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, Machine Learning, Explainable AI, Scientific Understanding

1 Opacity: one word, many things

The incredible success of artificial intelligence (AI) systems in recent years is considered mostly a consequence of the recent advancements in machine learning (ML) techniques, which make artificial agents able to extract information, learn knowledge and build models from data on their own. Unlike more traditional AI systems, those based on ML possess an impressive inferential power that allows them to analyse large amounts of data and identify patterns that neither the human eye nor traditional statistical methods would likely ever be able to discover (Alpaydin, 2021). Unfortunately, these systems suffer from the problem of being opaque, or, as they say, 'black boxes'. Roughly speaking, that a ML 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 (Zednik, 2019).

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 (Adadi & Berrada, 2018; Guidotti, Monreale, Turini, Pedreschi, & Giannotti, 2018). 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 (Burrell, 2016; Durán & Formanek, 2018; Héder, 2020; Miller, 2019; Zednik, 2019). In recent years, their efforts have led to the birth of *eXplainable AI* (XAI), a new area of research aimed at rendering ML systems less opaque and more humanly understandable (Samek, Montavon, Vedaldi, Hansen, & Müller, 2019). Despite the extensive technical and philosophical literature on the subject, however, both the exact meaning of *opacity* and the reasons leading users to consider certain systems more opaque than others still remain unclear. In the literature, there exist a sort of "received view"[1] that considers opacity a consequence of the high-complex and sub-symbolic design of certain ML systems, which prevents users from understanding their structure and functioning and interpreting their behaviour on various levels. This received view relies on the following simple considerations. At a very abstract level, we can describe the behaviour of a ML model in terms of a function f mapping specific features of the input into predictive outcomes (an example is a function f that maps the features "high bloody pressure" and "obesity" into the predictive outcome "high probability of a hearth attack"). In more traditional statistical ML systems (e.g., linear regression) f is typically a function defined in a low-dimensional space that can be easily turned into an analytical expression (e.g., an equation) or a graphical representation (e.g., a plane in a 3-D space) understandable by humans. When considering symbolic ML systems such as *decision-trees* (DT) or *rules-based systems* generated by inductive logic programming (IRBS), even if f might be difficult to turn into a representation that is understandable by humans, the

[1] See e.g. in Adadi and Berrada (2018); Arrieta et al. (2020); Baldi (2021); Doshi-Velez and Kim (2017); Guidotti et al. (2018).

behaviour of the system can be understood by reconstructing the inferential steps leading to the outcome in terms comprehensible by users. Differently, in models generated by highly-complex sub-symbolic ML systems (e.g. deep neural networks or support vector machines), f is a non-linear function over a high-dimensional space that is usually impossible to turn into a representation understandable by humans. Furthermore, the sub-symbolic nature of these models makes it hard to reconstruct in comprehensible terms the steps leading to an outcome. As a consequence, their inner structure and functioning remain obscure (Baldi, 2021).

Consistently with these considerations, highly-complex sub-symbolic systems such as deep learning neural networks (DNN) and support vector machines are usually described as 'black boxes', 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"(Guidotti et al., 2018, p.7) since these less-complex or symbolic models are easily understandable and interpretable for humans. Analogously, in (Arrieta et al., 2020, 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 model 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.

Although relevant, these considerations put all the emphasis on the intrinsic design of the ML systems, failing to recognize that also the context of use, the users' cognitive skills, and the purposes for which an ML system is involved may affect its perceived opacity, respectively, transparency. In fact, it is not uncommon that users deem a system transparent in one context but opaque in another. Consider for example DNNs. Consistently with the received view, such complex sub-symbolic systems are usually considered opaque. However, there exist contexts where DNNs are considered more transparent than symbolic models. Computational neuroscience is an example of one such context. There, DNNs are employed to model brain networks implementing high-level cognitive functions, such as human perception (Cichy & Kaiser, 2019). Contrary to traditional symbolic models, DNNs not only can accurately simulate high-level cognitive processes but are also able to explain how a purely sub-symbolic architecture, working similarly to the human brain, can implement them. At the same time, there are contexts in which both complex sub-symbolic models and symbolic ones are deem equally opaque. An example is the use of ML models to predict pathological phenotypes from the analysis of genome mutation. In this context, it is common to consider both DNNs and decision trees equally opaque since both do not shed light on either the molecular pathways or the mechanisms responsible for the predicted pathological phenotypes.

These and other similar examples highlight the fact that opacity does not have a single and well-defined meaning, but is rather a plural concept. Stated otherwise, kinds of opacity and reasons to deem a system opaque are many and different; their 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 opacity. Those proposed by Burrell (2016), Creel (2020) and Boge (2021) are among the most relevant.

In her analysis, Burrell (2016) identifies three different manifestations 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. Differently, Creel (2020) starts from the fact that the structure and functioning of AI systems can be described at three different levels of abstraction. Hence, she distinguishes between three different forms of opacity: 'run opacity', 'structural opacity', and 'algorithmic opacity', each one related to users' understanding of the structure and functioning of the system at a given level of abstraction. "Algorithmic opacity" is related to the abstract specification level and concerns users' understanding of the algorithm describing the overall system's behaviour; "structural opacity" is related to the implementation level and concerns the users' understanding of the program (code) implementing the algorithm; "Run opacity" is related to the physical execution level and concerns users' understanding of the physical process executing the program. Both these taxonomies make some progress in characterizing the plural nature of "opacity" but still miss a dimension of opacity, which, instead, is recognized by Boge (2021) and concerns the fundamental distinction between *understanding of a model* and *understanding with a model*. In the context of scientific research, models are not generally interesting *per-se* but only as they allow scientists to understand something about the world. Models generated by ML systems, however, are typically mere predictive models, i.e., they do not convey information relevant to understand *why* and *how* phenomena occur, but simply to predict their occurrence. Actually, this represents one of the major reason why scientists tend to consider such models *opaque*. However, the meaning they attribute to the term "opacity" is clearly different from the usual one in XAI. It does not refer to users' understanding *of* the ML model but to users' (in)ability to understand something about target-phenomena *with* a ML model. Despite Boge (2021) marks a fundamental progress over the previous taxonomies, in our opinion his analysis remains too broad and needs to be deepened. In fact, as discussed in the following sections, a further inquiry reveals that the two dimensions identified by Boge (2021) actually include many different specific forms of opacity. Furthermore, there exists kinds and sources of opacity that neither Boge (2021) nor Burrell (2016) and Creel (2020) have explored, which notwithstanding deserve a careful analysis.

Taking these considerations into account, we propose a more detailed analysis that starts from the identification of three macro-dimensions of opacity, and then refine further each of them by including specific forms, as depicted in Figure 1. The main dimensions of opacity are called, respectively, *access opacity*, *link opacity* and *semantic opacity*.

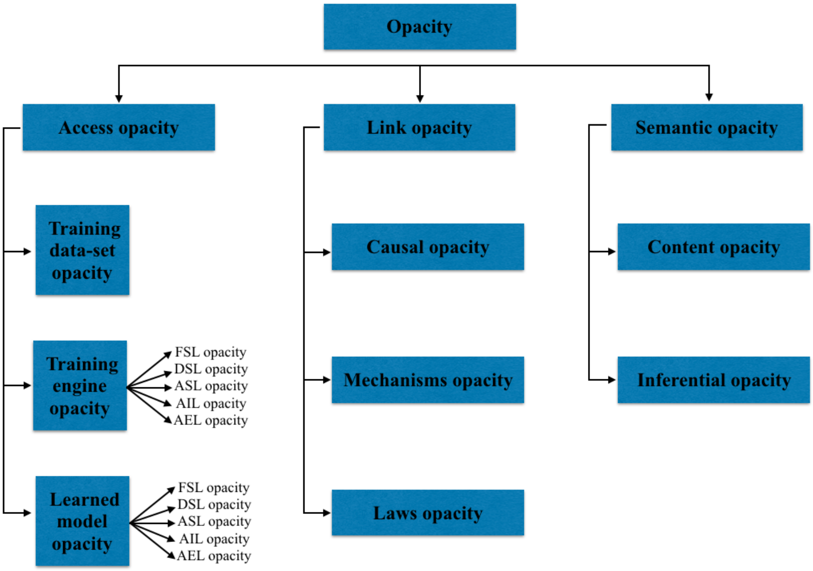


Fig. 1: Map of the different forms of opacity

2 Access opacity

Access opacity concerns the capability of understanding the structure and functioning of an AI system. It occurs when human users have limited epistemic access to elements that are relevant for explaining, predicting, and controlling the behavior of the considered system.[2] Notice that by "having an epistemic access to an element", we mean the ability 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.[3] 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[4] and format[5] (López-Rubio & Ratti, 2021). The intuition is that, as human users possess limited cognitive resources, their ability to explain, predict and control the

[2] The notion of 'epistemically relevant element' is borrowed from (Humphreys, 2009).
[3] They are related with the three forms of opacity described by Burrell (2016).
[4] I.e., the *number* of its elements and their mutual relations.
[5] I.e., the *type* of elements it includes and how they are related.

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:

1. the *training sample*: the data-set used to train the system.
2. the *training engine*: the computational process that allow the system to learn from data during the training process;
3. the *learned model*: the final model of data obtained after running the training engine 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 users 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 users 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.[6] 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 and biases (Marr, 2015). 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 the data model that influence the system behavior, therefore, is high.[7] 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

[6] For an overview of the different meanings of the term big data, see: Kitchin and McArdle (2016).

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

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 users. Furthermore, these transformations are usually irreversible, making impossible for users to go from the features space back to raw data (Bishop, 2007).

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*[8], 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[9].

According to a widely accepted tradition in the philosophy of computer science, the structure and functioning of computational artifacts may be 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 (Floridi & Sanders, 2004; Primiero, 2019). Here, we adopt the classification proposed by Primiero (2019), who distinguishes between five different LoAs:

1. The Functional Specification Level (FSL), which consists of an abstract-mathematical specification of the artifact's overall behaviour in terms of the function it computes;
2. The Design Specification Level (DSL), which specifies the procedures for computing the function identified at the FSL generally in terms of an abstract state-transition machine;
3. The Algorithm Design Level (ADL), which consists of the algorithmic (operational) specification, generally in terms of *rules*, of the procedures specified at the DSL;
4. The Algorithm Implementation Level (AIL), which consists of the translations in terms of high and low-level programs of the algorithms specified at the ADL;
5. The Algorithm Execution Level (AEL), which consists 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 users but insufficient or inadequate for others.[10] For instance, a molecular biologist interested in using AI to predict cancer will probably deem sufficiently detailed a description

[8] 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., Clustering methods) produce models that are nothing but compact descriptions of data, whereas others generate *patterns* that can be used to generate predictive outcomes.

[9] For a philosophical perspective on this concept, see Turner (2018).

[10] In this respect, our taxonomy extends that proposed by Creel (2020).

provided at the FSL. Differently, 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). Consequently, whether and to what extent an artifact results opaque will depend both on who the users are and which LoAs are accessible to them. In general, we can say that a given computational artifact A is opaque for a certain stakeholder S if and only if S has limited epistemic access to the LoAs of A suitable for their cognitive skills 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.[11]

In general, ML 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 (e.g., the fluctuations in financial markets in economics or gene regulation in biology) 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[12] with a data-centric approach based on the collection and the AI-supported analysis of observational data (Leonelli, 2016).

In reality, as we have recently argued in (Facchini & Termine, unpublished), 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 theory-free scientific research and the replacement of classical models with those generated by ML algorithms, which, however, are very different from the former. In fact, while models usually involved in scientific research represent causal pathways, mechanisms, and laws governing target-phenomena, ML models are nothing but *functions* correlating features of observational data

[11] 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 by Sullivan (2020), which concerns the extent "to which [a ML] model fails to be empirically supported and adequately linked to the target phenomena".

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

with predictive outcomes. Hence, although being powerful from a predictive point of view, they are often unable to provide sufficient information to explain *how* and *why* the target phenomena occur and to figure out ways for controlling them (Baldi, 2021). This point highlights a huge epistemic limitation of the fully data-driven approach. Regardless of whether one takes a realistic or instrumentalist stance towards scientific knowledge, in fact, scientific understanding requires more than mere statistical associations. It needs information about the causes of the phenomena, the mechanisms that produce them, and the laws that regulate their functioning. As argued by de Regt (2017), the lack of this type of information impedes our ability to explain, intervene on, 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 pathways leading to the occurrence of the target-phenomena. As argued by de Regt (2017), the identification of the causal pathways is essential for the understanding of target-phenomena 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,[13] grounded on the existence of actual cause-effect links, and spurious ones, which are the by-product of statistical paradoxes[14].

As pointed out in (Pearl, 2019; Pearl & Mackenzie, 2018), the ability of computational systems to recognize causal pathways strictly depends on their ability "to choreograph a parsimonious and modular representation of their

[13] 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.

[14] A famous example is the well-known Simpson's paradox, see (Pearl, Glymour, & Jewell, 2016; Pearl & Mackenzie, 2018).

environment, interrogate that representation, distort it by acts of imagination and finally answer 'What if?' kind of questions" (Pearl, 2019, p.1). These, in particular, may be either statistical, interventional, or counterfactual questions. The former concern the statistical regularities observed in the naked data and have the form "what if I see x?"; for example: "what if I see salt in the water?". The second one concern the consequences of intervention and has the form "what if I do x?"; for example: "what if I add salt to the water?". Finally, the latter concern some counterfactual state of affairs and have the form "what if I had done x?", for example: "what if I had added salt to the water?", or the contrastive form "what if I had done y instead of x?"; for example: "what if I had added sugar instead of salt?". 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" (Pearl, 2019, 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[15].

Available ML 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 the reason of 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." (Illari & Williamson, 2011, 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 (Bechtel & Richardson, 2010). 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" (Halina, 2017, p.217). Finally, it supports generalization because mechanisms "work in the same or similar way under the same or similar conditions" (Craver & Darden, 2013, p.19). Formulating a mechanistic explanation, however, needs much more than mere observational data.

[15] On this topic, see (Pearl, 2019; Pearl et al., 2016; Pearl & Mackenzie, 2018).

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 on mechanistic reasoning, Bechtel and Richardson (2010) identify two mains reasoning strategies followed by scientists to identify mechanisms' structure and functioning, which they name *decomposition* and *localization*. Roughly, the former consists 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. A different account of mechanistic reasoning is proposed by Craver and Darden (2013). According to these authors, mechanistic reasoning is a hypothesis-driven practice that combines scientific exploration, hypotheses-formulation and experimental manipulation. Loosely speaking, the search for mechanisms is an iterative process consisting of the iterated application of specific reasoning and experimental techniques that allow scientists to refine a raw hypothesis about the mechanism's structure and functioning, generally in the form of a sketch representation 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 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 to be 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 and the realist.[16] A detailed discussion 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 functions learned by ML algorithms and, indeed, a ML 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:

[16] On the debate about instrumentalists and realists, see (Psillos, 2005).

- 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[17], 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 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.

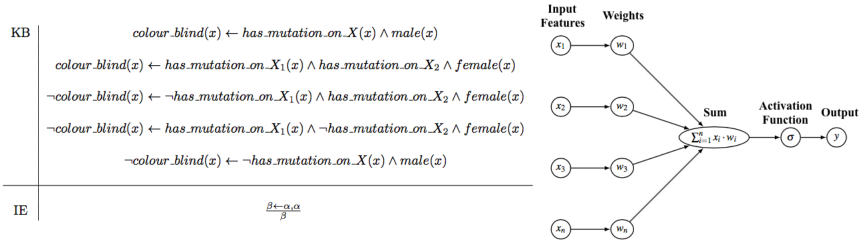
Unfortunately for AI systems, the identification of laws of the second kind is a purely theoretical work. It relies on the human mind's ability to go beyond the observable phenomena and figure out in what the supposed basic structure of reality might consist. ML-based AI systems do not possess this ability,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

4 Semantic Opacity

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 systems, the structural aspect coincides with the properties of the model that the system learns from data, which can be specified at different LoAs as explained in Section 2. These properties are relevant for understanding how the learned model works and, therefore, are connected with the problem of access opacity mentioned above. Stated otherwise, the opacity of the structural aspects of information is a form of access-opacity and, more specifically, an occurrence of access-opacity of the learned model.

Differently from structural aspects, semantic aspects coincide with the potential semantic contents of the information stored by the learned model. They are not directly relevant for determining the functioning of the model. Nevertheless, understanding these aspects is fundamental for users to grasp and interpret the information that the system learns and manipulates. In fact, it happens that if the format used to store and manipulate information prevents users from giving it a meaningful interpretation, then users deem the system opaque. This sense of "opacity", however, cannot be included into any of the kinds described so far. It represents a new form of opacity that we call *semantic opacity.*

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



KB

$$colour\_blind(x) \leftarrow has\_mutation\_on\_X(x) \land male(x)$$

$$colour\_blind(x) \leftarrow has\_mutation\_on\_X_1(x) \land has\_mutation\_on\_X_2 \land female(x)$$

$$\neg colour\_blind(x) \leftarrow \neg has\_mutation\_on\_X_1(x) \land has\_mutation\_on\_X_2 \land female(x)$$

$$\neg colour\_blind(x) \leftarrow has\_mutation\_on\_X_1(x) \land \neg has\_mutation\_on\_X_2 \land female(x)$$

$$\neg colour\_blind(x) \leftarrow \neg has\_mutation\_on\_X(x) \land male(x)$$

IE

$$\frac{\beta \leftarrow \alpha, \alpha}{\beta}$$

(a) System A

(b) System B

Fig. 2: Two expert systems

Semantic opacity can occur in three different circumstances. First, when the learned model lacks a clear, well-defined semantic interpretation that allows users 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 users' 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. 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 users from grasping its semantic content and using it for their purposes. Notice that each type of AI system adopts a peculiar format to represent the information learned from the data. In general, the choice of the format affects the users' ability to provide the stored information with an interpretation that allows them to grasp its semantic content.

By way of example, let us compare the two expert systems A and B reported in the Figures 2a and 2b.

A is an example of rules-based system generated through inductive logic-programming and used in the context of medical decision-making to predict whether a patient will suffer from colour blindness. ML systems such as A store the information learned from data by means of logical sentences stemming from a given formal language that are called *hypotheses*, e.g.:

$$colour\_blind(x) \leftarrow has\_mutation\_on\_X(x) \land male(x)$$

Hypotheses are collected in the knowledge base (KB) and manipulated through iterative applications of the rules included in the inference engine

(IE) in order to generate predictive outcomes. It is easy to see how a standard Tarskian semantics, which maps the syntactic elements (i.e., predicates, variables, quantifiers, Boolean connectives) to features relevant to the context (i.e., genetic mutations, sex, disease, patients), may easily provide the information stored in the KB with a meaningful interpretation that allow users to grasp its semantic content.

Things are different with a DNN such as B, which stores the information learned from data by means of "weights", i.e., numerical parameters connecting the various nodes in the multi-layer network. Providing these parameters with an interpretation is hard. In fact, they usually have a mere instrumental meaning, that is, their values are chosen simply as those values that allow the network to minimize the prediction error (Baldi, 2021). They lack any meaningful semantic interpretation that would allow users to grasp the content of the stored information.

4.2 Inferential opacity

This form of opacity occurs when the format of the inferences used by an AI system to manipulate information prevents users 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. For instance, a rules-based system such as A in the example above manipulates information by applying the rules included in the IE to the hypotheses stored in the KB. Conversely, a DNN such as B manipulates the information using analytical calculations that merge input data and learned parameters. Unfortunately, it is not always possible to provide inferences with an interpretation meaningful for the context of use that allows users to reconstruct the reasoning pathways followed by the system in humanly understandable terms. In some cases, inferences have a purely instrumental value, i.e., they allow the system to generate accurate predictive outcomes, but lack of any meaningful semantic interpretation. In other cases, they may posses a well-defined and meaningful semantics that, however, is 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 can 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 users' 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 users' 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 and Future Developments

Starting from the crucial observation that what users 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-dimensions 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

The authors would thank the participants of PT-AI 2021 conference and the reviewers for their constructive feedback. They would also thank Hajo Grief, Florian Boge, Giuseppe Primiero, and Fabio D'Asaro for suggestions and comments on earlier drafts of this paper.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Alpaydin, E. (2021). *Machine learning, revised and updated edition.* Cambridge, MA: MIT Press.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... others (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115.

Baldi, P. (2021). *Deep learning in science.* Cambridge: Cambridge University Press.

Bechtel, W., & Richardson, R.C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Cambridge, MA: MIT press.

Bishop, C.M. (2007). *Pattern recognition and machine learning, 5th edition.* Springer.

Boge, F.J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, 1–33.

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

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

Craver, C.F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences.* Chicago, IL: University of Chicago Press.

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

de Regt, H.W. (2017). *Understanding scientific understanding.* Oxford, UK: Oxford University Press.

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

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

Facchini, A., & Termine, A. (unpublished). *Beyond hypothesis-driven and data-driven biology through explainable ai: a proposal.*

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

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

Halina, M. (2017). Mechanistic explanation and its limits. S. Glennan & P. Illari (Eds.), *The routledge handbook of mechanisms and mechanical philosophy* (pp. 213–224). London: Routledge.

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

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

Illari, P., & Williamson, J. (2011). Mechanisms are real and local. P.M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (p. 818-844). Oxford: Oxford University Press.

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

Leonelli, S. (2016). *Data-centric biology: A philosophical study.* Chicago, IL: University of Chicago Press.

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

Marr, B. (2015). *Big data: Using smart big data, analytics and metrics to make better decisions and improve performance.* London: John Wiley & Sons.

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

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

Pearl, J., Glymour, M., Jewell, N.P. (2016). *Causal inference in statistics: A primer.* London: John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect.* London: Hachette.

Primiero, G. (2019). *On the foundations of computing.* Oxford: Oxford University Press.

Psillos, S. (2005). *Scientific realism: How science tracks truth.* London: Routledge.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (2019). *Explainable ai: interpreting, explaining and visualizing deep learning* (Vol. 11700). Cham: Springer Nature.

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

Turner, R. (2018). Computational artifacts. *Computational artifacts* (pp. 25–29). Cham: Springer.

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