Analogue Models and Universal Machines. Paradigms of Epistemic Transparency in Artificial Intelligence

Hajo Greif¹

Received: 18 May 2021 / Accepted: 17 February 2022 / Published online: 5 March 2022
© The Author(s) 2022

Abstract

The problem of epistemic opacity in Artificial Intelligence (AI) is often characterised as a problem of intransparent algorithms that give rise to intransparent models. However, the degrees of transparency of an AI model should not be taken as an absolute measure of the properties of its algorithms but of the model's degree of intelligibility to human users. Its epistemically relevant elements are to be specified on various levels above and beyond the computational one. In order to elucidate this claim, I first contrast computer models and their claims to algorithm-based universality with cybernetics-style analogue models and their claims to structural isomorphism between elements of model and target system (in: Black, *Models and metaphors*, 1962). While analogue models aim at perceptually or conceptually accessible model-target relations, computer models give rise to a specific kind of underdetermination in these relations that needs to be addressed in specific ways. I then undertake a comparison between two contemporary AI approaches that, although related, distinctly align with the above modelling paradigms and represent distinct strategies towards model intelligibility: Deep Neural Networks and Predictive Processing. I conclude that their respective degrees of epistemic transparency primarily depend on the underlying purposes of modelling, not on their computational properties.

Keywords Models in science · Computer models · Cybernetics · Epistemic opacity · Deep Neural Networks · Predictive processing · Scientific explanation and understanding

✉ Hajo Greif
mail@hajo-greif.net
<https://hajo-greif.net/>

¹ Philosophy of Computing Group, Faculty of Administration and Social Sciences, Warsaw University of Technology, Warsaw, Poland

1 Introduction

A frequently recurring theme in studies of contemporary Artificial Intelligence (AI) is the claim that the operations of AI systems, particularly in Machine Learning, are epistemically opaque. Opacity, the diagnosis continues, makes AI difficult to understand and govern. However, the various domains of AI development and use are differentially affected by this problem. In computer science, epistemic opacity is mostly perceived as an engineering problem that requires engineering solutions, but is considered acceptable in practical terms to some extent. In areas where users or citizens in general might be negatively affected by intransparent operations of AI, epistemic opacity is framed as a normative problem that should be resolved, regulated or altogether avoided. In the domain of scientific modelling though, epistemic opacity assumes a remarkably ambiguous status, variously figuring as an unwelcome constraint that can and should be overcome or as an ‘essential’ epistemic condition that affects computer models more generally.

This paper is an attempt to resolve some of the ambiguity in the notion of epistemic opacity in the context of AI-based scientific modelling. It does so from a history and philosophy of science perspective. In particular, this analysis takes issue with the dichotomy between the view that epistemic opacity is chiefly a matter of complex or otherwise intractable algorithms that can be resolved on the algorithmic level and the claim that opacity is an ‘essential’ characteristic of AI models. Against this dichotomy, I will argue that epistemic transparency comes in degrees and differentially applies to various levels of a model. It is a function both of the degrees of an epistemic agent’s perceptual or conceptual grasp of a given model and of the elements and relations embodied in that model, which may or may not enable such a grasp. This condition, which will be discussed here as ‘model intelligibility’, is not only or primarily affected by the complexity or tractability of algorithms. Instead, it is chiefly affected by an element of underdetermination that is proprietary to computational methods and that has to be resolved by other means than reducing or modelling their complexity and increasing their tractability.

After outlining the problem of epistemic opacity in AI in Sect. 2, I will systematically compare and contrast two modelling paradigms that had a bearing on AI from its very beginnings: ‘analogue models’, as described by Black (1962), and digital computer models as first envisioned by Turing (1936), in Sects. 3 and 4 respectively. A deeper understanding of the conditions of model intelligibility in either case is developed in a comparative discussion of two related-but-distinct examples from contemporary AI that connect to the previous modelling paradigms in contrasting ways (Sect. 5). These case studies will help to bring home the point that the degrees of epistemic transparency of AI models are a situated affair that only partly depends on their computational properties.

In terms of methodology, the present approach will be partial (rather than comprehensive) because it focuses on examples of two selected modelling paradigms rather than seeking to cover the entire field of AI. There might be some wider implications though. The present approach will be partial (rather than impartial)

because it is based on the premiss that the best way of understanding science lies in understanding its use of models in representing and explaining world affairs.

2 Parsing Epistemic Opacity

David Marr, one of the pioneers of AI modelling in cognitive science, distinguished between ‘Type 1’ and ‘Type 2’ theories. He stated that the Type 1 condition, which states that ‘before one can devise [suitable algorithms], one has to know what exactly it is that they are supposed to be doing’, barely ever holds in AI. Instead, the epistemic situation more often is a Type 2 one, in which ‘a problem is solved by the simultaneous action of a considerable number of processes, *whose interaction is its own simplest description*’ (Marr, 1977, p. 38, emphasis in original). In these situations, a scientific discipline does not have an axiomatic theory of its subject matter at its disposal. Although there have been numerous attempts at formulating axiomatic theories for some subdomains of AI, none of them gave rise to, and many were not even intended to provide, a fully formed, general and widely accepted theory that might count as AI theory proper.

In this straightforward sense, AI remains a model-based science. In the context of scientific inquiries, models are understood as material or conceptual structures that establish or improve epistemic access to empirical affairs or theoretical axioms that might otherwise remain fully or partly beyond the cognitive grasp of human observers. In the most general sense, ‘A model is an interpretative description of a phenomenon that facilitates access to that phenomenon’ (Bailer-Jones, 2009, p. 1). Models are means of making phenomena intelligible and, at least in part, first enable the formulation of explanatory theories (for early conceptions of models in science, see Hertz, 1899; Boltzmann, 1902; for classic philosophy of science accounts, see Black, 1962; Hesse, 1966; the most pertinent contemporary source is da Costa & French, 2003, who further develop many of the aspects first outlined by Black and Hesse; but see also, for example, Magnani et al., 1999; Morgan & Morrison, 1999; Suárez, 2009; Weisberg, 2013; for an authoritative overview, see Frigg & Hartmann, 2020).

AI might share the diagnosis of being model-based with other pre-paradigmatic disciplines—in the Kuhnian sense of not yet having aligned into a coherent field of investigation with an established set of theories, methods and shared assumptions. However, there is another, more specific way in which AI might serve as a paradigm—in the non-Kuhnian sense of being a prime example—of a specific kind of model-based science: Having co-originated with computer science and the very practice of computer modelling, and forming one of their earliest domains of application, it is the first and most prominent example of a discipline that involves algorithms as a core constituent of its models. Being the concrete implementations of Turing’s (1936) computational method, algorithms are finite sequences of unequivocally defined steps of effectively calculating a mathematical function (Markov, 1960; Kleene, 1967; Knuth, 1973). By virtue of these properties, they are able to reliably and effectively produce numerical solutions to any problem that can be mathematically formulated and is amenable to an effective solution at all. (In the specific case

of Machine Learning and Deep Neural Networks to be discussed in Sect. 5 below, one has to distinguish between the set of instructions that determine how the model is solved and the ‘training algorithms’ used to fit the model with the data; the former will be relevant here.)

In their principled openness, which I will discuss in more detail in Sect. 4, computer models depart from more traditional ways of doing science. They created what some scholars deem novel, unique, and uniquely problematic epistemic situations of ‘epistemic opacity’, which have been influentially defined by Humphreys (2009) as follows:

a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process. (p. 618)

Conversely, the epistemic transparency of a model is understood as its ‘analytic tractability’, defined as the ‘ability to decompose the process between model inputs and outputs into modular steps, each of which is methodologically acceptable both individually and in combination with the others’ (Humphreys, 2004, p. 148). Accordingly, attainment of epistemic transparency depends on the observer’s ability to discern the elements of a model, to grasp their interrelations and temporal sequence and to recognise their respective roles in the model.

The key to Humphreys’ analysis of epistemic opacity as being ‘essential’ in a relevant class of cases is his twofold assumption that, in the use of computer models in science, the goal of epistemic transparency is unattainable for human observers *in principle*, and that it is unattainable in principle by virtue of some of the specific characteristics of the computational method. At first glance, this method should be predestinated for epistemic transparency by virtue of the precision and determinateness of the algorithms it uses. However, they might give rise to epistemic opacity for either of two distinct reasons (discussed by Humphreys, 2004, pp. 148–151 in direct reference to Marr’s above-cited distinction between Type 1 and Type 2 theories, see p. 3): First, there might be an analytic solution in terms of a set of axioms from which it can be predicted, whereas the number of computational steps required for reaching that solution is too high to be accomplished by human beings in timely and practical fashion. Accomplishing these steps might well be within the scope of a computer’s capabilities though, as in the case of computer-assisted proofs of mathematical theorems. In these cases, opacity is not essential. Second, however, the problem to be solved might plainly have no analytic solution. The only way of solving it lies in computing every step of the respective system’s development over time, with no axioms from which those steps could be predicted and general solutions produced.

In the first class of cases, analytic tractability diminishes inversely proportional to the model’s increasing computational complexity. From a computer science perspective, Lipton (2016) supplements this account with the degrees of decomposability and predictability as related but partly independent variables. Each of these factors will make the operations and functions of a model and its

elements more or less difficult to discern for a human observer. In these cases of non-essential opacity, an opaque model can be made more transparent if its complexity can be reduced, if it can be decomposed into its operationally pertinent elements, if the elements' behaviours can be reliably predicted, or if one can devise other means of interpreting or meaningfully modelling the properties relevant to the model's interpretation in a given epistemic context. This class of cases, to which I will henceforth refer to as problems of 'computational tractability', is addressed in the Explainable AI (XAI) paradigm, broadly conceived (see, for example, Gunning, 2019; Adadi & Berrada, 2018; Guidotti et al., 2018 for computer science and Burrell, 2016; Creel, 2020; Zednik, 2021 for philosophical and social science perspectives). The twofold premiss of this paradigm is that reducing or modelling computational complexity or related factors in view of a human-understandable solution is possible in a given context of application, and that the algorithms or other computational elements involved are in fact the epistemically relevant elements in that context.

In the second class of cases highlighted in Humphreys' above-cited definition, the root of opacity lies in the application of computational methods to problems that are per se outside the domain of analytically solvable problems. Computational complexity and related factors are secondary to and partly independent of this more fundamental condition. Reducing computational complexity or using superordinate models to make the model more tractable cannot fully remedy this condition, as there is no analytic solution that could serve as their basis. Numerical approximations are the only possible solutions. Agent-based models are a classic example of methods used in this kind of situation (Goldberg, 1989; Holland, 1992; Wilensky & Rand, 2015). However, the domain of such solutions and of the problems to which they can be applied is vast (as I will discuss in Sect. 4). Only under the assumption of principled analytic intractability, the notion of 'essential' epistemic opacity becomes meaningful. In Humphreys as well as in other philosophical studies of computer models and simulations in science (Winsberg, 2010; Lenhard & Winsberg, 2010), this *prima facie* pessimistic diagnosis is counterbalanced by an expressly pragmatic stance, under which essential opacity can be alleviated by mobilising background knowledge about the subject matter, experimental and observational methods and epistemic context that allow for a 'holistic sanctioning' of models. Analytic intractability need not undermine the overall empirical validity of a scientific model, provided that such validity can be established on other, otherwise independent levels of that model.

This pragmatic treatment should make clear that, even if epistemic opacity is considered essential rather than reducible to analytic solutions, the respective diagnosis depends on a set of assumptions concerning an epistemic agent's perception and pre-existing knowledge of the processes in question. There are limits to what human beings can cognitively grasp not only in concrete situations and with a concrete level of formal and technical skills, but also in terms of constitution-based constraints on the volume, kind and complexity of information that they can process under all nomologically possible circumstances. However, determining opacity to be essential under such circumstances does not amount to a negative a-prioristic epistemology that would set out limits of knowability in-principle for all possible epistemic agents.

In fact, Humphreys (2009) explicitly considers the potential role of non-human, artificially intelligent epistemic agents in science.

To the extent that analytic tractability of a computer model can be accomplished under the first of the above interpretations, it is likely to be conducive to the model's overall epistemic transparency. However, it is not a necessary condition, first, because there might be other elements of the model potentially affected by problems of opacity of their own and, second, because there might be other ways of making a model empirically valid and useful even if and when it remains analytically intractable, as under the second interpretation. Accordingly, opacity on the computational level poses a systematic problem only if and only to the extent that it affects the epistemically relevant properties of the model. It poses a *fundamental* epistemic problem if and only if the computational properties of the model are either coextensive with or negatively affect its epistemically relevant properties. In what follows, I will demonstrate in exemplary fashion that this is neither necessarily nor usually the case. The conditions of the transparency of a computer model, understood as the intelligibility of its epistemically relevant elements and overall structure, interact with but are otherwise independent of its computational properties.

3 Analogue Models and the Cybernetic Paradigm

The conditions of intelligibility of the epistemically relevant elements of models can be analysed by comparative reference to an early and classic contribution to the modelling literature in the philosophy of science, Max Black's 'Models and Archetypes' chapter in (1962). In that chapter, Black introduces 'analogue' models by recourse to a modelling paradigm that is related to but, at the time, was competing with early AI and computer models: cybernetics. Black characterises analogue models as material models to which a certain set of formal conditions applies. These conditions will help to elucidate both the specific properties of computer models that render them epistemically opaque and a set of strategies of making them empirically valid and useful nonetheless.

In general terms, Black introduces his concept of analogue models as follows:

An analogue model is some material object, system, or process designed to reproduce as faithfully as possible in some new medium the *structure* or web of relationships in the original. [...] It] is a symbolic representation of some real or imaginary original, subject to rules of interpretation for making accurate inferences from the relevant features of the model. (Black, 1962, p. 222, emphasis in original)

More precisely, analogue models establish isomorphism relations, understood in the mathematical sense as 'point-to-point correspondence' between features of the model and features of the original or, in more contemporary parlance, target system. Isomorphism, thus understood, is an identity relation between structures in terms of a bijective function that individually pairs every element in one structure with exactly one element in the other, so that no element remains un-paired and no element is paired more than once. Besides these isomorphisms between the elements of

model and target system, and in a number of relevant respects, the *relations* between the respective elements on either side will have to be equivalent between model and the target system, so as to reproduce ‘the *structure* or web of relationships in the original’.

The isomorphism relations ($\longleftrightarrow$) that obtain between the elements (A, B, C) on each side (m, o) and the equivalences ($\equiv$) between their internal relations ($|p, q, [] r$) can be visualised in a simple graphical scheme:

$$\left[\begin{array}{ccc} A_m & \longleftrightarrow & A_o \\ |p_m & \equiv & |p_o \\ B_m & \longleftrightarrow & B_o \\ |q_m & \equiv & |q_o \\ C_m & \longleftrightarrow & C_o \end{array} \right]$$

$$r_m \quad \equiv \quad r_o$$

Not all possible relations between all elements of the model and the original will matter here, but only the subset of elements and relations that have been determined relevant to the explanatory or communicative functions of the model. In this sense, isomorphisms and their epistemic import are agent-relative. For the same reason, analogue models are characterised as ‘symbolic representations’ by Black (1962, p. 222): the set of isomorphism and correspondence relations included in the model are established by convention. Even though intuitive or even seemingly obvious similarities on some level of observation are likely to guide the design of an analogue model, the crucial relations of structural correspondence between model and target system are those which allow for ‘making accurate inferences from the relevant features of the model’.

Conversely, not every relation between model and target system that would formally qualify as isomorphic or structurally corresponding will be included in the model but only those which allow for ‘independent confirmation’, given that analogue models can only ‘furnish plausible hypotheses, not proofs’ (Black, 1962, p. 223). In this sense, agent-relativity is limited. When used in scientific contexts, an analogue model of a real system has to be viewed and evaluated as a constituent part of the methods of a concrete empirical inquiry that seeks to explain the structure or behaviour of that system.

According to this analysis, the epistemic enablers and constraints of analogue models have to be sought in the concrete inquiries of which they are part, and in the epistemic situation of their designers and users. In the second classic contribution to the philosophy of models from the 1960s, Mary Hesse’s *Models and Analogies in Science* (1966), there is a fictitious debate between a ‘Duhemist’, who considers ‘material’ models useful but in principle dispensable aids to scientific theorising that necessarily fall short of the rigour and precision of axiomatic theories, and a ‘Campbellian’, who accepts the local and context-bound character of models while still taking them to be either ‘essential to the logic of scientific theories’ or ‘logically essential for theories’ (two conflicting characterisations in Hesse 1966, 7 and

19 respectively). Some uses of models may remain indifferent to empirical inquiries altogether, such as mathematical models, as these might be self-sufficient as formal structures (Black, 1962, p. 225), or they may remain indifferent to scientific inquiries altogether, such as models used for demonstrative or other practical purposes. Analogue models, however, are equally concrete and bound to scientific endeavours. The elements and the overall properties of a model are designed in such a way that they allow human observers to perceptually or conceptually grasp a hypostasised set of properties of its target system. The model is designed under the assumption that its elements and overall properties mediate cognitive access, as faithfully and precisely as possible under a given set of circumstances, to those elements and properties of a target system which are required for its explanation or understanding.

Black's own examples of analogue models are 'hydraulic models of economic systems, or the use of electrical circuits in computers' (1962, p. 222). The hydraulic model referenced here is Phillips' (1950) 'Monetary National Income Analogue Computer' (MONIAC), which was influential at the time and innovative in its explicitly cybernetic modelling approach. Given Black's description of analogue models and given his choice of examples, it is probable that MONIAC and other cybernetic models directly informed his concept of analogue models. Broadly cognitively oriented cybernetic models that were widely discussed at the same time, such as Ashby's (1960) homeostat or Walter's tortoises (1950, 1951), would qualify as Blackian analogue models, too. They constituted a paradigm related to and interacting with, but explicitly and consciously distinct from, the emerging domain of computer science and AI (see Dupuy, 2009; Boden, 2006 for historical accounts).

The epistemically relevant elements of Phillips' MONIAC, Ashby's homeostat or Walter's tortoises are the valves, tubes, electric circuits and other material components whose operations and relations within the model are designed to partly represent and contribute to an explanation of the structure and behaviour of their respective target systems. The force of this exemplary type of analogue models rests on the assumption of real analogies between the systems on either side. This assumption is characteristic of cybernetic models but not made explicit by Black (1962). Models of this type provide partial explanations of their target systems under the particular premiss that model and target system in fact operate in accordance with one and the same set of 'circular causal and feedback mechanisms' (von Foerster et al., 1953). The analogies in structure and relationships between them are grounded in a shared set of functional principles. Designing the model means to articulate a hypothesis concerning those mechanisms by means of recreating the functional principles of the target system in the model. Whether this set of mechanisms is in fact shared between model and target system will be a topic for independent confirmation or refutation.

Conversely, if there is no presumption of shared mechanisms, any analogy will have to be independently established. For example, it might be grounded in phenomenal similarity, in a shared genealogy, or more abstractly in the possibility of describing part of either system's behaviour by the same set of equations. Either way, the nature, extension and depth of the analogy is merely suggested, not warranted by the perceptually or conceptually accessible properties of the model. Where the cybernetic approach carries a promise of detecting as-yet unknown

mechanisms shared between model and target system through exploring the properties of the model and then comparing them with the properties of the target system, an approach that does not operate under the hypothesis of real analogies and shared mechanisms will require more justification of the proposed analogies while carrying less explanatory force.

If these observations are to the point, the analogue models of cybernetics will provide paradigms both of a model-based science and of epistemically transparent modelling. First, they are pre-theoretic in the strong sense envisioned for models in science by Marr (1977) and, more systematically, by Hesse (1966). These models are pre-theoretic in that they are used, in straightforward material fashion, for generating and testing structures that provide first and tentative formulations of the theoretical propositions of a general theory of systems. Second, the cybernetic paradigm carried the promise of precisely and unequivocally identifying the model's epistemically relevant elements, their interrelations and their relations to their target systems in such way as to be graspable individually and as a whole by human observers. The structural correspondences proposed in the model are supposed to argue for themselves, both through their observable properties and through the formal descriptions that go into their design. Third, and conversely, those elements of the model for which no such correspondence relations are defined are either considered epistemically irrelevant or provide a domain for further exploration of possible shared properties. In Hesse's (1966) terminology, these elements of the model provide 'negative' and 'neutral' analogies respectively, besides the isomorphic 'positive' ones. By virtue of these properties, the cybernetic paradigm operates in a bottom-up and embodied fashion not merely in the conception of its subject matter, but first and foremost in its methodological approach.

4 Computer Models and Universal Machines

Although they emanated from the same intellectual context, computer models provide a striking contrast to the analogue models of cybernetics. They originated in Turing's (1936) work on computation, in which he proposed a method by which any mathematical problem that can be solved by an 'effective method' can be solved by a suitably programmed discrete machine, which he christened 'Logical Computing Machine' (LCM). On an abstract level, any system that adheres to the set of basic operational principles specified by Turing will be able to perform computations. These principles can be reconstructed as follows (I am following Copeland, 1996, 2017; Shapiro, 1999 here; Turing himself never provided a fully explicit definition of computation):

- c.1 *The domain condition* The domain of computable functions is exhausted by the functions that are 'effectively calculable' in such a way that they can be solved, in principle, by an LCM.

- c.2 *The effectiveness condition* An effective method of calculation consists in a finite set of exact instructions that produce a correct solution to a function in a finite number of steps.
- c.3 *The specification condition* An LCM comprises of a finite set of symbols, a finite set of possible states, a transition function and a potentially infinite memory, designed in accordance with c.2 and in pursuit of solutions within c.1.

The general method embodied in the LCM (c.2) was concretely modelled on the behaviour of human computers (Copeland, 1997, 2017), which helped to define the modern notion of computing: Their task was to accomplish complex calculations in a collective but centrally governed manner, in which higher-order logico-mathematical operations are broken down into elementary arithmetical routines that could be accomplished with only a modicum of mathematical skills and without any technological aids beyond paper and pencil.

Once the principles of the LCM were physically implemented along the specifications in c.3 in digital computers such as the Automatic Computing Engine (ACE), it became possible to claim that, ‘without altering the design of the machine itself, it can, in theory at any rate, be used as a model of any other machine, by making it remember a suitable set of instructions’ (Turing, 1946). If and when they succeed at this task, computer models may establish any kind of modelling relation that lies within the mathematically delimited domain of c.1, given appropriate time and computing resources. They do so with a high degree of determinacy and precision, as stated in c.2, and within the specifications described in c.3.

With respect to the two types of epistemic opacity discussed in Sect. 2 above, these properties of computer models imply, first, that they could in principle solve all problems that are analytically tractable *and* all analytically intractable problems that are amenable to solutions by numerical approximation instead, to the extent that these problems fall under the condition of c.1-computability. Second, the models’ computational elements can be identified and their behaviour analysed in principle, given conditions c.2 and c.3. On this level, one might encounter problems of an unwieldy multitude of elements and of the complexity of interactions between them. Hence, there might be a problem of *computational* tractability in practice, which first and foremost concerns properties internal to the model. However, given the determinateness and precision of algorithms, complexity and tractability are not an insurmountable problem. The algorithms used in a model might be fairly simple, while the model turns out to be complex and intractable. Nor do complexity and intractability per se indicate that there also is a problem of essential opacity, understood as the *analytic* intractability of the problem that the model is supposed to solve.

The reverse side of the specific properties of computer models described in c.1, c.2, c.3 lies in the *prima facie* absence of the very constraints that make analogue models meaningful in the first place. Computer models neither have to rely on isomorphism relations between concrete elements and relations in model and target system nor on the requirement that human observers may conceptually and perceptually grasp these relations. The only limiting condition is that the computer model effectively captures the structure and complexity of the phenomenon by whatever means

available under the above principles. With two notable but problematic exceptions to be discussed below, there is no straightforward way of telling what kinds of elements on what levels of the model are supposed to partake, *qua their structure and properties*, in what kind of modelling relation. The pertinent modelling relation and its elements will have to be established in other ways. This form of underdetermination is specific to computer models and conceptually distinct from the complexity and tractability problems, although it interacts with them in practice. At least in part, the universality condition for the domain of computable problems is a precondition of the complexity and tractability problems. They are expressions of the force and scope of the computational method.

If one seeks to identify the epistemically relevant elements of computer models under the condition of universality, a whole spectrum of candidates can be nominated: First, and most basically, the epistemically relevant elements might be the algorithms used in the model. However, their empirical bearing will remain unspecified as long as one's perspective is confined to the level of algorithms. What algorithms do is to specify computations. Even if they do so in the most transparent and tractable manner, they will not provide information on how the model of which they are part relates to some world affair. If an observer wants to learn more about what an algorithmic structure might represent beyond specifying a computation, he or she must move to higher-order properties of the model. Alternatively, the algorithms might offer a partial analogy of their target system in terms of a shared set of functional principles. They will do so to the extent that the target system is a computational system, of which a computer model will then be an analogue model.

Second, the higher-order properties of the model might be seen in the mathematical forms realised by the algorithms. However, if we follow Black's above-cited observation concerning mathematical models (see p. 9), these structures and functions might be self-contained and either have no empirical bearing on any real-world system, or that bearing might be too general or too remote for the phenomenon to be recognisable in the mathematical structure. For example, Monte Carlo simulations generate pseudo-random inputs for a certain domain of possible inputs, where the distribution of the inputs over that domain is then sampled and measured in order to numerically approximate a solution for a problem for which other methods, in particular analytic methods, are not available. Monte Carlo methods are applicable to a wide spectrum of phenomena, from the physical processes in the ignition of a nuclear bomb (which was the first application of that method in computer simulations, see Galison, 1996) to 'particle filtering' approaches in Bayesian inference (Bishop, 2006; Murphy, 2012) and evolutionary computing (Holland, 1975, 1992). There is no expectation though that these phenomena are connected, let alone unified, by a shared mechanism that would be described by the Monte Carlo method—unless probabilistic structures such as Markov blankets are assumed to be real entities (as under the 'Free Energy Principle'; see p. 19 below and the discussion in Bruineberg et al. 2021). Otherwise, the probabilistic nature of this method does not even convey information on whether the pertinent phenomena themselves are stochastic or deterministic in nature.

Third, and more specifically, the relevant higher-order properties of the model might be symbolic forms that are realised by the pertinent algorithms and described

by the pertinent mathematical forms (for these distinctions, compare Marr's, 1982 levels of analysis and their contemporary applications by Zednik, 2021; Creel, 2020). On this level, expectations concerning shared mechanisms are articulated. Classical AI in particular focused on computer models as symbol systems, where computation and cognition alike were seen as sets of algorithmic operations upon meaningful symbols (Newell, 1980; Newell & Simon, 1976). The concrete implementation of these symbol systems was explicitly disregarded in part (for example by Fodor & Pylyshyn, 1988), whereas the meaningfulness of the symbols was taken to be independently warranted. This kind of approach deliberately accepts epistemic opacity on the level of the concrete implementation of the low-level computations that realise those higher-order symbol systems. Even on a level below embodiment-based critiques of classical AI (from Dreyfus, 1979 onwards), symbolic AI remained disembodied in terms of the concrete operational realisation of its models. Classical connectionism may be considered as the exact opposite in this respect, as it focused on sub-symbolic processes that were modelled on the implementation level and were supposed to represent features of the operational structure of the human nervous system (Rumelhart & McClelland, 1986).

Fourth, and conversely, a more material relation might be pursued by taking the hard- and software components of a computer model to be its epistemically relevant elements. A certain computer architecture, combined with specific programs, will serve as an analogue of the properties of its target system. For example, one might ask whether parallel rather than serial processing will embody the best analogues of human cognitive processes (Rumelhart & McClelland, 1986), or whether quantum-physical phenomena need to be modelled by quantum computers (Hagar & Cuffaro, 2019). The overall modelling relation will be borne by computer systems as concrete, material mechanisms that implement a certain computational structure. However, analogies of these kinds might be begging the question of what warrants a structural match with real-world systems. If computer systems physically implement computer models that, qua the scope and versatility of the computational method, might have been programmed or implemented differently, the justification for choosing a certain type of computer system over another as a model will have to include additional arguments for relations of isomorphism, phenomenal similarity or other kinds that could optimally or exclusively be established by this model rather than others. This requirement cannot be satisfied on the previously discussed levels of algorithms and mathematical structures.

There are two distinct ways in which the definition of the epistemically relevant elements of computer models may avoid the indeterminacy between the previous levels of modelling. They delimit that spectrum on either side: At the highest-order level, there is one modelling relation that is unequivocally analogue in the sense envisioned by Black (1962). The structure and functions of digital computers display isomorphisms with the behaviour of human computers who execute routines of arithmetic. However, this analogy is limited to a narrow subset of human abilities and exclusively considers their behavioural side. It also reverses the direction of the modelling relation, as human computers serve as the model *for* digital computers. (Compare Wittgenstein, 1947/1980, §1096: 'Turing's 'Machines'. These machines are *humans* who calculate.') At the other end of the spectrum, there is the claim

that all relevant processes that can be modelled by computers are computational processes themselves, so that the model's algorithms reflect the fundamentally computational structure of its target system. Under this assumption, cognition *is* a form of computation (Pylyshyn, 1980), or all sorts of physical processes implement computations (Putnam, 1988; Chalmers, 1996). A claim for a shared set of functional principles between model and target systems is thereby established at the algorithmic level, but at the cost of inflating the extension, and possibly also the meaning, of the concept of computation.

Unless a pan-computationalist view is adopted, the basic algorithmic properties of a computer model are just one of the various aspects of that model. These properties will be epistemically relevant either to the extent that they are shared with or represent properties of the target system or to the extent that their computational tractability (as distinguished from analytic tractability, see pp. 5 and 11) is necessary for overall comprehension of the model. Computer models will meet the first requirement at most partly. They will have to meet the second requirement only to the extent specified as necessary in a given epistemic context. In some cases, they might simultaneously meet both requirements to some extent. In many cases though, awareness will be required of the negative analogies between the computational structure of the model and the structure of the target system, which might either be computationally distinct or non-computational altogether. There is no a priori way of inferring the appropriate level and mode of reference of modelling relations from the algorithmic properties of a computer model.

5 Case Studies of Model Intelligibility

If the argument in the preceding Sects. 3 and 4 is to the point, analogue models will appear as a paradigm of models that are geared towards epistemic transparency. They will be so by virtue of their embodied and purposefully constrained qualities that are designed for being perceptually or conceptually graspable. Where analogue models work by recreating the structure and web of relationships of the original in some a priori unspecified medium, computer models may create or recreate a wide variety of structures and relationships in one ab initio specified medium. The wide spectrum of possible modelling relations into which computer models may enter contrasts with a *prima facie* narrow conception of their epistemically relevant properties that do not offer human observers direct access. Such access will have to be specifically created on some level or levels of the model, where these levels are identified in the modelling process.

This contrast will be better understood by drawing an analogy to the practice-oriented and context-dependent notion of intelligibility of scientific theories and their representations (de Regt, 2017). According to de Regt's 'criterion of intelligibility', 'a scientific theory *T* (in one or more of its representations) is intelligible' if it enables its users to 'recognize qualitatively characteristic consequences of *T* without performing exact calculations' (2017, p. 33). On this account, models are one of the possible representations of *T* that make it intelligible to its users. Analogue models seem to be one prominent example of models that

accomplish this. However, in the very context in which Black (1962) introduced the concept, they are pre-theoretic in Hesse's (1966) sense (see p. 10), so there is no pre-existing T or representation of T on which to rely. Instead, analogue models often serve theory construction in fields that are still lacking axiomatic theories. They can do so because they are designed to be intelligible to their users, so that they can explore possible regularities and cause-effect relationships by manipulating the model or by observing its behaviour. Accordingly, in a modified criterion of intelligibility, models M may take the place of theories T in this specific class of cases.

In contrast, AI and computer models in general are not naturally geared towards intelligibility and therefore unsuitable for model-based theory construction, for a twofold reason: On one side, their aim of universality renders their elements and structure very malleable and indifferent to specific modelling problems. A specific model-to-world fit has to be created without much guidance from the properties of the model. On the other side, and by virtue of their universality, the problems that AI and computer models are supposed to solve frequently reside in domains where axiomatic theories are neither available nor can be expected to become available. Analytical intractability might prevail, so there is no viable path from T to M either.

Given this situation, there are three possible strategies of addressing model intelligibility in AI and computer models:

- i.1 The criterion of intelligibility for human users is dismissed. A computer model M 's numerical solutions will be deemed self-sufficient as long as they are considered useful in some context.
- i.2 There are ways of making M 's solutions intelligible by providing models of its algorithms and computational processes, which might have a limited bearing on representing or explaining a given phenomenon.
- i.3 There are representations of a computer model M that enable its users, in analogy to the criterion of intelligibility for T , to 'recognize qualitatively characteristic consequences of M without performing exact calculations'. An analogy to the function of analogue models is thereby established.

These strategies can be exemplified by a comparative discussion of two related-but-distinct case studies from contemporary AI-based modelling approaches.

Deep Neural Networks: Deep Learning or Deep Neural Network approaches in AI (DNN; see Goodfellow et al., 2016; LeCun et al., 2015; Krizhevsky et al., 2012; Schmidhuber, 2015) are a subclass of Machine Learning (ML) models. They are broadly based on connectionist models but move beyond both ML and connectionism in several respects. While DNN and ML share their general aims and outlook, only DNNs are expressly committed to connectionist modelling. Where classical connectionism sought to provide models of cortical information processing, DNNs, like ML methods, are mostly used for applied tasks, such as object or image recognition, but also for strategic problem solving tasks that have hitherto been considered beyond the reach of AI, such as playing games of

Go (Silver et al., 2017). To an increasing extent, they are also used in predictive modelling in science. In some of these domains, they exceed the scope of human abilities.

While taking a number of cues from the structure and functions of the mammalian brain, DNNs are typically not intended to resemble or be isomorphic to them in terms of providing models of their structure and functions. Instead, they are designed to solve classes of problems that are partly inaccessible to information processing in all known natural organisms. Without being provided with explicit models of the pertinent domain, DNNs use training data to extract structures from data sets on numerous levels of abstraction and to develop solution strategies generate models on their own.

DNNs diverge from the classical neural network paradigms not merely in their aims, but also in a number of methodological aspects (which are laid out in more detail in Buckner, 2019): First, they add a large number of hidden layers to the network, moving significantly beyond the number of layers found either in classical neural networks or in all known natural organisms. Second, DNNs purposefully restrict connections between layers to circumscribed areas rather than allowing all units in adjacent layers to become fully connected. Third, they allow the activation functions of network nodes to be differentiated rather than uniform while assigning them specific roles in the network. The two latter conditions make DNNs distinctly heterogeneous networks with a diversification of functions and a division of labour unknown to classical and natural neural networks. Fourth, DNNs are provided with methods of counterbalancing idiosyncratic classifications that would otherwise lead to ‘overfitting’ the data and impede the recognition of invariant characteristics of an object throughout variant conditions of presentation. This might result in incomprehensibly idiosyncratic structures—a problem that does not become acute in classical neural networks and that does not occur in natural perception.

The specific challenges for DNNs arise in part from their computational power, and in part from the architecture they keep sharing with classical connectionism. Their difficulties with generalising beyond the training data owe to the depth and memory of the networks. Adapting or correcting once established patterns to new data or recognition tasks require a specific effort that cannot be directly inferred from the architecture of the system. More immediately and obviously, the functional diversification, computational complexity and intractability of DNNs, like for ML more generally, make their classification decisions difficult to follow and predict both on the operational level and on the level of human interpretation. In fact, the primary aim of DNNs is to make the most of the computational complexity that can be technologically mastered at a given stage—which is, sometimes deliberately, bought at the cost of limitations on what a human observer will be able to learn from and about a model that successfully produces a solution to the problem in question.

Such acceptance of epistemic opacity might be partly grounded in DNNs’ reliance on mechanisms and architectures that are *ab initio* and purposefully biologically implausible, and therefore cannot possibly have a claim on naturalistic modelling. Whereas, for example, the recourse to back-propagation mechanisms in classical connectionism will constitute a negative analogy to the functions and structure of natural nervous systems and make its models less realistic in light of their

cognitive modelling aims, DNNs are less affected by this limitation to the extent that they typically do not seek to provide cognitive models, let alone propose shared mechanisms, in the first place.

Predictive Processing: The Predictive Processing paradigm (PP; see Clark, 2013; Dayan et al., 1995; Hohwy, 2013, 2020) uses connectionist models to explain the functional principles of cortical information processing in humans and other higher animals. Operating, like the majority of modern cognitive science, in a tradition that takes the problem of perceptual ambiguity as its starting point, it posits that an organism has to infer the structure and properties of the world from ambiguous and underdetermined sensory input. (Berkeley, 1709 was the first to provide an explicit account of how three-dimensional images of an object or scene arise from a two-dimensional retinal image.) However, in contrast both to the notion of symbolic information processing in classical AI and to classical and DNN connectionism, PP claims to realistically model cortical information processing as a bidirectional hierarchical process of prediction error minimisation between sensory input and higher-order dynamic world models. Instead of computing an array of symbols that individually represent object properties and assembling them into a comprehensive, image-like representation of world affairs, the brain seeks to estimate the degree to which incoming sensory information will diverge from its own current model of the world, and tries to create situations that minimise such divergence.

Sensory information and dynamic world models are understood in PP as the input and output layers of a neural network respectively. The higher levels produce generative models of the specific structures and events in the organism's environment that may have given rise to a certain pattern of sensory input, seeking to disambiguate that input. Being generative models, they do so without the benefit of explicit models of the world that would be separately provided. Instead, the brain keeps building and rebuilding models in probabilistic fashion from what sensory input provides. The benefit of this approach is that sensory information uptake is economically managed: It is restricted to the detection of unexpected variation in comparison to the brain's current prediction of how the world stands. The bulk of the contents of the brain's higher-order world model at a given stage is provided by the model itself, while it is being continuously updated with and adjusted to any unexpected variation that is transmitted from the input level.

This type of predictive and generative approach to modelling in the nervous system builds on probabilistic techniques of data compression in computer engineering, in which fully explicit representations of every data point, such as a pixel of a digital image, are replaced by predictions of whether and how a neighbouring data point might diverge from the current one. Only when there is a change in values, that difference is explicitly encoded. In return for this data processing analogy, the PP model of cortical information processing pictures the brain as an essentially Bayesian machine: It is a system that produces and acts on essentially probabilistic information and strives to increase the probability of its predictions to become true and reduce what the information theorist calls 'surprisal', namely unexpected divergence on the input level. Organisms thereby actively anticipate their own sensory input and the requisite responses before that input actually occurs.

The models introduced under the Predictive Processing paradigm propose a close match between the algorithmic processes in the model and the model's target system. They suggest a direct, relatively low-level and partially isomorphic analogy between algorithmic and cognitive processes. The isomorphisms between PP models and cortical information processing are supposed to hold both on the level of neural computations and on the level of the superordinate hierarchical computational structures within which they take place. These computations and structures are embodied in the models in such a way that a conceptual grasp of the model shall enable an understanding of cortical information processing and a (partial) explanation of its mechanisms.

In aims and methods, PP unreservedly puts its models in the service of a scientific explanation of cognitive and related processes. In a manner that combines computational theories of the mind with a cybernetic, isomorphism-based view of modelling, PP offers a set of in-principle testable hypotheses concerning the basic functional principles of cortical information processing. This set of hypotheses includes claims concerning possible shared mechanisms between cortical and the computer-based kinds of information processing designed under the PP paradigm. How fundamentally these mechanisms are supposed to shared varies between different PP-related approaches. The Free Energy Principle (FEP) in particular postulates that the brain's uncertainty-minimising activities reflect an overarching principle of self-organisation that it claims to apply to all living systems, down to the organism-environment distinction, and thereby assumes a quasi-metaphysical status (Friston, 2010, 2013; Friston & Stephan, 2007; Kirchhoff, 2018; Kirchhoff et al., 2018; Kirchhoff & Kiverstein, 2021). In doing so, the FEP explicitly if marginally acknowledges its cybernetic heritage.

Discussion: Even though independent verification of PP models has not been provided to date, they have a justified claim to offering biologically more plausible and scientifically more tenable models than either the back-propagation mechanisms proposed by classical connectionism or any of the distinctive features of DNNs. In terms of definition of the epistemically relevant elements of their models, PP is a computer modelling approach that postulates, more than many other 'computationalist' views, analogies of the isomorphic kind on the algorithmic level. In contrast, DNNs do while PP does not exploit the potential complexity and diversification of their connectionist systems when seeking to match the current body of knowledge on neural information processing. In part, however, PP and DNN approaches build on a fundamentally similar set of basic connectionist principles and have seen contributions from the same authors (Geoffrey Hinton in particular).

With respect to the aim and nature of the modelling relations involved, DNNs are mostly geared towards effective problem-solving rather than towards empirically adequate modelling, mechanistic explanation or even scientific understanding more generally. Nonetheless, DNNs are increasingly used to contribute to scientific endeavours. Unlike the very specific analogies proposed in PP models, DNNs are both more universal in terms of possible modelling relations and more indirect in terms of accomplishing them (for an insightful discussion of the possible contributions of DNNs to scientific understanding and their limits, see Sullivan, 2019). In cognitive science, DNNs that are tailor-made for biological plausibility

by restricting the number of layers and types of operations to what can be expected in natural organisms might provide for partial models of human visual recognition (Cichy et al., 2016). In that case, they will approximate the properties of classical, cognitive modelling-oriented connectionism rather than exploiting the full spectrum of computational possibilities available to DNNs. Alternatively, DNNs might be used to generate models of cognitive strategies of empirical generalisation and abstraction (Buckner, 2018). In this case, the analogy to human cognitive strategies remains on a general and higher level that provides only limited information on the mechanisms by which these strategies are pursued in human beings. However, ML-based approaches in cognitive neuroscience might also consciously accept a trade-off between the predictive power of DNN-based models and their neural plausibility (Stinson, 2020) or the depth of understanding phenomena that they might provide (Chirimuuta, 2021). In physics, DNNs are typically deemed unable to provide or help generating explanations of phenomena, for want of an ability to track existing or to outline new theoretical accounts of those phenomena. They might assume more exploratory functions though, such as in delimiting the search domain for new particles (Boge and Grünke, in press). In the biomedical sciences, DNNs are considered useful to the extent that they provide information on genomic patterns and correlations that would otherwise remain hidden, and thereby support more traditional mechanistic models, but they are also being used in fully data-driven approaches that accept a trade-off between predictive power and model transparency (López-Rubio & Ratti, 2021; Facchini & Termine, in preparation). There is no *a priori* way of determining a genuine set of possible roles that DNNs might play in these and other sciences, but the development and implementation of models that would directly contribute to scientific explanation and theory building does not seem to be among these roles.

To the extent that DNNs are used for scientific modelling, its epistemically relevant elements may indeed remain insufficiently recognisable for human observers. Apart from their computational complexity or intractability, the key contributing factor to their opacity, as I have sought to demonstrate on these pages, lies in the underspecification of the level on which a model-target relation would be supposed to hold. To the extent that DNN models are truly generative, such relations might be specified neither by human designers nor for human observers. There are no isomorphisms and analogies that human users could rely on in absence of an inclusion of their specific vantage points in the model. If human beings are only partly involved in defining the modelling relations, there will be only limited human use for the relations thus defined. Their use for DNNs themselves is a different matter. This is one way of making sense of Humphreys' 'post-anthropocentric epistemology' (2009): the epistemic agents for whom such models are useful might not be human after all.

Given these observations, the mapping of the two case studies on two of the three strategies concerning model intelligibility (see p. 16) is straightforward: PP models are supposed to work as intelligible models that ground their analogy relations in the computational architecture and algorithms of the connectionist networks they employ (see i.3). In contrast, DNNs remain at least *prima facie* indifferent towards the criterion of intelligibility (see i.1) because their computational elements are not made for being graspable as representing elements of a target system. Such relations

can only be established on the level of interpreting the model outputs, if at all. What remains to be elucidated are the merits and limitations of the third strategy: i.2 is the domain of XAI-related approaches that seek to render the algorithms and computational elements involved in a model more intelligible. The primary aim of these approaches is to make the algorithms of DNNs, ML or other complex AIs computationally tractable and to reduce or model their complexity to a degree that matches human cognitive capacities. XAI-based approaches are therefore concerned with the internal properties of the model.

If my previous analysis is to the point, XAI-based approaches are likely to face two related problems: First, they might fail to address the opacity problem that pertains to computer models in general—which is not chiefly their complexity but their universal applicability to the domain of computable problems. Second, resolving opacity problems at the complexity or tractability level is indeterminate in its implications for how a model relates to some world affair, and what its users might learn from it. It might have such implications to the extent that the model's relation to a phenomenon in fact becomes more graspable. It might fail to have such implications if and to the extent that a better understanding of the model's internal properties is powerless with respect to an opacity problem that is contained within the former: A better operational understanding of the numerical solutions to analytically intractable problems will not make those problems analytically tractable. It will not resolve essential epistemic opacity.

6 Conclusion

The argument on the previous pages can be summarised as follows: With respect to scientific modelling, epistemic transparency is a function of the cognitive grasp of the epistemically relevant elements of a model that its users and designers develop. The scientific usefulness of the model as a whole will depend on a proper definition and understanding of these elements. The more isomorphism relations can be established to hold between model and target system, and the more credible the case is that can be made for shared functional principles between them, the more empirical credibility accrues to the model. Computer models face a particular difficulty with respect both to defining its epistemically relevant elements and to establishing a cognitive grasp of these elements. The model's algorithms neither necessarily nor naturally constitute its epistemically relevant elements. Accordingly, making the algorithms computationally tractable might be necessary in many cases but will not be sufficient for establishing higher-level model intelligibility.

I have begun to develop an argument that computer models are not fundamentally different from DNNs in terms of epistemic opacity. The complexity and conceptual alienness of DNNs might compound the condition of opacity, but essential opacity is not their privilege, nor does it turn non-essential opacity into an essential one. Instead, the feature that distinguishes them from other computer modelling and AI approaches, including the otherwise related PP paradigm, lies in their essential indifference towards the endeavour of scientific explanation and understanding.

DNNs, like ML approaches more generally, are neither designed for this endeavour, nor can they be recruited for it in a similar way to analogue or more traditional computer models. Their ability to master complex cognitive tasks in ways that are in part beyond human comprehension actually testifies to that indifference. Their structure and operations do not offer naturally comprehensible analogues of natural systems, nor is there an attempt or even a pretence of resolving or compensating for the opacity that is its own and any other computer model's way of being. If science is the project of methodically and empirically adding to the body of human knowledge, it remains an open question whether DNNs may contribute to that project, or whether it will have to be redefined in method or purpose.

Acknowledgements The origin of this paper was a symposium on 'Deep Learning and the Philosophy of Artificial Intelligence' co-organised by the author at the GWP 2019 conference of the German Society for Philosophy of Science. I thank Cameron Buckner, Holger Lyre and Carlos Zednik for motivating and moving this project. I also thank the guest editors of this Special Issue for their encouragement and patience, Alessandro Facchini and Alberto Termine for their input in the process of preparing the manuscript, and two anonymous reviewers for their competent and constructive criticisms.

Funding Funding was provided by Narodowe Centrum Nauki (Grant No. 2020/37/B/HS1/01809).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>.
- Ashby, W. R. (1960). *Design for a brain. The origin of adaptive behaviour* (2nd ed.). Wiley.
- Bailer-Jones, D. (2009). *Scientific models in philosophy of science*. Pittsburgh University Press.
- Berkeley, G. (1709). *An essay towards a new theory of vision*. Rhames and Pepyat.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Black, M. (1962). *Models and metaphors*. Cornell University Press.
- Boden, M. A. (2006). *Mind as machine: A history of cognitive science*. Oxford University Press.
- Boge, F. J., & Grünke, P. (in press). Computer simulations, machine learning and the Laplacean demon: opacity in the case of high energy physics. In A. Kaminski, M. Resch, & P. Gehring (Eds.) *The Science and Art of Simulation II*. Springer.
- Boltzmann, L. (1902). Model. In D. M. Wallace, A. T. Hadley, & H. Chisholm (Eds.), *Encyclopaedia Britannica* (10th ed., Vol. 30, pp. 788–791). Adam and Charles Black, The Times.
- Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2021). The emperor's new Markov blankets. *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X21002351>
- Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, 195, 5339–5372. <https://doi.org/10.1007/s11229-018-01949-1>
- Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. <https://doi.org/10.1111/phc3.12625>

- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Chalmers, D. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108, 309–333.
- Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199, 767–790. <https://doi.org/10.1007/s11229-020-02713-0>
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1), 27755. <https://doi.org/10.1038/srep27755>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 1–73. <https://doi.org/10.1017/S0140525X12000477>
- Copeland, B. J. (1996). What is computation? *Synthese*, 108, 335–359.
- Copeland, B. J. (1997). The broad conception of computation. *American Behavioral Scientist*, 40(6), 690–716. <https://doi.org/10.1177/0002764297040006003>
- Copeland, B. J. (2017). The Church-Turing thesis. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*, spring 2020 edn, The Metaphysics Research Lab, Stanford. Retrieved from <https://plato.stanford.edu/archives/sum2020/entries/church-turing/>.
- Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>
- da Costa, N., & French, S. (2003). *Science and partial truth: A unitary approach to models and scientific reasoning*. Oxford University Press.
- Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7(5), 889–904. <https://doi.org/10.1162/neco.1995.7.5.889>
- de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.
- Dreyfus, H. L. (1979). *What computers can't do. A critique of artificial reason*. Harper & Row.
- Dupuy, J. P. (2009). *On the origins of cognitive science: The mechanization of the mind*. MIT Press.
- Facchini, A., & Termine, A. (in preparation). Beyond hypothesis-driven and data-driven biology through explainable AI: A proposal. *Venue* tbd.
- Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 8, 3–71.
- Frigg, R., & Hartmann, S. (2020). Models in science. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*, spring 2020 edn, Metaphysics Research Lab, Stanford. Retrieved from <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.
- Friston, K. (2010). The free energy principle: A unified brain theory? *Nature Reviews: Neuroscience*, 11(2), 127–138.
- Friston, K. (2013). Life as we know it. *Journal of The Royal Society Interface*, 10(86), 20130475. <https://doi.org/10.1098/rsif.2013.0475>
- Friston, K., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417–458.
- Galison, P. (1996). Computer simulations and the trading zone. In P. Galison & D. Stump (Eds.), *The disunity of science: Boundaries, contexts, and power* (pp. 118–157). Stanford University Press.
- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 93. <https://doi.org/10.1145/3236009>
- Gunning, D. (2019). DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, IUI '19, p. ii, <https://doi.org/10.1145/3301275.3308446>.
- Hagar, A., & Cuffaro, M. (2019). Quantum computing. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*, winter 2019 edn, Metaphysics Research Lab. Retrieved from <https://plato.stanford.edu/archives/win2019/entries/qt-quantcomp/>.
- Hertz, H. (1899). *The principles of mechanics. Presented in a New Form*. Macmillan. With an introduction by H. V. Helmholtz. Translated by D. E. Jones and J. T. Walley.
- Hesse, M. B. (1966). *Models and analogies in science*. University of Notre Dame Press.
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, 35(2), 209–223. <https://doi.org/10.1111/mila.12281>
- Holland, J. H. (1975). *Adaptation in natural and artificial systems*. MIT Press.

- Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267(1), 66–73.
- Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.
- Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169, 615–626.
- Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life-mind continuity thesis. *Synthese*, 195(6), 2519–2540. <https://doi.org/10.1007/s11229-016-1100-6>
- Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, 199, 4791–4810. <https://doi.org/10.1007/s11229-019-02370-y>
- Kirchhoff, M. D., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15, 20170792. <https://doi.org/10.1098/rsif.2017.0792>
- Kleene, S. C. (1967). *Mathematical logic*. Wiley.
- Knuth, D. E. (1973). *The art of computer programming* (2nd ed., Vol. 1). Addison-Wesley.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.) *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* (Vol. 1, pp. 1097–1105). Curran Associates.
- LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, 521, 436–444.
- Lenhard, J., & Winsberg, E. B. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 41(3), 253–262. <https://doi.org/10.1016/j.shpsb.2010.07.001>
- Lipton, Z. C. (2016). The mythos of model interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)* (pp. 96–100).
- López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, 198(4), 3131–3156. <https://doi.org/10.1007/s11229-019-02271-0>
- Magnani, L., Nersessian, N. J., & Thagard, P. (Eds.). (1999). *Model-based reasoning in scientific discovery*. Springer.
- Markov, A. (1960). *Theory of algorithms*. American Mathematical Society.
- Marr, D. (1977). Artificial Intelligence—a personal view. *Artificial Intelligence*, 9, 37–48.
- Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. MIT Press.
- Morgan, M. S., & Morrison, M. (Eds.). (1999). *Models as mediators: Perspectives on natural and social science*. Cambridge University Press.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4, 135–183.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126.
- Phillips, A. W. H. (1950). Mechanical models in economic dynamics. *Economica, New Series*, 17(67), 283–305.
- Putnam, H. (1988). *Representation and reality*. MIT Press.
- Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundations of cognitive science. *The Behavioral and Brain Sciences*, 3, 111–169.
- Rumelhart, D., & McClelland, J. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). MIT Press.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Shapiro, S. (1999). Effective procedure. In R. Audi (Ed.), *The Cambridge dictionary of philosophy* (p. 254). Cambridge University Press.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359. <https://doi.org/10.1038/nature24270>
- Stinson, C. (2020). From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science*, 87(4), 590–611. <https://doi.org/10.1086/709730>
- Suárez, M. (Ed.). (2009). *Fictions in science. Philosophical essays on modelling and idealisation*. Routledge.

- Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, ahead of print. <https://doi.org/10.1093/bjps/axz035>
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42, 230–265.
- Turing, A. M. (1946). Letter to W. Ross Ashby of 19 November 1946 (approx.). The W. Ross Ashby Digital Archive. Retrieved from <http://www.rossashby.info/letters/turing.html>.
- von Foerster, H., Mead, M., & Teuber, H. L. (Eds.). (1953). *Cybernetics: circular causal and feedback mechanisms in biological and social systems. Transactions of the ninth conference, March 20–21, 1952*. Josiah Macy, Jr. Foundation.
- Walter, W. G. (1950). An imitation of life. *Scientific American*, 182, 42–45.
- Walter, W. G. (1951). A machine that learns. *Scientific American*, 185, 60–63.
- Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
- Wilensky, U., & Rand, W. (2015). *An introduction to agent-based modeling: Modeling natural, social, and engineered complex systems with NetLogo*. MIT Press.
- Winsberg, E. B. (2010). *Science in the age of computer simulation*. University of Chicago Press.
- Wittgenstein, L. (1947/1980). *Remarks on the philosophy of psychology*. Blackwell.
- Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34, 265–288. <https://doi.org/10.1007/s13347-019-00382-7>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.