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

Mel Andrews

August 26, 2024

Abstract

This paper contends with the widespread belief that the methods of machine learning (ML) have the capacity to radically disrupt the nature of scientific knowledge or practice on the grounds that these methods enable a form of theory-free inductive inference. Such views about scientific ML flow directly from what I term a *theory-free ideal* in science: a scientific meta-narrative according to which the influence of theory on scientific knowledge-production should be minimised, if not altogether eliminated. By means of two case studies, I argue that this theory-free ideal, like its normative corollary, has a deleterious effect on the epistemic standing of ML-based science.

1 Introduction

The prospects of machine learning (ML) for science have opened wide in the last decade, in which time ML-based methods were adopted in the Large Hadron Collider at CERN for sorting the significance of particle collision events (Duarte et al., 2018) and DeepMind released its AlphaFold, AlphaFold 2.0, and AlphaFold 3.0 (Jumper et al., 2021), capable of predicting tertiary and quaternary protein structure from amino acid sequence data, effectively solving one of biology's most complex and enduring open problems. The rapidity and ubiquity of machine learning uptake across all sectors of public life, in particular, science, has sparked an onslaught of speculation concerning its nature and the downstream consequences of its widespread use.

Such speculation has issued from cultural commentators, journalists, and media personalities, from the researchers and engineers producing the tools of ML and the scientists deploying them and from philosophers, in both academic and popular venues. Responses focussed on the epistemic status of ML and its projected impact on science have echoed statements to the effect that machine learning differs radically from prevailing modelling, statistical, or scientific methods in ways that are projected to change the landscape of scientific discovery or the nature of the epistemic fruits of scientific enterprise.

These interlocutors predict 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). Call this view the *disruption claim*. According to this perspective, ML methods are seen as holding the potential to retire or else displace the role of theorising 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." Some of these statements regarding the scientific usage of ML echo proclamations that were once made of classical statistical method: that big data analytic tools promise to allow the raw data to "speak for themselves" (Levins & Lewontin, 1985).

These claims of disruption could be understood as instances of ML or AI *hype*—they issue from spokespeople swept up in a wave of drastically overselling the capabilities of presently existing ML techniques [1]. Indeed, Hansen & Quinon (2023) argue that AI hype is principally responsible for belief in the possibility of theory-free science. While cultural misapprehensions of AI no doubt play a role, I argue that the root of such beliefs runs far deeper, and is in fact grounded in a conception of scientific objectivity.

Dating back to the first articulations of the modern scientific method, generally located in the writings of Francis Bacon (1878), the notion of *objectivity* has reigned supreme. Bacon advocated a kind of empiricism which, according to modern scholarship, sought to minimise the role of both normative values and theoretical considerations on scientific knowledge-production. The view that the influence of normative values should be eliminated from the scientific enterprise has been dubbed a *value-free ideal*, and critiqued on the grounds of both its in-principle untenability and its negative influence on science in practice Douglas (2009). In the present work, I will argue that its twin, what I dub a *theory-free ideal*, is both epistemologically ill-founded and pernicious in its practical influence.

This theory-freedom is intended as a negation of theory-mediation, theory-drivenness, theory-involvement, and theory-ladenness. It is also, we will see from an examination of the source literature, a denial that methods rest essentially on domain-knowledge or prior conceptualisation of the target phenomena, or that they should be understood as representing features of target systems in any epistemically salient respect. "Theory" is hence to be understood in a broad and colloquial sense, as incorporating domain knowledge or conceptualisation 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 conceptualisation of target phenomena play an essential role in shaping an em-

[1] Often, though not always, because such individuals stand to materially benefit from this widespread cultural misperception

pirical research programme.

Claims about the theory-freedom and revolutionary potential of scientific ML have been, in part, motivated by a concern for the future of science. If the scientific process becomes automated, purged of theory, and overtaken by uninterpretable black-box algorithms with human domain experts pushed out of the loop, this third category worries that the epistemic products of science may cease to be accessible to human interpreters. Interestingly, whether motivated by optimism or pessimism for the future of science, assessments of the role of ML in science have converged upon the same essential thesis: science will undergo drastic change with the advent of ML-based methods, because such methods are theoretically unmoored or conceptually impoverished in a way that sets them fundamentally apart from existing methods in statistics or applied mathematics. I dub this second claim the *distinctness claim.*

If, indeed, the procedure of science or the status of knowledge produced in science are set to radically change, this merits serious engagement by scientists and philosophers of science. If, instead, as I will argue, these are misplaced beliefs—which has established a foothold in not only the public consciousness but in communities of relevant experts—this narrative ought to be challenged, for it will lead scientists and the public astray. 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 attempt to construct the most unassailable version of these theses of disruption and distinctness before turning to assail them with the realities of scientific applications of ML. In the interlude, I will delve into what I diagnose as the root cause of these misapprehensions: the theory-free ideal. The perniciousness of this theory-free ideal is illustrated via two case studies in the scientific application of ML.

2 Disruption & distinctness

2.1 The beliefs of working scientists

Although talk of AI or ML bringing about "revolutions in science" receives its most hyperbolic declarations from journalists, it echoes—or is echoed in—the words 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). Summarising 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.

In their "AI for Science" technical report, Stevens et al.. (2020) similarly collect the wisdom of scientists from a diverse array of disciplines about the integration of AI/ML tools in their research processes. The scientists quoted in the technical report broadly endorse the disruption claim, writing that "grand challenges have emerged in the earth, environment, and climate disciplines that could be revolutionized through application of AI methods," p.28 that "AI will revolutionize the development of process scale models" in earth and environmental sciences p. 34, that "AI can revolutionize synthetic biology" p. 38, that AI has the capacity for "revolutionizing human health" and "AI has the potential to extend the average human life" p.43, that AI "promises a revolutionary understanding of complex materials and chemical processes across the entire hierarchy of relevant length and time scales" p.135, that "AI techniques that can optimize the design of complex, larger scale experiments could completely revolutionize the way experimental nuclear physics is done" p.63, and that "[t]]he already successful ongoing efforts using AI in cosmology along with new—and possibly unexpected—approaches will come together in the next 10 to 15 years to revolutionize our understanding of the universe and help answer some of the deepest questions in physics" p. 47.(Stevens et al., 2020).

A second report from The Royal Society, this one spanning a full 108 pages, is entitled "Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research." The monograph details avenues of impact on all fronts of scientific research, from epidemiology to materials science (Leontidis, 2024). While not all of the outlined pathways to disruption rest on the supposition of the epistemic distinctness of machine learning methods, several do. The adoption of deep learning methods in the scientific workflow is argued to be "transforming data analysis and knowledge generation" in the way it is used "to automatically extract and learn features from raw data" (Leontidis, 2024).

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. Scientists across disciplines widely conceptualise of the work of "normal" science as conforming to a roughly hypothetico-deductive template. 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 involving) research strategies has been a part of how scientists conceptualise of their work for decades. The picture these assembled voices paint, however, seems to point to a lessened overall need for theoretical input within scientific discovery.

Articulations of distinctness and disruption claims 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 new research centers, as well as industry sponsorship for conferences, awards, and similar 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"—what Desai et al. (2022) refer to as "the epistemically revolutionary new frontier raised by data science: the so-called 'theory-free' paradigm in scientific methodology" (Desai et al., 2022). 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).

Rafael Alvarado and Paul 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 modelling, where explicit theories and models are used” (Rowbottom et al., 2024, 172).

In a 2021 paper, Eamon Duede writes that philosophers and scientists alike have widely made claims of the epistemic distinctness of ML and its disruptive potential for science. 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 disruption and distinction claims, 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 empirical research strategies is ultimately responsible for philosophical endorsement of disruption and distinctness claims.

Sreckovic, Berber, and Filipovic (2022) differentiate machine learning techniques from standard practices in statistical modelling, arguing that statisticians employ theoretical assumptions, while machine learners do not (Srećković et al., 2022). Sreckovic, Berber, and Filipovic (2022) evaluate what they hold to be the key differences between traditional modelling approaches and machine learning methods in terms of the explanatory capacity of both and their capacity to elucidate causal relationships. Sreckovic 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,” Sreckovic 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 characterisations 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. Sreckovic et al. merely provision a philosophical exposition and justification thereof.

In a similar vein, Florian 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 modelling 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 modelling practices in the sciences, Boge claims, ML modelling 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 modelling or computer simulation in science and the application of machine learning methods is that the former procedure begins with a conceptualisation 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 modelling contexts, the lack of background theory or conceptualisation of the target phenomenon is taken as an 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.

Mieke Boon (2020) signs on to the thesis of the epistemic distinctness of ML, but on this basis denies disruption. Boon argues against the thesis that machine learning methods will obviate the need for auxiliary or intermediary human conceptual apparatus in the generation of scientific knowledge. She argues that the reason that we grant any sort of a priori plausibility to statements to the effect that big data will usher in a scientific revolution flows from a shared implicit view of how science works—one which she argues to be in error. She labels this erroneous conception of science a "strict empiricism." Her goal is to "make plausible that on an empiricist epistemology the elimination of any human contribution to scientific knowledge is in fact already built in as a normative ideal...strict empiricist epistemologies indeed support the claim that objective, although opaque, data-models produced in machine learning processes can replace and may even be preferable to human-made scientific knowledge" (Boon, 2020, 46).

Boon advocates for the necessity of human capacities for conceptualisation, abstraction, and interpretation in every aspect of collecting, preparing, and manipulating data: "not only when setting up the data-generating instrumentation and seeing to its proper functioning, but also in assessing and interpreting the data, drawing relationships between data from different sources, and for making the distinction between 'real' phenomena and artifacts" (Boon, 2020, 59). Further, "[t]he necessity to prepare data that are about something in the real world also implies that phenomena are crucial in scientific practices, even when only aiming at the generation of data for machine-learning processes" (Boon, 2020, 57). As evidenced in these passages, Boon clearly takes data provenance and processing to be an interpretive affair. But for her argument against would-be empiricist dogma to work, she must take it axiomatically that data and data models are objective and worldly. This is part and parcel of the misconception of scientific process and products which I believe Boon seeks to argue against—the misconception which I am, in this paper, chiefly arguing against. Namely, the misconception of data as being raw, objective, and worldly—unmediated by human theorising and conceptual grasp on the target.

If we banish the idea that data is objective and worldly from the start, instead viewing data collection, cleaning, processing, and interpretation in an inference-licensing capacity as a fundamentally theory-mediated affair, Boon's contentions with empiricist epistemologies appear to dissipate. Perhaps the stumbling block is most easily seen in Boon's in-passing characterisation of the role of idealisation in mathematical representation. Boon claims that "machines are not confined by the kinds of idealizations and simplifications humans need to make in order to fit data into comprehensive mathematical formalisms" (Boon, 2020, 51). The idea that the role of idealisation in scientific representation ultimately serves the human-interpretability of our representations—and that idealisations are evitable or eliminable—is not, of course, novel or idiosyncratic to Boon. It is, however, revelatory of her commitments to the representational properties of applied mathematics. Mathematical representation is conceptual work. Idealisation is essential to it. Use of ML-based tools in science thus cannot escape the necessity of idealisation.

Boon is a vocal proponent of a theory-laden conception of data. Yet her analysis of the prospects for machine learning in science appear to reveal inconsistencies in her view. Like Boge and Sreckovic et al.., Boon concludes that applications of ML in science will fall short of providing understanding or explanation in virtue of being conceptually impoverished. This, on her view, sets applications of ML to scientific research intrinsically apart from "real science." "'[R]eal science' and machine learning technologies," she writes, "operate in very different domains and must not be regarded as competing" (Boon, 2020, 58). If data is necessarily theory-laden and conceptually-mediated, however, then it cannot be the case that ML-facilitated science is a theory-free or concept-free epistemic activity, because the use of ML in science will be necessarily inflected by the theoretical and conceptual commitments inherent to the data. This account, therefore, appears to conflate potential misuse with the necessary operation of ML in science.

Boon, Boge, and Sreckovic 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 modelling methods in science nad traditional statistics. Boge and Sreckovic et al. further contend that the widespread adoption of ML methods will catalyse 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.

3 Conceptions of scientific objectivity

The concept of objectivity is central to modern science, both as this denotes an abstract construct or pursuit-worthy ideal and as set of human practices spanning several centuries. A philosophical debate concerning the variety of objectivity scientists ought to strive for is as old as modern science itself, and has remained active throughout its history. 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 doctrine concerning the ultimate 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.

3.1 The value-free ideal

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 corrollary 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 minimise the impact of human values and to constrain their influence to appropriate venues and junctures. A mostly implicit ideal of total value-freedom, however, is widespread and influential in scientific practice, as well as in the public's reception of the outputs of science, in science education, and in the interplay of science and public policy (Douglas, 2009). Various incarnations of both value-freedom qua philosophical doctrine and meta-narratives of value-freedom have emerged over the history of modern science.

The nineteenth century German social scientist Max Weber is widely held to be history's strongest proponent of a scientific objectivity rooted in value-agnosticism (Douglas, 2009; Proctor, 1991). Weber argued that the sciences—at least the social sciences, with which he was most intimately in contact—should strive for *wertfreiheit* or *werturteilsfreiheit* (value-neutrality or freedom from value judgement). This was the end of the era of gentlemen scientists and the beginning of the professionalisation of science with the ascendance of the German university. Sociology in 1800s Germany was deeply culturally-biased. The ideal of value-neutrality Weber sought to promote was intended to make scientists aware of their positionality, their cultural background, and their normative commitments, so that they might, to the extent possible, reduce or compensate for their biases in relation to the subject matter. Thus the positive articulation of objectivity as value-freedom holds up value-neutrality as a means of excising from scientific practice or the interpretation of scientific results *inappropriate* value-impingement.

Twentieth and twenty-first century philosophers of science, however, have argued that the end goal of total freedom from normative influence is unachievable on both practical and in-principle epistemic grounds. Denying the necessary influence of values on science, moreover, merely cements them, lends them an air of objectivity, and renders them unavailable to critical scrutiny (Douglas, Longino, Elliott). Heather Douglas (2009) delivers a careful analysis of two major twentieth century disputes over science's credentials and the appropriate role of science in public life: the Science Wars, which concerned whether scientific knowledge was discovered in the world or socially constructed, and the dispute over the appropriate role of science in governing (United States) public policy, which revolved centrally around the validity of the distinction between sound science and pseudoscience. Douglas concludes that what she refers to as the value-free ideal of science is pernicious, threatening both the credibility and utility of science. The meta-narrative that tells us that science is value-free effectively serves to conceal loci of normative input and reifies the implicit norms of uncritical scientists. Value-freedom, however, is not the only criterion invoked in philosophical conceptions of objectivity, nor is it the only conception of objectivity which holds sway over scientific practice in the form of meta-narratives.

Francis Bacon's *Novum Organum* (1620) is taken to be the first thorough articulation of the methods, aims, and scope of modern science. It also contains the *loci classici* for our modern conceptions of scientific objectivity. In the exposition of his idols of the mind, Bacon writes that "human understanding is like a false mirror, which, receiving rays irregularly, distorts and discolors the nature of things by mingling its own nature with it" (Bacon, 1878). One approaches objectivity by casting off these distorting "idols." Dear's (1992) history of the early-modern conception of objectivity begins by claiming that the concept must be understood as antonymous to the "distorting mirror" of subjectivity, a clear nod to the Baconian passage (Dear, 1992). Bacon's conception of objectivity is widely taken to have involved minimising the influence of human values and cultural biases on the production of scientific knowledge (**?**)(Proctor, 1991). Indeed, Proctor attributes to Bacon the genesis of the ideal of value-freedom in science.

3.2 The theory-free ideal

Of no lesser import to the attainment of scientific objectivity, Bacon urged minimisation of the influence of preconception, or theory. Wilson (1998) summarises the Baconian view of science as the "gathering of large numbers of facts and the detection of patterns. In order to obtain maximum objectivity, we must entertain only a minimum of preconceptions" (Wilson, 1998). In a history of the Baconian conception of Objectivity, Daston (1994) writes that "[w]hat Baconian facts seemed to promise was neither consensus nor freedom from all bias, but simply freedom from theoretical bias" (Daston, 1994). Indeed, "theory-freedom" or "alleged neutrality with respect to theory" is taken to be the defining feature of Bacon's conception of scientific objectivity and, hence, the genre of empiricism he endorsed (Daston, 1994)..

Many debates in the intervening centuries, in natural philosophy, epistemology, and philosophy of science have chiefly concerned the extent to which observation is necessarily theory-laden, empirical knowledge necessarily shaped by preconception, and strong assumptions or conceptual infrastructure required to get inductive inference off the ground. Working scientists in various disciplines have, too, in their own ways, debated the proper role of theory in science. There have even been moments in scientific history during which scientists took the doctrine of expunging theoretical bias too far, resulting in an ill-conceived empiricism.

Theory-freedom appears to have been a motivating factor in the specific methodologies chosen by early phrenologists and eugenecists. Francis Galton, pioneering figure of the eugenics movement, believed that good research practice should consist in "gathering as many facts as possible without any theory or general principle that might prejudice a neutral and objective view of these facts" (Jackson et al., 2005). Karl Pearson, statistician and fellow purveyor of eugenicist methods, approached research with a similar ethos: "theorizing about the material basis of heredity or the precise physiological or causal significance of observational results, Pearson argues, will do nothing but damage the progress of the science" (Pence, 2011). In collaborative work with Pearson, Weldon emphasised the superiority of data-driven methods which were capable of delivering truths about nature "without introducing any theory" (Weldon, 1895).

Beyond isolated and intermittent episodes like this, the meta-narrative of theory-freedom has not historically held much sway over the practices of working scientists. This is, at least, until the last 50 years. Theory-freedom, I argue, was only to become a widespread motivating factor in scientists' conceptions of their work in the 20th century. The acceptance of such an ideal hinges on the ubiquity of domain-generic statistical methods and "data-driven" research in the special sciences. However it would not become prevalent and pernicious until the age of "big data" and with the adoption of machine learning in the sciences. Until ML-assisted science became a reality or, at least, an imminent potential, the idea that science should or could be rendered free from theory was not seriously entertained in either philosophical or scientific communities.

With the advent of ML-assisted science, however, 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 levelled against theory-free reasoning over the centuries" (Leonelli & Zalta, 2020).

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 6 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.

4 The necessity of theory

Even the most simplistic of experimental designs reveals the nature and extent to which data, and scientific practice at large, are "theory-laden." 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 representationa and analysis. How we choose to measure and analyse 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. "Data" refers to abstract, formalised 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 representation. 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 parameterised.

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, and 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, idealisations, 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, 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.

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, 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, as well as 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. echoe 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.

The conclusion that an inference procedure cannot be rendered theory-neutral is overdetermined when we consider accounts from philosophy of science and theoretical computer science of constraints on inductive generalisation. Inductive inference is the procedure of gaining knowledge by extrapolating from a limited number of instances to a more general class—the fundamental task of ML. According to Norton (2003), an account which he dubs the *material theory of induction*, 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 line of research (Norton, 2003). There is no one inductively-valid formula to rule them all. Learning theory has itself independently discovered the impossibility of a universally valid domain-generic inference rule: the no free lunch theorems (Wolpert & Macready, 1997). While these results obtain only in a very artificial setting, the moral they deliver is an important one for ML in practice: inductive inference only works in virtue of having learned domain-specific inductive biases.

5 Theory-ladenness in scientific ML

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

Far and away the most impressive result that ML methods have achieved for science is AlphaFold 2.0. 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 straightforward, as is predicting the polypeptide chains these amino acid sequences will form. What mess of three-dimensional spaghetti those amino acid chains will assume once synthesised, however, is another matter entirely. This is an essential problem for the biomedical sciences. The three-dimensional anatomy of protein structure is determinative of 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—I have to recognise that protein structure is understood at four levels. DNA is a string composed of four alternative 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 of 9 atoms, which will bond together to form the backbone of the polypeptide chain. From this molecular backbone extends the R-group or side chain, the determinant of the amino acid's "flavour." 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 flavours. 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 several 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.

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. These associations of covarying amino-acid substitutions lend us what is known as a *protein contact map* which further lends us a *multiple sequence alignment* (MSA).

The AlphaFold team created their own database of protein structures—now the largest existing database of its kind—by scraping[2] existing publicly-available databases. DeepMind's AlphaFold 2.0 runs queries on an amino acid sequence in its pre-processing stage to obtain a multiple sequence alignment (MSA). Any modern approach to predicting protein structure begins with an amino acid sequence as input. As we have noted, given the state of modern biological knowledge, it is trivial to determine amino acid sequences given the protein's genetic blueprint. To construct the inputs, AlphaFold queries protein structure databases to assemble an MSA. In addition to the primary amino acid sequence and MSA, AlphaFold was also supplied as input database-derived *templates*—three-dimensional atomic maps—for a small number of sufficiently similar homologous protein structures. The templates and the MSA are rendered together to create what the AlphaFold team dubs a *pair representation*.

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 architecture which is easier to train and outperforms competing architectures by

[2] I.e., automatically extracting web data.

parallelising and better attending to higher-level contextual factors in the training data. AlphaFold passes both the MSA and the pair representation back and forth through the trunk of the model for a set number of iterations (48 blocks), 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 generative neural network which produces a plausible candidate 3-D protein structure. The 3D protein structure 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. Our aim is to show that theoretical considerations are playing an essential role 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.

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. Taking on board the notion of theory-laden measurement, we understand that the data on which AlphaFold is trained is richly structured by existing empirical knowledge of the target domain 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. 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 utilised 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 utilised 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 is then back-propagated through the network to update the model's parameters). 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. We can likewise only put the results of such a modelling effort to *use* when we have accomodated them within a theoretical framework.

5.2 Transcriptomics

Single-cell transcriptomics is a method for inferring cellular-level gene expression. The technique is utilised for identifying cell populations, modelling 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 massive 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; for this reason, researchers typically employ dimensionality reduction techniques. Dimensionality reduction is a (unsupervised ML) method of mapping a high-dimensional dataset to a lower-dimensional space—or *embedding* higher-dimensional data in a lower-dimensional space. Dimensionality reduction techniques are used to distill essential patterns from large datasets, make analyses tractable, and isolate signal from noise.

A now well-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: UMAP or t-SNE. The method produces visualisations for exploratory data analysis. A scientist cannot very well eyeball a 250,000-dimensional manifold and distill from it useful and meaningful information (or eyeball it at all). A two-dimensional embedding, however, might very well reveal visually intuitive information about cell populations and trajectories. However, as Chari and Pachter (2021) demonstrate, this now standardised procedure in single-cell transcriptomics lacks theoretical motivation, represents poor statistical practice, and is effectively incapable of providing meaningful biological information; instead it creates the opportunity for erroneous interpretation (Chari & Pachter, 2021).

In a series of analyses, the researchers demonstrated that the practice of repeated application of dimensionality reduction techniques introduced heavy distortions and was incapable of preserving the interpretively salient features of the datasets under investigation: local and global structure, distance, and continuousness (Chari & Pachter, 2021). Interpretive practices surrounding the resultant visualisations, they concluded, led to erroneous or conflicting conclusions. Chari and Pachter (2021) found that the combined use of supervised and unsupervised ML methods in single-cell transcriptomics was haphazard:

> “[T]he same k-nearest neighbor (knn) graph constructed from the higher dimensional PCA space is passed to both the clustering algorithm and the embedding algorithm...the embedding is then not an independent assessment of clustering results and is likely to form clusters that match the knn graph even if that graph does not represent the ‘original’ underlying manifold. Together, the use of such embeddings to imply or infer continuous relationships then becomes an arbitrary endeavour, with a user unable to trust seemingly dramatic connections or isolated populations, and likely to choose what seems most appealing” (Chari & Pachter, 2021, 14).

In one particularly striking example, Chari and Pachter projected transcriptomics datasets onto arbitrary shapes (a flower, von Neumann’s elephant) and found that they preserved the interpretively salient features—local and global structure, distance, and continuousness—commensurate with, or better than, the resultant embeddings from PCA → t-SNE or PCA → UMAP workflows (Chari & Pachter, 2021).

These techniques fall under the heading of what I term *Rorschach research methods* or *intuition laundering*: interpreters of the results of these data analysis methods are free to cast upon them whatever intuitive interpretation appeals to them, wielding the graphics to lend supposed empirical support to their claims. Such poor practices are likely to emerge anywhere the methods of ML are employed without adequate theoretical grounding and statistical literacy. Chari and Pachter (2021) demonstrate that semi-supervised learning methods and targeted embeddings for specific featural dimensions are capable of elucidating far more than the naïve methods they critique. Such approaches, however, require domain expertise, critical thinking, and being able 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

AlphaFold is a case of resounding success; perhaps the greatest win for ML in science to date. No other application of ML to science has achieved quite so stark an advantage over pre-existing techniques. Many applications of the tools of ML to science, by contrast, 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 what sort of knowledge could be gained by their use. Many scientists have also faced great frustrations in incorporating computational tools into their research paradigms, either because they were attempting to utilise ML in an untenable, "theory-free" manner or because they faced difficulty in their attempts to imbue ML-based tools with the requisite theory or domain knowledge.

Researchers in the biomedical sciences bemoan the fields' recent infatuation with the tools of ML in operation with its longstanding "theory-aversion"—what I have termed a theory-free ideal in science (Coveney et al., 2016). Incorporating theoretical principles into ML-assisted and big data-fueled research can prove difficult, and is unlikely to happen when institutional and publishing incentives overwhelmingly favour the collection of higher volumes of data and the adoption of novel computational tools over critical thinking and principled research design (Coveney et al., 2016). In fundamental physics, by contrast, the need for theoretically-informed models is more apparent and is met with less resistance. Karniadakis et al. review methods of incorporating physical principles into applications of DL in physics (Karniadakis et al., 2021). Incorporating theory into ML-assisted scientific practice is no simple matter, but work of this kind reveals both its possibility and its necessity.

The use of unsupervised learning in physics is now concentrated on "physics-informed" architectures—forcing the model to conform to the form of a known e.g., physical, principle. These methods, unlike "naïve" unsupervised cluster or regression techniques, which can only occupy relatively simplistic intermediary calculational or bookkeeping roles, can play a far more central role in research. This is precisely *because* they have the conceptual resources to serve a meaningful role in empirical research. In this sense, the conclusion I reach aligns with some of the reasoning in Boge, Srećković et al., and Boon: theory-involvement is a requisite feature of our conceptual instruments in science for them to be able to elucidate previously unknown features of our natural world from data. The issue is that the perfectly theory-free vision of ML (or DL, or unsupervised DL) in science, which is the target of these scholars' critiques, either singles out a strawman or a failure case.

6 Conclusion

It is widely believed of the methods of ML—as reflected in the texts countenanced in this paper—that if loosed on enough data, they are capable of discovering meaningful patterns, natural joints, or mind-independent truths of their own accord, sans input from human theorising or conceptualisation of the target system. There is a naïve version of this view that only those unfamiliar with the ins and outs of applying ML methods could hold onto. However, a more sophisticated version of the thesis also exists and is commonly held even by engineers, researchers, and practitioners building and deploying ML-based tools. This is the idea that unsupervised learning tools are capable of discovering mind-independent natural patterns or boundaries in a "principled" manner without arbitrarity or human input. If we believe this, and if we also believe the techniques of ML to be "opaque" or "uninterpretable" in some novel way—in a way that sets them fundamentally apart from existing conceptual tools or instruments in science—then what is learned via these tools will be inscrutable to human scientists. If these tools are then adopted widely in scientific practice, this will then entail radical change to the varieties of epistemic outputs science is capable of generating. 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 and much at stake in its appropriate use. Where to localise theoretical considerations in DL-based scientific workflows appear to differ substantively, along various dimensions, from a certain canonical mode of scientific modelling. On a received, roughly hypothetico-deductive view of experimental science and statistical modelling, 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 represents the target are specific to the epistemic concerns of a particular scientific/modelling exercise. In big data analysis and 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 we are ever in contact with the data. This practice stands in stark defiance of Bogen and Woodward's (1988) claim that data are 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 of a model architecture and specification of loss, and in training. Theoretical considerations further come in at the level of model evaluation, in our formal assessments of the success of the 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*. 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.

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 specialisation of science, the proceduralisation of science, its automation, the progressive increase in the distribution of intellectual labour it involves, the extraction of the knowledge of domain experts and its mechanisation 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 runs at least as much from the automation of scientific practice to the adoption of the tools of ML in science as it does in the reverse.

To make sense of the present day landscape of science and the directions in which it is evolving, we will require a philosophy of science of machine learning. This must, however, be a philosophy of science willing to cast off an outdated, monolithic, and overly-restrictive conception of scientific methods and the epistemic outputs of science. This must be a philosophy of science willing to weigh-in on debates and draw boundaries between admissible practices and the pseudo-scientific or pseudo-statistical. It must be a philosophy of science that is not conned by the hyperbolic narrative of the inscrutability of machine learning methods; that is willing and able to comprehend the techniques and how they are wielded to empirical ends.

7 Acknowledgements

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, Florian Boge, Daniel Burnston, David Childers, Clark Glymour, Nic Fishman, Leif Hancox-Li, Zachary C. Lipton, Arya McCarthy, Amira Moeding, John D. Norton, Angela Potochnik, and Jennifer L. Whyte.

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.

Bacon, F. (1878). *Novum organum*. Clarendon press.

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

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.

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*.

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.

Coveney, P. V., Dougherty, E. R., & Highfield, R. R. (2016). Big data need big theory too. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2080), 20160153.

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

Daston, L. (1994). Baconian facts, academic civility, and the prehistory of objectivity. *Annals of scholarship*, *8*(3), 337–363.

Dear, P. (1992). From truth to disinterestedness in the seventeenth century. *Social Studies of Science*, *22*(4), 619–631.

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.

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

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

Jackson, J. P., Weidman, N. M., & Rubin, G. (2005). The origins of scientific racism. *The Journal of Blacks in Higher Education*, *50*(50), 66–79.

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.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

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.

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).*

Leontidis, G. (2024). Science in the age of ai: How artificial intelligence is changing the nature and method of scientific research.

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

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.

Pence, C. H. (2011). “describing our whole experience”: The statistical philosophies of wfr weldon and karl pearson. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 475–485.

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.

Proctor, R. (1991). *Value-free science? purity and power in modern knowledge*. Harvard University Press.

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.

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.

Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D. (2020). Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science. *Report from the US Department of Energy*.

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

Weldon, W. F. R. (1895). Attempt to measure the death-rate due to the selective destruction of carcinus moenas with respect to a particular dimension. *Proceedings of the Royal Society of London*, *57*, 360–379.

Wilson, E. O. (1998). *Consilience: The unity of knowledge* (Vol. 31). Vintage.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82.