

A logical approach to algorithmic opacity

Mattia Petrolo¹, Ekaterina Kubyshkina² and Giuseppe Primiero²

¹University of Lisbon, CFCUL, Alameda da Universidade, 1649-004 Lisbon, Portugal

²University of Milan, Philosophy Department, Via Festa del Perdono, 7 20122, Milan, Italy

Abstract

In [1], we introduced a novel definition for the epistemic opacity of AI systems. Building on this, we proposed a framework for reasoning about an agent's epistemic attitudes toward a possibly opaque algorithm, investigating the necessary conditions for achieving epistemic transparency. Unfortunately, this logical framework faced several limitations, primarily due to its overly idealized nature and the absence of a formal representation of the inner structure of AI systems. In the present work, we address these limitations by providing a more in-depth analysis of classifiers using first-order evidence logic. This step significantly enhances the applicability of our definitions of epistemic opacity and transparency to machine learning systems.

Keywords

Transparent AI, epistemic opacity, epistemic logic, evidence models, neighborhood semantics

1. Introduction

Recently, the use of computational algorithms has significantly increased across different areas of human life, leading to the development of explainable AI and other human-centered approaches aimed at understanding the nature of AI models. One common issue discussed in these approaches is the epistemic opacity problem, that is, a problem about the epistemic accessibility and reliability of algorithms. In this study, we aim to provide a novel epistemological and logical analysis of this problem in order to identify the conditions under which this form of opacity can be eliminated.

This work builds upon the insights we introduced in [1]. In this instance, we present a new and more expressive formal framework that holds promise for accurately representing an agent's epistemic attitude regarding the opacity of an AI system.

2. An epistemological definition of opacity


To characterize the epistemic opacity of algorithms, we follow the methodology proposed by Durán and Formanek [2] and adapt Humphreys' definition of epistemically opaque process:

2nd Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-23, co-located with AIXIA 2023, Roma Tre University, Rome, Italy, 2023

✉ mpetrolo@fc.ul.pt (M. Petrolo); ekaterina.kubyshkina@unimi.it (E. Kubyshkina); giuseppe.primiero@unimi.it (G. Primiero)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

[A] process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process (Humphreys [3], p. 618).

This characterization of opacity relies on the fact that an agent “X does not know,” which, in turn, requires a definition of what knowledge is. However, Humphreys’ account leaves this question unanswered, since he does not touch upon the question of what “knowledge of an algorithm” is. Traditional epistemology defines knowledge as corresponding to justified true belief, but this analysis has been challenged by Gettier [4] famous counterexample, which prevents one to consider luck-dependent cases as cases of genuine knowledge. To avoid this problem, our characterization of opacity must carefully consider the justificatory component involved in the analysis. Additionally, in order to specify which elements of algorithms are epistemically relevant, we must take into account their specific structure. Cormen et al. [5] describe an algorithm as “any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output” (p. 5). This can be seen as a minimal characterization of algorithm, containing the minimal elements one has to take into account. Although informal and brief, this characterization highlights three key elements of an algorithm: its input, procedure, and output. We argue that a sound definition of epistemic opacity (and transparency) for algorithms must take these elements into account. Our proposal is as follows:

An algorithm is *epistemically opaque* relative to an epistemic agent A at time t just in case at t, A does not have

- an epistemic justification for I,
- or an epistemic justification for P,
- or an epistemic justification for O;

where I, P, O express the algorithm’s input, procedure, and output, respectively.

The previous definition has an important feature: the components I, P, and O of the algorithm are related, but irreducible one to another. As a consequence, the lack of epistemic justification for any component constitutes a sufficient condition for epistemic opacity. In theory, the three conditions can occur independently, but in most real-world algorithms, these forms of opacity are interconnected, which complicates the epistemic opacity problem. Based on this definition, in most cases, the algorithms we interact with on a daily basis are epistemically opaque. Our aim is here is to present a formal framework to reason about an agent’s epistemic attitudes towards opaque algorithms and examine the conditions required for achieving epistemic transparency.

The aforementioned epistemological definition of algorithmic opacity is not new; we introduced it in [1]. In that work, we also outlined a logical framework to formally represent opacity. However, the framework we presented in [1] has two primary limitations. First, the formal representation of epistemic justification for input and output is overly idealized and simplified. Second, the relationship between input, procedure, and output is left implicit due to the absence of a formal representation of their inner structure. Consequently, connecting these three elements in a unified setting is challenging. In our current work, we address both of these

limitations, enhancing the applicability of our definitions to machine learning (ML) systems and providing a more detailed analysis of classifiers using first-order evidence logic.

3. Formal framework

To reason formally about an agent's epistemic attitudes towards opacity, it is necessary to reformulate the definition of opaque algorithm in logical terms. In order to do so, we utilize tools from epistemic logic and express the epistemic justification for each component by using a specific modality.

Before delving into the formal semantics, let us fix some terminology to bridge the characterization of an algorithm used in the epistemological definition with the actual architecture of machine learning systems. In particular, from now on, we refer to the model of a given classifier trained using supervised learning. The input is usually called *test set* and it consists of data points, which are evaluated on the basis of the labels of the training set. The procedure can be divided into the *training set* and the *model*. The training set consists of labelled data points. The model consists of *parameters* (features taken into account), *hyperparameters* (the 'significance' attributed to each feature), a *target* (the predicate one wants to classify for), and a *mathematical function* corresponding to the chosen learning algorithm, which evaluates the input with respect to the target. Finally, the output consists of the *data points* of the test set with labels attributed according to the criteria of the trained model.

The semantics we are proposing is a first-order extension of a neighborhood semantics for evidence logic provided by van Benthem et al. [6].

The language \mathcal{L} is defined as follows:

$$\phi := x \mid P(x) \mid \neg\phi \mid \phi \wedge \psi \mid \forall x\phi \mid \Box\phi \mid B\phi \mid K\phi$$

Definition 3.1. A first-order evidence model is a tuple $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$, where W is a non-empty set of worlds, $E \subseteq W \times \mathfrak{p}(W)$ is an evidence relation, D is a non-empty set, for each w , V_w is a function that to each n -place predicate symbol assigns a subset of D^n . We write $E(w)$ for the set $\{X \mid wEX\}$. Two constraints are imposed on the evidence sets: For each $w \in W$, $\emptyset \notin E(w)$ and $W \in E(w)$.

Definition 3.2. A w -scenario is a maximal collection $\mathcal{X} \subseteq E(w)$ that has the finite intersection property: for each finite subfamily $\{X_1, \dots, X_n\} \subseteq \mathcal{X}$, $\bigcap_{1 \leq i \leq n} X_i \neq \emptyset$.

Definition 3.3. Let $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$. An assignment g is a function that to each variable assigns an element of D . Given assignments g and g' , $g' \sim^x g$ means that g' agrees with g on all variables save possibly x . The relation $\mathcal{M}, g, w \vDash \phi$ is defined by induction, where w is a world, g is an assignment, and ϕ is a formula of first-order modal logic.

- $\mathcal{M}, g, w \vDash P(x_1, \dots, x_n)$ iff $(g(x_1), \dots, g(x_n)) \in V_w(P)$
- $\mathcal{M}, g, w \vDash \phi \wedge \psi$ iff $\mathcal{M}, g, w \vDash \phi$ and $\mathcal{M}, g, w \vDash \psi$
- $\mathcal{M}, g, w \vDash \neg\phi$ iff $\mathcal{M}, g, w \not\vDash \phi$
- $\mathcal{M}, g, w \vDash \forall x\phi$ iff for any $g' \sim^x g$, $\mathcal{M}, g', w \vDash \phi$

- $\mathcal{M}, g, w \models \Box\phi$ iff there exists X such that wEX and for all $w' \in X$, $\mathcal{M}, g, w' \models \phi$
- $\mathcal{M}, g, w \models B\phi$ iff for each w -scenario \mathcal{X} and for all $w' \in \cap\mathcal{X}$, $\mathcal{M}, g, w' \models \phi$
- $\mathcal{M}, g, w \models K\phi$ iff for all $w' \in W$, $\mathcal{M}, g, w' \models \phi$

A formula ϕ is said to be true at the world w if $\mathcal{M}, g, w \models \phi$; otherwise it is said to be false at w . If $\mathcal{M}, g, w \models \phi$ for every world w , $\mathcal{M}, g \models \phi$. If $\mathcal{M}, g \models \phi$ for any assignment g , $\mathcal{M} \models \phi$.

In what follows, we consider that the worlds of W represent epistemic possibilities for an agent. The evidence relation E associates each world with sets of evidences considered by the agent. The domain D ranges over data points of the possible test set. Finally, V_w is a usual valuation function.

Now we are able to define epistemic justification (EJ) for I, P, and O.

Definition 3.4 (EJ for I). Let a_1, \dots, a_n be data points of an input I , P_1, \dots, P_m are mutually exclusive and exhaustive parameters of a classifier C . An agent has an epistemic justification for input I to the classifier C in a world w iff for all $a_i \in I$, $\mathcal{M}, g, w \models \neg K\neg(\pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i))$, where $\pm P_{j \in [1, m]}(a_i)$ stands either for $P_j(a_i)$, or $\neg P_j(a_i)$, depending whether a_i satisfies the property checked by the parameter P_j .

This definition can be explained as follows. We consider that an input is constituted of data points a_1, \dots, a_n . These data points can have (or not) the features evaluated by the parameters of the classifier. We say that an input is epistemically justified iff the agent considers as an epistemic possibility the fact that data points of the input can be correctly evaluated by the parameters of the classifier. In other words, there exists an epistemic state of the agent such that it validates the matching (or not) between the data points and the parameters. Formally, there exists w' s.t. for all $a_i \in \{a_1, \dots, a_n\}$ $\mathcal{M}, g, w' \models \pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i)$, where $\pm P_{j \in [1, m]}(a_i)$ stands either for $P_j(a_i)$, or $\neg P_j(a_i)$.

Definition 3.5 (EJ for P). Let R be a process which transforms an input I into an output O , a_1, \dots, a_n are data points of the input I , P_1, \dots, P_m are mutually exclusive and exhaustive parameters of a classifier C , T is the target. An agent has an epistemic justification for process P in a world w iff $\mathcal{M}, g, w \models \forall x \Box R(x, \pm P_1(x) \wedge \dots \wedge \pm P_m(x), T(x))$, where $\pm P_{j \in [1, m]}(x)$ stands either for $P_j(x)$, or $\neg P_j(x)$, depending whether a_i satisfies the property checked by the parameter P_j .

As for the definition of the epistemic justification for the procedure, we say that there is such a justification iff for any input the agent has an evidence for a process which transforms it into the output (the matching of the data points' features with the parameters) with respect to the target. Formally this amounts to say that for any x there is a neighbourhood, in which all worlds validate $R(x, (\pm P_1(x) \wedge \dots \wedge \pm P_m(x)), T(x))$. Notice, that this does not mean that R is the only process which can transform the input into the output.

Definition 3.6 (EJ for O). Let R be a process which transforms an input I into an output O , a_1, \dots, a_n are data points of the input I , P_1, \dots, P_m are mutually exclusive and exhaustive parameters of a classifier C , T is a target. An agent has an epistemic justification for output O in a world w iff for all $a_i \in I$ $\mathcal{M}, g, w \models \neg B\neg R(a_i, \pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i), T(a_i))$, where $\pm P_{j \in [1, m]}(a_i)$ stands either for $P_j(a_i)$, or $\neg P_j(a_i)$, depending whether a_i satisfies the property checked by the parameter P_j .

We say that an agent has an epistemic justification for the output, once she does not disbelieve the result of processing the input (matching its features with the parameters) with respect to the target. Formally, this means that there exists a world w' in the intersection of all w -scenarios, such that the result of the process $R(a_i, \pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i), T(a_i))$ is valid in w' .

As a consequence, we obtain the following definition of epistemic opacity for classifiers.

Definition 3.7. *Let R be a process which transforms an input I into an output O , a_1, \dots, a_n are data points of the input I , P_1, \dots, P_m are mutually exclusive and exhaustive parameters of a classifier C , T is a target. A classifier C , defined as before, is epistemically opaque ($\mathcal{O}C$) relative to an epistemic agent in a world $w \in \mathcal{M}$ if:*

$\mathcal{M}, g, w \models \mathcal{O}C$ iff for all $a_i \in \{a_1, \dots, a_n\}$ $\mathcal{M}, g, w \models K \neg (\pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i)) \vee \neg \forall x \Box R(x, \pm P_1(x) \wedge \dots \wedge \pm P_m(x), T(x)) \vee B \neg (a_i, \pm P_1(a_i) \wedge \dots \wedge \pm P_m(a_i), T(a_i))$.

Now let us apply our definition to a simplified example of an agent interacting with a classifier which distinguishes photos of dogs from other images. An agent inserts a photo of a cat, that is, the data point of the test set is a . The classifier processes the images by evaluating them on three parameters: the form of the head (P_1), the form of the nose (P_2), and the length of the whiskers (P_3). Let the inserted image match the parameter P_1 (that is, the form of the head of the cat on the image corresponds to the form of the head considered as suitable for a dog in the training set), but not the parameters P_2 and P_3 . The target is to determine whether the input is an image of a dog.

Example 3.1 (Transparent classifier).

In accordance with the Def. 3.7, the classifier is considered to be transparent for an agent once she has an epistemic justifications for all the three components: I , P , and O . In our example, this means that: (1) $\mathcal{M}, g, w \models \neg K \neg (P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a))$, (2) $\mathcal{M}, g, w \models \forall x \Box R(x, \pm P_1(x) \wedge \pm P_2(x) \wedge \pm P_3(x), T(x))$, and (3) $\mathcal{M}, g, w \models \neg B \neg (a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a))$.

Let us provide an example of such model¹. Let $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$, such that $W = \{w, w'\}$, $E(w) = \{\{w, w'\}, \{w'\}\}$, $D = \{a, b\}$, $V_w(P_1) = \{a, b\}$, $V_w(P_2) = \{b\}$, $V_w(P_3) = \{b\}$, $V_w(T) = \{a, b\}$, $V_w(R) = \{(a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a)), (b, P_1(b) \wedge P_2(b) \wedge P_3(b), T(b))\}$.

In this model, (1) is satisfied because $\mathcal{M}, g, w' \models P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a)$. Clause (2) is satisfied because $\{w'\} \in E(w)$, $\mathcal{M}, g, w' \models R(a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a))$, and $\mathcal{M}, g, w' \models R(b, P_1(b) \wedge P_2(b) \wedge P_3(b), T(b))$. Clause (3) is satisfied because there exists a w -scenario $\{\{w, w'\}, \{w'\}\}$ such that $w' \in \cap \{\{w, w'\}, \{w'\}\}$ and $\mathcal{M}, g, w' \models R(a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a))$.

Intuitively, (1) means that the agent considers all the parameters of the classifier and considers as an epistemic possibility the correct matching between the information presented on the photo of a cat and these parameters. Clause (2) means that for all possible inputs the agent has an evidence that the process evaluates them with respect to the target. Clause (3) means that the agent does not disbelieve the result provided by the classifier.

Example 3.2 (Lack of EJ for I).

Let $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$ where $W = \{w, w'\}$, $D = \{a, b\}$, $V_w(P_1) = V_{w'}(P_1) = \{b\}$. In this case $\mathcal{M}, g, w \models K \neg (P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a))$ because none of the worlds satisfies $P_1(a)$. Intuitively,

¹We define only relevant aspects of this model. The omitted definitions are not essential for the purposes of this example.

this corresponds to the situation in which the agent does not consider the correct matching between parameters and the photo as possible. This can be so, for instance, because the agent is unaware that the form of the head (P_1) is a parameter for the classifier, or, simply, because she is convinced that the form of a cat's head cannot match the form of the dog's head.

Example 3.3 (Lack of EJ for P).

Let $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$, where $W = \{w, w'\}$, $E(w) = \{\{w, w'\}, w'\}$, $D = \{a, b\}$, $V_w(R) = \{(a, \neg P_1(a) \wedge P_2(a) \wedge P_3(a))\}$. In this model, $\mathcal{M}, g, w \models \neg \forall x \Box R(x, \pm P_1(x) \wedge \pm P_2(x) \wedge P_3(x), T(x))$ because $\mathcal{M}, g, w' \models \neg R(b, \pm P_1(b) \wedge \pm P_2(b) \wedge \pm P_3(b), T(b))$. Intuitively, this corresponds to a situation in which an agent does not have evidence for a process to evaluate a possible input b , independently of her evidences about evaluating the image a . For instance, in our example, the agent might have evidences for the process to match the relevant data point of the image of a cat with corresponding parameters with a target of determining that it is not an image of a dog. However, she might not have evidences for evaluating another image of a wolf.

Example 3.4 (Lack of EJ for O).

Let $\mathcal{M} = \langle W, E, D, \{V_w\}_{w \in W} \rangle$, where $W = \{w, w'\}$, $E_w = \{\{w, w'\}, \{w'\}\}$, $D = \{a, b\}$, $V_w(R) = V_{w'}(R) = \{(b, \pm P_1(b) \wedge \pm P_2(b) \wedge \pm P_3(b), T(b))\}$. In this model, $\mathcal{M}, g, w \not\models \neg B \neg R(a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a))$, because for each w -scenario \mathcal{X} , for all $w'' \in \cap \mathcal{X}$, $\mathcal{M}, g, w'' \models \neg R(a, P_1(a) \wedge \neg P_2(a) \wedge \neg P_3(a), T(a))$. Intuitively, this corresponds to a situation in which an agent disbelieves the output of the classifier. For instance, the image of the cat inserted by the agent is evaluated as a dog. This contradicts the agent's belief, thus indicating that the classifier is opaque for her.

4. Conclusion

In [1], we dubbed the analysis provided by the modal framework we introduced the *IPO model of algorithmic opacity*. This framework provides an original epistemological definition of algorithmic opacity based on a tripartite analysis of algorithms. On this foundation, it conceives epistemic opacity as a (non-primitive) modal operator. It is true that the framework we are presenting in this current work is more intricate and might seem less intuitive compared to the one outlined in [1]. However, we believe that the complexity of these definitions in the current setting stems from the intricate relationships between I, P, and O and their corresponding epistemic justifications. Consequently, this complexity is a trade-off necessary for our analysis to be effectively applied to real-world ML systems. In future work, our aim is to further develop the IPO model, focusing on both its epistemological and formal aspects. Epistemologically, we plan to compare the definitions we introduced with various forms of opacity examined in the literature (e.g., as discussed by Burrell [7], Creel [8], Boge [9], Facchini and Termine [10]). This comparison will help ascertain if our definition is comprehensive enough to encompass all possible forms of opacity. From a formal standpoint, the next natural step is to introduce a logical system for reasoning about opacity and prove its completeness with respect to the evidence models. Additionally, we aim to compare our framework with other logical formalisms that appear to offer flexibility in representing the fundamental concepts of the IPO model (see, e.g., [11] and [12]).

Acknowledgments

The authors acknowledge the support of the Project PRIN2020 BRIO - Bias, Risk and Opacity in AI (2020SSKZ7R) awarded by the Italian Ministry of University and Research (MUR). The research of Ekaterina Kubyshkina is funded under the “Foundations of Fair and Trustworthy AI” Project of the University of Milan. Giuseppe Primiero and Ekaterina Kubyshkina are further funded by the Department of Philosophy “Piero Martinetti” of the University of Milan under the Project “Departments of Excellence 2023-2027” awarded by the Ministry of University and Research (MUR). Mattia Petrolo acknowledges the financial support of the FCT – Fundação para a Ciência e a Tecnologia (2022.08338.CEECIND; R&D Unit Grants UIDB/00678/2020 and UIDP/00678/2020) and the French National Research Agency (ANR) through the Project ANR-20-CE27-0004.

References

- [1] E. Kubyshkina, M. Petrolo, Reasoning about algorithmic opacity, in: G. Boella, F. A. D’Asaro, A. Dyoub, G. Primiero (Eds.), Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, AIXIA Series, CEUR Workshops Proceedings, 2023, pp. 39–45.
- [2] J. Durán, N. Formanek, Grounds for trust: Essential epistemic opacity and computational reliabilism, *Minds and Machines* 28 (2018) 645–666.
- [3] P. W. Humphreys, The philosophical novelty of computer simulation methods, *Synthese* 169 (2009) 615–626.
- [4] E. Gettier, Is justified true belief knowledge?, *Analysis* 23 (1963) 121–123.
- [5] T. Cormen, C. E. Leiserson, R. Rivest, C. Stein, *Introduction to Algorithms*, Cambridge: MIT Press, 2009.
- [6] J. van Benthem, D. Fernández-Duques, E. Pacuit, Evidence logic: A new look at neighborhood structures, *Advances in Modal Logic* (2012).
- [7] J. Burrell, How the machine ‘thinks’: Understanding opacity in machine learning algorithms, *Big Data & Society* 2 (2016) 1–12.
- [8] K. A. Creel, Transparency in complex computational systems, *Philosophy of Science* 87 (2020) 568–589.
- [9] F. J. Boge, Two dimensions of opacity and the deep learning predicament, *Minds and Machines* 32 (2022) 43–75.
- [10] A. Facchini, A. Termine, Towards a taxonomy for the opacity of ai systems, in: V. C. Muller (Ed.), *Philosophy and Theory of Artificial Intelligence 2021*, volume 63 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, 2022, pp. 73–89.
- [11] X. Liu, E. Lorini, A logic of “black box” classifier systems, in: A. Ciabattoni, E. Pimentel, R. de Queiroz (Eds.), *Logic, Language, Information, and Computation. WoLLIC 2022*, volume 13468 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 158–174.
- [12] S. Artemov, The logic of justification, *The Review of Symbolic Logic* 1 (2008) 477–513.