of problem formulation, data collection and curation, model design, model training, and model evaluation. When, in contrast, investigators ignore or suppress theoretical assumptions in pursuit of an ideal of theory-freedom, methodological pathologies emerge: statistical artifacts are mistaken for structure, arbitrary modeling choices or starting assumptions are mistaken for empirical discoveries, and downstream inference is rendered unreliable. A case study explored in Section 2.5.2 illustrates this contrast class. I attribute the prevalence of the theory-free conception of ML to a meta-narrative concerning ideals of scientific objectivity: the *theory-free ideal.* I conclude with recommendations for scientists in working explicitly with theoretical resources, and recommendations for philosophers in their engagement with ML and the claims surrounding it.

2 Distinctness

2.1 The beliefs of working scientists

A monograph titled "The AI revolution in scientific research," released jointly by The Royal Society and the Alan Turing institute, offers scientists' own assessments of anticipated changes to scientific practice spurred by the involvement of ML (Society & Institute., 2019). Summarizing the opinions of the assembled scientists, the authors write that "AI" is set to have "a disruptive influence on the conduct of science"(Society & Institute., 2019, p.10). Such pronouncements appear to be underpinned by a conception of the workings of ML in science as a theory-free enterprise, given the authors' description of the normal function of ML and data scientific methods. The standard way to apply ML in science, they write, is "to start from a large data set, and then apply machine learning methods to try to discover patterns that are hidden in the data—without taking into account anything about where the data came from, or current knowledge of the system" (Society & Institute., 2019, p.9). The authors explicitly contrast this use case with the potential for more theory-driven research techniques, à la PINNs (physics-informed neural networks). However, it is clear from the exposition that a theory-agnostic conception of the typical function of ML models informs the authors' predictions of disruption.

Chubb, Cowling, and Reed (2022) conducted a survey of identified leaders across various scientific fields concerning the adoption of AI/ML based methods within their research practices. A consistent theme amongst the researchers surveyed was the sentiment that "AI could prompt 'unforeseen' outcomes, potentially leading to a reframing of disciplines, modes and methods of knowledge production" (Chubb et al., 2022, 1442), and that "AI could be used in the near future to bypass traditional means of knowledge production" (Chubb et al., 2022, 1445). One interviewee explained the difference between "traditional" and AI-based methods as follows: "[n]ormally the scientific progress goes like this, so you have a hypothesis and then you collect data and try to verify or falsify the hypothesis, and now you have the data and the data, so to say, dictates you what hypothesis you can find. So, this is how methodologies, scientific methods are changing" (Chubb et al., 2022, 1446).

These overviews of scientists' perceptions of the place of AI in science, and its potentially transformative role, paint a relatively coherent picture. Machine learning, or "AI," enables scientists to carry out their work in a far more data-driven, and far less theory-driven capacity. Certainly, some research paradigms (or stages within a research pipeline) are more exploratory than others. A distinction between exploratory (broadly, data-driven) and explanatory (broadly, theory-driven or theory-involving) research strategies is popularly held by working scientists. The picture these assembled voices paint, however, seems to point to a lessened overall need for theoretical input within scientific discovery.

Articulations of the distinctness of ML emanating from science journalists à la Anderson (2008), (Hey et al., 2009), Mayer-Schönberger & Cukier (2013), and (Spinney, 2022) no doubt represent far more sensationalist visions for the role of ML in science than most working scientists would assent to. The average scientist would likely deny that AI/ML will soon altogether obviate the need for theory, preconception, or domain-expertise within scientific knowledge-production. Nevertheless, the overarching perception that the methods of science can or should be rendered free from theory exerts a force on the research practices of working scientists. Funding for grants and for institutes and centers, as well as industry sponsorship for conferences, awards, and the like often hinges on scientists conveying the novelty and disruptive potential of their methods which, increasingly, is tied to an ideal of theory-freedom.

2.2 A philosophical defense

Philosophers have been quick to respond to assertions that the rising tide of ML-adoption will enable a "post-theory science" (Spinney, 2022, 1). Some philosophers have critiqued this vision of ML-infused science, some endorsed it, while others have simply acknowledged its ubiquity (Alvarado & Humphreys, 2017; Beisbart & Räz, 2022; Boge et al., 2022; Boon, 2020; Creel, 2020; Desai et al., 2022; Duede, 2023; Hansen & Quinon, 2023; Kawamleh, 2021; Kitchin, 2014; Leonelli & Zalta, 2020; Pietsch, 2021, 2022; Pigliucci, 2009; Rowbottom et al., 2024, 2023; Sullivan, 2022; Srećković et al., 2022).

Alvarado and Humphreys (2017) take stock of observations on ML and big data from scholars hailing from a range of disciplinary backgrounds. These scholars describe the widespread adoption of ML and "big data" analytic methods resulting in "a common epistemological effect" (Alvarado & Humphreys, 2017, 739). The primary manifestation of this "epistemological shift" being that "[t]heory...at the level of how knowledge is produced and structured...[has] been replaced by information stored in databases too large to read and processed by algorithms too complex to understand"(Alvarado & Humphreys, 2017, 739). If ML or big data analytic methods are indeed "interpretation-free," Alvarado and Humphreys write, this will entail "a permanent change in the way that science is pursued"(Alvarado & Humphreys, 2017, 744). In a treatment of the representational status of ML in science and its relation to the scientific realism debate, Rowbottom, Curtis-Trudel, and Peden (2023) begin from the premise that scientific ML "contrasts with traditional scientific modeling, where explicit theories and models are used" (Rowbottom et al., 2024, 172).

In a 2021 paper, Duede writes that philosophers and scientists alike have widely made claims of the epistemic distinctness of ML and its disruptive potential. Duede observes that "to scientists and science funding agencies alike, artificial intelligence both promises and has already begun to revolutionize...science" and that "nearly every empirical discipline has already undergone some form of transformation as a result of developments in and implementation of deep learning and artificial intelligence"(Duede, 2023, 1089). But, as Duede notes, philosophers and scientists, while agreeing on the revolutionary potential (or actuality) of AI/ML in science, have made separate meaning of it. Duede sets out to address these discrepancies, attributing what he perceives as philosophical pessimism concerning the role of ML in science, in large part, to "a failure on the part of philosophers to attend to the full range of ways that deep learning is actually used in science"(Duede, 2023, 1090). In his critique of philosophical reactions to claims of the novelty of ML, however, Duede leaves these theses unchallenged. I will argue that the failure Duede documents on the part of philosophers to account for how ML might actually be implemented in scientific practice is ultimately responsible for philosophical endorsement of claims of ML's distinctness.

Srećković, Berber, and Filipović (2022) differentiate machine learning techniques from standard practices in statistical modeling, arguing that statisticians employ theoretical assumptions, while machine learning practitioners do not. "ML models," the authors write, "are constructed based on data instead of theoretical assumptions about the target system.(Srećković et al., 2022, 166).

Srećković, Berber, and Filipović (2022) evaluate what they hold to be the key differences between traditional modeling approaches and machine learning methods in terms of the explanatory capacity of both and their ability to elucidate causal relationships. Srećković et al. diagnose the methods of machine learning as uninterpretable, and not resting on theoretical considerations. This, according to the authors, prevents the practice from getting at underlying causes and furnishing explanations of natural phenomena. The ability of ML techniques to provide prediction in the absence of explanation is projected by the authors to alter the landscape of how we conduct science.

"In contrast to explanatory-focused statistical models," Srećković et al. argue, "ML models reach predictions without the theoretical backup that supplements the correlations found in the data with a potential causal interpretation" (Srećković et al., 2022, 160). Machine learning, they argue, is "theory-agnostic" in that "there are no a priori assumptions concerning the mechanism of the target phenomenon" (Srećković et al., 2022, 165). While the authors acknowledge a sort of disappearing line between ML and traditional statistical techniques, their emphasis is on drawing out broad characterizations of the two disciplines and what separates them. Whereas for "traditional statistics, standard models rely on the representation of underlying causal mechanisms, and they are used for retrospective testing of an already existing set of causal hypotheses...ML models are constructed based on data instead of theoretical assumptions about the target system. The purpose of such models is primarily forward-looking, i.e. to predict new observations" (Srećković et al., 2022, 166). Here, the contrast the authors draw between broadly "data-driven" and "theoretically-motivated" methods is telling. This distinction is not one the authors have introduced: such a divide between theory-driven or hypothesis-driven research and data-driven research is held widely among engineers and scientists. Srećković et al. merely provision a philosophical exposition and justification thereof.

In a similar vein, Boge (2022) speculates that a revolution in either scientific practice or its epistemic footing may be in store owing to the adoption of machine learning—specifically deep learning—methods. Boge's argument rests on the idea that deep learning is both instrumental in an idiosyncratic sense among modeling approaches in the sciences, and that it exhibits a novel kind of epistemic opacity to its deployers. These identifying facets of deep learning pose an impediment to understanding and explanation (in the scientific sense), especially when deployed in exploratory settings where the successful results of scientific enquiry will require novel concept-formation. Owing to their divergence from standard mathematical modeling practices in the sciences, Boge claims, ML modeling techniques "have the potential to profoundly 'change the face of science'" (Boge et al., 2022, p.71).

Boge urges that the distinction between the procedure of classical mathematical modeling or computer simulation in science and the application of machine learning methods is that the former procedure begins with a conceptualization of the target phenomenon under investigation, while this step is absent in the use of ML. "The difference," Boge writes, "between CS [computer simulation] and DL [deep learning] may be summarized as follows: The former begins with a conceptualization of the target, and from that predicts 'hypothetical data'. The latter begins with a conceptualization of data" (Boge et al., 2022, p.59).

Especially in exploratory modeling contexts, the lack of background theory or conceptualization of the target phenomena is taken, by Boge, as a potentially serious impediment to understanding. While Boge grants that DL models might represent, he holds that they fail to be explanatory for lack of theoretical context and conceptual content, writing that a "DL model...is conceptually too poor to provide an understanding of underlying mechanisms" (Boge, 2022, 57).

Boge takes after de Regt in his stance on the relation between representational status and explanatory status: "for representational models to explain, they must also be constructed under the principles of an intelligible theory, where a theory is intelligible if it has certain qualities that 'provide conceptual tools for achieving understanding' (de Regt, 2017, p. 118)"(Boge, 2022, 54). Boge predicts profound changes to the practice and epistemic products of science because ML-based tools will fail to provide understanding or explanations due to their lack of theoretical or conceptual motivation and content.

Boge and Srećković et al. both appear to sign onto the thesis that ML methods are theory-free or devoid of some essential variety of conceptual content which enables them to serve classical explanatory or inferential roles in science. The methods of ML are, hence, understood as distinct from canonical modeling methods in science and traditional statistics. Boge and Srećković et al. further contend that the widespread adoption of ML methods will catalyze disruptive change in science, while Boon argues that the theory-freeness of ML methods rules them out as viable tools for science. These scholars take the perceived differences between "normal science" or even "real science" and machine learning to amount to the degree to which they are theory-laden, theory-driven, or conceptually rich. As I will demonstrate in the subsequent sections, no use of ML in science is "theory-free," and those that aspire to this ideal tend to result in poor scientific practice.

Boon (2020) signs onto the distinctness thesis, maintaining that machine learning methods are in a category apart from classical statistical or scientific methods owing to their theory-agnosticism. "Machine learning," Boon writes, "is different from computer simulations, which utilize scientific knowledge to build mathematical models...[t]he machine-learning process does not draw on scientific models that are constructed by means of theories, laws, mechanisms and so forth. No theory or mechanism or law needs to be fed to the machine-learning process" (Boon, 2020, 47). Interestingly, she diverges from the norm in taking the distinctness thesis as a reason to *reject* the notion that ML will have a transformative influence on the conduct of science. Boon denies that machine learning methods will obviate the need for human conceptual apparatus in the generation of scientific knowledge, arguing science to be an essentially theory-involving activity. "[S]cience," Boon writes, "...cannot be replaced by machine learning technologies whatsoever since incomprehensive, opaque data-models do not tell us anything meaningful about the world. Therefore, 'real science' and machine learning technologies operate in very different domains and must not be regarded as competing" (Boon, 2020, 58).

Boon, Boge, and Srećković et al. each sign onto the idea that ML methods are in some sense theory-free or devoid of conceptual content, and hence distinct from canonical modeling methods in science and traditional statistics. Boge and Srećković et al. further contend that the widespread adoption of ML methods will catalyse disruptive changes to science, while Boon argues that it is precisely the theory-freeness of ML methods which rules them out as capable of unseating existing modes of knowledge-production in the sciences. These scholars take the perceived differences between "normal science" (or "real science," as Boon puts it) and machine learning to amount to the degree to which they are theory-laden, theory-driven, or conceptually rich. As I will demonstrate in the subsequent sections, no use of ML in science is “theory-free.” Scientific applications of ML that aspire to this ideal of theory-freedom tend to result in poor scientific practice.

3 Conceptions of scientific objectivity

The concept of objectivity is central to modern science, both as abstract ideal and as a set of human practices. What variety of objectivity scientists ought to strive for has been contested territory for centuries. One thread of this debate concerns the extent to which scientific practices and the knowledge that they produce are ineliminably structured by human values. Another concerns the extent to which such practices and outputs are necessarily structured by theory, in the sense of conceptual content, or prior commitment to the nature of the subject-matter.

Philosophical conceptions of objectivity are rooted in accounts of the nature and possibility of empirical knowledge. They are highly abstracted from on-the-ground empirical practices and have little direct influence on them. But scientists in modernity have operated with their own, albeit often implicit, conceptions of scientific objectivity. These have permeated public conceptions of science which, in turn, feed back into scientists’ self-conceptions of their work and its epistemic foundations. Thus philosophical conceptions of scientific objectivity and meta-narratives of scientific objectivity come apart.

A recent literature on values in science has offered an extensive treatment of the philosophical conception of objectivity as freedom from normative influence and its corollary meta-narrative: the value-free ideal. Few historical interlocutors have put forward explicit defense of objectivity as total value-agnosticism. Instead, it has been argued that science strives to minimize the impact of human values or to constrain their influence to appropriate venues and junctures. A mostly implicit ideal of total value-freedom, however, is widespread and influential. Its influence extends to scientific practice, to public reception of science, science education, and to the interplay of science and public policy (Douglas, 2009). Philosophers of science have argued that the end goal of total freedom from normative influence is unachievable on both practical and in-principle, epistemic grounds (Douglas, 2009; Elliott & McKaughan, 2014; Longino, 1990). Denying the necessary influence of values on science, it is argued, merely cements them, lends them an air of objectivity, and renders them unavailable to critical scrutiny.

So then, there is a broadly philosophical doctrine which conceives of scientific objectivity as a minimization of the undue influence of normative values on science. This doctrine comes apart from a scientific meta-narrative which pushes for the total elimination of values in science. This meta-narrative, according to feminist philosophers of science, is neither achievable nor desirable. I take it to be important to make the distinction here between the well-reasoned doctrine which has had actual historical proponents[1] and the harmful meta-narrative. Strains of the values in science literature have had a habit of slurring these together.

For our purposes, we can remain agnostic as to the truth of either the philosophical doctrine on values in science or its parallel metanarrative. Here we are concerned, instead, with a conception of scientific objectivity centered on the appropriate purview of theory. Much like the value-centered conception, there is both a reasonable doctrine and an extremefied meta-narrative. The doctrine calls for scientists to constrain the influence of theory to appropriate venues. After all, we do not want our research efforts to be artificially constrained by what we think we know about the phenomena, what we have conjectured about the phenomena, or simply the limitations of our theoretical apparatus. Essentially, this doctrine calls for our research practices to be *non question-begging*.

What I am calling the *theory-free ideal* is the corresponding meta-narrative. According to the theory-free ideal, science should strive to minimize, or even eliminate theoretical input from empirical research. With the advent of ML-assisted science, belief in the narrative of theory-freedom has become commonplace. Leonelli (2020) observes that one of the dominant responses to the rise of ML and big data analytic methods in science is to see it as a championing of what I have here dubbed the theory-free ideal: "[one] way to interpret the rise of big data is as a vindication of inductivism in the face of the barrage of philosophical criticism leveled against theory-free reasoning over the centuries" (Leonelli & Zalta, 2020, Sec.6, Par.4). The meta-narrative that tells us that science can be rendered theory-free is not innocuous: it effectively serves to conceal loci of theoretical input and reifies the implicit beliefs of uncritical scientists.

No doubt, the deep incorporation of ML methods into empirical research pipelines brings about changes to where domain knowledge and theoretical considerations come to bear on the scientific process and its outputs. The case studies reviewed in Section 2.5 are revelatory of some of these differences. Fundamental changes to the nature and loci of theory-impingement, however, have occurred continuously throughout the history of science. The development of computer simulation, sampling methods, or the formal apparatus for statistical analyses essentially shifted where theoretical considerations came into play in the inferential process. So, too, for that matter, did the Newtonian style of mathematical thought-experimentation and his method of fluxions. Novel conceptual tools entail novelty to the nature of conceptual influence on the brute work of empirical inference. None can obviate the need for conceptual infrastructure, nor can they open up novel pathways to knowledge of the world.

[1] For instance, Max Weber.

4 The necessity of theory

4.1 In induction

Inductive inference is a form of inference grounded in empirical observation. Induction is contrasted against deduction, and lacks the guarantees of deductive inference. The 18th century philosopher David Hume's several treatises on the subject lend us our modern philosophical conception of induction through a skeptical appraisal—what has come to be known as the problem of induction. As Hume motivated the puzzle, there is no frequency of occurrences of an identical phenomenon that would justify inference to an inductive generalization with certitude. No matter on how many occasions we have seen the sun rise, we are not licensed to the certain knowledge that it will rise again tomorrow. Inductive generalization, then, is a means of arriving at knowledge requiring infrastructure that goes beyond a mere spate of observations. Inductive generalization requires background conceptual infrastructure, or theory, to get off the ground. Indeed, the very ability to categorize several experiences as "instances of the same phenomenon" requires a concept or "theory" of the phenomenon. Such classifications are, of course, essential to inductive generalization. That conceptual or theoretical resources must be brought to bear on an inference procedure to license inductive generalization is essential to our modern philosophical understanding of induction.

4.2 In science

The work of science produces empirical knowledge by means of inductive inference.[2] Induction, in turn, rests on theoretical resources. Sellars (1956)'s characterization of the "myth of the given" offers an account of the necessity of conceptual frameworks in the act of observation and inductive generalization, without which science could not generate empirical knowledge. Norton (2003)'s "material theory of induction" describes how successful inductive inference is never licensed by universal, domain-generic formal rules, but always proceeds by the application of local rules warranted by hard-won empirical—in Norton's words, "material"—facts tied to a specific scientific research context (Norton, 2003).

Turning to scientific practice, even simplistic experimental designs reveal the nature and extent to which scientific observation and inference are theory-inflected. The very act of investigation involves commitment to the existence and in-principle measureability of some phenomenon. If we are making measurements and performing quantitative analyses thereon, we are further committed to the phenomenon being amenable to quantitative representation and analysis.

[2] While certain 20th century philosophers of science, including Hempel and Popper, made cases for the role of deductive reasoning in scientific inference, these projects are generally considered to have failed by their own lights. E.g., Popper's admission that some hypotheses could receive greater or lesser evidential corroboration pushes him to an inductive view of scientific inference.

How we choose to measure and analyze records of a phenomenon generally includes a commitment to its quantitative ontology, e.g., is it categorical, ordinal, or cardinal? Measurement cannot be total, and therefore there is always a commitment as to what to look at experimentally and what to exclude. The very design of our instruments of measure and their calibration includes various commitments to the nature of the worldly phenomena under investigation. There is always, for instance, a commitment to the appropriate level of abstraction at which to study the phenomenon in question, which manifests in settings on instruments of measure, such as degree of magnification or periodicity of sampling. In fundamental physics, when we cool our instruments to reduce the contamination of our measurements by thermal noise, it is our prior theoretical grasp on the target phenomena, the physical systems under study, that motivates us to do so.

Crucially, "data" does not refer to physical phenomena.[3] "Data" refers to abstract, formalized representation of the results of direct observation or measurement. Data must be capable of serving an evidential role in licensing inferences about natural phenomena. Given that data is a form of mathematical representation, it does not intrinsically hold semantic meaning or refer to empirical phenomenon. The meaning that data holds for scientific inference exists in virtue of human interpretation and empirical grounding. For the use of any mathematical analysis—including the modes of analysis enabled by ML—to ground any scientific inference, it must be given conceptual content. This is already an essential form of theory-ladenness. The parameters of any machine learning model and its outputs are a step removed from input data, but are likewise mathematical representations. The data-derived parameter weights of a neural network, for instance, capture salient statistical patterns in the training data which are then leveraged to regress or classify the data on which they are tested or deployed. They represent abstract features of the training data. The representational status of neural network models is derivative of the representational status of the data on which they are parameterized.

A number of philosophers have provided strong rationales for rejecting the possibility of theory-free science. Leonelli (2012, 2018, 2020) stresses the essential theory-ladenness of data, decrying the popular conception of data as "raw" and "objective." Leonelli (2018) investigates "the different extents to which theory—understood broadly as a set of theoretical commitments and goals—impinges on inferential processes from data" (Leonelli, 2019b, 22). In several book-length treatments of the use and interpretation of data in scientific practice (e.g., (Leonelli, 2018, 2019a; Leonelli & Tempini, 2020; Leonelli & Beaulieu, 2021)), Leonelli concludes that there is no place in scientific practice in which we have data that is not already, to some degree, shaped by our existing conceptual or theoretical grasp on the phenomenon, commitments to epistemic goals and questions to be answered, idealizations, and auxiliary assumptions.

This view is a rejection of "[t]he naïve fantasy that data have an immediate relation to phenomena of the world, that they are 'objective' in some strong,

[3] See Sections 3.3.2 and 3.3.3

ontological sense of that term, that they are the facts of the world directly speaking to us" (Longino, 2020, 391). Bogen (2016) argues that it is the very fact that data is not raw, that it is, in a sense, "impure" that makes it able to serve the meaningful epistemic role it does. Boyd (2018); Boyd & Bogen (2009) argues further that it is not in spite of, but owing to the theory-ladenness of data that empirical science garners us its epistemic results.

4.3 In machine learning

Inductive inference is the procedure of gaining knowledge by extrapolating from a limited number of observations to a more general class. The fundamental task of ML is the extraction of statistical patterns from a training dataset and the extrapolation of this pattern to prediction or classification tasks on unseen instances. ML is therefore, straightforwardly, a class of formal methods for inductive inference (Bergadano, 1993; Harman et al., 2007; Sterkenburg & Grünwald, 2021). It is worth noting here that this position is not altogether uncontested. Buchholz & Raidl (2025) argue that, while general consensus holds that "ML algorithms inductively infer general prediction rules from observations," a falsificationist appraisal of statistical learning theory reveals that ML "combines the methodological approaches of deduction and induction" (Buchholz & Raidl, 2025, p.2). Even a Popperian perspective, however, must admit to an essential inductive component to ML enabled inference.

If inductive generalization writ large cannot be accomplished without theoretical input, then no specific formal inference scheme can accomplish inductive generalization without theoretical input. Any ML-enabled inference, therefore, requires theoretical input—whether explicit or implicit. One of the primary places this theoretical input comes into play is in *problem formulation*,, which includes the articulation of an inference task, the conceptualization of input data and learning objectives, and the election of success criteria.

We have discussed that inductive inference is data-driven, that scientific inference is data-driven, and that data, in these contexts, must be shaped by human concepts or theoretical resources in order to scaffold these inferences. The data that serves to support inference in ML is no different. Several accounts detail the role of theory in the variety of data on which ML is trained and deployed—often referred to as "big data."

Kitchin (2014) echoes that features of data collection and processing render data essentially theory-laden, in light of culturally-shared and ubiquitous background theoretical understanding of phenomena. Further, as Kitchin argues, data deprived of all semantic meaning would be uninformative, that is, unable to serve their essential epistemic role of scaffolding inference. In a similar spirit, Frické (2015) argues that theory must guide the selection of data to scaffold algorithm-assisted inference. Hansen & Quinon (2023) argue that ML-assisted science can never be made theory-free, as theoretical considerations necessarily enter in at the junctures of problem-formulation, data collection and curation, data pre-processing, and model-selection and validation. Desai et al. (2022) note that the theory-ladenness of observation makes it impossible to make observations or take measurements without the guidance of background theory. Desai et al. echo common sentiments among philosophers about the prospects of a wholly predictive science: such a view of the process of arriving at empirical knowledge is a naïve one, and ignores that one of the primary aims of science is explanation or understanding of the world.

Interestingly, it is not only philosophical considerations that lead us to the conclusion that ML-guided inferences cannot be rendered free from theory. Formal results from statistics and ML independently reveal the reliance of such methods on inductive biases or a priori constraints on the hypothesis spaces through which they search. Statistical learning theory (SLT) is the branch of theoretical computer science that looks to supply a theoretical basis (and formal guarantees) for inference with ML. The relation between ML/SLT as formal inductive method and the philosophical study of induction has not gone unnoticed: a number of scholarly works have treated the intersection of these subjects (Bergadano, 1993; Harman et al., 2007; Sterkenburg & Grünwald, 2021). These texts have drawn out the philosophical relevance of learning theoretic results for both the study of induction and the practice of ML. The most salient results for our purposes are the no free lunch (NFL) theorems (Wolpert, 1996).

The no free lunch theorems are a set of results in statistical learning theory demonstrating the impossibility of a universally valid and purely data-driven inference rule. Though the philosophical implications of the theorems are vexed (see (Sterkenburg & Grünwald, 2021)), with some artistic license, we may think of the no free lunch theorems of supervised learning as a formalization of the idea that inductive inference would be impossible if the reality we inhabited (and, hence, the data we learn from) did not exhibit some (learnable) regularity (Wolpert, 1996).[4] In learning from data we must, therefore, make assumptions about the friendliness of that data to our epistemic intentions; empirical data alone is not enough to get induction off the ground (Sterkenburg & Grünwald, 2021). Learning from data is possible only with the incorporation of theoretical assumptions, in the form of prior selection of a model class, hypothesis class, or priors. As Gillies writes in *Artificial Intelligence and Scientific Method*, "inductive rules of inference. . . do not generate hypotheses from data alone. . . but from data together with some background knowledge (or assumptions)" (Gillies, 1996, p.39).

5 The role of theory in scientific ML

I have argued that the notion that widespread adoption of the methods of ML in science will obviate the need for theorizing is 1. widespread, 2. symptomatic of a theory-free ideal in science, and 3. untenable. In the final section of

[4] More technically, the NFL theorems are an impossibility result for the existence of any learning algorithm that outperforms others on generalization to all learning environments, given the assumption of a uniform probability distribution over all possible learning environments. This uniformity assumption is, in a sense, an inversion of Hume's "uniformity of nature," which is the stipulation of structuredness to reality/experience/data.

this paper, I will attempt to illustrate its perniciousness by means of two case studies, which concern instances of actual application of modern ML methods in scientific practice. The first case study concerns a use case for ML in science that is deeply theory-laden and self-aware in its theory-ladenness. This use of ML in science has marked a scientific breakthrough, and been a resounding epistemic success. The second study concerns a use case for ML in science that is marketed as bypassing the need for theory. This application of ML has been decried as statistical malpractice, its results at best uninformative, at worst, dangerously misleading. With these cases I aim to show the unavoidability of theoretical work in scientific applications of ML, and the deleterious effects of the ideal of theory-freedom on scientific practice.

5.1 The unreasonable effectiveness of AlphaFold

AlphaFold 2.0 is heralded as the most impressive result that ML methods have achieved for science to date. To appreciate the unprecedentedness of the AlphaFold results, we must first appreciate the scientific problem it is confronted with. The problem of protein folding is notoriously difficult. There is very little that we can say from the genotypic specification of a particular protein about how it will fold. Mapping from sequences of adenines, cytosine, guanines, and thymines to a menagerie of amino acids is relatively straightforward, as biological problems go; so is predicting the polypeptide chains these amino acid sequences will form. What mess of three-dimensional spaghetti those amino acid chains will assume once synthesized, however, is another matter entirely. This is an essential problem for the biomedical sciences. The three-dimensional anatomy of protein structure determines its function and is thus a crucial object of scientific inference.

To truly comprehend the difficulty of the protein folding problem—and how the methods of machine learning were able to get around it—we first have to recognize that protein structure is understood at four levels. DNA is a string composed of four alternating base pairs. It encodes information in sequence. When proteins are assembled, that DNA is read, codon by codon, and a polypeptide chain is built up from twenty amino acids on the basis of these instructions. These amino acid sequences are dubbed the "primary structure" of a protein. All amino acids are composed of the same base molecular structure, which will bond together to form the backbone of the polypeptide chain, containing an α-carbon, an amino group, a carboxyl group, and a hydrogen. From this molecular backbone extends the R-group or side chain, the determinant of the amino acid's "flavor." The secondary structure of a protein refers to the morphology that polypeptide chains take on on their own, owing to bonding patterns in the backbone. The morphology of these peptide chains results from local interactions between adjacent and semi-adjacent molecules in the backbone of the peptide chain. Owing to the periodicity of the placement of amino acids with certain valences (and other molecular-bond determining features) in the chain, they will typically either form what are known as α helices or β sheets. Up until this point things have remained relatively straightforward, as biological problems go: we have a basic, repeated molecular structure and its self-interaction in the form of hydrogen bonding.

The tertiary structure of a protein is determined by the R-groups of the amino acids. Recall that these come in twenty flavors. Recall that virtually all forms of non-covalent bonding are available to these molecules now. Recall that amino acids can exhibit hydrophobic and hydrophilic proclivities. If a protein is composed of more than one polypeptide chain, it will have a quaternary structure as well. At the tertiary and quaternary levels of protein structure, we have advanced from assembling text from bit strings to attempting to predict all of the ways in which many distinct kinds of spaghetti thrown together in a pot can cohabitate, given six dimensions along which spaghetti substructures may or may not like to interact. Also, the spaghetti exhibits quantum behaviors.

At first blush, this seems like an unsolvable problem. The initial trick—the trick that gets existing bioinformatic solutions off the ground—lies in noting that when we have a variant in one amino-acid we can see what *non-local* variants tend to co-vary along with it. This begins to tell us something about what might be touching what in the tertiary and quaternary protein structures. Still a difficult problem, but more manageable.

The AlphaFold team began by creating their own sequence database—now the largest existing database of its kind—by large-scale clustering of existing sequence repositories. DeepMind's AlphaFold 2.0 takes an amino acid sequence as input in its pre-processing stage and derives a multiple sequence alignment (MSA). An MSA encodes evolutionary information, usually highlighting relationships of homology. In addition to the primary amino acid sequence and MSA, AlphaFold is also supplied as input database-derived *templates*—three-dimensional atomic maps—for a small number of sufficiently similar homologous protein structures. Distance and orientation features and sequence features are derived from preexisting template representations to form what the AlphaFold team dubs a *pair representation*, encoding relationships between pairs of amino acid residues.

AlphaFold treats the prediction of 3-dimensional protein structure from these pair representations and MSAs as a graphical problem, rendering the representations in the primary trunk of the model architecture into gradated bitmaps. The problem formulation for the DeepMind team was to "view the prediction of protein structures as a graph inference problem in 3D space in which the edges of the graph are defined by residues in proximity" (Jumper et al., 2021, 585). The core structure of AlphaFold 2.0 is a transformer: a form of DNN which exploits parallelization and attention mechanisms to incorporate contextual factors in the data at multiple levels of abstraction simultaneously—and, importantly, enabling information from these levels to 'talk to' each other. AlphaFold passes both the MSA and the pair representation (separately) back and forth through the trunk of the model for a set number of iterations (48 blocks) per recycle, progressively refining the representations, and allowing the two distinct representations (MSA and pair representation) to influence one another as each is refined. The output of this refinement procedure is then, in the final stage, fed to a structure module that uses Invariant Point Attention to output atom coordinates. The predicted coordinate map is then passed, with MSA and pair-representations, back through the trunk. This is repeated for three iterations until a final predicted 3D protein structure is achieved.

Let us draw out what is salient about this scientific procedure for our analysis, the aim being to examine the role played by theoretical considerations. The AlphaFold team explicitly incorporate theoretical resources at the stage of data provenance and engineering, the stage of architecture design, hyperparameter selection and model training, and at the stage of model evaluation and interpretation.

In the first place, theory integration comes in at the level of the data in terms of what the data ultimately represents and how it is imbued with that representational content. The data on which AlphaFold is trained is richly structured by existing empirical knowledge of the target domain (the molecular structure of proteins and their evolutionary trajectories) and our theoretical understanding thereof. AlphaFold sits atop a wealth of domain knowledge about the form and function of proteins. Theory also comes into play in how the data is handled for the specific task in question and how it is made to serve as evidence in this task. AlphaFold is, at its core, an instance of (semi-)supervised learning.[5] The exercise is premised on the idea that the rules of association between amino acid sequences and three dimensional protein structure lie latent in cross-taxa protein structure data. It is further premised on the supposition that the systematic breakdown in protein structure and function resultant from certain amino acid substitutions can be leveraged to learn the complex bonding affinities governing 3-dimensional protein structure. Part of what is noteworthy in this case study is the insight to take the publicly available data and turn it into novel representational forms in multiple places: combining MSAs and templates to create pair representations, and projecting those into effective heatmaps of sequence-structure associations so that the inference task could be treated like a graphical problem.

The architecting of the various model components utilized in AlphaFold 2.0 was similarly bound to theoretical considerations. AlphaFold is not a domain-generic model; the model architecture is hand-tailored to the specific task of learning to predict three dimensional protein structure from MSAs and pair representations—a novel representational form for the task. AlphaFold 2.0 employs a transformer network that is designed to iteratively refine progressively more accurate guesses at the true protein structure. The transformer trunk utilized in AlphaFold was created to combine and refine representations of the specific form it is fed in a novel training and deployment procedure. Perhaps the most strikingly theory-laden aspect of AlphaFold 2.0 is the engineering of specially tailored loss functions. In training a DNN, a loss function governs how the distance metric is calculated between present output and desired output of the model. In a typical neural network training regime, the error term is then

[5] Supervised learning methods train a model to approximate a human categorization or decision, known, in ML, as a label. Unsupervised learning methods, by contrast, work to discover patterns from unlabelled data. Semi-supervised learning incorporates both labelled and unlabelled data.

backpropagated through the network, layer by layer, updating the weighting of the model's parameters so as to minimize the gradient of the loss. In specifying the loss function, machine learners are able to express precisely what it is that they are interested in learning for a particular task. In AlphaFold 2.0, the loss function is heavily tailored to the problem of predicting folded protein structure from amino acid sequences. The researchers employed "a loss term that places substantial weight on the orientational correctness of the residues" (Jumper et al., 2021, 585). Loss terms specific to the learning of various structural features of protein folding along a number of dimensions were employed at all stages of training and fine-tuning: "satisfaction of the peptide bond geometry is encouraged during fine-tuning by a violation loss term" (Jumper et al., 2021, 586-587).

Finally, model-evaluation, that is, judging the success of the trained model and interpreting its results requires integrating the resulting predictions of AlphaFold into existing biological knowledge. We can only judge the success of such a model when it is understood against the backdrop of our prevailing scientific accounts. AlphaFold 2.0's success was only legible in the CASP14 (Critical Assessment of Structure Prediction) experiment in achieving at or near the performance of theory-guided and experimentally-obtained protein structures. We can likewise only put the results of such modeling efforts to *use* when we have accommodated them within a theoretical framework.

5.2 Transcriptomics

A new research pipeline involving the application of multiple dimensionality reduction transformations in sequence to transcriptomics data has become popular in recent years. These methods take datasets originally expressing hundreds of thousands of dimensions and successively transpose them down to lower and lower manifolds, ultimately outputting a colorful 2-dimensional plot which researchers interpret both visually and via quantitative metrics. These plots are taken to represent meaningful groupings of samples according to gene-expression. Computational biologists Tara Chiari and Lior Pachter investigated the credentials of these methods, finding them to fall short of good scientific and statistical practice in key respects (Chari & Pachter, 2021). These dimensionality reduction methods are motivated over more traditional techniques in transcriptomics by appeal to their data-driven nature and their relative lack of human input. As the critical review by Chari & Pachter (2021) reveal, this way of selling the methods and their capabilities obscures 1. How they distort or discard data and the informative patterns it is intended to reveal, 2. Aspects of the methods under a researcher's control, 3. The human role in interpreting (or misinterpreting) exploratory visualizations of this nature.

Single-cell transcriptomics offers an approach to inferring cellular-level gene expression. The technique is utilized for identifying cell populations, modeling transcription dynamics, inferring the developmental trajectories of cellular populations, and monitoring changes in cell populations relative to health status. Single-cell transcriptomics emerged with the availability of mass quantities of high throughput RNA sequencing and expression data. It is typical in such exercises to be working with datasets which possess hundreds of thousands of feature dimensions; expression data for thousands of genes across millions of cells (Kobak & Berens, 2019). For this reason, researchers typically employ dimensionality reduction techniques. Dimensionality reduction is a method of mapping a high-dimensional dataset to a lower-dimensional space—or *embedding* higher-dimensional data within a lower-dimensional *embedding space*. Dimensionality reduction techniques are used to distill essential patterns from large datasets, make analyses tractable, and isolate signal from noise. Dimensionality reduction is a method of unsupervized ML, meaning it works by extracting the contours of data, rather than working to extrapolate predefined success criteria encoded in labeled data as in supervised learning.

A now established workflow in single-cell transcriptomics involves applying dimensionality reduction techniques sequentially to high-throughput RNA expression data; first linear methods which reduce the dataset to tens of dimensions using principle component analysis (PCA) or analogous techniques of dimensionality reduction, followed by one of two purpose-built two-dimensional nonlinear reductions: t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). These methods produce visualizations for both qualitative and quantitative exploratory data analysis.

The intuition behind using dimensionality reduction techniques in this way is as follows. The data on which analyses are run has a certain number of features, which determine the dimensionality of the dataset. The data we are looking at can then be thought of as points lying within a space of that many (n-)dimensions. The specific datapoints we are seeking to characterize lie along a (likewise n-dimensional) manifold within that space. Advanced statistical methods like t-SNE and UMAP are designed to be capable of transposing very high-D data down to low-D embedding spaces while preserving informative structure and discarding irrelevant dimensionality. These methods try to distill the relationships between datapoints and their neighbors and preserve these in projecting the data down to lower-D spaces.

In the transcriptomics workflow under critique, PCA is first run on the original data, creating a linear transformation down to a space with tens of dimensions. This is followed by one of two nonlinear reductions: t-SNE or UMAP. The t-SNE method works within the high-D embedding to calculate pairwise similarities using a t-distribution and then runs a distance-preserving embedding into a lower-D space. UMAP works according to the same general principle, but has more advanced mathematical underpinnings and stronger theoretical guarantees. UMAP first builds what is called a fuzzy simplicial complex and from this constructs a low-D graph optimized to express the salient distances represented therein. UMAP outperforms t-SNE in preservation of global structure.

In order to be useful tools for the kinds of inferences transcriptomics aims for, these methods must faithfully preserve both local and global structure. It can be difficult to independently verify the success of such methods when used in exploratory data analyses with complex data, where ground truth is unavailable. One available means of validating these methods is to draw a very simple shape in high-D, replicate the dimensionality reduction workflow, and determine whether the original structure can be recovered. In their 2021 probe of this now-standard transcriptomics workflow, Chari and Pachter do just this: using UMAP to produce a 2-dimensional representation of a manifold whose structure is already known. The researchers found that not only was the original structure so obscured as to not be inferrable from the resultant embedding, but erroneous groupings of the original structure were introduced in the transformation.

This procedure was one among a series of analyses carried out by Chari & Pachter (2021). Taken together, their assessments demonstrated that the practice of repeated application of dimensionality reduction techniques introduced heavy distortions to the original manifold representation. Critically, Chari and Pachter's analyses reveal that the now-standardized PCA plus t-SNE or UMAP workflow is incapable of preserving the interpretively salient features of the datasets under investigation: local structure, global structure, distance, and continuousness (Chari & Pachter, 2021). What is more, interpretive practices surrounding the resulting visualizations led to erroneous or conflicting conclusions.

Exploratory, data-driven methods like the transcriptomics workflow under scrutiny are motivated on the grounds of their supposed empiricism; they are claimed to be untainted by human bias and unconstrained by existing hypotheses. These applications of dimensionality reduction techniques, in particular, are lauded as handling "all data and all relationships;" however, this common practice "distorts data in obscure ways, attempts to pack the capabilities of many different analyses into one space, and is easily manipulated" (Chari & Pachter, 2021, p.15). Chari & Pachter (2021) refer to the use of dimensionality reduction techniques in transcriptomics as a "blind application" of "heuristic procedures" (p.13), arguing that "there is little theoretical support for this practice" (p.1). As substitute for such (supposed) atheoretical practices, Chari and Pachter endorse "targeted analyses" and "hypothesis-driven biological discovery" (p.15).

While the heart of Chari and Pachter's critique is a demonstration of poor practice involving the combined t-SNE/UMAP workflow, they point clearly to alternative approaches, including semi-supervized learning methods and targeted embeddings for specific featural dimensions. The ambitions of such analyses are, necessarily, more constrained than those of sequential embedding pipelines. They require specifying in advance what features of the data are under investigation. They also require making explicit what assumptions go into the analysis: "it is possible to construct embedding spaces which more explicitly control and improve nearest-neighbor structure and retention," write Chari and Pachter, "[h]owever, such optimizations require making an assumption regarding the appropriate distance/similarity metric, as is generally the case with the neighborhood-based analysis methods ubiquitous across the tasks [on which t-SNE/UMAP are applied]" (p. 14).

The authors of the critique endorse the methods of dimensionality reduction in more limited, principled, and theory-informed applications to transcriptomics: “By targeting the objective of an embedding...one can take advantage of prior knowledge/annotations and more directly determine the necessary dimensionality for a given question” (Chari & Pachter, 2021, 15). Such alternatives require incorporation of domain expertise, critical thinking, and the ability to both identify and (statistically) articulate what you are looking for—characteristics markedly absent from the t-SNE/UMAP workflows under scrutiny.

5.3 Takeaways for ML in scientific practice

The vast majority of applications of the tools of ML to science have been run of the mill: automating laborious processes, achieving minor gains in efficiency or accuracy over human classification or “analogue” statistical techniques without notable breakthroughs in the variety of knowledge gained by their use. The accomplishments of AlphaFold 2.0 are a striking departure from these more quotidian uses. The scientific community has acknowledged AlphaFold as a resounding success; perhaps the greatest win for ML in science to date, if the Nobel Prize is any measure. No other application of ML to science has achieved quite so stark an advantage over pre-existing techniques.

Paralleling the mounting successes of ML in scientific application are a growing number of instances of scientific ML gone wrong: ML-involving research practices deemed deficient by the scientific communities in which they are embedded, e.g., (Goddard et al., 2018; Andrews et al., 2024; Bowers et al., 2023). Some of these replicate the same pattern of errors observed in our transcriptomics case study, and can be traced to the same root cause.

Goddard et al. (2018), for example, offer a critique of a strikingly similar misuse of dimensionality reduction techniques in neuroscience. These are, like the use of PCA, t-SNE, and UMAP in transcriptomics, unsupervised, exploratory uses of dimensionality reduction techniques. Like the transcriptomics case, these are ML methods typically reserved for rote data transformation that have been ambitiously repurposed for inference to the structure of the target system; in this case, not gene expression dynamics, but the neural representation of visual features. Goddard et al. (2018) reveal the inadequacies of these methods by an approach similar to that employed by Chari & Pachter (2021): applying the methods to areas with known ground truth, revealing, in the process, that the methods are incapable of recovering the most basic features of encoding.

The methods of ML are increasingly saddled with more and more ambitious tasks in science: extending far beyond mere signal processing to playing a fundamental role in inferring the structure of our natural world. When ambitious scientific applications of ML, like AlphaFold, have succeeded, it is in virtue of the conceptual resources they have incorporated. In this sense, the conclusion I reach aligns with reasoning expressed in Boge, Srećković et al., and Boon: theory-involvement is a requisite feature of our conceptual instruments in science for them to be capable of elucidating previously unknown truths of our natural world from data. The issue is that the perfectly theory-free vision of ML in science, the primary object of these scholars’ concerns, is neither the normal nor necessary operation of these methods. It singles out either a strawman or a failure case.

6 Conclusion

It has been alleged by representatives from science and engineering communities as well as philosophers of science—reflected in the texts handled in this paper—that the methods of ML, loosed on sufficient data, are capable of discovering meaningful patterns, natural joints, or mind-independent truths of their own accord. This is believed to be possible in the absence of input from human theorizing or conceptualization of the target system. Inductive inference, however, is understood by philosophers to rest essentially on theory. A dilemma emerges, forcing us to elect between a refutation of the thesis of theory-freedom or a revision to our standing conception of induction. As I have argued in this text, the ideal of theory-free learning via ML from "raw data" is a confused one. Incorporation of domain expertise is crucial for epistemically responsible deployments of ML, within and without science proper.

Advancing the state of the discourse away from false dichotomies and misdirected concerns is essential, for there is both much that is interesting and potentially novel about ML—DL in particular—and much at stake in its appropriate use. Where to localize theoretical considerations in DL-based scientific workflows appear to differ substantively, along various dimensions, from various canonical modes of scientific or statistical modeling.

On a conventional view of experimental science—one held widely by modern scientists across disciplines—we are typically formulating hypotheses and going out to collect data capable of adjudicating between our hypotheses. Thus the ways in which our conceptual grasp on the target phenomena come into play in how the data represent the target are specific to the epistemic concerns of a particular scientific or modeling exercise. In applied ML, we are often handed data corpora or else construct them from amalgamations of preexisting datasets. This means that a significant amount of the interpretive work—the work of mapping the data onto target phenomena, imbuing it with representational status and content—is work done before the modeler ever comes in contact with the data. This practice seems to defy Bogen and Woodward's (1988) claim that data are intrinsically limited to serving an evidentiary role in a particular experimental context (Bogen & Woodward, 1988).

Theoretical or interpretive work typically comes in again in the problem formulation, in the engineering or choice of model architecture, and in model training regimes, including choice of hyperparameters and loss or cost functions, as seen in the case of AlphaFold. Theoretical considerations further come in at the level of model evaluation, in our formal assessments of the success of the modeling exercise. Finally, such considerations come into play in what we take ourselves to have learned from the model output and, effectively, in *how the model is wielded*, or the interventions predicated thereon. Undoubtedly, the accelerating adoption of ML-based methods will bring about changes to on-the-ground research practices, including changes to the loci of theoretical input thereon. Such changes, however, will have to be not only domain-specific, but specific to the role with which ML methods are saddled in particular applications.

The landscape of science is also undergoing significant changes today, which are worthy of philosophical scrutiny in their own right. Changes to the social, institutional, governmental, and economic infrastructures that support science, and to the knowledge economies it results in, are a rich philosophical subject. These include the fragmentation and specialization of science, the proceduralization of science, its automation, the progressive increase in the distribution of intellectual labor it involves, the extraction of the knowledge of domain experts and its mechanization and codification into operational formulae. Reactions to the adoption of ML in science have largely framed ML as catalyst to these changes. I wish to counter that we can instead view ML as symptomatic of a much older and deeper trend in the development of scientific practice, one which often replicates the form of the society in which scientific practice is embedded in its social structure, its economic model, and its governance. The causal arrow, therefore, may run as much from the automation of scientific practice and the balkanization of scientific expertise to the adoption of the tools of ML as it does in the reverse.

Acknowledgments

The author is indebted to the following scholars for detailed feedback, critical engagement, and illuminating discussion without which this paper would never have come to fruition: Colin Allen, Robert Batterman, Zvi Biener, Daniel Burnston, Tony Chemero, Rachel Childers, Nic Fishman, Clark Glymour, Joyce Havstad, Leif Hancox-Li, Zachary C. Lipton, John D. Norton, Angela Potochnik, Jan-Willem Romeijn, and Jennifer L. Whyte.

7 Compliance with Ethical Standards

The author discloses no conflicts of interest and the research conducted for this manuscript was not empirical in nature.

References

Alvarado, R., & Humphreys, P. (2017). Big data, thick mediation, and representational opacity. *New Literary History*, *48*(4), 729–749.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), 16–07.

Andrews, M., Smart, A., & Birhane, A. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, *5*(9).

Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, *17*(6), e12830.

Bergadano, F. (1993). Machine learning and the foundations of inductive inference. *Minds and Machines*, *3*, 31–51.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43–75.

Boge, F. J., Grünke, P., & Hillerbrand, R. (2022). *Minds and machines special issue: Machine learning: Prediction without explanation?* Springer.

Bogen, J. (2016). *Empiricism and after.* Oxford University Press.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The philosophical review*, *97*(3), 303–352.

Boon, M. (2020). How scientists are brought back into science—the error of empiricism. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 43–65.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Boyd, N. M. (2018). Evidence enriched. *Philosophy of Science*, *85*(3), 403–421.

Boyd, N. M., & Bogen, J. (2009). Theory and observation in science. *Stanford Encyclopedia of Philosophy*.

Buchholz, O., & Raidl, E. (2025). A falsificationist account of artificial neural networks. *The British Journal for the Philosophy of Science*.

Chari, T., & Pachter, L. (2021). The specious art of single-cell genomics. *BioRxiv*, 2021–08.

Chubb, J., Cowling, P., & Reed, D. (2022). Speeding up to keep up: exploring the use of ai in the research process. *AI & society*, *37*(4), 1439–1457.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, *87*(4), 568–589.

Desai, J., Watson, D., Wang, V., Taddeo, M., & Floridi, L. (2022). The epistemological foundations of data science: a critical review. *Synthese*, *200*(6), 469.

Douglas, H. (2009). *Science, policy, and the value-free ideal.* University of Pittsburgh Pre.

Duarte, J., Han, S., Harris, P., Jindariani, S., Kreinar, E., Kreis, B., ... others (2018). Fast inference of deep neural networks in fpgas for particle physics. *Journal of Instrumentation*, *13*(07), P07027.

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, *90*(5), 1089–1099.

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science*, *81*(1), 1–21.

Frické, M. (2015). Big data and its epistemology. *Journal of the association for information science and technology*, *66*(4), 651–661.

Gillies, D. (1996). *Artificial intelligence and scientific method.* Oxford University Press.

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., & Carlson, T. A. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *NeuroImage*, *180*, 41–67.

Hansen, J. U., & Quinon, P. (2023). The importance of expert knowledge in big data and machine learning. *Synthese*, *201*(2), 35.

Harman, G., Kulkarni, S., & Roeper, T. (2007). *Reliable reasoning: Induction and statistical learning theory.* Bradford Books.

Hey, A. J., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . others (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589.

Kawamleh, S. (2021). Can machines learn how clouds work? the epistemic implications of machine learning methods in climate science. *Philosophy of Science*, *88*(5), 1008–1020.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & society*, *1*(1), 2053951714528481.

Kobak, D., & Berens, P. (2019). The art of using t-sne for single-cell transcriptomics. *Nature communications*, *10*(1), 5416.

Leonelli, S. (2018). La ricerca scientifica nell'era dei big data.

Leonelli, S. (2019a). *Data-centric biology: A philosophical study.* University of Chicago Press.

Leonelli, S. (2019b). What distinguishes data from models? *European journal for philosophy of science*, *9*(2), 22.

Leonelli, S., & Beaulieu, A. (2021). Data and society: A critical introduction. *Data and Society*, 1–100.

Leonelli, S., & Tempini, N. (2020). *Data journeys in the sciences.* Springer Nature.

Leonelli, S., & Zalta, E. N. (2020). Scientific research and big data. *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition)*.

Levins, R., & Lewontin, R. (1985). *The dialectical biologist*. Harvard University Press.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press Princeton, NJ.

Longino, H. E. (2020). Afterword: Data in transit. *Data journeys in the sciences*, 391–399.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647–670.

Pietsch, W. (2021). *Big data*. Cambridge University Press.

Pietsch, W. (2022). *On the epistemology of data science*. Springer.

Pigliucci, M. (2009). The end of theory in science? *EMBO reports*, *10*(6), 534–534.

Rowbottom, D. P., Curtis-Trudel, A., & Peden, W. (2023). Evidence, computation and ai: why evidence is not just in the head. *Asian Journal of Philosophy*, *2*(1), 11.

Rowbottom, D. P., Peden, W., & Curtis-Trudel, A. (2024). Does the no miracles argument apply to ai? *Synthese*, *203*(5), 1–20.

Sellars, W. (1956). Empiricism and the philosophy of mind.

Society, T. R., & Institute., T. A. T. (2019). The ai revolution in scientific research.

Spinney, L. (2022). Are we witnessing the dawn of post-theory science. *The Guardian*, *9*, 2022.

Srećković, S., Berber, A., & Filipović, N. (2022). The automated laplacean demon: How ml challenges our views on prediction and explanation. *Minds and Machines*, *32*(1), 159–183.

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

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

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