The Devil in the Data:
Machine Learning & the Theory-Free Ideal

Mel Andrews
June 1, 2023

Abstract

Machine learning (ML) refers to a class of computer-facilitated methods of statistical modelling. ML modelling techniques are now being widely adopted across the sciences. A number of outspoken representatives from the general public, computer science, various scientific fields, and philosophy of science alike seem to share in the belief that ML will radically disrupt scientific practice or the variety of epistemic outputs science is capable of producing. Such a belief is held, at least in part, because its adherents take ML to exist on novel epistemic footing relative to classical mathematical or statistical modelling approaches utilised in science. Namely, they take modelling with ML to be a “theory-free” enterprise, in the sense of not resting essentially on input from human conceptual grasp on the target phenomenon and domain expertise. I take this view to arise from the further, and more deeply entrenched belief that data is worldly and objective; i.e., data is viewed as recapitulating or representing with perfect fidelity the structure or properties of the systems in nature it is sampled from. Yet most contemporary philosophers of science take on board, in one version or other, the thesis of theory-ladenness or theory-mediation of observation or measurement. From this, it follows that most philosophers of science (and, I will venture, many scientists) hold irreconcilable views on the nature of data, an internal tension which threatens the integrity of their appraisals of ML in science. Taking the thesis of theory-ladenness on board—and its implications for the nature of data seriously—it follows that there is no reason to believe that ML differs fundamentally in its epistemic footing from established mathematical modelling approaches in the sciences. I show that usage and interpretation can differ from “standard practice” in scientific projects which wield the tools of ML without accepting a difference in the epistemic or representational status of such tools.

1 Introduction

The field of artificial intelligence—or, perhaps more accurately, fields, as there have been a number of distinct research communities and bodies of scholarship under this denomination which bear little in common beyond the title—has been
beset by waves of popular attention and ebbs and flows of funding since its incipience. It is worth noting that there have been (at least) two epochs of so-called “hype” in the past twenty years which have compounded together to emerge in the present form of obsession over “automated science.” In the late 1990s and early 2000s there occurred an early internet data-mining craze, in which it was discovered, for the first time, that rudimentary statistical techniques, unleashed on sufficiently large datasets, could reveal complex and meaningful patterns. Between 2010 and 2013 there occurred what has since been called a deep learning revolution, in which breakthroughs in neural network approaches to image classification (alexnet) and natural language processing (NLP) (word2vec) renewed public interest in the methods of machine learning. We may be on the cusp of a third wave at present with the sudden ascension of transformer models, a form of deep learning which is capable of converging more reliably and efficiently owing to inbuilt “attentional” mechanisms which hone in on areas of highest salience in the representations they are presented. Publicly accessible beta releases of large language models (LLMs), as well as diffusion-based and related techniques for generation graphics have, in the last year, taken an unsuspecting world by storm.

The prospects of machine learning for science have, in particular, opened wide since 2018, when machine learning techniques were adopted in the large hadron collider at CERN for sorting the significance of particle collision events (Duarte et al., 2018) and 2021, when DeepMind released its AlphaFold 2.0 (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. Suffice it to say that 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.

This response has issued from cultural commentators, journalists, and media personalities, from the mathematicians and engineers producing the tools of ML and the scientists deploying them and, of course, from philosophers, in both formal and informal venues. Responses focussed on the epistemic status of ML and its projected impact on science have echoed in major part and by a wide margin 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 scholars prophesy a data-driven or ML-driven scientific revolution (Hey, Tansley, Tolle, et al., 2009; Mayer-Schönberger & Cukier, 2013; Anderson, 2008; Spinney, 2022). They foretell an end to hypothesis testing (Anderson, 2008; Mayer-Schönberger & Cukier, 2013; Spinney, 2022). They envision ML methods retiring or else displacing the role of theorising in science (Anderson, 2008; Mayer-Schönberger & Cukier, 2013; Spinney, 2022; Srečkovič, Berber, & Filipović, 2022). They predict that ML will obviate the need for domain knowledge or expertise in science (Mayer-Schönberger & Cukier, 2013). The predictive success of ML is projected to replace the need for insight into causal
mechanisms or the data-generating process (Mayer-Schönberger & Cukier, 2013; Spinney, 2022). Some of these statements 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).

Each of these claims reveals an anxiety that science will radically change owing to a mass turn over in the sorts of epistemic tools wielded in sciences. There exists a widely-held conception of science as capable of building a veridical picture of the world because data, the primary currency of science, is itself veridical—in the sense of being worldly and objective, uncompromised by the messiness and arbitrariness of mediation by the conceptual schema of epistemic agents. The use of machine learning methods poses a dilemma to such a picture. Their efficacy seems to force us to accept one of two conclusions: either the status of scientific practice and knowledge is poised to radically change, or we will need a radical revision to our established depiction of how science works. Curiously, philosophers of science confronted with this puzzle have thus far near-unanimously elected to fall upon the first horn, offering narratives of future sciences, radically transformed by a new radically uninterpretable, radically instrumentalist paradigm. I take the advent of machine learning methods in science instead to be a clarion call to revise our understanding of data and its place in scientific practice.

The remainder of this text will proceed as follows. I first define what I mean by data, by theory, and by the conception of data as theory-mediated or theory-laden. I then illustrate some of the diversity of uses of mathematics in science, machine learning inclusive. Following this, I delve into several assessments of machine learning for science by philosophers of science and attempt to show that their claims of distinctness rest on misconceptions about the ontology and/or epistemology of data and are therefore unfounded.

2 Theory-Ladenness

Some definition of terms is in order. By data, I refer to the results of measurement or observation which serve an evidential role in an inference procedure. I take theory to be conceptual commitment to the nature of worldly phenomena. 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 measurability of some phenomenon and, if we are making measurements and performing quantitative analyses thereon, commitment to its quantitative nature. How we choose to measure a phenomenon includes generally a commitment to what I would term the “quantitative ontology” of the phenomenon, e.g., is it scalar or ordinal? Measurement cannot be total, and therefore there is always a commitment as to what to look at experimentally and what to exclude. There is always a commitment to the appropriate level of abstraction at which to study the phenomenon in play in terms of such things as instrument settings like degree of magnification or periodicity of sampling. The very design of our in-
stirns of measure and their calibration includes various commitments to the nature of the worldly phenomena under investigation. 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. “Data” is not physical phenomena. “Data” is abstract representation of the results of direct observation or measurement which is capable of serving an evidential role in licensing inferences about physical phenomena.

In contemporary philosophy of science, it is accepted widely that data is theory-mediated (Boyd & Bogen, 2009; Bogen & Woodward, 1988; Gitelman, 2013; Leonelli, 2019b). Philosophers of science have largely overcome “[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) and accept now that there is “no pristine separation of model and data” (Lloyd, 2018, 397).

Bogen and Woodward (1988) (Bogen & Woodward, 1988) have been read (e.g., by (Schindler, 2007)) as arguing for a bottom-up view of the inference from data to phenomena, unmediated by theory. As (J. F. Woodward, 2011) points out, the intent in (Bogen & Woodward, 1988) was only to rule out that theory is predictive or explanatory of data. Bogen and Woodward are amenable to theory-mediation understood as mediation by conceptual grasp of epistemic agents over the phenomena and the need for “substantive empirical assumptions” (J. F. Woodward, 2011). Bogen (2016) has himself argued 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 (Bogen, 2016). Boyd (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.

Philosophers of science now popularly profess allegiance to a theory-laden conception of data; certainly none nowadays outright defend an account of data as raw and objective. But the philosophers of science most attuned to how data is understood in the context of scientific practice note that in many philosophical accounts, data is still treated as raw and objective, “as reliable information source—a mere “input” into processes of modelling” (Leonelli, 2019b, 4). Leonelli (2018) argues that mainstream accounts from within philosophy of science—though they might profess otherwise—tend to treat data as representational, and understand this representation relation in terms of the recapitulation of worldly structure in data. “Philosophers tend to assume that data have some sort of representational content, in the sense of instantiating some of the properties of a given target of investigation in ways that are mind-independent” (Leonelli, 2019b, 4).

Although no philosophers of science today explicitly endorse an interpretation of data as objective and worldly, unfortunate relics of this view remain widespread, often in the form of a conception of data as mere “empirical input for modelling” hence “implicitly accepting a view of data as intrinsically
reliable representations of the world” (Leonelli, 2019b, 4). Leonelli (2018) investi-
gates “the different extents to which theory—understood broadly as a set of theo-
etrical commitments and goals—impinges on inferential processes from data” (Leonelli, 2019b, 22). In several book-length treatments of the use and in-
terpretation 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.

3 Species of Applied Mathematics

Let us take a step back from our philosophical analysis to illustrate what we mean by applied mathematics, scientific models, and machine learning for science. Philosophers of science, as we have seen, contrast theoretical models from data models (Suppes, 1966). Physicists, in turn, have traditionally drawn a line between theoretical models and phenomenological models (Cartwright, 1983). Machine learning, a form of statistical modelling, is contrasted with “traditional” statistical modelling, a contrast often attributed to the “athe-
etorical” nature of machine learning (Breiman, 2001; Srećković et al., 2022; Shmueli, 2010). Such distinctions all rest on the idea that some usage of ap-
plied mathematics in science is theoretically-driven and ontologically-committed while other applied mathematics is not. I think it will be illustrative to trot through a few examples, running the gamut from most canonically theoretically-motivated modelling work in physics to most “instrumental” and “opaque,” as seen in modern applications of deep learning for science.

3.1 Mathematical Modelling in “Normal Science”

Phenomenological models concern entities and quantities which are available to our (instrument-mediated) observation. Phenomenological models are typically contrasted with theoretical models, concerning unobservables, i.e., entities and quantities which are inferred on the basis of the wider web of mathematical-conceptual infrastructure. Take as examples of phenomenological models the harmonic oscillator or the Ising model. The harmonic oscillator, along with its damped and driven variants, is a simple equation of classical mechanics describing sinusoidal oscillations around an equilibrium point. The equation, in its basic (undamped, undriven) form, relates a restoring force $F$, a mass $m$, a posi-
tion $x$, and a spring constant $k$. Force, within a classical mechanical framework, if we are to take Newton seriously, is not to be lent much ontological independence. It is an abstract object, a calculational device, or a heuristic (De Gandt, 1995; Cohen & Whitman, 1999). But, for classical mechanics, force is an indis-
pe nsable construct. Mass and position are measurables. The spring constant $k$ might be read as a referring variable which references a real measurable, this
being the stiffness or elasticity of the spring. In actuality, it is better read as a non-referring term, as it bundles together various things which express variable degrees of ontological independence and measurability, including elasticity, the resisting medium, and elements of convention and convenience.

The Ising model, when applied to magnetic polarisation, reveals how phase transitions occur in magnets under varying conditions of temperature and pressure. The model consists of a lattice of variables or nodes which are canonically taken to represent atomic spin (magnetic dipole moments). The spin states of the particles are Boolean. Spin states interact but only locally, obeying a Markov condition. The model reveals how global state transitions can emerge from confined local interactions among spin states. In both the case of the Harmonic oscillator and the case of the Ising model, we see that there are some seemingly arbitrary (arbitrary with respect to the world) or conventional elements to the formal models, but that some elements of the model are directly representational of things that we can observe or measure. We can also witness, in both cases, how idealisation renders these models explanatory.

The theoretical-phenomenological distinction is, of course, itself a fuzzy boundary. There are arbitrary, historical reasons why certain bits of applied maths are taken to fall into one and not the other of these categories. Science, even physics, is a social enterprise, and some of what we do there we do by convention. The Ising model of ferromagnetism rests fundamentally on particle spin—this is the crucial, i.e., difference-making, variable in the model. But spin is a theoretical entity; we can only indirectly infer its existence and valence from relations of higher level variables in an experimental context.

This moves us to our prime example of a theoretical model in physics. We can zoom-in, conceptually, on the magnetic dipole moment which is at the heart of the Ising model. Particle states and interactions like spin state are theorised under the Standard Model of particle physics. We have mathematical representations of all varieties of particle states, including spin, under perturbation theory in quantum mechanics, which deals with interactions of the fundamental forces. The classical notational standard for this was once bra-ket notation. [Drop in the bra-ket equivalent notation for a simple Feynman diagramme and the Feynman diagramme.] It is a dizzying notation. It is uninterpretable. Even for those most fluent in this formalism, it is unclear what, if anything, the individuated syntactic elements of the theory are meant to “refer” to. It is next to impossible to visualise. This is where Feynman diagrammes come in. Feynman diagrammes are a visual notational form which allow us to represent and reason about features of fundamental force and particle interactions which we will never be able to directly observe. When we utilise a cluster of circles to represent the nucleus of an atom and draw some little dots following dotted-line orbits around it to represent the electrons, we know that the diagramme is, in a sense, lying to us. But we take the theoretical entities represented therein and the spatial relationship (mis)represented therein to be, in a sense, real. In a Feynman diagramme, we are still, in a sense, representing particles and their interactions, which we hold to be part of the physical makeup of the world. But the configuration of the diagramme itself is not mappable onto nature in the
way a diagramme of an atom is, the Feynman diagramme rather gives a pictorial
depiction of how the mathematical entities are to be reasoned about. The Feyn-
man diagramme is not interpretable in the sense that we ought not to attempt
to map what it depicts onto something that we presume to exist in nature.
However, the Feynman diagramme renders other mathematical representations
in physics more interpretable.

Both what physics call “theoretical modelling” and “phenomenological mod-
elling” fall under the heading of “theoretical modelling” for philosophers of sci-
ence, as contrasted with “data modelling.” Data modelling is identical to simple
statistical modelling. The canonical example of this is linear regression, in which
we fit a linear model to two variables by parameter estimation from data. We
might employ such a technique if we want to assess the relationship between
rainfall and frog population, or smoking and cancer. We can also employ linear
regression with algorithmic parameter estimation. This is an application of ma-
chine learning to scientific modelling. It is not functionally different from “ana-
logue” linear regression in what sorts of inferential problems it can be applied
to or in what sorts of “epistemic goods” it is capable of producing (knowledge,
understanding, explanation, etc.).

The use cases for machine learning methods in science are very broad. When
scholars make pronouncements about the epistemic novelty of machine learn-
ing methods, it appears that they are referring (wittingly or no) exclusively to
deep learning methods. It is worth noting here that machine learning properly
denotes much, much more than mere deep learning, and that various machine
learning methods are already widely deployed across the sciences—as they have
been for some time. Even restricting our analysis to the methods of deep learn-
ing alone, however, we still find both the methods and the use cases to be quite
diverse. Potential applications of DL techniques in scientific enquiry range from
anomaly discovery in cosmological surveys and sifting signal from noise in par-
ticle collision events, to nanomaterials discovery, biomaterials discovery, drug
discovery, and protein folding, to serving as a model of image processing in the
primate visual system. Generative Adversarial Networks (GANs) alone, for in-
stance, can be useful in materials discovery or drug discovery, in preprocessing
data or images for later automated classification, or in “filling in the blanks” of
discrete time step images—for instance, neuroimaging or cell development—to
create a continuous time evolution. We can utilise the methods of DL to approx-
imate solutions to stochastic PDEs. We can use DL to probe the latent space of
biomaterials for as-yet-unimagined cold-tolerant hydrophilic protein structures.
We can use DL to represent how the mammalian visual cortex engages in ob-
ject recognition. We can use DL to draw a line between particle collision events
likely to be interesting and those which are uninformative. There remain such a
plurality of use cases for deep learning in science that any two have little more to
do with one another than any two arbitrary instances of applied mathematics.
The techniques of machine learning can be utilised in science for problems which
are as basic as linear regression or image pre-processing. The more impressive
results, however—the results commanding the attention of the broader scientific
and philosophical community—are far more sophisticated than these.
3.2 The Unreasonable Efficacy of Alphafold

Far and away the most impressive result that ML methods have achieved for science is AlphaFold. To appreciate the impressiveness and 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 tell a priori from the genotypic specification of a particular protein 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. A single base variant can lead to amino acid deletions, substitutions, and frameshifts which can result in misfolded proteins. Some of these can be lethal.

To truly comprehend the difficulty of the protein folding problem—and how the methods of machine learning were able to get around it—we have to recognise that protein structure is understood at four levels. Deoxyribonucleic acid, DNA, is a string, composed of four alternative base pairs. It encodes information in sequence. When proteins are assembled, that DNA is read, in sequence, 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 core 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—canonically, α chains and β sheets. 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 a fussili-like structure (α helices) or a linguini-like structure (β sheets). This is, of course, an over-simplification, but up until this point things remain relatively straightforward. There is 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 stages 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. You have 20000 guests over for dinner and you have to sit them around a table in such a way as to make none of them unhappy with their nearest neighbors. Their neighbor-preferences fall along twelve dimensions. The table is a spiralling, self-intersecting roller coaster. You have exceptionally noisy access to each one of these factors.

At first blush, this seems like an unsolvable problem. The first 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. We have gone from attempting to picture, how, exactly, a ball of noodles will be intricately folded and knotted together from a box of uncooked spaghetti to attempting to picture a ball of noodles from a box of uncooked spaghetti on which narrow bands upon individual noodles are color-tagged and can be matched to corresponding colored bands on other noodles. 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 existing publicly-available databases. DeepMind’s AlphaFold 2.0 runs two parallel queries in its pre-processing stage. 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 then 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 core structure of AlphaFold 2.0 is an attention network—commonly termed a transformer. A transformer is a subset of neural network architectures which owes its efficiency and accuracy to its ability to single out the most salient features on the input (this can be accomplished by encoder-decoder architectures, multiple attention heads, etc) and the ability to hierarchically model context in its inputs via embeddings. 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 to influence one another as each is refined. The

---

1 I.e., automatically extracting publically-available web data.
output of this refinement procedure is then fed to a generative neural network which produces a plausible candidate 3-D protein structure. The 3-D protein structure is then passed, with MSA and pair-representations, back through the trunk. This is repeated for three iterations.

What is striking about this scientific procedure? For one, all of the data that was available to AlphaFold was preexistent. This is to say that all of the approaches aimed at solving or bypassing the protein folding problem which predated AlphaFold had access to that same data, the same preexisting scientific representations. All approaches in computational biology to this problem begin with the same big-data-derived representations: templates and MSAs. What AlphaFold 2.0 is doing that is remarkable is a progressive refinement of existing representations to draw out what is most salient for the generative task. AlphaFold is effectively engaged in a highly mathematically sophisticated enhancement of existing scientific representations from data. We ought to understand AlphaFold as taking preexisting scientific representations which were both uninterpretable and unuseful and rendering them inferentially useful (without rendering them human-interpretable).

In “classical” statistical modelling, we are typically formulating hypotheses and going out to collect data capable of adjudicating between our hypotheses. This means that 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 good deal 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. The remainder of the interpretive work comes in, typically, in what we take ourselves to have learned from the model output and, effectively, in how the model is wielded. This is a fairly pivotal break from standard scientific practice and merits serious consideration. It cannot mean—or so I will argue—that machine learning as a way of doing applied mathematics or scientific modelling puts itself on novel epistemic footing, or that the representations it leverages are of a novel kind. Nor can it mean that there is no representation, no content to the modelling exercise with ML. The content is merely, as it were, “coming in the back door”—which is to say, the totality of the data collection, cleaning, curating, labelling, and pre-processing procedure, much of which falls outside of our hands in the typical case of deployed ML, is an interpretive act.

AlphaFold is a case of resounding success—indeed, I think that it can be called, without controversy, 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 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. But 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, Dougherty, & Highfield, 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. 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.

4 Philosophy of Science on ML

Various philosophers of science have sought to weigh-in on the advent of machine learning techniques and what they portend for the future of scientific practice and empirical knowledge. This section provides a critical engagement with this scholarship and attempts to show that the beliefs that ML exists on novel epistemic footing or augures scientific revolution emerge from lower-order commitments to the objectivity of data.

4.1 Boge (2022)

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, Grünke, & Hillerbrand, 2022, p.71).

Boge differentiates between what he holds to be two distinct aspects in which models employed in science can be instrumental. The first axis of instrumentality Boge characterises is one in which the modeller makes idealisations that are counter to fact. The second axis of instrumentality refers to the contentlessness
of features of the model. The proposal is that the contentlessness, or meaninglessness, of e.g., parameters and hyperparameters in deep learning models makes these models instrumental in a special way, although Boge acknowledges that phenomenological models in physics and statistical models (when not theoretically generated) can exhibit this quality. There is a second quality of deep learning models which, Boge contends, in combination with the first, poses a unique puzzle. Boge acknowledges that traditional modelling and computer simulation can be opaque in the sense of the mechanisms by which the model reaches its final results being inscrutable. This first dimension of opacity can be thought of as a model-relative opacity. It concerns the complexity of the internal workings of the model and is present in a situation in which we do not know how a model or simulation arrives at its particular conclusions. In other words, the internal mechanisms of the model are uninterpretable. The second dimension of opacity is meant to be indicative of a lack of insight into what sorts of regularities are indeed being picked up on by the model or a lack of insight into what features in the data the model is fed drive its outputs. In other words: “[w]hen a DNN learns to approximate a desired function, it is hence not only opaque how, precisely, it achieves this goal: It is also opaque what it is about the data that drives this process” (Boge, 2022, p.62). This facet of opacity is, in Boge’s estimation, “ultimately, the distinctive factor which sets DL apart from all traditional models and, eventually, impairs our ability to acquire scientific understanding in a special way” (Boge, 2022, p.61).

Exploratory modelling contexts are those which Boge defines to be unguided by “background theories” or strong commitments about the target phenomena under study; ones in which the conceptual schema which serve to make sense of the phenomena have yet to be defined. In such contexts, the usage of models which are both instrumental in the sense of some of their (formal) features being contentless and opaque in the sense that one cannot read off from the results of the modelling procedure what features of the data were decisive in lending that result (or even perhaps what worldly features captured in that dataset were decisive) Boge takes to be an impediment to understanding.

Boge offers us an illustration of what he takes to be the key differences in the interpretation of what he calls “classical mathematical models” and machine learning models (represented in terms of the function a neural network approximates to). In the case of the classical mathematical model, what is crucial is that the constituent variables of the model are each, in themselves, mappable to some feature of the projected function—which presumably, themselves, have clear worldly correspondents.
Of this illustration, and of the discrepancy between “classical” modelling and deep learning, Boge writes that “it is not immediately clear what (hyper)parameters should be taken to represent about the curve in Fig. 2b, and certainly even less so as to what they represent about the system whose behavior is in turn represented by that curve” (Boge, 2022, 51). This “apparent meaninglessness of the parameters” in neural networks, in contrast to “traditional mathematical models” is what Boge takes, first and foremost, to differentiate the two classes. The uninterpretability of DL models and their outputs is effectively predicated on the inscrutibility of the model in what it represents. Taking the model structure in this exercise to be the neural network model in deployment, what Boge wants is for this to represent something that can clearly be plotted and which can further be clearly taken to stand-in for features of the target system that the modeller is leveraging the model to license inferences about. In other words, Boge appears to want the model structure to be fully decomposable into discrete representational elements. But, as we have seen from our detour through the diversity of uses of applied mathematics in science in section 2, this ideal of decomposability and interpretability is simply untrue of entire swathes of existing ‘ML-free’ modelling approaches in the sciences.

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. Especially in exploratory modelling contexts, the lack of background theory or conceptualisation of the target phenomenon is taken as an impediment to understanding. Both of these concerns are undercut by an understanding of data as prestructured by our theoretical or conceptual grasp on the target
phenomena. And, indeed, most researchers doing epistemically well-positioned scientific work with ML-based tools engage in laborious processes of labelling or feature-engineering; a theory-heavy activity. In the scientific application of DL techniques that have broken away from the need for feature engineering, the cutting edge work is heavily focussed on how to build in theoretical constraints (e.g., (Karniadakis et al., 2021)).

4.2 Sreckovic, Berber, & Filipovic (2022)

In a similar vein, 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 (Srecković 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 to be uninterpretable and not to rest 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” (Srecković 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” (Srečković 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” (Srecković et al., 2022, 166). Here, the contrast the authors draw between broadly “data-driven” and “theoretically-motivated” methods is telling. For even unlabelled data is not “theory-agnostic” in the sense that would be needed for such a distinction to be meaningful.

4.3 Boon (2020)

As I have argued, most philosophers of science grappling with the existence and scientific uptake of tools from ML have accepted the premises that machine learning is, in the first place, fundamentally different from existing approaches in applied mathematics and, in the second, that it will usher in sweeping changes
to the epistemic products or practices of science. One noteworthy exception is a 2020 chapter by philosopher of science Mieke Boon.

Boon (2020) 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). There are three facets of this empiricist epistemology which Boon perceives to contribute to the misplaced ideal of automated science. These commitments are to: 1. the deductive-nomological model of scientific explanation, 2. the collapse of the distinction between data and phenomena, and 3. the semantic view of theories. Boon contests each of these positions on the basis that they grant no epistemic role for human understanding or interpretation, or intermediary conceptual instruments such as theories, laws, or causation, over and above mere patterns in data. I will briefly recapitulate her arguments against these positions.

Boon claims that on the Hempelian deductive-nomological and inductive-statistical accounts of scientific explanation, the relata that make up a scientific explanation can be substituted for, traced back to, or subsumed by, mere patterns in data, or data models. In the second place, there is a perennially recurring debate in philosophy of science over whether “phenomena” refer to that which can be known of the mind-independent things in the world or whether it is reducible to mere “patterns in data” (Boyd & Bogen, 2009; Bogen & Woodward, 1988; Bogen, 2016; J. F. Woodward, 2011). The empiricist standpoint, according to Boon, holds that “phenomena are nothing more than statistically justified mathematical structures in data” (Boon, 2020, 53). If we take the aims of science to be the facilitation of knowledge or understanding of phenomena, and if phenomena may be collapsed into mere patterns in data then, as Boon argues, there is no reason to preserve theories or theoretical models, or even scientific concepts aimed at facilitating human comprehension of phenomena. Lastly, Boon targets the semantic view of theories. On this account, theories are understood as abstractly-formulated logical sentences specifying relations among phenomena. Models of theory exist at a lower level of abstraction, effectively numerically filling in the blanks of the variables specified under a theory. Lastly, data models elucidate patterns in the data, enabling the relation of the “raw data” or “raw observation” to the theory via the theory model. Boon critiques the semantic view in not granting sufficient epistemic value and epistemic autonomy to theory models over data models.
Unlike the other philosophers of science I have surveyed in this paper, Boon clearly apprehends that there is something amiss in the assessments that machine learning is on novel epistemic footing and will replace existing approaches to science. She also traces this back to a shared, latent false depiction of scientific practice—in particular, of the operation of models or applied mathematics in science. I am, of course, sympathetic to the position as I myself argue here along similar lines. However, in her conclusions, I worry that Boon herself falls prey to the fallacy that is at the heart of her three critiques. Namely, in taking data to be in a sense of the world or capable of recapitulating worldly structure.

In order for Boon’s assessment of this would-be implicit empiricist epistemology to do the work that she seems to want it to, we would have to get on board from the start with the fundamental premise that the data which scientists derive from their instruments is merely sampled from the properties of the worldly phenomena under study. Data modelling then merely cleans, homogenises, renders interpretable, or ekes out patterns in this data. Crucially, the edifice of Boon’s argument hangs on the fact that it is the move from models of data to models of theory—the move represented in the gap between boxes 2 and 3, per Boon’s diagramme—that brings in human conceptual or interpretive faculties. The data collection and handling—the data itself, even—is a purely objective affair up until the relation between data model and model of theory is introduced.

Once again, summarising the ills of empiricism Boon perceives:

“When taking experimental laws...to be data-models, this implies that no additional epistemic value is gained by theories over data-models, especially when data-models accurately represent large data-sets achieved by machine learning technologies. Hence...the epistemic value of theories is to adequately represent data-models, where ‘represent’ means ‘structural similarity,’ i.e. being (partially) isomorphic. In turn, data-models represent the measured data. If we assume that representational relationships in science are transitive, this implies that from an epistemological point of view empirically adequate theories do not add anything to empirically adequate data-models—as empirically adequate data-models already allow for adequate predictions of ‘real-world’ data, theories and models become unnecessary” (Boon, 2020, 57).

Boon is putting ‘real-world’ data in scarequotes, presumably, because she understand the idea to be fallacious; however, she utilises the phrase here be-
cause her argument against “strict empiricism” rests essentially on a conception of data as worldly. Data is not an abstract, conceptual instrument but a mere sampling of worldly structure. Boon goes on in this very chapter to expound 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 and data modelling 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 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 of course 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, once again, 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.

5 Conclusion

Philosophers of science—alongside scientists, engineers, journalists, and laypeople alike—have alleged that machine learning differs substantively from prevailing mathematical tools in science along several dimensions, and its widespread adoption by scientists is projected to radically alter the landscape of scientific practice or the epistemic products it is capable of lending us. These proclamations generally take the form of one of the following claims:

1. There is a radical qualitative difference in how machine learning models latch onto dependencies in the natural world relative to the operation of “classical” statistics or “classical” mathematical modelling in the sciences.

2. The successful production of scientific knowledge (or understanding, or explanation, or whatever we take the intended epistemic products of science to be) with ML does not require theorising or hypothesising.

3. ML-based science is a blind search through a space of associations; it is unprincipled, bottom-up, or ad hoc in some unprecedented way.

4. Successful science with ML does not require thought about how the way that we have gone about carving up the space of dataset features and learning objectives represents features of the target systems we are attempting to learn about, or the questions we are attempting to answer.

5. Successful science with ML does not require importation of domain knowledge/expertise or rest on previously established facts in the domain of inquiry.

6. Successful science with ML does not lend us causal knowledge, does not lend us mechanistic knowledge, and does not lend us knowledge about the data-generating process.

7. Machine learning is only capable of generating predictions but not of generating explanations.

8. Unlike “classical” statistics or “classical” mathematical modelling in the sciences, ML is “opaque,” “uninterpretable,” or a “black box,” in that we lack insight into the means by which the model produces a given output or the features of the data on which it is trained or deployed or the dependencies in nature it is picking up on to reach that output.

9. Because of any one of the above or some proper subset thereof, the rising adoption of ML across the sciences will entail a disruptive change in the methodological practices or epistemic aims and standards of science at large.
I take it that many, if not all, of these alleged differences between ML in science and classical scientific methods dissipate once we appropriately appreciate the theory-ladenness of data and the diversity of ways in which ML and mathematics generally gets applied in scientific practice. In closing, I will address these claims in order.

1. **ML exists on distinct epistemic footing.** I have attempted to demonstrate in this paper, through examples of applied mathematics in science and through engagement with existing philosophical literature on ML that mathematics (and ML) can serve a plethora of roles in scientific practice but that, when done right, all of it is capable of bearing the same epistemic fruits.

2. **ML does not require hypothesising or theorising.** Only certain branches of science in certain stages of their development have proceeded by hypothesis testing. Plenty of scientific practices fail to conform to the hypothetico-deductive model. What is more, and as I have attempted to demonstrate in section three of this text, the tools of machine learning can be plugged into an existing scientific procedure in all the same places that “traditional” mathematical or statistical modelling techniques can—inclusive of those involving traditional hypothesis testing. Thus failure to conform to the familiar hypothesis-generation-and-test model of science is not true of all applications of ML to science; what is more, not all existing scientific practices adhere to this model, so the lack of hypotheses cannot represent a break between traditional scientific practice and ML-assisted scientific practice. As to the involvement of theorising: in general, the aim of science is to generate knowledge or understanding or some other product that is necessarily useful and available to human epistemic agents (Potochnik, 2017). As such, any practice that can be properly termed “scientific” requires “theorising” at least in the minimal sense of conceptualisation of the phenomena under study. This holds true for any application of ML—conceived of as epistemic technology—as well. See (Boon, 2020) for a more detailed argument as to why human epistemic agents cannot be automated out of the scientific procedure. All informative uses of ML involve theorising, though theorising in ML does not always come into play in the places in which we are used to looking for it.

3. **ML-facilitated science, in contrast to “normal science,” is merely fishing for patterns in an ocean of data. It is bottom-up or unprincipled in some novel way.** Some approaches that we could imagine to doing “science with big data” certainly conform to such a picture. But so do many uses of classical statistical methods, especially in the social, behavioural, and psychological sciences. One could argue that such a blind and unprincipled search for significance is pseudoscientific. I am, myself, amenable to such a view; that scientific practices which involve throwing off-the-shelf statistics willy-nilly at data, culling for patterns, and slapping unmotivated, post-hoc interpretations thereon can little tell us about the world. This is
because such an endeavor constitutes poor statistical practice. We can certainly also engage in poor statistical practice in ML-facilitated science. This is neither a difference between ML and normal scientific practice (poor statistical practice exists in science generally, and has for nearly as long as scientists have had statistical tools at their disposal) nor does it encompass the totality of applications of ML to science—as I have sought to illustrate, there are plenty of more principled and epistemically well-founded ways to go about using ML in science.

4. Conducting science with the tools of ML obviates the need for critical thought about data provenance, data handling, how our data and models represent the world, and our epistemic aims. Such expressions have taken two forms. One of these has been the form of ML hype. The salesman of AI snake oil proclaims that while once carrying out science required a great deal of critical thinking and conceptual labour on the part of the researcher, the tools of ML free us up to explore the world unhindered. This supposed lack of critical thought going into the use of ML is also heralded as an epistemic virtue: certain proponents of ML for science conceive of it as freeing us from the unfortunate constraints and biases of human knowledge and conceptual grasp on the phenomena. Call this the “theory-free ideal” of science. The second form that this expression takes is one of doomsaying for the practice and products of science. If science no longer rests fundamentally on human input, will it go on producing knowledge or understanding that is usable or interpretable to us human cognitive agents? I will repeat here, as I have gone to some length to demonstrate over the course of this paper, that the data upon which a model is trained and deployed necessarily embodies human conceptualisation of the system from which the data is drawn and the phenomenon which that data facilitates our learning about—as does the use of such epistemic tools. We can always let theoretical commitments “in the back door” in this way, by relying on data of unknown and unquestioned provenance or by wielding instruments which we do not understand. This is common in scientific practice even without the use of ML, especially in less mature and more data-centric scientific paradigms. Campaigns of disinformation and hype surrounding ML do indeed threaten to make the phenomenon worse. Perpetuation of the idea that ML works differently from existing scientific or statistical techniques only stokes these flames. The pessimists about ML in science would better serve their own aims by refusing to buy into narratives of radical departure from existing techniques and instead focussing on adapting existing best practices in science and statistics to the ML era.

5. The use of ML in science does not rest on previously amassed knowledge in the field or the domain expertise of scientists. In our appraisal of the nature and epistemic role of data in scientific practice, we have established that data is not “raw,” “objective,” and “worldly,” but encodes commitments (in a multitude of ways and to varying degrees) to the nature of
the target phenomenon and the nature and aims of the investigative procedure. This constitutes both the practical knowledge of scientists about effective scientific practice and the empirical knowledge of scientists about their subject matter. The thesis is overdetermined when we consider the material theory of induction. Inductive inference is the procedure of gaining knowledge by extrapolating from a limited number of instances to a more general class. This is the nature of the learning in machine learning. According to Norton (2003), 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 (material) facts tied to a specific line of research (Norton, 2003). There is no one inductively valid formula to rule them all. The way in which the methods of ML and their epistemic aims are portrayed by those buying into and perpetuating ML “hype” is as though they seek to (and, indeed, will eventually) discover such a universally valid formal principle of inductive inference. Such a project is a doomed one. Fortunately for the field of ML, its own epistemological theory (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.

6. **ML does not deliver causal knowledge.** One can only securely recover causal relationships by means of an intervention or an observation of a “natural experiment” (J. Woodward, 2003). I take it that the idea behind the claim that ML-facilitated science will not furnish us causal knowledge rests on the idea that scientific methods involving ML will not involve intervention or experimentation but merely trawling through large volumes of pre-collected, even auto-collected data. Not all extant scientific practice or scientific modelling rests, however, on intervention or experimentation. Biological anthropologists study the fossil record—something we could not possibly hope to intervene on—for insights into human evolutionary history. Astronomers study patterns of light emanating from galaxies thousands of lightyears away for insight into spatially and temporally distant phenomena: the births and deaths of stars long ago in far-away galaxies. Further, we can, and often do, put ML to work in conjunction with experimental procedure—plenty of the data that we utilise in analysis with ML is of direct experimental provenance. Lastly, two of the essential sub-disciplines of machine learning are devoted to learning causal relations. All varieties of mathematical modelling, ML included, have the potential to reveal causal structure if wielded carefully and appropriately and used in conjunction with the right sort of data. All modelling also, in principle, has the ability to fall short of providing real causal knowledge, and even the capacity to be deceptive in how it represents potential causal relations. How best to uncover causal relations, achieve inductive robustness,
generalise to unseen instances, and how to avoid deceptive practices are the subject of ongoing work at the forefront of research in ML.

7. **ML cannot lend us explanations of phenomena.** This follows if one thinks that ML is not capable of lending us causal or mechanistic knowledge about target phenomena. But as I have said, ML is no worse poised to lend this sort of knowledge than any other statistical method employed in science. Machine learning has often been dismissed as “mere curve-fitting” or “mere prediction.” However, the interesting applications of ML in science, those which are creating a stir—here one need only look to AlphaFold—are clearly curve-fitting scaled up in such a way as to be highly theoretical and intended to be explanatory. The interesting questions about ML in science are not whether or not it can be explanatory but how it can be explanatory. To this, there will no doubt be many answers, for the methods of ML and their potential use cases in science are exceptionally diverse.

8. **ML is uninterpretable or opaque in some novel way.** As I have gestured at in my analysis of various case studies of the use of ML and of “classical” mathematical models in science, interpretability (inasmuch as we grant that this is a well-defined concept—we have plenty of reason to believe that it is not (Lipton, 2016)) is not a feature we can take for granted of the formal tools in existing scientific practice—even those formal tools most canonically “theoretical” in nature. What is more, the “uninterpretability” or “opacity” of ML (especially DL/DNN) models often inherits from the opacity or uninterpretability of the representations they are fed, or even of the very problem they are being wielded to solve or the phenomenon in nature they are being utilised to investigate, as revealed in an analysis of AlphaFold. A more thorough exposition of this feature of science conducted via ML—and its broader implications for science—will be the subject of a subsequent paper. That opacity is not universally true of ML-assisted science, nor a unique feature thereof, I think can readily be seen from the examples provided in section three.

9. **The widespread adoption of ML will entail radical change to scientific practice or the epistemic standing of science.** Science is changing. Science has been changing for centuries, sometimes slow and plodding, sometimes in starts and fits, more readily characterised by punctuated equilibria than by gradualism. There are many very real and substantive changes to scientific practice, to the social, institutional, state, and economic infrastructures that support it, and to the knowledge economies it results in occurring right now—and at a dizzying clip. 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 or formalisation. Reactions to the adoption of ML in science have largely framed ML as catalyst to these changes. I want to
counter that we ought to 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 from the automation of scientific practice to the adoption of the tools of ML in science, not the reverse.

In closing, I will linger a moment on this last point. There is a parallel to be drawn here between ML for science and ML ethics, in particular, criticism of ML tools as utilised in the judicial system. Many of the most left-leaning critics of the use of ML-based tools for crime prediction, for sentencing, and for bail-setting urge abstention from the use of such tools in contexts of socially-consequential decision-making. They study the tools exhaustively and thereby reach the conclusion that no form such a tool could take could be acceptably useful and ethical. But this is because the judicial system we have, the judicial system in which the use of such tools is ineluctably embedded, itself embodies conflicting aims and virtues.

ML-based tools—in governance as in science—are here, and are here to stay. Rather than asking whether or not we agree with the use of such tools, or fixating on which aspects of them are to blame for problems we foresee them bringing about, we should be asking what a vision of these tools being utilised to good effect would look like and, importantly, what a vision of the world and of science would look like in which these tools could in principle be wielded to good effect.

Most of the assessments of ML from philosophers of science I think amount to the observation that we might do shoddy science with ML—or that ML even increases the likelihood of doing shoddy science. I am here to say that we are already doing shoddy science—at least as shoddy in precisely the same ways (if marginally slower) as shoddy science might be done with the tools of ML. We are already rearing generations upon generations of scientists to solve nature’s puzzles by “throwing math at it” without teaching them to speak the language of math (or, for that matter, to speak the language of nature).

It is already the case that people who are not trained in the scientific method (e.g., government contractors and industry practitioners) are adopting tools from ML—much as they once adopted the tools of classical statistics—to carry out what is essentially social and behavioural science. I am speaking here of 22-year-olds fresh out computer science Bachelors who are suddenly tasked with undertaking consequential, large scale social-scientific research in virtue of having proficiency with ML-based research tools. Social and behavioural scientists, who are meticulously trained in the nuance of their subject matter and the complexity of the requisite research methods, already face difficulties in carrying out informative and epistemically well-situated research in their respective areas. When the work is left to individuals, collectives, or institutions who know little about the functioning of ML tools, little about the mathematical principles which undergird them, little about scientific best practice, and little about the target phenomena they are using ML tools to investigate, as one might reason-
ably expect, the results are bad—bad qua science, bad qua mathematics, and bad qua ethics.

As philosophers of science, we can go on touting the radical differences between ML and the established tools of science, further inflating the already overblown bubble of ML hype and lending credence to the unprincipled usage of ML. Or we could get serious and begin to grapple with how best to use these tools. Better yet, we can take a step back and take stock of the direction science is evolving in. We can get to work envisioning a system of science that is better positioned to deliver us secure knowledge about the natural world. Then we can begin to ask how machine learning fits into this picture.

References


