Towards a Taxonomy of Pragmatic Opacity*[*long abstract*]

Alessandro Facchini[1] and Alberto Termine[2]

[1]IDSIA USI-SUPSI
[2]U. Milano

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 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 propose to view opacity as a pragmatic concept. Based on this, we then explicit the distinction between access opacity, link opacity and informational opacity, hence providing the groundwork for a conceptual taxonomy of the concept of opacity for AI systems.

1 The pragmatic nature of opacity in AI

Developed to provide tools and methods for reducing the opacity of AI system and making them more humanly understandable, the research program of *explainable AI* (xAI) has captured the attention of several scholars (Guidotti et al., 2018; Liao et al., 2021; Doshi-Velez and Kim, 2017; Zednik, 2019).[1] Nevertheless, some of the fundamental concerns motivating this program remain unanswered. The relevant xAI literature provide in fact very few systematic attempts to answer questions such as: what kind of forms opacity may take? or, which features make an AI system opaque? in which sense?[2] Instead the majority of xAI scholars simply classify systems as more or less opaque depending on their

*Both authors have contributed equally to this work. They are thus listed in alphabetic order.
[1]For a survey of current xAI literature, see e.g. (Adadi and Berrada, 2018).
[2]See, for instance (Burrell, 2016)

similarities with traditional rules-based systems, which are assumed to be the prototype of transparent systems (Liao et al., 2021; Doshi-Velez and Kim, 2017). Unfortunately, such "naive" approach does not really contribute at clarifying the notion of opacity. On the contrary, it might undermine the whole xAI program; after all, how can opacity be reduced without really understanding its nature?

Here, we take a different approach. The starting observation is that, in the field of AI, opacity is perceived as concerning the use of a system by a given stakeholder, in a certain context and with a certain purpose. It occurs when the information the system conveys is inadequate or insufficient to fulfill the desired uses and objectives attributed to it. It has therefore to be understood as a *pragmatic*, or *contextual*, concept, which may take different forms and characteristics precisely depending on the nature of the given context, the given stakeholder and its purposes.

2 A first small taxonomy

In what follows, we propose a taxonomy that distinguishes three main forms of pragmatic opacity, each being prone to a further deeper analysis.[3]. We refer to them as *access opacity*, *link opacity* and *informational opacity*.

2.1 Access opacity

Access opacity is about the capability of understanding the structure and functioning of a system. It manifests when human stakeholders have limited access to the epistemically relevant elements (EREs) within the system's inner structure that allow them to *understand*, *predict* and *control* the computational behaviour of the considered system. Here, by *epistemically relevant* we denote those elements that allow human stakeholders to *understand*, *predict* and *control* the computational behaviour of the considered system.[4] Three main factors may be limiting the access to those EREs.[5] The first is represented by the transparency policies adopted by the system's designers, as they might deliberately obscure some relevant details of the system's structure for either commercial, competition or privacy reasons. The second factor is the stakeholder's background knowledge. Intuitively, the more a stakeholder is familiar with a given AI system, the more they are able to understand, predict and control the system's behaviour. The third factor is the

[3]This deeper analysis is presented in the extended (full) version of this work.

[4]The notion of epistemically relevant element is borrowed from (Humphreys, 2009).

[5]The three factors corresponds to the three kind of opacity described by Burrell (2016): "opacity as intentional corporate or state secrecy", "opacity as technical illiteracy", and finally "opacity as the way algorithms operate at the scale of application".

complexity of the system's structure.[6] As humans have limited cognitive resources, their ability to understand, predict and control the system's behaviours decreases as the number of relevant elements increases. Usually, the bigger is a system, the more it results opaque to access because the higher is the number of EREs included in its structure.

Access opacity may occur in different forms, the reason being that a given description of an AI system depends on the chosen level of abstraction (LoA) (Primiero, 2019). Each LoA having its own proper structure that includes different epistemically relevant elements, for each one of them a different form of access opacity can then be identified.[7]

2.2 Link opacity

Link opacity concerns the use of AI systems in scientific research. It manifests when a system is used to model a given phenomenon but conveys inadequate or insufficient information about the elements that are relevant for explaining, predicting and controlling the considered phenomenon.[8] In this context, link opacity represents a serious problem especially for the employment of machine learning (ML) models in sciences. As a matter of fact, because of their incredible accuracy in predicting phenomena by learning directly from data (Alpaydin, 2021; Baldi, 2021), in recent years ML systems have been replacing more traditional scientific models (Anderson, 2008; Leonelli, 2016). Unfortunately, as they generate mere associative models, ML systems are usually unable to inform about either the laws, the causes or the mechanisms behind the predicted phenomena, and therefore they do not fulfill the desiderata of a scientific model, namely explaining, predicting and controlling (or performing intervention) (López-Rubio and Ratti, 2021).

Similarly to access opacity, link opacity also occurs in different forms. This is because the elements that are relevant for explaining, predicting and controlling a given phenomenon vary depending on the nature of the considered phenomenon.

Notice that link opacity and access opacity are logically independent concepts, in the sense that their definitions are mutually independent[9]. However, some form of access transparency may be necessary to obtain some form of link transparency.[10]

[6] See also (López-Rubio and Ratti, 2021).

[7] As explained in the full version of this work, the three forms of opacity described by Creel (2020), and called, respectively, run opacity, structural opacity and algorithmic opacity, constitute three different instances of access opacity differing precisely by the considered level of abstraction in the case where the stakeholder is a computer scientist and what is at stake is the understanding of the behaviour of a computing artefact.

[8] The concept of link opacity is close to link uncertainty, a notion introduced by Sullivan (2020).

[9] The concept of link opacity is not needed to define access opacity, and viceversa.

[10] This point is explained in the full version of this work.

2.3 Informational opacity

Informational opacity relates to the format, or setup, adopted by a system for storing and manipulating information. In rules-based systems, for instance, information is stored by means of logical formulae that are manipulated through the application of syntactic rules. On the contrary, a deep neural networks represents the information it learns by modifying the weights associated with its connections.

It is crucial to notice that formats influence the stakeholder-system interaction. They may indeed prevent the stakeholder to access the stored information or to reconstruct the inferences thorough which the system manipulate information. In general, how and to what extent the stakeholder-system interaction is affected is function of both the nature of the format and the stakeholder's cognitive abilities and purposes. We thus say that a system manifests informational opacity for a given stakeholder (in a given context and with certain purposes) when, because of the used format and the stakeholder's cognitive abilities and purposes, either the stakeholder cannot get access to the information embedded in the system, or they are unable to reconstruct the inferences that manipulate such information.

Acknowledgments

We thank the participants of the 4th Conference on Philosophy and Theory of Artificial Intelligence for their constructive feedback.

References

Adadi, A. and M. Berrada (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*. MIT Press.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine 16*(7), 16–07.

Baldi, P. (2021). *Deep Learning in Science*. Cambridge University Press.

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

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

Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning.

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

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

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

Liao, Q. V., M. Singh, Y. Zhang, and R. Bellamy (2021). Introduction to explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–3.

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

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

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

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