Models, Algorithms, and the Subjects of Transparency

Hajo Greif

Philosophy of Computing Group, Faculty of Administration and Social Sciences,
Warsaw University of Technology, Plac Politechniki 1, 00-661 Warsaw, Poland
hans-joachim.greif@pw.edu.pl, https://hajo-greif.net/

Abstract Concerns over epistemic opacity abound in contemporary debates on Artificial Intelligence (AI). However, it is not always clear to what extent these concerns refer to the same set of problems. We can observe, first, that the terms 'transparency' and 'opacity' are used either in reference to the computational elements of an AI model or to the models to which they pertain. Second, opacity and transparency might either be understood to refer to the properties of AI systems or to the epistemic situation of human agents with respect to these systems. While these diagnoses are independently discussed in the literature, juxtaposing them and exploring possible interrelations will help to get a view of the relevant distinctions between conceptions of opacity and their empirical bearing. In pursuit of this aim, two pertinent conditions affecting computer models in general and contemporary AI in particular are outlined and discussed: opacity as a problem of computational tractability and opacity as a problem of the universality of the computational method.

Keywords: Artificial Intelligence, Epistemic Opacity, Black Box Problem, Explainable AI, Computer Models in Science

The problem: AI co-originated with computer science and the practice of computer modelling and formed one of the earliest domains of application of computer modelling methods. It is the paradigm of a discipline that involves stored programs and digitally encoded information as core constituents of its models. By virtue of these properties and in the course of increasing the scope and depth of computational methods, AI gave rise to epistemic situations that are arguably novel, unique, and uniquely problematic. These situations are being referred to as 'epistemic opacity' or the 'Black Box Problem'. Most generally speaking, an AI is epistemically opaque if an observer cannot adopt a position from which to discern either the operations of the system or their bearing on some world affair, or both. There is a heterogeneous array of definitions and interpretations of this problem, among which I identify three distinct classes:

I. There are approaches that consider epistemic opacity a fundamentally technical problem that is in principle resolvable within the framework of computational methods. This is the domain of 'Explainable AI' (XAI), which proposes methods of making the operations of AI systems better discernible.

II. There are approaches that consider epistemic opacity an essential problem of AI systems that cannot be resolved within the framework of computational methods. These views fall into two distinct sub-classes:
   a. Opacity might be essential for reasons that universally apply to the computational method.
   b. Opacity might be essential for more specific reasons that apply to more specific types of AI systems.

The argument: I argue that the key to understanding the differences between the previous approaches I. and II. lies in how they perceive, first, the nature and the role of the epistemic agents to which an AI system could be epistemically opaque or transparent. Opacity might be a property that mostly or exclusively pertains to an AI system and its operations, or it might be more of a relational property that depends on the means of epistemic access to the system that an agent has at his or her disposal. Second, the differences between the above interpretations I. and II. depend on how they perceive the relation between a set of algorithms, data structures and computer architectures on the one hand and the models of which they are part on the other. More precisely, the difference between the 'explainable' and the 'essential' accounts lies in how a model's respective epistemic properties are perceived to rest upon each other. Epistemic problems may arise either from the degrees of complexity or tractability of computational processes and structures, or from the ways in which they are applied in the construction of models. Which of these factors is considered most relevant in turn partly depends on the previously mentioned epistemological commitments that often remain implicit but shall be briefly explicated in what follows.

Defining opacity: While a consensual and unified definition of epistemic opacity in AI is missing in the literature, the earliest and most frequently cited definition is the one proposed by Paul Humphreys (2009):

> [. . .] 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 an epistemic agent's '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, 148). The notable features of this twofold definition are, first, that opacity in general and essential opacity are defined with respect to epistemic agents. A process is epistemically opaque or transparent to a concrete agent in a concrete epistemic situation. Second, Humphreys' definition refers to processes rather than systems, where these processes are in turn specified in terms of computational processes that concern models, and where the degrees of epistemic transparency versus opacity are framed in terms of analytic tractability as a mathematical concept. Let me discuss these two features in turn.

Agent-relativity: The degrees of epistemic transparency versus opacity are agent-relative on an individual level because they depend on an agent's conception of the epistemically relevant elements of a given model or process. On the one hand, the degrees of knowledge of that model or process might be at variance between different agents, which makes a difference with respect to their abilities of recognising the relevant elements. On the other hand, the agents' various interests in and expectations towards that model or process might be at variance, too, with no benchmark or universal standard being available in many cases of what counts as and needs to be recognised as epistemically relevant.

However, the degrees of opacity are agent-relative on a general level, too, in terms of depending on the constraints on the volume, kind and complexity of information that a specific epistemic agent or population of agents can process. If only for the qualifier 'given the nature of X', 'essential' opacity is neither to be understood as being firmly grounded in the properties of the process that is being perceived or conceived of as opaque nor as a condition that holds in all or all nearby (logically) possible worlds. Philosophical presuppositions of these kinds will invite misunderstandings of a claim that is empirically more pertinent and philosophical in a different way, namely that cognition is 'bounded' and 'situated' (see, for example, Robbins and Aydede, 2009). Epistemic opacity is a condition that applies to concrete epistemic agents in relation to to concrete processes in concrete real-world contexts. It is an essential condition precisely to the extent that there are no realistic means for an epistemic agent of transcending a given set of real-world constraints.

The agent-relativity of epistemic opacity might not be universally recognised but is well-reflected in the literature, where one can detect various stages of specificity of the argument: Most broadly, Humphreys (2009) developed his concept of opacity in view of human and other epistemic agents and their cognitive limitations. In particular, he considered the possibility of forms of computer-based science that remain inaccessible to human agents while being accessible to AI systems. In contrast, Beisbart (2021) formulates an explication of epistemic opacity on the premiss that a process is opaque to the degree that it is difficult for humans to know and to understand why its outcomes arise, with no other epistemic agents being admitted to consideration. Here, the cognitive constitution of human beings is viewed as exclusively relevant. On the most specific level, several authors identify the cognitive constitution of situations of opacity as specifically pertaining to concrete 'stakeholders', their knowledge and their intentions in a given situation (Zednik, 2021; Páez, 2019; Langer et al., 2021). Tomsett et al. (2018) consider agent-relativity the main reason why there is no consensus among AI practitioners about the meaning of epistemic opacity in the first place. Even though the authors discussed here might not agree whether there is such a thing as essential opacity or whether there are means of resolving opacity instead, they would all agree that it is an *epistemic* condition, that is, a condition that affects the availability or acquisition of knowledge.

Models: There is a certain ambiguity in the literature between an understanding of epistemic opacity as a problem with tracking the internal properties of an AI model, that is, its algorithms, its computational structure more generally or its complexity, and an understanding of opacity as a problem with how the model relates to a given world affair, that is, how and by what criteria data are classified, how and on what grounds a model generates a prediction. Boge (2021) distinguishes between these forms of opacity as 'h-opacity' as a class of situations where the observer has no insight into how a model operates, versus 'w-opacity' as a class of situations where the observer has no insight into what it represents. He considers w-opacity as a specific characteristic of AI of Machine Learning-based models in particular, to the extent it arises from unsupervised learning processes in which the features learned by the model cannot be traced back to the processes by which they were learned, and thus to the 'how' aspect. In similar fashion, Facchini and Termine (2022, in this volume) distinguish between, on the inward-looking side, 'access opacity' as a problem with 'understanding the structure of a system' and 'informational opacity' as an issue concerning 'the format, or setup, adopted by a system for storing and manipulating information'. On the outward-looking side, they characterise 'link opacity' as 'insufficient information about the elements that are relevant for explaining, predicting and controlling the considered phenomenon' (see also Sullivan, 2019). Situations of opacity of either kind are often de facto connected, but they are not related by necessity. In boundary cases, an internally opaque model might result in intelligible predictions, while gaining information about the inner workings of a w-opaque or link-opaque model might not suffice to make it intelligible.

In order to better understand why these two forms of epistemic opacity are partly independent and why w- or link opacity is particularly pertinent to AI, it will be important to recall the nature and function of models in scientific inquiry more generally (classical and still valid philosophical accounts of models in science include Black 1962; Hesse 1966; more contemporary ones include Morgan and Morrison 1999; da Costa and French 2003; for an authoritative overview, see Frigg and Hartmann 2020). According to Ludwig Boltzmann's first modern definition of of models, a model is 'a tangible representation [. . .] of an object' (Boltzmann, 1902), where that representation might exist purely 'in the head', but typically assumes the shape either of verbal or formal descriptions or of material objects. Since the mid-20th Century, implementing models in computers has become an additional and increasingly important part of the scientific method. The unifying and most general characteristic of models in the empirical sciences is that 'A model is an interpretative description of a phenomenon that facilitates access to that phenomenon' (Bailer-Jones, 2009, 1). Models are typically designed to be observed or manipulated in such a way that they provide information about a phenomenon in situations where direct observation or experimental manipulation of that phenomenon is not possible, either in principle or due to practical constraints.

Models accomplish this aim by bearing a variety of types of material and formal representational relations to a given phenomenon, namely similarity, structural isomorphism or sameness of properties. In order to establish these representational relations, models are designed to address properties of the phenomena in a variety of ways, most notably approximation, abstraction and idealisation. As far as the empirical sciences are concerned, models of all kinds mediate perceptual or conceptual access to their pertinent phenomena by postulating relations between selected elements of the model and selected elements of the phenomenon, where the kinds of relations fall under one of the previously mentioned classes. The selected elements are those which are considered most relevant to the explanation or understanding of the phenomenon in question. These 'positive' analogies (in Hesse's 1966 parlance) are to be distinguished from 'negative' analogies – relations that are known *not* to hold between them – and 'neutral' ones, where possible relations are yet to be explored and might turn out to be informative at a later stage of inquiry. There are two distinct philosophical views of the role of models in science: They might either serve an ancillary and subordinate function to axiomatic theories, in terms of making such theories more intelligible, or they might be an independent and necessary precondition for the formulation of axiomatic theories (cf. the dispute between Hesse's fictional 'Duhemist' and 'Campbellian', 1966). On the latter view, models, despite being idealising, approximative or in part even fictional, are foundational to science.

Given their purposeful partialness, models in science are designed in such a way that their complexity is limited to a degree that matches human cognitive skills while providing enough empirically adequate detail to enable an understanding of the phenomena. As the flip side of the same coin, the use of models in science also involves some degree of acceptance of epistemic opacity concerning those elements of a model which, at a given stage of inquiry, are deemed irrelevant to an empirically adequate representation of the phenomenon under investigation. However, as to the first point, simplification is not per se the aim of modelling in science (unlike, for example, the discussion in Sullivan 2019 seems to suggest). It is one among several possible ways of facilitating access to a phenomenon among others, and only in some types of models it takes centre stage. The relevant qualities of models are tailored to the aims and abilities of human cognitive agents by whatever means suitable. As to the second point, epistemic opacity is considered acceptable in a model only with respect to some of its internal aspects, and only with respect to those aspects which are not expected to interfere with proper recognition of the epistemically relevant elements, and therefore the representational qualities of the model. Under the interpretations outlined here, models are never supposed to be w-opaque or link-opaque, although they might be h-opaque to a certain, methodically circumscribed extent.

Computer models: Computer models are in important ways at variance with the kind of models described in the previous paragraphs. They rely on the distinctive feature of the digital computer 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, 1). Accordingly, computer models may establish any kind of modelling relation that lies within the domain of functions that are 'effectively calculable' or computable (Turing, 1936). Functions of this kind can be solved using a finite set of symbols, a finite set of possible states, a transition function and a potentially infinite memory.

A function is computable if and when these means jointly suffice to produce a correct solution in a finite number of discrete steps. With respect to its use in scientific modelling, the computational method is remarkable in at least three ways:

c.1 It is precise and determinate in method and specification;
c.2 It is simple and uniform in its basic elements and principles;
c.3 It is 'universal' with respect to all phenomena that are amenable to the requisite computational procedures.

In a relevant subset of cases, computational methods can be used to produce numerical solutions to problems that are altogether unsolvable by the analytic means of mathematics, in such a way that both the solutions and the paths towards those solutions remain beyond the reach of human cognitive agents (Humphreys, 2009).

If one goes by determinateness characteristic (*c*.1), one might expect computer models to be paradigms of epistemic transparency. On a naive conception, computer models should be more precise and more amenable to proof than the analogy-based, expressly partial and sometimes deliberately distorted models of pre-computational modern science. Given that epistemic opacity of AI is often framed as opacity of algorithms, whereas algorithms are at the same time defined as finite sequences of unequivocally defined discrete steps of effectively calculating a mathematical function (Markov, 1960; Kleene, 1967; Knuth, 1973), such a naive conception might look problematic.

If one goes by the characteristic of simplicity and uniformity (*c*.2), which hold on the level of basic operations, one might expect them to foster transparency in principle. In practice, however, they might give way to various dimensions and degrees of complexity and related problems with computational tractability. The degrees of complexity vary with the degrees of attainable algorithmic sophistication and computer power that affect the interactions between those simple and uniform elements. Under this view, opacity problems are particularly pertinent to computer models and complex AI applications that are not specifically designed to be intelligible, interpretable or explainable to human observers. In Machine Learning, the demands of mathematical optimisation and of human interpretability are expressly at cross purposes (Burrell, 2016). If and when computational complexity of a model increases, its mathematical tractability diminishes, so that the algorithm's operations and functions become more difficult or practically impossible to discern. However, if complexity can be reduced or modelled, or if one can devise other means of interpreting or otherwise meaningfully structuring that complexity, an opaque algorithm can be made more transparent. This is the premiss of the Explainable AI (XAI) paradigm (Gunning, 2019). Moreover, AI systems are designed under that paradigm to either facilitate or directly provide explanations of their concrete paths to a given solution.

If, however, one goes by the characteristic of universality (*c*.3), epistemic opacity starts at a more basic level of computer models and is harder if not impossible to resolve. Universality creates a specific epistemic problem that manifests even if and when the previous condition of simplicity and uniformity is fulfilled on the level of elementary operations of the computer model. Accordingly, higher-level complexity is not the beginning of the problem. If a multitude of types of phenomena can be modelled by the same limited set of computational tools, provided that the phenomenon in question is amenable to computational modelling at all, and if there is no similarity, isomorphism or sameness relation that one can reasonably expect to hold between that limited set of elements and the phenomenon, and if understanding a model requires a grasp of such relations, epistemic opacity is inherent to computer models as such. With one notable exception that I will discuss in the next paragraph, the computational method's indifference towards what human observers might learn from it is more fundamental under the condition of universality than in the case of complexity. To obtain modelling relations that rely on isomorphism or similarity as in non-computational models, the computer model's output will need to be interpreted, through visualisations or in other forms suitable to human epistemic agents. Thus, there is no representational relation between the elements of the model and elements of the target system that could be specified already on the basic computational level and that would match non-computational modelling relations.

In computer models, the epistemically relevant modelling relations can therefore be and need to be established only on the higher levels of interpreted output, which leads to two forms of underdetermination that are widely discussed in the literature on computational methods (Turing, 1936; Putnam, 1967, 1960): In the first case, 'Turing Universality', one can decompose various complex, higher-order operations into sets of computational elements that, as such, are type-identical. This is the type of relation indicated in the Turing (1946) quote on p. 5 above. In the second case, known as 'Multiple Realisability', one can decompose one complex, higher-order operation or set of operations into various sets of distinct computational elements that jointly perform the same functions. In either case, the properties of the computational elements of the model do not unequivocally determine the model's higher-level properties. An at least partial exception to these two underdetermination problems holds if and when some of the computational operations in the model are supposed to bear isomorphism, similarity or sameness relations to a given phenomenon. In this kind of case, the assumption is that the modelled phenomenon itself is in a relevant sense computational or that some of its key elements have key properties in common with computational processes.

One might object against this line of reasoning that it will be equally difficult to infer the properties of an analogue model from the properties of its elements, but the relevant difference is this: An analogue model and its elements are chosen or designed in light of a perceived or expected similarity, isomorphism or sameness relation to a phenomenon. The elements are selected because some of their properties are expected to do some specific part of the representational work of the model as a whole. In contrast, there is no selection for such representational qualities in the computational elements of a computer model, that is, the algorithms, data sets or computer architecture. Their relevant representational qualities can only be and must be entirely derived from the overall structure of the model. To use an analogy: Where the elements of an analogue model play a representational role that is similar to that of the elements of an image, where one stroke of a brush might ideally suffice to represent an object or feature, so that those elements individually carry some meaning and jointly shape the meaning of the overall image, the role of the computational elements of a computer model is more like that of the letters of an alphabet, which individually carry no such meaning. Only the words and sentences in a language that uses these letters do so. Only from there, the role of the individual exemplars of letters in carrying that meaning can be reconstructed. Otherwise, there is little about the language that one could infer from the limited set of letters of an alphabet alone.

The connection: If we map the previous analysis on the question of epistemic opacity, it seems to fall into two distinct but related problems (see I. and II., p. 1–2 above):

I′. Complexity: If opacity is mainly a problem of complexity or computational tractability more generally, it concerns the internal properties of a model. To the extent that complexity can be reduced or modelled, opacity can also be resolved on the level of internal properties. The expectation is that if and to the extent that problems with the internal properties of the model (h-opacity) can be resolved, opacity on the representational level (and therefore w-opacity) can be resolved, too.

II′. Universality: If opacity is mainly a problem of the universal applicability of the computational method, a different kind of epistemic uncertainty is intrinsic to computer models. It manifests on the level of a model's internal properties, but it directly affects the representational properties of the model as a whole (and therefore is w-opaque). If and to the extent that this type of opacity is inherent to computer models, it can either be resolved by recourse to external resources for interpreting the model's output and its relation to some phenomenon, or it cannot be resolved at all.

The universality condition is conceptually independent of the complexity condition. In practice, they frequently interact though. The basic diagnosis is this: Under both I′. and II′., each computational element of a model might individually be or be made epistemically (h-) transparent. It might be possible to know how and by what rules that element operates, and how it interacts with the other computational elements. However, this knowledge alone does not ensure the model's overall epistemic transparency under an internal perspective. According to I′., the interactions between the elements might be too complex. Nor does that knowledge suffice to infer the representational relations of the model as a whole. According to II′., this condition is imposed by the model's computational nature and cannot be resolved by internal means.

The most pertinent case of interactions between the complexity and the universality variable are Deep Neural Network approaches in AI (DNNs; Goodfellow et al. 2016; Krizhevsky et al. 2012; LeCun et al. 2015; Schmidhuber 2015), where complex and intentionally biologically implausible connectionist architectures are used to generate classifications or predictions from large data sets. DNN methods accomplish this in two distinctive ways: First, their models are 'generative', that is, produced by the networks themselves rather than being explicitly provided to them. Second, DNN architectures purposefully abstain from plausible analogies to natural forms of neuronal information processing. By virtue of these features, above and beyond their complexity, DNNs provide little guidance to human observers as to how they arrive at a certain classification or prediction, and how it is supposed to represent aspects of the phenomenon in question. For being maximally complex on top of being exquisitely computational, they appear to be inaccessible to the external means of fixing representational relations that make other forms of computer models empirically meaningful. Reducing or modelling their complexity in view of a certain epistemic aim is a difficult task if that epistemic aim itself remains uncertain.

Conversely, the most pertinent case of an approach that assumes an analogy-based modelling relation between some of its computational elements and the phenomenon under investigation is the Predictive Processing paradigm (PP; Clark 2013; Dayan et al. 1995; Hohwy 2013, 2020). It uses connectionist models that are in terms of computational architecture closely related to DNNs in order to explain the functional principles of cortical information processing in humans and other higher animals. In the PP context, cortical information processing is modelled as a bi-directional hierarchical process of prediction error minimisation between sensory input and higher-order dynamic world models. Sensory information and world models are respectively understood as the input and output layers of a neural network. In similar fashion to DNNs, PP networks involve a generative model-building process, but unlike DNNs, their network architectures and their elements are geared towards biological realism. In turn, the error minimisation strategies in PP networks are modelled on probabilistic techniques of data compression in computer engineering. Accordingly, cortical information processing is modelled as a *sui generis* computational process, with pertinent analogies holding, if not on the level of elementary computations, then on higher levels of computational architectures and operational principles.

Conclusion: Models are entities that work for concrete epistemic agents. Their purpose is to make phenomena intelligible for those agents. Whereas the computational method is supposedly universal, its very universality makes computer models prima facie indifferent to the needs and aims of concrete epistemic agents. Where PP is an approach that postulates partial analogies between computational architectures and cortical information processing, and thereby propose a model that seeks to facilitate epistemic access to a certain circumscribed domain of phenomena in cognitive inquiries, DNNs, along with other Machine Learning approaches, are the paradigm of an approach to computer modelling where indifference towards the needs and aims of human epistemic agents is elevated to an operational principle. If we go by the complexity condition outlined in I′. above, all their internal opacity-qua-complexity might be resolvable in principle, but the requisite solutions might not be reachable or useful for human epistemic agents. If we go by the universality condition in II′. above, human epistemic agents might not be the natural addressee of those models. In limiting cases, the solutions offered are per se beyond the scope of human-made analysis. In some cases and under certain conditions, the models might be tailored to human epistemic purposes by establishing some form of analogy relation and thereby be made more transparent and scientifically useful. In other cases, this aim will remain stubbornly elusive.

Acknowledgements

This work is dedicated to the memory of my dear and trusted PW colleague Helena Bulińska-Stangrecka, who tragically and prematurely died when I was finalising my manuscript. In terms of content, this paper owes a lot to my collaboration with Alessandro Facchini and Alberto Termine, who tried hard to rid me of my philosophers' naiveté concerning how AI works. The same goes for Cameron Buckner, Holger Lyre, Jan Passoth and Carlos Zednik, who worked towards that goal earlier. Any remaining naiveté will not be the fault of any of those helpful minds.

References

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science.* Pittsburgh University Press, Pittsburgh.

Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, 199(3):11643–11666.

Black, M. (1962). *Models and Metaphors.* Cornell University Press, Ithaca.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines.*

Boltzmann, L. (1902). Model. In Wallace, D. M., Hadley, A. T., and Chisholm, H., editors, *Encyclopaedia Britannica*, volume 30, pages 788–791. Adam and Charles Black, The Times, London, 10 edition.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):1–12.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):1–73.

da Costa, N. and French, S. (2003). *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning.* Oxford University Press, Oxford/New York.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, 7(5):889–904.

Facchini, A. and Termine, A. (2022). A first contextual taxonomy for the opacity of ai systems. In Müller, V. C., editor, *Philosophy and Theory of Artificial Intelligence 2021.*

Frigg, R. and Hartmann, S. (2020). Models in Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*, page html. Metaphysics Research Lab, Stanford, spring 2020 edition.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning.* MIT Press, Cambridge.

Gunning, D. (2019). DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii, New York. ACM.

Hesse, M. B. (1966). *Models and Analogies in Science.* University of Notre Dame Press, Notre Dame.

Hohwy, J. (2013). *The Predictive Mind.* Oxford University Press, Oxford.

Hohwy, J. (2020). New directions in predictive processing. *Mind & Language*, 35(2):209–223.

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method.* Oxford University Press, Oxford.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169:615–626.

Kleene, S. C. (1967). *Mathematical Logic.* Wiley, New York.

Knuth, D. E. (1973). *The Art of Computer Programming*, volume 1. Addison-Wesley, Reading, 2 edition.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, volume 1, pages 1097–1105, Lake Tahoe. Curran Associates.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)? – a atakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296:103473.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, 521:436–444.

Markov, A. (1960). Theory of algorithms. *American Mathematical Society Translations*, 15.

Morgan, M. S. and Morrison, M., editors (1999). *Models as Mediators. Perspectives on Natural and Social Science.* Cambridge University Press, Cambridge.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Putnam, H. (1960). Minds and machines. In Hook, S., editor, *Dimensions of Minds*, pages 138–164. New York University Press, New York.

Putnam, H. (1967). Psychological predicates. In Capitan, W. H. and Merrill, D. D., editors, *Art, Mind and Religion*, pages 37–48. University of Pittsburgh Press, Pittsburgh.

Robbins, P. and Aydede, M., editors (2009). *The Cambridge Handbook of Situated Cognition.* Cambridge University Press, Cambridge.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*, online first:1.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv*, 1806.07552.

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.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34:265–288.