OPENING THE PANDORA'S (BLACK) BOX OF AI MADE IN EUROPE : INTERPRETABILITY - EXPLAINABILITY - COMPREHENSIBILITY

Christophe Denis, Sorbonne University, LIP6 [75252] Paris, also at ERIAC, [76821], Mont-Saint-Agnan

Anaëlle Martin, CEIE, University of Strasbourg [67046] Strasbourg, France

Abstract

In the already extensive literature connecting « black-box effect » and explainability, it is well established that some *deep learning methods* (Dl), although successful from the accuracy point of view, are opaque in terms of understanding how they make decisions. While the lack of explicability of *machine learning* (ML) techniques raises operational, ethical and legal problems, the technical and political solutions provided by researchers and policy-makers to make algorithms more « transparent » may also appear problematic in terms of clarity. We argue that there is a strong paradox in listing numerous requirements that are conceptually indeterminate[1], in order to address the challenges posed by « black box algorithms ». We suggest that before any prescriptive or normative ambition in the field of AI ethics and regulation, some epistemological prolegomena are required to distinguish the epistemic functions and uses of descriptive, predictive and explicative models.

PROBLEM STATEMENT

"*In the modern system it should appear as though everything were explained*", Wittgenstein, Tractatus Logico-Philosophicus (6.372)

"*Our civilisation is characterized by the word progress. (. . .) Its activity is to construct a more and more complicated structure. And even clarity is only a means to this end and not an end in itself. For me on the contrary clarity, transparency, is an end in itself. I am not interested in erecting a building but in having the foundations of possible buildings transparently before me. "* Wittgenstein (MS 109 200: 5.11.1930)

External properties like explainability and fairness are commonly associated to deep neural network to try to overcome its "black box" effects. It is therefore necessary to clarify the particularity of the black box associated with deep learning, which is not always negatively connoted. For example, in software engineering, the use of black boxes facilitates maintenance and reduces the code programming time. The concept of black box is not the prerogative of deep learning as it is used in science and software engineering. In everyday life, an institution can also be considered as a black box since we do not know its internal functioning, for example when we ask for an administrative act. From a functional point of view, a black box, represented by Figure 1, compute output values from inputs ones without any knowledge of is internal workings.

Input
Black Box
Output

Figure 1: Functional point of view of a black box

[1] Indeed, it is not always clear if the rules that are mentioned are legal or ethical and if the concepts refer to binding principles or « values ».

TRANSPARENCY AND ITS COROLLARIES AS SOLUTIONS TO THE "BLACK BOX" PROBLEM: TRACEABILITY, AUDITABILITY AND COMMUNICATION

Ethical principles & requirements set out by the Guidelines for Trustworthy AI

The aim of the Guidelines is to promote Trustworthy AI (lawful, ethical and robust) and provide guidance on how such principles can be operationalized in socio-technical systems.

The ethical principles are as follows:

1. respect for human autonomy;
2. prevention of harm;
3. fairness;
4. explicability.

The ethical requirements are as follows:

1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination and fairness;
6. environmental and societal well-being;
7. accountability.

Explicability is viewed as a « principle » — closely linked with the rights relating to Justice — while transparency is considered as a « requirement ». The latter encompasses transparency of elements relevant to an AI system: data, system and business model. It includes traceability, explainability and communication. The principle of explicability means that « processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected ». Traceability involves that data sets and processes should be documented to allow for an increase in transparency. This enables identification of the reasons why an AI-decision was erroneous. Communication implies that AI systems should not represent themselves as humans to users. Beyond this, the AI system's capabilities and limitations should be communicated. Explainability requires that whenever AI has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI's decision-making process. It should be timely and adapted to the expertise of the stakeholder.

Legal obligations laid down by the AI Act

In its preamble, the *AI Act* asserts that the proposal lays down obligation that will apply to providers and users of high-risk AI systems. It promotes public trust in the use of AI by facilitating audits of the AI systems with new requirements for documentation, traceability and transparency. In accordance with Article 13, "transparency and provision of information to users" are required for high-risk AI systems. High-risk systems should be designed in such a way "to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately".

The proposal calls for « an appropriate type and degree of transparency » and provision of « concise, complete, correct and clear information that is relevant, accessible and comprehensible to users ». The proposal states that « to address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output. It is also recalled that the exercise of important procedural fundamental rights could be hampered, « where such AI systems are not sufficiently transparent, explainable and documented ». In case of fundamental rights violation, effective redress will be made possible « by ensuring transparency and traceability of the AI systems ». According to Article 12, high-risk AI systems should be designed with capabilities enabling the automatic recording of events while the systems is operating.

THE NEED FOR CONCEPTUAL CLARIFICATIONS

After having proposed a clarification and an adaptation of the notions of interpretability and explainability such as one encounters them in the already abundant literature on the subject, we recall in this article the interest of implementing the epistemological distinctions between the *different epistemic functions* of a model, and between the *epistemic function* and the *use* of a model[2]. We argue that systematically explaining *deep learning* to all its users is not always justified, could be counterproductive and even raises ethical issues. For example, how to assess the correctness of an explanation that could even be unintentionally permissive or even manipulative in a fraudulent context? There is therefore a need to revisit the theory of information (Fisher, Shannon) and the philosophy of information (Floridi) in the light of *deep learning*. This information will allow certain users to produce their own reasoning (surely an abductive one) rather than receiving an explanation. Last but not least, should we trust a *machine learning* model ? Trust means handing over something valuable to someone, relying on them. The corollary is that "the person who trusts is immediately in a state of vulnerability and dependence", and all the more so on the basis of an explanation whose correctness is difficult to assess. We believe that using human relationship terms, like trust or fairness in the context of machine learning, necessarily induces anthropomorphism, whose bad effects could be addiction (Eliza effect) and persuasion rather than information. In contrast, our philosophical and mathematical research direction tries to define conviviality criteria in machine learning based on Ivan Illich's thought.

According to Illich, a convivial tool must have the following properties:

- it must generate efficiency without degrading personal autonomy;
- it must create neither slave nor master;
- it must widen the personal radius of action.

[2] C. Denis; F. Varenne. Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine. ROIA, Volume 3 (2022) no. 3-4, pp. 287-310.

As presented in the last part of the talk, neural differential equations, by providing trajectories rather than predictions, seem to be an efficient mathematical formalism to implement convivial deep learning tools.

Clarification of the notion of explainability

According to the High-Level Expert Group on AI, « explainability » is contained in the requirement of transparency, as well as « traceability » and « communication »[3]. These concepts are described in the guidelines as « other explicability measures » (e.g. traceability, auditability and transparent communication on system capabilities) and are required when an explanation as to why a model has generated a particular output is not always possible. These cases are referred to as « black box algorithms ». Explainability which seems to be a kind of « sub-requirement » concerns the ability to explain « both the technical processes of an AI system and the related human decisions »[4]. As for « explicability », it is an « ethical principle » that is, according to the High-Level Expert Group, « crucial for building and maintaining users' trust in AI systems ».

It appears that the concepts of explicability and explainability, although terminologically distinct, are not easily distinguished. The experts refer to both explicability and explainability, which is quite confusing given that the doctrine debates the status and definition of these notions. According to Floridi, for example, explicability is a richer notion than explainability, as it is combining demands for intelligibility and accountability[5]. As a result, it enables both people working with and those affected by AI systems to understand and challenge outcomes. Herzog on his part considers that explicability means both more and less than explainability understood in a sense that considers only mechanistic explanations. It means more because explicability demands explanatory interfaces tailored to the recipient and use-case that focus on putting the respective stakeholders in a position to take responsibility. It means less because the principle remains flexible enough to not strictly demand mechanistic explanations at every level of usage and may even allow for none if not available[6]. Others challenge the very existence of a principle of explicability, regardless of the term used: explainability, transparency, understandability.

We suggest to apply here a « grammatical investigation » in the sense that Wittgenstein understood it. In his time, Wittgenstein denounced in a radical way the

[3] Ethics Guidelines for Trustworthy AI, p. 14.

[4] Trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability).

[5] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recom- mendations. Minds Mach. 28, 689–707 (2018). https://doi.org/10. 1007/s11023-018-9482-5

[6] Herzog, C. On the risk of confusing interpretability with explicability. *AI Ethics* **2**, 219–225 (2022). https://doi.org/10.1007/s43681-021-00121-9

syncretism which reigned in science, in ethics (and philosophy) and in epistemology. According to the Viennese philosopher, this attitude encouraged scholars to provide justifications and explanations for phenomena that science had difficulty to understand (what he called the « mythological power of explanation »). To counter this metaphysical drift, Wittgenstein developed a « grammatical-therapeutic philosophy ». Taking into account the wittgensteinian epistemology, we will adopt the definitions proposed by C. Denis and F. Varenne to define the notions of interpretability[7], explainability[8] and understandability[9] (comprehensibility) and to determine their mutual relationship. We claim that before enshrining a principle of explicability, whether at the ethical or legal level, or both, a clear (and unambiguous) definition of the concept must be provided. Making AI explainable must be an epistemic requirement before a moral obligation and a binding principle (principle of explicability).

In the field of AI, the epistemological question is fundamental because the validation of a « black box » differs from that of the mathematical and causal modeling of a physical phenomenon. Indeed, contrary to the previous one, machine learning methods do not pretend to represent a causality between the input and output parameters, despite the use of misleading terms, derived from the statistical theory : the so-called « explanatory » variables[10].

The explanation could also be considered as a « colossus with feet of clay » on the methodological level. Indeed, learning methods are often used when it is difficult or impossible to define the functional specifications of a process. In particular, one interprets the question posed to the machine learning algorithm, and one then wishes to obtain an explanation of the prediction obtained on a question which is not the one solved by the algorithm. There is a famous example commonly used to underline the need for explainability. Suppose you want to implement an algorithm that detects on an image that an animal is a wolf or a husky. The machine learning method uses half images of wolves or husky. The results obtained are spectacular until the day a wolf without a snowy backdrop is detected as a husky. A more in-depth study provides the following explanation: "The learning algorithm did not "learn" to recognize the

[7] Interpretability of a model: set of symbols having the property of being composed of elements (signs, figures, concepts, data, etc.) that each have a meaning for a human subject. A model is interpretable when all its symbols are interpretable

[8] The explainability of a model is the ability to deploy and explain the outputs of the algorithm in a series of steps that are linked together by what a human being can meaningfully interpret as causes or reasons.

[9] The notion of understandability must be defined on the basis of the notion of interpretation (and not the opposite), and must be distinguished from the definition of explanation. There is comprehension of a phenomenon when the human being has the possibility of grasping the whole of it and of unifying its successive manifestations under a single representation easy to conceive and to recall.

[10] Christophe Denis, Franck Varenne. Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine. Revue Ouverte d'Intelligence Artificielle, Association pour la diffusion de la recherche francophone en intelligence artificielle, 2022, 3 (3-4), pp.287-310. ff10.5802/roia.32ff. ffhal03640181

difference between a wolf and a husky but to recognize the presence or absence of snow on the image".

As a corollary, the first reflex is to blame the machine learning method, which deceived us. On the contrary, we believe that the algorithm has done its job correctly, that is to say finding a robust criterion for distinguishing the images. It is therefore not a search for an explanation but the definition of the functional specifications of the black box without degrading its performance that one could not expect with a transparent model.

Last but not least, does a more transparent model facilitate access to the world of knowledge? We do not think so, on the contrary. Following the work of Herbert Simon and cyberneticians, intelligence is obtained by a feedback loop by acting on a black box taking into a simpler world representation. The deep neural network allows, without our yet knowing how to explain it mathematically, to find regularities and symmetries in the complexity of the world.

To conclude, we believe it is a little schizophrenic to think that "opening" the machine learning black boxes would permit us to access to knowledge about our physical world. The functional description, as fine as it is, of the Galilee's telescope does not make it possible for us to understand the ins and outs of the theory of heliocentrism. Our current research consists in placing deep neural networks in this appropriate scientific and epistemic paradigm, the cybernetic one, for which the notion of black box is an asset. Indeed, from signals measured in the visible domain, deep learning manages an informational structure that an human abductive-inductive reasoning allows;