Towards a Taxonomy for the Opacity of AI Systems

Alessandro Facchini[1] and Alberto Termine[2†]

[1]Dalle Molle Institute for Artificial Intelligence, USI-SUPSI, Lugano-Viganello, 6962, Ticino, Switzerland.
[2]Department of Philosophy Piero Martinetti, University of Milan, Milan, 20122, Lombardy, Italy.

Contributing authors: alessandro.facchini@idsia.ch; alberto.termine@unimi.it;
[†]Both authors contributed equally to this work.

Abstract

The research program of eXplainable AI (XAI) has been developed with the aim of providing tools and methods for reducing opacity and making AI systems more humanly understandable. Unfortunately, the majority of XAI scholars actually classify a system as more or less opaque by confronting it with traditional AI systems such as linear regression models or rules-based systems, which are usually assumed to be the prototype of transparent systems. In doing so, the concept of opacity remains unexplained. To overcome this issue, we view opacity as a concept whose meaning depends on the context of application, and on the purposes and characteristics of its users. Based on this, in this work, we distinguish between access opacity, link opacity and semantic opacity, hence providing the groundwork for a taxonomy of the concept of opacity for AI systems.

Keywords: Opacity, Explainable AI, Taxonomy, Scientific Understanding

1 The many faces of opacity

The incredible success of artificial intelligence (AI) systems is mostly a consequence of the recent advancements in machine learning (ML)[1]. Unlike more traditional AI, ML systems possess an impressive inferential power that allows them to analyze large amounts of data and identify patterns that neither the human eye nor traditional statistical methods would likely ever be able to discover [5]. Unfortunately, these systems suffer from the problem of being opaque, or, as they say, of being 'black boxes'. Roughly speaking, that an AI system is opaque means that it is difficult for users to know how it works, as well as to interpret its decisions at various levels and evaluate its behaviour against scientific and ethical norms [6].

Given its impact on various spheres of contemporary society, the opacity problem has recently caught the attention of many scholars, both from engineering, philosophy, and the social sciences. In general, engineers have directed their efforts towards the development of methods and tools to mitigate opacity and obtain *explainable* AI systems [7, 8]. Philosophers and social scientists, on the other hand, have focused on analysing the concept of opacity, as well as its epistemological, ethical, social and legal implications [6, 9–12]. Together, their joint efforts have led to the birth of *eXplainable AI* (XAI), a new area of research aimed at rendering AI systems less opaque and more humanly understandable [13].

Despite the extensive technical and philosophical literature on the subject, a complete and in-depth analysis aimed at clarifying what it means for an AI system to be opaque is still lacking.[2] In the technical literature, there exist a sort of 'received view' that is implicitly adopted by the majority of XAI scholars[3]. This view tends to consider opacity as an *intrinsic* property of ML systems. More specifically, it sees opacity as a by-product of their inner complexity [4] and of their sub-symbolic nature. Consistently with these considerations, a deep learning neural network (DNN) is thus commonly described as a 'black box'[5] whereas "[i]n the state of the art, a small set of existing interpretable [transparent] models is recognized: decision tree, rules [rule-based systems], linear models"[8, p.7] since these 'traditional' models are considered easily understandable and interpretable for humans. Analogously, in [14, Sec. 2.5.1], the authors claim that "[t]ransparent models convey some degree of interpretability by themselves", that is their users can immediately understand the process followed by the models to produce any given output, each of their parts can be explained, and the models can be "simulated or thought about strictly by a human", as it is precisely the case with linear/logistic regression models, decision trees, and rule based systems.

[1] ML can be defined as the study of how to make artificial agents able to extract information, learn knowledge and build models from data by themselves. It constitute a broad field, that includes a wide variety of tools and techniques, ranging from logical [1], kernel [2] and graph-analysis methods [3] to deep learning [4].

[2] See e.g. [11].

[3] See e.g. in [7, 8, 14–16].

[4] e.g., the number of nodes in a circuit.

[5] This is the term used to refer to opaque AI systems in the debate.

The problem with the received view is its inability to go beyond these 'naive' considerations and explicitly recognize the relevance of the specific contexts and purposes for which an AI system is used. In fact, it is not uncommon that stakeholders deem a system transparent in one context but opaque in another. Consider for instance DNNs. Consistently with the received view, in most contexts such complex sub-symbolic systems are considered opaque. In the computational cognitive neuroscience of human vision though, contrary to the received view, scholars regard DNNs as being more transparent than rules-based or symbolic models. The reason is that, contrary to the latter, the former not only simulates high-level processes related to human vision but are also able to explain how a purely sub-symbolic architecture, working similarly to the human brain, can implement them [17]. At the same time, there are contexts in which both complex sub-symbolic models and symbolic ones are deem equally opaque because neither is able to provide information relevant for their purposes. Consider, for instance, a molecular biologist interested in understanding the occurrence of a given pathological phenotype starting from the analysis of genome mutations. They will probably hold both a decision tree and a DNN to be equally opaque since neither sheds light on the mechanisms leading genome mutations to cause the occurrence of the pathological phenotype.

At this point, a question spontaneously arises: why do many XAI scholars follow the received view and tend to regard a given system as intrinsically more opaque than another one? The reason may be due to the fact that most of XAI scholars are computer scientists or work in the context of information technologies (IT). Their specific purposes usually consist of understanding the internal functioning and checking the reliability of AI systems, both tasks easier to perform on simple symbolic models than complex sub-symbolic ones. However, nowadays AI systems are deployed and employed in contexts very different from those of IT and by users whose background is not necessarily in computer science. The reasons and meanings of 'opacity' change depending on the considered context and stakeholders. In particular, what stakeholders mean by saying that an AI system is opaque in a given context depends on the nature, extent and characteristics of their purposes, their background knowledge, and their cognitive abilities.

Opacity is thus a plural concept, admitting many different forms, whose clarification and characterisation constitutes a crucial philosophical work. Nonetheless, at the moment there are very few attempts available to carry out a conceptual analysis of the notion of opacity. To our knowledge, the most relevant are those proposed by [9] and by [18]. Specifically, in her analysis, [9] identifies three forms of opacity: 'opacity as intentional corporate or state secrecy', 'opacity as technical illiteracy', and 'opacity as the way algorithms operate at the scale of application', each related to a different source. On the other hand, [18] distinguishes between 'run opacity', 'structural opacity', and 'algorithmic opacity'. Each one related to a given level at which she deems possible to describe the structure and functioning of AI systems. More specifically,

run opacity arises with the physical execution level, structural opacity arises with the implementation level, and algorithmic opacity arises with the abstract level. In our opinion, both of these taxonomies are not detailed enough. In a certain sense, they still suffer from the perspectivism of the received-view. By focusing on the inner structure and functioning of an AI system, they omit to explore aspects that stakeholders working in contexts other than IT and having a background other than computer science can deem essential, as, for example, the quality and the format of the information that an AI system may learn from data.

Taking these considerations into account, in this work we propose a taxonomy that identifies three principal forms of opacity, each being prone to a further deeper analysis. We call them, respectively, *access opacity*, *link opacity* and *semantic opacity*.

2 Access opacity

Access opacity concerns the capability of understanding the structure and functioning of an AI system. It manifests itself when human stakeholders have limited epistemic access to elements that are relevant for explaining, predicting, and controlling the behavior of the considered system.[6] Notice that by "having an epistemic access to an element", we mean to figure out the location of the element and the functional role it plays in the overall structure and functioning of the system.

We identify three main factors that may limit epistemic access and thus cause access opacity.[7] The first coincides with the transparency policies adopted by the system's designers, who might deliberately obscure some relevant details of the system's structure and functioning for either commercial, competition, or privacy reasons. The second is related to the stakeholder's background knowledge and skills. Intuitively, the more a stakeholder is familiar with a given AI system, the more they can understand, predict and control the system's behavior. Finally, the third arises with the complexity of the system's structure, conceived as a function of both the system's size [8] and format [9] [20]. The intuition is that, as human stakeholders possess limited cognitive resources, their ability to explain, predict and control the system's structure and functioning decreases as the complexity of the system increases.

Once clarified these general aspects, we are ready to deepen some details. In doing so, we will analyze the different forms in which access opacity may occur and identify the specific causes related to each of them.

First of all, we should note that an AI system based on machine learning techniques is a complex computational architecture that includes distinct components, the main ones being:

1. the *training sample*: the data-set used to train the system.

[6] The notion of 'epistemically relevant element' is borrowed from [19].
[7] They are related with the three forms of opacity described by [9].
[8] I.e., the *number* of its elements and their mutual relations.
[9] I.e., the *type* of elements it includes and how they are related.

2. the *training engine*: the set of procedures that allow the system to learn from data during the training process;
3. the *learned model*: the final model obtained from data after running the training process on a specific training sample;

Each component plays a fundamental role in determining the overall behaviour of the AI system and, as we will clarify in the following, it is related to a specific form of access opacity.

2.1 Opacity of the training sample

This form of access opacity occurs when stakeholders have limited epistemic access to the data included in the samples used to train the model. There are several circumstances where this may happen. A first circumstance is when the system's constructors decide to not adopt data transparency policies, and therefore do not provide or partially hide the training sample, generally because of ethical or commercial reasons. A second circumstance is when stakeholders have difficulties to interpret the training sample and check the reliability of the data it contains because of its complexity. This scenario is very common when dealing with big data samples.[10] The large size and the variety of data-formats these contain, in fact, makes it hard to check their reliability and identify potential sources of mis-training [22]. A third circumstance occurs when the training process takes place in an open environment, such as the web or the specific part of the world an autonomous robot or a self-driving car is interacting with. In general, to determine in retrospect what data influence the training process in an open environment is practically impossible. The risk that training a system in an open environment produces undetectable biases in its beliefs and behavior is therefore high.[11] Finally, a fourth circumstance is when a stakeholder cannot make sense of the data included in the training samples because the transformations applied during the construction of the training samples bring them into an incomprehensible format. This scenario is very common when dealing with DNNs. In fact, the DNNs specific training procedures cannot generally be applied to raw data but require these to be mapped in an adequate space, technically called the *features space*. In many cases, the transformations applied to map the raw data into the feature space alter its format so much that they eventually result incomprehensible to stakeholders. Furthermore, these transformations are usually irreversible, making impossible for stakeholders to go from the features space back to raw data.

[10] For an overview of the different meanings of the term big data, see [21].

[11] The Google tool *Quickdraw* provides a good example of an AI system trained in an open environment.

2.2 Opacity of the training engine and of the learned model

In technical language, *training engine* refers to the computational architecture that allows an AI system to learn from data. *Learned model* [12], instead, refers to the model of data obtained by training the AI system on a given sample through a proper engine. Differently from the training sample, which is technically a database, the training engine and the learned model are computational artifacts[13].

According to a widely accepted tradition in the philosophy of computer science, the structure and functioning of computational artifacts are usually described and understood at different *levels of abstraction* (LoA for short), namely collections of interpreted type variables, each one modeling an entity or activity relevant to characterizing the structure and functioning of the artifact [24, 25]. [26], in particular, distinguish between five different LoAs:

1. The Functional Specification Level (FSL), consisting of a very general description of the artifact's architecture that includes a specification of the various *functions* it computes and that are responsible for its overall functioning
2. The Design Specification Level (DSL) specifying, generally in terms of *state-transition systems*, the procedures for computing the functions identified at the FSL
3. The Algorithm Design Level (ADL) consisting of the algorithmic (operational) specifications, generally in terms of *rules*, of the procedures specified at the DSL
4. The Algorithm Implementation Level (AIL) consisting of the translations in terms of programs of the algorithms specified at the ADL
5. The Algorithm Execution Level (AEL), consisting of the physical executions, on hardware, of the programs specified at the AIL

Each LoA provides a different description of the artifact's structure and functioning, which may be suitable and relevant for some stakeholders but insufficient or inadequate for others.[14] For instance, a molecular biologist interested in using AI to predict cancer will probably deem sufficiently detailed a description provided at the FSL. Whereas a computer scientist in charge of checking the reliability of the training procedures, or the learned model, will probably be interested in a more fine-grained description that may also include details about the algorithms (ADL), the programs (AIL), and even the hardware (AEL). Accordingly, whether and to what extent an artifact is opaque depends both on who the stakeholders are and which LoAs are accessible to

[12] Notice that, here we use the term *model* in a very broad sense. In fact, the specific nature of the learned model varies depending on the AI system and the kind of ML methods applied, some methods (e.g., Kernel methods) produce models that are nothing but *compact descriptions of data*, whereas others generate *predictive models* that can be used to predict phenomena from data.

[13] For a philosophical perspective on this concept, see [23].

[14] In this respect, our taxonomy extends that proposed by [18].

them. In general, stakeholders deem opaque an artifact if they have limited epistemic access to the LoAs suitable for their background knowledge and relevant to their purposes. This reasoning holds both for the training engine and the learned model as both are computational artifacts.

3 Link opacity

Link opacity concerns the use of AI systems to model phenomena in scientific research. It occurs when a system that is used to model a given target phenomenon conveys inadequate or insufficient information about the elements that are relevant for explaining, predicting, and controlling such a target phenomenon.[15]

In general, ML-based AI systems are very good at extracting information from large amounts of data and generating highly accurate predictive models without the necessity of background knowledge or human intuition. This ability confers them a clear advantage over more traditional tools in the study of highly complex phenomena [16] that represent the target of much contemporary science. For this reason, these systems have quickly spread in several sectors of scientific research, leading to a progressive replacement of the standard scientific methodology[17] with a data-centric approach based on the collection and the AI-supported analysis of observational data [28].

In reality, as we have recently argued in [29], we can distinguish between two different kinds of data-centric approaches to scientific research: a data-informed one, which preserves the classical models and ways of scientific explanation despite the intensive use of AI systems to perform statistical analyses, and a fully data-driven one, characterized by the full replacement of classical models with models learned from data. In this latter approach, models learned by AI systems are not considered mere statistical tools but representations of the target phenomena, hence acquiring a fundamental role in understanding and predicting them. Unfortunately, although being powerful from a predictive point of view, models learned by ML-based AI are often unable to provide sufficient information to explain *how* and *why* the target phenomena occur and to figure out ways for controlling them [16]. The reason is that, while AI is very good at learning how to correlate observational data with predictions, it is also typically unable to figure out the *causes*, the *mechanisms* and the *laws* beyond observed phenomena. This point highlights a huge epistemic limitation of the fully data-driven approach. In fact, regardless of whether one takes a realistic or instrumentalist stance towards scientific knowledge, scientific understanding usually requires more than mere statistical associations. It needs information

[15] We call this form of opacity 'link opacity' to emphasise the fact that it undermines our ability to establish a link between the model and the phenomenon it is intended to represent. Stated otherwise, it undermines our ability to establish whether a model is an 'actual' representation of the target phenomenon or just a possible one. In this regard, the notion of 'link-opacity' resembles that of 'link-uncertainty'. Introduced in [27], link-uncertainty concerns the extent "to which [a ML] model fails to be empirically supported and adequately linked to the target phenomena".

[16] E.g., the fluctuations in financial markets in economics or gene regulation in biology.

[17] By *standard scientific methodology* we mean the approach based on the formulation and the experimental evaluation of hypotheses explaining the observable facts.

about the causes of the phenomena, the mechanisms that produce them, and the laws that regulate their functioning. As argued by [1], the lack of this type of information impedes our ability to explain, predict and control the target phenomenon, and thus to achieve what he calls *pragmatic understanding*. For this reason, when an AI system is unable to provide scientists with information that is essential for the pragmatic understanding of a phenomenon, they tend to consider it as opaque.

Notice that, similar to access opacity, link opacity also occurs in different forms as the elements that are relevant for explaining, predicting, and controlling a given target phenomenon vary depending on the nature of the phenomenon under consideration. In general, we may identify three main forms of link-opacity, each related with one of the three fundamental notions that were previously mentioned: cause, mechanism and law. We refer to these forms, respectively, as *causal opacity*, *opacity of the mechanisms* and *opacity of the laws*.

3.1 Causal Opacity

Causal opacity occurs when an AI system cannot reconstruct the causal chains beyond the predicted phenomena. The identification of the causal chains leading to the occurrence of a given phenomenon is essential for its understanding because it allows scientists:

- to distinguish between the variables necessary and sufficient for the occurrence of the target phenomenon from those that are merely related to it,
- to predict the effects generated by external interventions and, therefore, to understand how to control the target phenomenon by acting on the variables related to it,
- to distinguish between genuine statistical correlations,[18] grounded on the existence of actual cause-effect links, and spurious ones, which are the by-product of statistical paradoxes[19].

As pointed out in [31, 32], the ability of computational systems to recognize causal chains strictly depends on their ability "to choreograph a parsimonious and modular representation of their environment, interrogate that representation, distort it by acts of imagination and finally answer 'What if?' kind of questions"[32, p.1]. These, in particular, may be either statistical, interventional, or counterfactual questions. The former arise with the statistical regularities observed in the naked data and have the form "what if I see x?"[20]. The second one arises with the consequences of intervention and has the form "what if I do x?"[21], while the latter arises with some counterfactual state of

[18] We do not mean here that a statistical correlation between a variable x and a phenomenon y is genuine if and only if x is a (proximal or distal) cause of y. Instead, the correlation between x and y is genuine even if x is related to y because of a common cause or effect.
[19] A famous example is the well-known Simpson's paradox, see [30, 31]
[20] For example, "what if I see salt in the water?"
[21] For example, "what if I add salt to the water?"

affairs and have the form "what if I had done x?"[22] or the contrastive form "what if I had done y instead of x?"[23]. Causal information is classifiable in terms of the kind of *what-if* questions it can answer. The classification generates a three-layers hierarchy where "questions at the level i (with $i = 1, 2, 3$) can be answered if and only if information from level $j \geq i$ is available"[32, p.1]. The three layers are respectively the *association layer* (AL), the *intervention layer* (IL) and the *counterfactual layer* (CL). Information about statistical regularities is enough for answering questions at the AL and can be inferred directly from the observational data using conditional expectation. At the IL the information requested no longer concerns only what we observe but what we can observe if we perform a certain action. At the CL, it concerns what we would have observed if a certain condition that did not occur had occurred. We can infer this information by using particular inference engines called *Structural Causal Models* (SCM), which, however, require more than naked data. In particular, they require some background hypotheses usually encoded in the form of a graphical diagram[24].

Available ML-based AI systems usually work at the AL. They do not possess imagination and thus cannot figure out hypotheses beyond the observed data. This inability prevents them from learning causal models and is a reason for their link opacity.

3.2 Mechanisms Opacity

In many fields of science, it is common to understand phenomena in terms of mechanisms, i.e., "entities and activities organized in such a way that they are responsible for the phenomenon." [33, p. 120]. The reason is that thinking in terms of mechanisms presents some clear epistemological advantages. It permits to manage with complexity and lead highly-complex phenomena back to simpler, more fundamental facts [34]. It allows us to provide an explanation by stating a description, as "[by] providing a description of the mechanism responsible for a phenomenon, one provides an explanation for *why* that particular phenomenon occurs and *why* it has the proprieties it does" [35, p.217]. Finally, it supports generalization because mechanisms "work in the same or similar way under the same or similar conditions" [36, p.19]. Formulating a mechanistic explanation, however, needs much more than mere observational data. It requires to hypothesize what simpler, more fundamental entities and activities may produce the target phenomenon by interacting with one another. The reason is that mechanistic thinking relies on heuristics that are very different from those used to train AI systems. Actually, the nature of these heuristics is a matter of debate. In their famous work, [34] identify two mains reasoning strategies followed by scientists to identify mechanisms' structure and functioning, which are *decomposition* and *localization*. Roughly, the former consists

[22] For example, "what if I had added salt to the water?"
[23] For example, "what if I had added sugar instead of salt?"
[24] On this topic, see [30–32]

of decomposing the overall phenomenon into low-level activities while the latter consists of localizing these activities in components of the system identified as responsible for producing the target phenomenon. According to [36] instead, scientists apply several different techniques to refine a raw hypothesis about the mechanism's structure and functioning, generally in the form of a sketch representation of it full of black boxes, until obtaining a sufficiently clear and detailed description. Regardless of the details, in both cases, the information necessary to understand a mechanism requires hypotheses that cannot be inferred from mere observational data but need a fundamental contribution of the imagination. Since the heuristics implemented in AI systems are unable to formulate this type of hypotheses, these systems cannot generate mechanistic explanations, and are eventually considered link-opaque.

3.3 Laws Opacity

Since its birth, discovering the laws that govern phenomena has represented a fundamental aim of science. From an epistemological point of view, scientific laws are essential to the understanding of phenomena. They allow scientists to explain why phenomena occur in a way rather than another, to predict under what circumstances they occurs, and to figure out how to act for controlling their occurrence. Philosophers of science have long debated the nature of laws taking sides on two opposing positions, the instrumentalist position and the realist one.[25] A discussion of the details of these specific positions is beyond the scope of this work. Here we simply note that, in scientific practice, the term 'scientific law' may refer to different things. In some cases, 'law' denote sentences that describe mere patterns of regularities between observable variables. An example is Charles' law in thermodynamics, which shows the relationship between the volume and the temperature of a gas. These kinds of laws do not substantially differ from the models learned by AI and, indeed, a ML-based AI system might easily infer Charles' law by analyzing a sufficiently large sample of data. In other cases, a law is instead a description of the structural relationships between observable variables and variables that:

- denote unobservable entities, whose existence scientists theoretically hypothesize but cannot statistically infer from observational data,
- scientists consider the main causes of the target class of phenomena.

Gauss's law, which relates the electric charge and the magnitude of the electric field [26], is an example of the latter.

Both types of law coexist in scientific practice, but scientists tend to consider laws of the second type epistemologically more relevant. Interestingly, the reason is not that they believe in the actual existence of unobservable entities, but because these laws allows them to bring the observed phenomena back into a single representation of reality and figure out how to control their

[25] On the debate about instrumentalists and realists, see [37].

[26] The electric field is an unobservable entity theoretically hypothesized to explain remote interaction among particles

occurrence. The epistemological value of these laws is therefore independent from the 'realists vs instrumentalists' debate and have pragmatical roots. For this reason, a science including only the first type of laws is very difficult, and maybe impossible, to imagine.

Unfortunately for AI systems, the identification of laws of the second kind is a purely theoretical work. It relies on the human mind's ability to go beyond the observable phenomena and figure out the supposed basic structure of reality might consist of. ML-based AI systems do not possess this ability, and as a result scientists tend to regard them as opaque.

4 Semantic Opacity

Semantic opacity concerns the semantic aspects of both the information that an AI system learns from data and the inferences the system performs to manipulate the it. Notice that, in information theory, it is common to distinguish between a *structural* and *semantic* aspect of information. The former concerns the mathematical and physical properties of information, whereas the latter concerns its meaning. In the case of ML-based AI systems, the structural aspect coincides with the mathematical and physical properties of the model that the system learns from data. This aspect is relevant for understanding how the learned model works and, therefore, it is closely related to the problem of the access opacity of the learned model.

The semantic aspect, instead, coincides with the semantic interpretation of the learned model. Contrary to the structural one, it is not directly relevant for determining the functioning of the model. It is however fundamental for the ability of the stakeholders to interpret the information learned by the AI system from the data and how it manipulates the information. Semantic opacity is precisely linked to this latter aspect of information. It occurs because the format used to store and manipulate information prevents stakeholders from giving a meaning to the information learned from data.

Semantic opacity can occur in three different circumstances. First, it can manifest itself when the learned model lacks a clear, well-defined semantic interpretation that allows stakeholders to make sense of both the information it stores and the inferences it performs. Second, it may take place when a semantic for the learned model is available but it is not comprehensible because of the stakeholders' limited cognitive resources, inadequate background knowledge, or lack of relevant epistemic skills. Third, it can arise when the semantics of the learned model provides the stored information with a meaning that is inadequate for the context.

In what follows we distinguish between two forms of semantic opacity. The first one concerns the content of the information learned by a model, whereas the second one concerns the inferences used to manipulate such information. Consequently, we call these two forms of opacity 'content opacity' and 'inferential opacity' respectively.

4.1 Content opacity

This form of opacity occurs when the format used by an AI system to store information prevents stakeholders from grasping its semantic content and use it for their own purposes. Notice that each type of AI system adopts a peculiar format to represent the information learned from the data. For example, a rule-based system stores information by using string of literals stemming from some formal language . A deep neural network, instead, stores information using statistical functions and parameters [16]. In the former example, a Tarskian semantics mapping the appropriate elements to the entities in the domain they represent easily provide the information stored by string of literals with a clear, well-defined, meaning. This is not the case for the latter example. In a DNN, in fact, the values of numerical functions and parameters have a mere instrumental meaning (they are those values that allow the network to minimize the prediction error). They lack any standard semantic interpretation, making DNNs prone to this form of semantic opacity.

4.2 Inferential opacity

This form of opacity manifests itself when the format of the inferences used by an AI system to manipulate information prevents stakeholders from making sense of the reasoning paths it follows. As for the format used to represent the information learned from the data, each type of AI system uses a specific kind of inferences to manipulate information. A rules-based system for instance manipulates information by applying syntactic rules to strings of literals, whereas a DNN uses numerical calculations applied to statistical functions and parameters. Unfortunately, inferences do not always come provided with a semantic interpretation that allows stakeholders to reconstruct in understandable terms the reasoning followed by the system. In some cases inferences are meaningless because they have a purely instrumental value, lacking any semantic interpretation. In other cases, they may posses a well-defined semantics that is however incomprehensible to a stakeholder because of their limited cognitive resources, their inadequate background knowledge or their lack of fundamental skills. In all these circumstances, we say that the inferences performed by the system under consideration are semantically opaque.

5 Dependencies between forms of opacity

This last section briefly explores the mutual dependencies among the forms of opacity that have been introduced.

First of all, notice that the three macro-forms of opacity are conceptually and logically independent. That is, none of them is definable in terms of another, and none of them represents a necessary or sufficient condition for the occurrence of another. Nevertheless, there may be circumstances in which different forms of opacity can influence each other. In what follows we summarize some of them.

The learned model is usually the part of an AI system that provides scientists with the information they need to understand a given phenomenon. Accordingly, having limited epistemic access to the inner structure and behaviour of the learned model may prevent scientists from obtaining enough information to understand the target phenomenon and thus contribute to the system's link-opacity. More specifically, the access opacity of a learned model may cause link-opacity whenever the stakeholders' epistemic access to the LoA providing the information that is relevant for the understanding of the target phenomenon is limited. For similar reasons, the access opacity of the learned model may cause semantic opacity.

As already mentioned, semantic opacity is strictly related to the stakeholders' ability to give a semantic interpretation to the LoAs of the learned model that are relevant to their purposes. This ability may be compromised by a limited epistemic access to the concerned LoAs and therefore cause semantic opacity.

Finally, there exists a fundamental relation between semantic opacity and link-opacity. In particular, semantic opacity causes link-opacity whenever a stakeholder cannot provide a clear and well-defined semantic interpretation to the LoAs of the learned model that are relevant for understanding the target phenomena.

6 Conclusions

Starting from the crucial observation that what stakeholders mean by saying that an AI system is opaque in a given context depends on the nature, extent and characteristics of their purposes, their background knowledge, and their cognitive abilities, we identified three conceptually and logically independent macro-forms of opacity: access opacity, link opacity and semantic opacity, and analysed their possible specific instantiations, as well as dependencies. As a result, we provided a first, albeit partial, taxonomy for the opacity of AI systems, considered as a contextual, plural concept. As such, the taxonomy goes beyond the received view that focuses on the inner structure and functioning of an AI system.

However, much still needs to be done. In particular, in addition to broadening the proposed taxonomy and deepening its analysis, it would be interesting, for example, to associate relevant existing XAI methods and tools with each of its members and specific context. It would also be interesting to apply the taxonomy to shed lights on the impact of machine learning in data-centric sciences, and in particular on the scientific understanding of phenomena. In fact, from this perspective, ultimately our goal is to show that contemporary XAI methods and tools can help reduce relevant forms of opacity that are limiting the integration of data-driven approaches with established standards of scientific explanation and understanding.

Acknowledgment

We thank the participants to PT-AI 2021 and the reviewers for their constructive feedback.

References

[1] De Raedt, L.: Logical and Relational Learning. Springer, Cham (2008)

[2] Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. The annals of statistics **36**(3), 1171–1220 (2008)

[3] Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence. CRC press, Boca Raton, FL (2010)

[4] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, Cambridge, MA (2016)

[5] Alpaydin, E.: Machine Learning, Revised And Updated Edition. MIT Press, Cambridge, MA (2021)

[6] Zednik, C.: Solving the black box problem: a normative framework for explainable artificial intelligence. Philosophy & Technology, 1–24 (2019)

[7] Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

[8] Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

[9] Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society **3**(1), 2053951715622512 (2016)

[10] Héder, M.: The epistemic opacity of autonomous systems and the ethical consequences. AI & SOCIETY, 1–9 (2020)

[11] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)

[12] Durán, J.M., Formanek, N.: Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds and Machines **28**(4), 645–666 (2018)

[13] Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R.: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning vol. 11700. Springer, Cham (2019)

[14] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)

[15] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

[16] Baldi, P.: Deep Learning in Science. Cambridge University Press, Cambridge (2021)

[17] Cichy, R.M., Kaiser, D.: Deep neural networks as scientific models. Trends in cognitive sciences **23**(4), 305–317 (2019)

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

[19] Humphreys, P.: The philosophical novelty of computer simulation methods. Synthese **169**(3), 615–626 (2009)

[20] López-Rubio, E., Ratti, E.: Data science and molecular biology: prediction and mechanistic explanation. Synthese **198**(4), 3131–3156 (2021)

[21] Kitchin, R., McArdle, G.: What makes big data, big data? exploring the ontological characteristics of 26 datasets. Big Data & Society **3**(1), 2053951716631130 (2016)

[22] Marr, B.: Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance. John Wiley & Sons, London (2015)

[23] Turner, R.: Computational artifacts. In: Computational Artifacts, pp. 25–29. Springer, Cham (2018)

[24] Floridi, L., Sanders, J.W.: The method of abstraction. Yearbook of the artificial. Nature, culture and technology. Models in contemporary sciences, 177–220 (2004)

[25] Primiero, G.: On the Foundations of Computing. Oxford University Press, Oxford (2019)

[26] Fresco, N., Primiero, G.: Miscomputation. Philosophy & Technology **26**(3), 253–272 (2013)

[27] Sullivan, E.: Understanding from machine learning models. The British Journal for the Philosophy of Science (2020)

[28] Leonelli, S.: Data-centric Biology: A Philosophical Study. University of Chicago Press, Chicago, IL (2016)

[29] Facchini, A., Termine, A.: Beyond Hypothesis-driven and Data-driven Biology Through Explainable AI: a Proposal

[30] Pearl, J., Glymour, M., Jewell, N.P.: Causal Inference in Statistics: A Primer. John Wiley & Sons, London (2016)

[31] Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Hachette, London (2018)

[32] Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Communications of the ACM **62**(3), 54–60 (2019)

[33] Illari, P., Williamson, J.: Mechanisms are real and local. In: Illari, P.M., Russo, F., Williamson, J. (eds.) Causality in the Sciences, pp. 818–844. Oxford University Press, Oxford (2011)

[34] Bechtel, W., Richardson, R.C.: Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. MIT press, Cambridge, MA (2010)

[35] Halina, M.: Mechanistic explanation and its limits. In: Glennan, S., Illari, P. (eds.) The Routledge Handbook of Mechanisms and Mechanical Philosophy, pp. 213–224. Routledge, London (2017)

[36] Craver, C.F., Darden, L.: In Search of Mechanisms: Discoveries Across the Life Sciences. University of Chicago Press, Chicago, IL (2013)

[37] Psillos, S.: Scientific Realism: How Science Tracks Truth. Routledge, London (2005)