Deep Learning Applied to Scientific Discovery:
A Hot Interface with Philosophy of Science

06.07.2021

Louis Vervoort[(1)], Henry Shevlin[(2)], Alexey A. Melnikov[(3),(4)], Alexander Alodjants[(5)]

[(1)] *School of Advanced Studies, University of Tyumen, Russian Federation*
[(2)] *Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK*
[(3)] *Valiev Institute of Physics and Technology, Russian Academy of Sciences, Russian Federation*
[(4)] *Terra Quantum AG, Switzerland*
[(5)] *ITMO University, St-Petersburg, Russian Federation*

Abstract. We scrutinize publications in automated scientific discovery using deep learning, with the aim of shedding light on problems with strong connections to philosophy of science, of physics in particular. We show that core issues of philosophy of science, related, notably, to the nature of scientific theories; the nature of unification; and of causation loom large in scientific deep learning. Therefore advances in deep learning could, and ideally should, have impact on philosophy of science, and vice versa. We suggest lines of further research, and highlight the role 'theory-driven' AI could have in future developments of the field.

1. Introduction.

From one point of view, it is obvious that there must be a strong overlap between philosophy of science and artificial intelligence (AI) research. AI, as a branch of computer science, is in search of machine-based cognition, if possible high-grade cognition, for instance scientific cognition. In this respect, AI scientists could stand to benefit from a detailed understanding of what scientific knowledge is, which is a central question of philosophy of science. Some well-known computer scientists and AI researchers have indeed emphatically noted the key role philosophy of science could and should play for

making progress in their field (Deutsch 2012). Conversely, general AI has attracted the interest of philosophers of science for several decades. But deep learning, based on multi-layer artificial neural networks (ANN) and currently often considered the most powerful approach to AI, has only gotten into the focus of philosophers since recently (e.g. Buckner 2018, 2019, López-Rubio 2018, López-Rubio and Ratti 2019, Schubbach 2019). It is this sub-discipline of AI, more specifically a niche that we term 'automated (or machine-based) scientific discovery', which we will scrutinize in this article, in the realm of physics. Indeed, recently a variety of research efforts using deep neural networks have been reported that are ultimately geared at creating 'artificial scientists', more precisely at executing different tasks that are done by scientists, such as solving quantum many-body problems (Carleo and Troyer 2017), discovery of latent properties (Zheng et al. 2018), generation of quantum experiments (Melnikov et al. 2018), and many more (review texts are, for instance, Dunjko and Briegel 2017, Roscher et al. 2020).

Our main goal in this article is to show that, from the conceptual (or logical or strategic) point of view, some key issues and problems arising in deep learning research involve typical problems of philosophy of science, of physics in particular. We will focus on deep learning as applied to automated discovery in physics. Note it is not extravagant to assume that automated scientific discovery will be spearheading deep learning (and possibly AI) in general: the cognitive tasks that are performed in scientific discovery are arguably among the most sophisticated. If indeed philosophy of science and scientific deep learning are as strongly related as we believe, philosophers of science would be in the position to contribute to deep learning and AI by assessing or perhaps even guiding some of its strategic research projects. Conversely, if next-generation deep learning research will be successful in emulating, to some meaningful level, capacities of human scientists, it should be expected to have a significant impact on key assumptions and theories of philosophy of science – in the most optimistic scenario (i.e., if researchers will succeed in creating a genuine 'automated scientist'), it could profoundly change them, or settle long-standing debates. Our main goal is to identify some of those debates and present to the philosophy public some of the state-of-the-art research problems that physicists using deep learning encounter nowadays. We will of course not attempt here to solve these problems, but suggest some lines of further work. We emphasize that this article wishes in the first place to attract the attention of philosophers to a domain that we feel is still underexplored, and to identify important interdisciplinary problems, rather than to propose solutions.

Recently Buckner (2019) has made a similar call to investigate topics of philosophical interest in deep learning. This author highlights three[1] philosophical topics of interest, including the 'Explanation Problem' we will mention below (this reference also gives a solid introduction to the functioning of deep learning networks). The Explanation Problem is for obvious reasons most relevant for a branch of AI and deep learning called 'explainable AI' or XAI (Doshi-Velez and Kim 2017, Gilpin et al. 2018, Barredo Arrieta et al. 2019). Interest in XAI has soared in the last decade or so, with research programs expanding rapidly particularly in the last three years (Barredo Arrieta et al. 2019, p. 3). This interest arises because most contemporary AI approaches, especially deep learning methods that use many hidden neural layers, obtain results – typically, a classification – in a way that is not tractable in its details by humans. For instance, an advanced ANN may classify a single image using 10^{10} computations and $5x10^7$ learned parameters (Gilpin et al. 2018). Since philosophers have scrutinized the philosophical aspects of XAI in quite some detail (e.g. Lepri et al. 2018, Buckner 2019, Páez 2019, Sullivan 2019a, Zednik 2019), we will not focus on it here.

Potential interactions between deep learning and philosophy of science may occur across a wide array of topics, but we will focus here on three classes of issues that seem particularly urgent to us: we term them 'P-T', 'P-U', and 'P-C' in the list below. Each problem in the list actually represents a wide cluster of problems; we will highlight only the aspects that are most relevant for our argument:

> **P-T) or the 'Theory Problem'**: What are (scientific) theories? How does one discover/build theories? How do scientists go from inventing/discovering concepts (variables) to laws to theories?;
>
> **P-U) or the 'Unification Problem'**: In what precise sense do theories unify, and how is this unification realized?;
>
> **P-C) or the 'Causation Problem'**: What is the precise nature of causation/causality? What role does it play in scientific theories? How can one track causes?

As said, there is a fourth problem that plays an important role in particular in XAI, namely P-E or the 'Explanation Problem' (Lepri et al. 2018, Buckner 2019, Páez 2019, Sullivan 2019a, Zednik 2019). Let us define it here for completeness:

[1] The three questions are the following, in brief (Bruckner 2019). Does deep learning favor empiricism or rather nativism (a debate in philosophy of mind)? Do deep ANN learn the way humans do? What kind of explanation do such networks provide? We will highlight here three different questions or problem clusters in scientific AI used in physics.

P-E) or the 'Explanation Problem': What is the nature of (scientific) explanation and understanding?

These problems are interrelated, but it will make sense to disentangle them here, following a long tradition in philosophy of science.

The problems P-T, P-U, P-C, P-E are well-worn debates in philosophy of science and of physics, but many aspects of them are hotly debated today. Arguably the thorniest cluster of problems is P-T, the Theory Problem. In physics, it is often held that theories condense the essence of scientific knowledge[2]. The Theory Problem has a definitional component, as well as a component relating to theory genesis or theory discovery and the logic of scientific theory construction[3] (Popper 1959, Nagel 1979). In one sense, P-T is logically prior to the other clusters, since these all refer, explicitly or implicitly, to the concept of theory (or at least they can be so construed in physics). For the sake of definiteness, we will assume here from the start that 'theories' are hypothetic-deductive systems of propositions closed under deduction, ideally axiomatized, in any case based on a (small) number of base hypotheses, laws, principles, axioms (Popper 1959, Hempel 1965, Nagel 1979). Just a few examples from the last three years of the numerous articles studying aspects of P-T in different disciplines are (Bloch-Mullins 2018, Dorst 2019, Poth and Brössel 2020).

In the course of drawing comparisons between AI research and philosophy of science, we will also encounter the Unification Problem (P-U). Classic references on the merits of unification in science are (Kitcher 1989, Sober 2003). This topic is again intensively debated today in various disciplines (for a few recent examples, see Colombo and Hartmann 2017, Nathan 2017, Blanchard 2018). The third problem we will consider, the Causation Problem (P-C), has perhaps the longest history, and involves a cluster of issues concerning the concept of cause. Classic accounts of causation are the 'functional' account of Mach, Schlick and others (Schlick 1949), the counterfactual account of Lewis (1973), and the more recent manipulability account of Woodward (2003). Particularly relevant for our discussion are theories of probabilistic causation (Suppes 1970, Hitchcock 2018) and their link with causal modelling/tracking techniques (Spirtes et al. 2000, Pearl 2009): we will argue that these theories gain cogency in view of the deep learning results we review. Just a few examples of work from the last three years related to causation

[2] For simplicity, we will here include 'models' under the umbrella concept 'theory'. That is enough for a first exploration of the interface between automated discovery in physics and philosophy of physics.
[3] Few philosophers believe that the topic of theory discovery/construction is fully covered by logic alone.

in philosophy of science are (Landes et al. 2018, Eva et al. 2019, Sullivan 2019b, Livengood and Sytsma 2020).

This article is organized as follows. To highlight the relevance of the problems P-T, P-U, P-C, we present in some detail a small number of publications in automated scientific discovery in physics (Section 2). We believe that in these articles the problems at the intersection of deep learning and philosophy of science arise particularly decisively. Thus, in Section 2 we will scrutinize the recent work by Iten, Metger, Wilming, del Rio and Renner (2018, 2020), and by Wu and Tegmark (2019) (a fuller discussion of this work is given in Vervoort et al. 2021). Even just judged by the number of citations of the works of these groups, and by the interest they receive in mass-media, they belong to the most visible in their community. Our goal in this Section is to present the content of these articles to a general, non-expert philosophy of science public. (We will not go into technical details; for in-depth introductions to the formal aspects of AI, deep learning and ANN we refer e.g. to Russell and Norvig 2020, LeCun et al. 2015, Silver et al. 2016.)

Having outlined these projects, we make in Section 3 our central claims, expounding in detail how P-T, P-U, and P-C arise in the mentioned works. We will also discuss in this Section some hints for promising lines of future research, also of interdisciplinary research. We will touch on the question of whether scientific deep learning favours some accounts of causation over others. We will analyze in some detail how far state-of-the-art deep learning comes in theory discovery, and offer our verdict. More speculatively, we will suggest that further research on the role that abstraction mechanisms play in theory building might have a pivotal significance for the development of deep learning (we term this the 'theory-heavy' conjecture). In Section 4, we will briefly contextualise our main claim (as to the relevance of P-T, P-U, and P-C) within contemporary research in cognitive science and philosophy of mind, notably by pointing to work by Lake, Ullman, Tenenbaum and Gershman (2017). These cognitive scientists have claimed that theory mastery is a central goal for future AI, and we will offer in Section 3 an independent argument for this idea from the point of view of philosophy of physics. Combining insights from Section 3 and the work by Lake et al., we will conjecture that a desirable, but perhaps elusive, paradigm change in AI would be that from 'data-driven' AI to 'theory-driven' or 'theory-based' AI. Clearly, Section 4 is of a more speculative nature; our main theses are presented in Section 3.

2. Automated scientific discovery. Iten et al. (2018, 2020), and Wu and Tegmark (2019).

In this Section we describe a few recent highlights in deep learning research in physics. For technical details, we refer to the original articles.

2.1. Iten et al. (2018, 2020).

As a first example, Iten et al. published in January 2020 an article, entitled "Discovering physical concepts with neural networks", which received wide attention in specialized and broad-public journals (Iten et al. 2020; a more complete version of this article is Iten et al. 2018). The authors begin by asking whether "the laws of quantum physics, and other physical theories more generally, [are] the most natural ones to explain data from experiments if we assume no prior knowledge of physics" (Iten et al. 2018, p. 1). Though the authors acknowledge that this question is unlikely to be answered in the near future, they attempt to make initial steps in this direction by exploring whether neural networks are capable of independently discovering key concepts in classical and quantum mechanics via exposure to experimental data.

The authors' original approach is to try to model the human scientific reasoning process. They claim to achieve this goal by implementing a neural network, called *SciNet*, that mimics a physicist answering questions about the future behaviour of a system. (Systems are typically encoded by a series of experimental trajectories: for instance a time series $\{t_i, x(t_i)\}$, i = 1, 2, …N, describing the motion of a particle.) In other words, *SciNet* aims at mimicking a physicist who deduces from observations the corresponding equation, e.g. in a simple case of constant motion at speed v: $x(t) = x_0 + vt$. Indeed, a physicist knows that the variables fully describing this system are position x, time t, initial position x_0 and constant speed v; and she knows the lawful relation linking these variables. Knowing x_0 and v, she can predict the position at any later time. Now, prediction can be seen as answering questions about the future: *SciNet* mimics this process of asking/responding to questions. The functioning of *SciNet* is schematized in Fig. 1, reproduced from Iten et al. (2018) (we have also reproduced the legend of the figure).

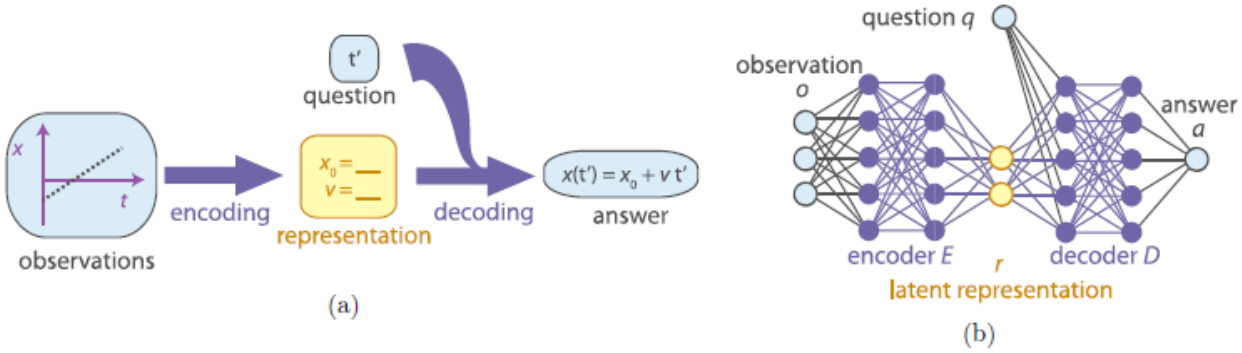

FIG. 1. Learning physical representations. (a) Human learning. A physicist compresses experimental observations into a simple representation (encoding). When later asked any question about the physical setting, the physicist should be able to produce a correct answer using only the representation and not the original data. We call the process of producing the answer from the representation "decoding." For example, the observations may be the first few seconds of the trajectory of a particle moving with constant speed; the representation could be the parameters "speed v" and "initial position x_0" and the question could be "where will the particle be at a later time t_0?" (b) Neural network structure for SciNet. Observations are encoded as real parameters fed to an encoder (a feed-forward neural network, see […]), which compresses the data into a representation (latent representation). The question is also encoded in a number of real parameters, which, together with the representation, are fed to the decoder network to produce an answer. (The number of neurons depicted is not representative.)

Fig. 1. From Iten et al. (2018), open-access.

Formally, this physical modeling process can be understood as a so-called 'encoder' function E: O → R, which maps the set of possible observations O to representations R, followed by a 'decoder' function D: R × Q → A, which maps the sets of all possible representations R and questions Q to answers A. In the spirit of ANN, the functions E and D are realized by neural nets, cf. Fig. 1, in this case two classic feed-forward neural networks. A key ingredient is the middle layer of neurons implementing the 'latent representation' (R): the output values of these neurons should represent *a simpler representation of the input data* (a human physicist would take x_0 and v in the above example), allowing one to answer questions about the future. In *SciNet* a so-called 'disentangling variational autoencoder' (β-VAE, see e.g. Kingma and Welling 2013) is used, a type of network that has proven to be capable, in recent years, to extract condensed representations from large data sets, notably variables that are *stochastically independent*; so the latter are encoded in the neurons of the representation layer. (The authors admit however that finding such minimal uncorrelated representations in a reliable way is still a problem of present research; usually results improve when more latent neurons are used.)

One result of (Iten et al. 2018, 2020) is the simulation of the one-dimensional damped pendulum with equation of motion: $m\ddot{x} = -kx - b\dot{x}$, which has the solution: $x(t) = A_0\, e^{-bt/2m} \cos(\omega t)$. Here k is the

spring constant and *b* the damping factor (ω is a function of these parameters). *SciNet* is given three latent neurons in this case and is trained with 50 correct positions $x(t_i)$ with equally spaced t_i (amplitude A_0 and mass *m* are fixed; *k* and *b* are varied between training samples). It learns correctly to answer the question "what will be the position at time *t*", with a root mean square error below 2% of the amplitude. Thanks to the β-VAE technique, *SciNet* stores *k* and *b* in two of the latent neurons, and does not store information in the third latent neuron. In other words, *SciNet* can identify the two parameters that are also used by real physicists, and recognizes that these are sufficient for its task. The authors classify this result as a form of so-called 'unsupervised learning', because *SciNet* extracts the independent variables *k* and *b* itself, without human labelling.

Further achievements of *SciNet* are, notably, 1) that it can determine the dimension of an unknown quantum system and decide whether a set of measurements is tomographically complete, and 2) that it can extract from geocentric planetary coordinates the heliocentric representation that a modern human scientist would also use.

Clearly these achievements are impressive, but to what extent do they draw upon or make assumptions about the nature of theories, unification, and causation? We will come back to this question in Section 3, where we will relate the above results to the problems P-T, P-U, and P-C.

2.2. Wu and Tegmark (2019).

Just months before the article of Iten et al., in September 2019, an article by Wu and Tegmark (2019) was published, with the title: "Toward an AI Physicist for Unsupervised Learning". Wu and Tegmark set themselves the task to program an ANN named *AI Physicist* that can, to some degree, analyze and interpret physical data in the same way as a human physicist, notably by extracting laws from given data and applying these to new data to predict the evolution of a system. It is worth quoting the authors' goals for the project at length (Wu and Tegmark 2019, p. 1):

> "We investigate opportunities and challenges for improving unsupervised machine learning using four common strategies with a long history in physics: divide-and-conquer, Occam's razor, unification and lifelong learning. Instead of using one model to learn everything, we propose a novel paradigm centered around the learning and manipulation of *theories*, which parsimoniously predict both aspects of the future (from past observations) and the domain in which these predictions are accurate. Specifically, we propose a novel generalized-mean-loss to encourage each theory to specialize in its comparatively advantageous domain, and a differentiable description length objective to downweight bad data and 'snap' learned theories into simple symbolic formulas. Theories are stored in a 'theory hub', which continuously unifies learned theories and can propose

> theories when encountering new environments. We test our implementation, the toy 'AI Physicist' learning agent, on a suite of increasingly complex physics environments."

Particularly relevant for the present purposes is the authors' intention to use real world strategies for scientific discovery, having "a long history in physics." The project again involves great know-how from computer science, but we will mostly set aside these technical aspects here and focus instead on what seems to us to be the findings that are most important from a conceptual point of view.

In slightly more detail, then, *AI Physicist* learns to recognize trajectories that are governed by (simple) equations describing a particle in a 2-dimensional potential energy[4] $V \sim (ax + by + c)^n$ for some constants a, b, c, with n = 0 (no force), n = 1 (uniform electric or gravitational field), n = 2 (spring obeying Hooke's law) or n = ∞ (ideal elastic bounce). Contrary to preceding work, this ANN is not specialized to highly accurately describe only one physical system (corresponding for instance to one type of potential with one fixed coefficient $n \in \{0, 1, 2, \infty\}$); it is trained to recognize and predict the evolution of several 'environments' each with their respective coefficient n (or superpositions of such environments, i.e. potentials). It is even trained to extract the *symbolic laws* (the law statements we would find in a physics book) out of data sets $\{x(t_i), y(t_i)\}$ describing these systems, and to recognize when these laws have simple forms with integer or rational coefficients – as physical laws for these environments indeed have. This is the 'Occam's Razor' strategy: the search for laws that have simple coefficients.

In practice, *AI Physicist* is trained, for each environment n, by feeding it with data sets generated with coefficients a, b, c that are randomly picked in intervals; therefore the symbolic laws for (x, y) – here difference equations – have coefficients that lie in intervals. For each environment, *AI Physicist* learns to parametrize these varying numerical coefficients, i.e. to recognize a given law statement as an element of a *family of laws* (a 'master theory') characterized by a continuous parameter[5], say p. In this manner it can store a parameter-family of laws in the 'theory hub' and use this family to recognize/describe new unknown data more efficiently. The parametrization step is called 'unification' (in a master theory) by Wu and Tegmark. Using an analogy to make the point, the authors note that "Newton's law of gravitation can be viewed as a master theory unifying the gravitational force formulas around different planets by introducing a parameter p corresponding to planet mass" (Wu and Tegmark 2019, p. 4).

[4] Or combinations of these potentials.

[5] For instance, various trajectories of a particle in a gravitational field of a large mass, can be parametrized by a parameter p = m = the large mass.

The general result is that *AI Physicist* learns faster with mean-squared prediction errors orders of magnitude smaller (until 10^6 times) than a standard feedforward neural net of comparable complexity; at the same time it can often retrieve integer and rational theory parameters exactly. As said, according to the authors the ANN achieves this by mimicking four strategies used also by seasoned real physicists, defined by them as follows (cf. Table 1, Wu and Tegmark 2019):

- Divide-and-conquer: Learn multiple theories each of which specializes to part of the data very well.
- Occam's Razor: Avoid overfitting by minimizing description length, which can include replacing fitted constants by simple integers or fractions.
- Unification: Try unifying learned theories by introducing parameters.
- Lifelong Learning: Remember learned solutions and try them on future problems.

The authors believe that these strategies could provide the basis for a paradigm that could work for a wide range of other scientific problems, enabling the discovery of accurate symbolic expressions for more complicated physical systems too.

But to what extent do these aspirations connect to debates in philosophy of science? It is to this question that we now turn.

3. Relevance of philosophy of science. Potential for cross-fertilization.

Let us now scrutinize if and how P-T, P-U, P-C arise in the research presented above. The articles summarized in the preceding Section make clear that some of the most advanced research efforts in deep learning aim at constructing ANNs that can learn, discover, and master (use) concepts, laws, and ultimately theories – in the cases studied, *physical* concepts, laws, and theories. This is evident in (Iten et al. 2018, 2020): here the networks are trained to discover concepts and numerical laws, and use these to answer questions about the future evolution of specific systems. This is also clear in (Wu and Tegmark 2019, recall the quote above). These authors end their article thus: "We hope that building on the ideas of this paper may one day enable AI to help us discover entirely novel physical theories from data" (p. 10). Thus, in a slogan, the goal of both (Iten et al. 2018, 2020) and (Wu and Tegmark 2019) is *unsupervised theory discovery* (and theory mastery) starting from experimental data. Note that the ANNs of these articles are trained to also *use* the discovered laws to predict the future evolution of systems – we use the term 'theory mastery' in this sense. As further argued in Section 4, it is a natural assumption that this

might count as an *ideal but possibly elusive* goal for future deep learning (as applied to physics problems): this is what we termed in Section 1 our theory-heavy conjecture.

Indeed, it has been a recurrent theme among philosophers of science that theory discovery/mastery is a central pillar of scientific knowledge (Popper 1959, Hempel 1965, Nagel 1979). Especially in physics, some authors do not hesitate to call theory construction/mastery the apex of scientific knowledge[6]. *But how close do these ANN really come to discovering and using full-blown theories?* What, from an epistemic or conceptual point of view, are the main strategic hurdles to overcome in order to bring this type of AI research to its next level?

We can of course not give a definite answer to this highly complex question, but we can at least make a start with scrutinizing the problem in more detail; this brings us on the terrain of arguably the hardest of our classic philosophical problems, P-T. One could, obviously, start by noting that the ANN can only learn to recognize and use certain numerically specified laws *after they have been massively trained by a human* to recognize and use precisely these numerically specified laws. That in itself seems not the key limitation: after all, human scientists also learn first from others before becoming creative (an analysis of the philosophical relevance of this point is given by Buckner 2019, Section 4.2). The essential limitation of the present ANN is that they can handle only *laws*, not genuine *theories* in the sense of philosophy of science. It is one thing to recognize that a given numerical environment can be described e.g. by the law of the damped pendulum or of planetary motion (as *SciNet* can) or of a particle in an electromagnetic field or Hooke's field (as *AI Physicist* can); it is a wholly different matter to understand that all these laws resort under Newton's theory of mechanics and/or Maxwell's theory of electromagnetism. *AI Physicist* seems, prima facie, to go further, since it can recognize *sets* of (numerically specified) laws that differ only in the value of a certain parameter. For example, it can recognize trajectories described by the second law of Newton *as applied to the special case of gravity, and for a specific range of values for the gravitational constant*. Such a parameter-family F of laws, in this case $F = \{(\ddot{x} = g_x, \ddot{y} = g_y)\}$ with parameters (g_x, g_y), is termed 'master theory' by Wu and Tegmark. But F is certainly not a theory in the sense of a full-blown hypothetic-deductive system as Newton's theory, based on symbolic higher-order laws (Nagel 1979).

[6] Physics can, it seems, be considered an idealized element of the set of natural sciences (surely because it deals with inanimate, relatively simple objects, and possesses, concomitantly, relatively mature theories). In other disciplines, the situation is not as clear-cut.

From a conceptual and computational point of view, the difference in the level of abstraction between laws and theories is tremendous. For constructing theories, the needed abstraction (or 'unification' as it is called by Wu and Tegmark, and in physics in general) amounts, notably, to recognizing that a potentially infinite set of particular laws, having different formal descriptions or law statements[7], can be understood as special cases of one master theory. But it appears that research on this cognitive capacity is in its infancy.

Indeed, a broader literature search shows that no concrete ideas seem to exist on how to formalize the crucial abstraction/unification step from specific laws to a genuine theory. In other words, the 'discovery' part of P-T remains largely an open question. Thus, the hardest of our philosophical problems appears to be, in hindsight perhaps not surprisingly, at the core of progress in scientific deep learning, at least in physics. This suggests a first possible avenue for research in philosophy of science, albeit a particularly challenging one: to investigate which steps in abstraction/unification are required to go from one level to the other. This (ideally interdisciplinary) research program would require engagement with P-T but also P-U, on unification, as recalled below. From here, the next program might involve the formal sciences and aim at formalizing and coding this gradual process of abstraction. For results in computer science that could indicate first steps in this direction, see e.g., besides the mentioned articles, Battaglia et al. (2018) and Khalili (2020).

In considering the prospects of such a project, it is interesting to dwell a little more on the question of how scientific theories are discovered by humans. Is there a recipe for theory discovery, a theory for theory construction? The history of science and of philosophy of science teaches us that, with high likelihood, the answer is 'no'. As is well known, in all traditional disciplines, say physics, chemistry and biology, the discovery of new high-grade scientific theories is a very rare event. What might be a guiding idea for further research, is the observation that in physics the highest level of theorizing often proceeds via *comparing existing theories*, and applying some *meta-laws* to these theories in order to *unify* them – and thus to come to a higher level of abstraction and a wider domain of application. These meta-laws are for instance invariance principles. A salient example is Einstein's construction of special relativity theory, which he derived using only two invariance principles: the light postulate (stating that the speed of light is identical for all inertial reference frames and independent of the speed of the source), and the principle of relativity (stating that the laws of physics are identical in all inertial reference frames). Note that in

[7] Note that each of such law statements, say Hooke's law, is de-multiplied in a nondenumerably infinite set of particular laws by the different values that its parameters, here mass m and spring constant k, can assume.

deriving his theory, Einstein compared classical mechanics and electromagnetism, realizing that only the latter complies with the light postulate. General relativity crucially builds on the principle of general covariance, allowing the unification of inertial and accelerated motion. And modern quantum field theories are governed by a growing set of fundamental symmetry/invariance principles (such as the permutation, CPT, and gauge symmetries, cf. e.g. Brading, Castellani and Teh 2017).

Thus, it appears that the cluster of topics concerning P-U and unification in science is of immediate relevance for scientific AI (a reference work is Kitcher 1989; for a discussion of different types of unification, see Morrison 2000). Wu and Tegmark explicitly state that they are inspired by the 'unification strategy' scientists use (cf. previous Section). At the same time it is clear that the unification realized by high-level scientific theory building is of a fundamentally different order than the unification realized in (Wu and Tegmark 2019). Arguably, only the former type of unification or abstraction or generalization is tantamount to a qualitatively higher-level understanding of the world[8]. Thus, a series of philosophical questions arise in applying our understanding of unification to this field of AI; answering these may have a bearing on this research. For instance: does theory building always proceed by applying meta-laws (unifying principles) to a set of existing theories? If so, of which precise nature are they? Should future neural nets (in the niche considered) include not so much a 'theory hub' than a 'principle hub'? Interestingly, this possibility has actually been proposed by computer scientists (Khalili 2020).

As a next AI-related issue with an obvious philosophical dimension, let us mention the VAE technique used in (Iten et al. 2018, 2019). We saw above that, in the case of the damped pendulum, *SciNet* can store k and b in the latent nodes (and then use these to answer questions about future behaviour); and that it needs to be extensively trained through similar enough trajectories with stochastically varying k and b. We noted that this is still far from recognizing the *law statement* describing the fed-in data (i.e., $m\ddot{x} = -kx - b\dot{x}$), let alone from deriving Newton's second law, which is one more step higher in abstraction. Still, these results are interesting, also from the point of view of philosophy of science, in particular the P-C cluster. Indeed, as we saw above, a key calculation ingredient is the disentangling variational autoencoder (β-VAE, Kingma and Welling 2013) that can identify variables (encoded in the neurons of the representation layer) that are *stochastically independent,* in this case k and b. Philosophy of science has, since a long time, recognized that stochastic independence is a signature of *causal* independence; probabilistic (in)dependence provides information about the causal connections between variables.

[8] Only the former type involves the construction of more general theories, while Wu and Tegmark's unification involves the construction of more general laws, which is only the first step in theory-building.

Indeed, standard works on causal modeling have shown that the key heuristic tool for hunting for causes is the concept of correlation, or if one prefers Reichenbach's principle (Spirtes et al. 2000, Pearl 2009). Concomitantly, philosophers have made a convincing case that causes can be formalized as conditional variables in conditional probabilities (Suppes 1970, Hitchcock 2018). Then, if ANN can automatically discover causally related and unrelated variables, they can indeed discover the essential relata (variables) in causal relations; and philosophy of science teaches us that causal relations are key ingredients of scientific theories.

To make the discussion more concrete, *SciNet* can identify, via the β-VAE technique, the parameters k and b, which could be seen as the essential causes of the damped pendulum system. This claim can readily be brought into harmony with a typical counterfactual construal of the concept of cause: the actual position x of the pendulum at time t would be different if k or b would be different; in this sense it is legitimate to call k and b 'causes' of the position[9]. So one may conjecture that the β-VAE technique could be a promising automated tool for discovering the causal variables in a system, a strategic ingredient of theory building. Of course, this ingredient is not enough: the equations still have to be formed, which might be the harder part.

We submit that there are interesting and essential questions lingering here for philosophers of science/physics, for instance those wishing to delve into the technicities of the VAE technique. E.g.: how universal is the VAE technique for causal discovery? Next, the nature of causation is a much debated topic in philosophy of science. We surmise that the importance of this technique in scientific deep learning is actually an argument in favour of the account of authors such as Suppes (1970) and Hitchcock (2018). But this matter should surely be investigated in a more systematic way.

These are the issues and questions raised by Iten et al. (2018, 2020) and Wu and Tegmark (2019) that connect, in our view, most interestingly to philosophical queries, in particular the problems P-T, P-U, P-C. In the following Section we succinctly relate some of these findings to cognitive science and philosophy of mind, to place them in a broader context.

4. Discussion. Link with Cognitive Science and Philosophy of Mind.

[9] More precisely, on token level, the event that the position x assumes a value X at time t = T is caused, among other events, by the event that k assumes the value K and b the value B. The causal relation is more simply interpreted or expressed here on a type level: x is 'caused' by (its values depend on) k and b (as well as on the mass of the pendulum and its initial amplitude, itself typically caused by an external force).

In the previous Section we suggested that among the clusters P-T, P-E, P-U, P-C, it is theory construction, P-T, that is arguably the hardest. We highlighted that this cluster arises in a pivotal way in the examples of deep learning in physics we studied. Interestingly, the key role of theory mastery for ANN-based cognition has actually already been highlighted by research in cognitive (neuro)science, and one can even find converging ideas in philosophy of mind. Both these disciplines are interested in natural but also artificial cognition and consciousness, and in robots (using deep learning/AI) that could emulate or even surpass human intelligence in some tasks. Let us start with cognitive science.

In an influential review article by Lake, Ullman, Tenenbaum, and Gershman (2017), entitled "Building machines that learn and think like people", the authors claim (p. 1):

> "Despite their biological inspiration and performance achievements, these [AI] systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn, and how they learn it. Specifically, we argue that these machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned […]".

The authors provide convincing results and interpretations from their discipline for the idea that, indeed, high-grade cognition, both of humans and machines, necessarily involves mastering of models and theories. For an all-round AI-system to be able to 'think like a human', it should, according to the authors, start with 'understanding' basic physical concepts and models of the environment and of human interaction and psychology. In short, it should possess, among others, a rudimentary 'theory of physics' – and it is this type of theory mastering, the authors suggest, that future ANN should learn in order to approach or surpass human cognition. This resonates with our claim that research related to the P-T cluster in philosophy of science could be relevant for indicating how machines can imitate human learning[10].

In philosophy of mind, too, there is research on artificial consciousness that points in the same direction. Several philosophers have argued that consciousness is strongly related to forms of cognition relying on abstraction and model-building (Shoemaker 1994, Shevlin 2020, Vervoort et al. 2021). It is widely recognized that mastering models/theories of (aspects of) the world, is the road to counterfactual deliberation, the capacity to envisage goal-oriented alternatives of action[11], key to free-willed agency and

[10] Said in passing, the research in cognitive science reviewed by Lake et al. (2017) is just rife with concepts and problems directly stemming from philosophy of science.

[11] To illustrate this in a science context: an agent who masters (at a reasonable level) a model/theory of the functioning of say an 'electromagnetic device' as a radio, can ponder rationally over (counterfactual) options as to how to repair the apparatus. A

consciousness (Vervoort and Blusiewicz 2020). Thus, while consciousness science is still a vexed and complex field to say the least, we nonetheless believe there is good reason to think that 'rational' consciousness – natural or artificial – is associated with capacities for flexible behaviour and abstract generalisation, capacities largely absent in current AI systems. Continuing this speculative note, one may argue that a key step in addressing this deficit will be building systems that do not merely apply shallow statistical heuristics to data but can learn and apply more generalizable conceptual frameworks to their environments.

Turning now somewhat from the immediate aim of this article, we wish to close by asking whether the claims presented thus far, stemming from AI, philosophy and cognitive science, can provide any broader insights regarding the future evolution of AI. We would indeed suggest that a general trend can be distilled from the above interdisciplinary analysis. In the language of philosophy of science, we would assert that AI with high-grade human-like intelligence cannot simply remain 'data-driven', as is now the case: it should become 'theory-driven' or 'theory-based' (while still analyzing data, of course). This is a direct extrapolation from the theory-heavy conjecture we argued for in the preceding Section. This idea comes very close to claims from (Lake et al. 2017), for instance when these authors assess the future prospects of classic, purely data-driven AI, also termed the 'purely predictive' or 'cues-all-the-way-down' approach by these authors[12]. Indeed, Lake et al. negatively assess the potential of pure-data AI, both for AI capable of physical reasoning and psychological reasoning (p. 12):

> "[I]t seems to us that any full formal account of intuitive psychological reasoning needs to include representations of agency, goals, efficiency, and reciprocal relations. As with objects and forces, it is unclear whether a complete representation of these concepts (agents, goals, etc.) could emerge from deep neural networks trained in a purely predictive capacity. Similar to the intuitive physics domain, it is possible that with a tremendous number of training trajectories in a variety of scenarios, deep learning techniques could approximate the reasoning found in infancy even without learning anything about goal-directed or social-directed behavior more generally. But this is also [just as in intuitive physics] unlikely to resemble how humans learn, understand, and apply intuitive psychology […]".

person who does not possess model/theory-based knowledge, has not the same chances of rational and efficient conduct. This idea can be generalized to more mundane contexts.

[12] Lake et al. provide several examples of recent methods that could perhaps be classified in-between what we call the 'data-driven' and 'theory-driven' approach, such as 'generative models of action choice', as in the Bayesian inverse planning (or 'Bayesian theory-of-mind') models of Baker, Saxe, and Tenenbaum (2009) or the 'naive utility calculus' models of Jara-Ettinger, Gweon, Tenenbaum, and Schulz (2015). Actually, a wide variety of other formal approaches could be seen as first steps to what we term 'theory-driven' AI; other references can be found in Lake et al. (2017).

This coheres well, it seems, with the idea that high-grade cognition, artificial or natural, must involve the use of axiological, social, psychological and other theories, by which agents (inevitably) must guide their (natural or artificial) life[13].

5. Conclusion.

The main goal of this article was to present to the philosophy of science public a few articles of a niche of deep learning, automated discovery in physics, and, especially, to identify three problem-clusters with strong connections to philosophy of science. These clusters involve problems related to theory discovery and mastery (P-T), unification (P-U), and causation/causal modeling (P-C). To the best of our knowledge, this link between the two disciplines along the three mentioned axes has not yet been reported. Since these problems are of pressing importance both in automated scientific discovery and in philosophy of science, progress in any of the two disciplines could, or should, have impact on the other one. Thus, this article is a call to philosophers to have a close look at this rich playground for research.

Besides this main goal, we also suggested some further lines of research. We noted that some of the reviewed AI research identifies independent physical variables as causes in a way that is in line with the account of causation of authors as Suppes (1970) and Hitchcock (2018), thus giving support to this theory. We analyzed in some detail how far state-of-the-art deep learning in physics comes in theory discovery, and concluded that much work remains to be done here. As a consequence, we conjectured that research on the role that abstraction mechanisms play in theory building might be important for the development of (scientific) deep learning. Finally, we highlighted the coherence of this latter claim with findings by Lake et al. (2017) in cognitive science. Based on this interdisciplinary comparison, we were in the position to speculate about a general trend in future deep learning research: it seems to us that the greatest leap in AI would be from purely data-driven to theory-based AI.

Clearly, the latter claims need to be substantiated by much more research. Whether deep learning systems and machines in general will ever be capable of the formidable cognitive achievement of discovering and/or using sophisticated conceptual frameworks, is an open question. And yet, in view of

[13] Indeed, we cannot elaborate this claim here, but the "representations of agency, goals, efficiency, and reciprocal relations" that Lake et al. mention can well be construed, we believe, essentially as hypotheses, rules or theories of axiology, of psychology, of sociology etc.

the tremendous manpower involved in AI research, one is tempted to believe that progress in that direction is inevitable. That is one more reason for philosophers to keep a close eye on this field of study.

References.

Baker, C. L., Saxe, R., Tenenbaum, J. B. (2009). "Action understanding as inverse planning", Cognition 113:3, 329-349.

Barredo Arrieta, A. et al. (2019). "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI", arXiv preprint arXiv:1910.10045.

Battaglia, P. W., Hamrick, J. B., Bapst, V., et al. (2018). "Relational inductive biases, deep learning, and graph networks", ArXiv Preprint ArXiv:1806.01261.

Blanchard, T. (2018). "Bayesianism and Explanatory Unification: A Compatibilist Account," Philosophy of Science 85:4, 682-703.

Bloch-Mullins, C. (2018). "Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts", British Journal for the Philosophy of Science 69:3, 605–632.

Buckner, C. (2018). "Empiricism without magic: Transformational abstraction in deep convolutional neural networks", Synthese, 195:12, 5339–5372.

Buckner, C. (2019). "Deep learning: A philosophical introduction", Philosophy Compass 14:10, e12625.

Brading, K., Castellani, E. and Teh, N. "Symmetry and Symmetry Breaking", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/symmetry-breaking/>.

Colombo, M., Hartmann S. (2017). "Bayesian Cognitive Science, Unification, and Explanation", British Journal for the Philosophy of Science 68:2, 451–484.

Daniels, B. C., Nemenman, I. (2015). "Automated adaptive inference of phenomenological dynamical models", Nature Communications 6, 8133.

Deutsch, D. (2012). "Philosophy will be the key that unlocks artificial intelligence", The Guardian (03.10.2012), retrieved (30.12.20) from https://www.theguardian.com/science/2012/oct/03/philosophy-artificial-intelligence?CMP=share_btn_tw

Dorst, C. (2019). "Towards a Best Predictive System Account of Laws of Nature", British Journal for the Philosophy of Science 70:3, 877–900.

Dunjko, V., Briegel, H. J. (2017). "Machine learning and artificial intelligence in the quantum domain" arXiv preprint arXiv:1709.02779.

Eva, B., Stern, R., Hartmann S. (2019). "The Similarity of Causal Structure", Philosophy of Science, 86:5, 821-835

Hempel, C. (1965). *Aspects of scientific explanation*, New-York: Free Press.

Hitchcock, C. (2018). "Probabilistic Causation", The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/>.

Iten, R., T. Metger, H. Wilming, L. del Rio, R. Renner (2018). "Discovering Physical Concepts with Neural Networks", arXiv:1807.10300v2.

Iten, R., T. Metger, H. Wilming, L. del Rio, R. Renner (2020). "Discovering Physical Concepts with Neural Networks", Physical Review Letters 124, 010508, 1-5.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., Schulz, L. E. (2015). "Children's understanding of the costs and rewards underlying rational action", Cognition 140, 14-23.

Khalili, A. (2020). "Artificial General Intelligence: A New Perspective, with Application to Scientific Discovery", engrXiv preprint, https://engrxiv.org/duz8g/

Kingma, D., Welling M. (2013). "Autoencoding Variational Bayes", arXiv:1312.6114.

Kitcher, P. (1989). "Explanatory Unification and the Causal Structure of the World", in *Scientific Explanation*, P. Kitcher and W. Salmon, 410–505. Minneapolis: University of Minnesota Press.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J. (2017). "Building machines that learn and think like people", Behavioural and Brain Sciences 40, e253 1-72.

Landes, J., Osimani, B., Poellinger, R. (2018). "Epistemology of causal inference in pharmacology", European Journal Philosophy Science 8, 3-49.

LeCun, Y., Bengio, Y., Hinton, G. (2015). "Deep learning", Nature 521, 436-444.

Lepri, B., Oliver, N., Letouzé, E. et al. (2018). "Fair, Transparent, and Accountable Algorithmic Decision-making Processes", Philosophy & Technology 31, 611–627.

Lewis, D. (1973). "Causation", Journal of Philosophy 70, 556–67.

Livengood, J., Sytsma, J. (2020). "Actual Causation and Compositionality", Philosophy of Science 87:1, 43-69

López-Rubio, E. (2018). "Computational functionalism for the deep learning era", Minds and Machines 28:4, 667–688.

López-Rubio, E., Ratti, E. (2019). "Data science and molecular biology: prediction and mechanistic explanation", Synthese, First Online, https://doi.org/10.1007/s11229-019-02271-0

Melnikov, A. A., Nautrup, H. P., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., Briegel H.J. (2018). "Active learning machine learns to create new quantum experiments", Proceedings of the National Academy of Sciences 115:6, 1221.

Morrison, M. (2000). *Unifying Scientific Theories*. Cambridge: Cambridge University Press.

Nagel, E. (1979). *The Structure of Science*. Indianapolis, Hackett Publishing Co.

Nathan, M. (2017). "Unificatory Explanation", British Journal for the Philosophy of Science 68:1, 163–186.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines, 29(3), 441-459

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, Cambridge.

Popper, K. (1959). *The Logic of Scientific Discovery*, London: Hutchinson.

Poth, N., Brössel, P. (2020). "Learning Concepts: A Learning-Theoretic Solution to the Complex-First Paradox," Philosophy of Science 87:1, 135-151.

Roscher, R., Bohn, B., Duarte, M. F., Garcke, J. (2020), "Explainable Machine Learning for Scientific Insights and Discoveries," in IEEE Access, vol. 8, pp. 42200-42216, and arXiv preprint arXiv: 1905.08883.

Russell, S., Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*, Hoboken, NJ: Pearson.

Schlick, M. (1949). "Causality in Everyday Life and in Recent Science", in Feigl, H. and Sellars, W., Ed., *Readings in Philosophical Analysis*, New York: Appleton.

Schubbach, A. (2019). "Judging machines: Philosophical aspects of deep learning", Synthese, First Online, https://doi.org/10.1007/s11229-019-02167-z

Shevlin, H. (2020, forthcoming). "General intelligence: an ecumenical heuristic for artificial consciousness research?", Journal of Artificial Intelligence & Consciousness.

Shoemaker, S. (1994). "Self-Knowledge and 'Inner-Sense'", Philosophy and Phenomenological Research, 68:2, 249–314.

Silver, D. et al. (2016). "Mastering the game of Go with deep neural networks and tree search", Nature 529, 484–489.

Sober, E. (2003). “Two Uses of Unification”, in F. Stadler (ed), *The Vienna Circle and Logical Empiricism–Vienna Circle Institute Yearbook 2002*, Dordrecht: Kluwer, 205-216.

Sullivan, E. (2019a). “Understanding from Machine Learning Models”, The British Journal for the Philosophy of Science, First Online, axz035, https://doi.org/10.1093/bjps/axz035

Sullivan, E. (2019b). “Universality caused: the case of renormalization group explanation”, European Journal Philosophy Science 9:36, 1-21

Spirtes, P., Glymour, C., and Scheines R. (2000). *Causation, Prediction, and Search*, 2nd Ed., Cambridge, MA: MIT Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*, Amsterdam: North-Holland Publishing Company.

Vervoort, L., Blusiewicz, T. (2020). “The CMT Model of Free Will”, Dialogue, Canadian Philosophical Review, 59:3, 415–435

Vervoort, L., Melnikov, A., Chauhan, M., Nikolaev, V. (2021). “Artificial Consciousness, Superintelligence and Ethics in Robotics: How to Get There?”, accepted for publication in Mind and Matter.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*, Oxford: Oxford University Press.

Woodward, J. (2019). “Scientific Explanation”, The Stanford Encyclopedia of Philosophy (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/scientific-explanation/>.

Wu, T., Tegmark M. (2019). “Toward an AI Physicist for Unsupervised Learning”, Physical Review E 100, 033311, with an open-access version on arXiv: arXiv:1810.10525v4.

Zednik, C. (2019). “Solving the black box problem: A general-purpose recipe for explainable artificial intelligence”, ArXiv:1903.04361 [Cs].

Zheng, D., Luo, V., Wu, J., Tenenbaum, J. B. (2018). “Unsupervised Learning of Latent Physical Properties Using Perception-Prediction Networks”, arXiv preprint arXiv:1807.09244.