Explainable AI as a consequence of target system ignorance in Machine Learning.

Clément Arlotti

IRT SystemX, 2 Boulevard Thomas Gobert, 91120 Palaiseau, France

Gist of the contribution : Machine Learning models work within an agnostic paradigm that integrates ignorance of target systems. We articulate its implications in the shaping of the explainability problem in AI.

1- Introduction and position of the problem.

In Machine-Learning (ML), the explainability and interpretability problems are most often considered as stemming from the “black-box” character of opaque algorithms leveraging millions (or even billions) of parameters [1], [2], [3], [4]. Two broad approaches of eXplainable Artificial Intelligence (XAI) are used to circumvent this pitfall [1], [5]. On the one hand, complex black boxes are analyzed with post-hoc techniques aimed at describing input-output mappings of the models without addressing their internals [1], [6]. On the other hand, the search for transparency is divided into different levels, e.g. constraining model complexity to ensure interpretability by design [7], or seeking mechanistic decompositions of the inner structures, features and components of complex models to obtain more direct understanding of their behavior [8], [9], [10].

In this article, we argue that both approaches (which we will term model-centric) tend to draw attention towards ML models as isolated entities, at the expense of describing the relation with their target systems in the world.

In the majority of studies, the target of the explanation is the ML model itself [1], [2]. Seminal XAI articles [1], [3] focus on these model-centric approaches without discussing this methodological choice. Furthermore, existing work aiming at providing philosophy-grounded analyses to explainability and interpretability problems generally seek to explain ML models’ outputs and functioning [2], [4] or the behavior of their underlying mathematical function [5]. As such, they do not consider ML models as representations of a target system and do not provide further insights about their relationship with constituting entities and behavior of the phenomena to predict.

In these approaches, the problem is considered as essentially stemming from ML models’ epistemic opacity, due to the complexity of their computations. Albeit relevant, we argue that this analysis would be incomplete without diving into the agnostic modelling paradigm at the core of ML methods.

2- The agnostic modelling paradigm.

These model-centric approaches indeed take place in a specific scientific investigation paradigm known as “agnostic science”, that uses “blind methods” [12]. Importantly, it remains unaddressed in most technical realizations in spite of its important implications in practical aspects, such as model deployment and trustworthiness.

It has been shown that when blind methods are used to make inferences and predictions, they do not rely on, nor provide further understanding of the phenomena [12]. They are agnostic to the real-world entities that constitute the studied phenomena. These methods do not build on prior structured knowledge or hypotheses about the (potentially causal) relations between relevant variables and mainly exploit data correlations to make their inferences. Rather, their strength is to leverage the adaptability of mathematical techniques to organize large amounts of complex data and automatically extract patterns in the models' latent spaces.

Blind methods allow machines to learn from data without being *explicitly* programmed for each step of the tasks they are intended to solve. Consequently, they do not decompose the phenomena into simple elements to provide an analytical, explicit mechanistic model of their parts and interactions. *Explicit* and *explaining* share a common Latin etymology (*explicare*) which corresponds to deploying or unraveling something that is self-contained or folded on itself.

In this article, we draw from the epistemology of models [6], [7], [8] and leverage the notion of target system to further specify what ML models remain "agnostic" or "blind" to, and what should be unraveled and unfolded in XAI. We argue that, in addition to minimizing the role of expert knowledge, target system ignorance impedes proper articulation of the explainability problem, alongside with other important phenomena like hallucinations in Deep Generative Models (DGMs) [9].

3- Target systems in the representational account of modeling.

We recall that the general function of a model is to facilitate the mediation between an epistemic agent and an object (e.g. a real-world phenomenon) as part of a specific questioning [6]. When a model is deemed to be a representation of a phenomenon, it focuses on salient aspects considered relevant for scientific practice, which necessitates a selection activity. The product of this selection is called a target system [10]. It is "carved out" from the phenomena, according to the epistemic agent's background knowledge, goals and interests. The model/target-system relationship sets up which spatio-temporal aspects of the phenomena should be represented and which are discarded as irrelevant. Hence, explanations of the phenomena by the model are built upon the constitutive elements of the identified target system. Furthermore, if these real-world elements can be related back to the model's parameters then the model is said to be explainable "by transitivity of representation" [7].

However, because they take place in the agnostic paradigm, ML models do not readily define their target system. They are built from massive amounts of data agnostically put together and often set up without structuring hypotheses about the phenomena of interest. This contravenes direct target system definition, which consists in identifying a given number of predictive variables, carved out from the multiplicity of potentially available data using expert knowledge, to identify influent factors and provide simple, relevant explanations.

4- The situation of ML models.

Aside from their general function (i.e. facilitating mediation), models can be endowed with a diversity of more specific epistemic functions (e.g. prediction, explanation, data analysis and reduction) enabled by a diversity of means to achieve a diversity of goals [6], [11], [12]. Acknowledging this diversity, we emphasize that ML methods iteratively build and refine inductive models whose epistemic function may be different (both in their construction and goals) from hypothetico-deductive ones. But they are no less models.

We show that ML methods seek to circumvent the intrinsic ignorance at the core of the agnostic paradigm, by defining their target system through repeated iterations of the ML pipeline steps, until model validation (data exploration, pre-processing, feature engineering and selection, model selection and training, evaluation, new iteration or validation). In doing so, a target system is progressively disentangled and unraveled from the agnostic mass of data, and “higher” epistemic functions such as explaining and facilitating understanding [6] can be leveraged.

During the execution of the ML pipeline this can be performed with *post-hoc* methods making use of :

1) The local vs. global dichotomy of XAI [13], to understand the impact of specific data on the explanations.

2) The model-agnostic vs. model-specific dichotomy of XAI [14], so the explanations are not entirely dependent on the model’s inner mechanisms.

However, this methodology still presents important pitfalls. When *post-hoc* explanations deploy surrogate models of the initial model, their underlying (simpler, more interpretable) mechanism is substituted to the original one, offering no guarantee of an adequate representation of the target system (e.g. SHAP and its game-theoretic value-allocation mechanism between ML features [15]). On the other hand, when *post-hoc* explanations simulate feature removal, they do not substitute any mechanism for the original model. They only highlight input-output relations *within* the agnostic framework used to build the model, which by construction may not convey relevant structured knowledge about the target system. One could argue that ML methods are used precisely to cope with the absence of prior knowledge of the input-output relations and such a function has to be built empirically from the data [16]. Because the form of this function is unknown, the outputs of a ML model must be explained in a *post-hoc* fashion. Here again, we would like to stress that not only the form of the input-output function is unknown, but also what this function *represents in the world*, the entities and interactions it leverages to build a target system and explanations.

5- Conclusions

ML models work within an agnostic paradigm that integrates ignorance of target systems. Hence, because explanations of real-world phenomena are made up of target system elements, the explainability problem does not only stem from the computational complexity and opacity of ML models. Although ML pipelines aim at mitigating this pitfall with *post-hoc* techniques and iteratively carve out target systems from the data, they still suffer from intrinsic limitations. Notably because increasing ML models’ complexity and number of parameters initially aims at handling the *a priori* ignorance of the target system’s features. Conversely, focusing on the inner workings of ML models only does not entail defining and elucidating their relationship with the target system. Yet, knowing how a ML model represents the real-world phenomena is key to make it explainable [7], [12]. Unless non-representational accounts of modelling [11] are successfully applied to ML methods.

References

[1] A. Barredo Arrieta *et al.*, « Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI », *Inf. Fusion*, vol. 58, p. 82-115, juin 2020, doi: 10.1016/j.inffus.2019.12.012.

[2] A. Erasmus, T. D. P. Brunet, et E. Fisher, « What is Interpretability? », *Philos. Technol.*, vol. 34, n° 4, p. 833-862, déc. 2021, doi: 10.1007/s13347-020-00435-2.

[3] Z. C. Lipton, « The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. », *Queue*, vol. 16, n° 3, p. 31-57, juin 2018, doi: 10.1145/3236386.3241340.

[4] C. Zednik, « Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence », *Philos. Technol.*, vol. 34, nº 2, p. 265-288, juin 2021, doi: 10.1007/s13347-019-00382-7.

[5] S. Buijsman, « Machine Learning models as Mathematics: interpreting explainable AI in non-causal terms », 2024, *https://philsci-archive.pitt.edu/23201/*.

[6] F. Varenne, *From models to simulations*. in History and philosophy of technoscience, no. 14. Abingdon New York: Routledge, 2019.

[7] C. Denis et F. Varenne, « Interprétabilité et explicabilité pour l'apprentissage machine : entre modèles descriptifs, modèles prédictifs et modèles causaux. Une nécessaire clarification épistémologique », in *National (French) Conference on Artificial Intelligence (CNIA) - Artificial Intelligence Platform (PFIA)*, Toulouse, France, juill. 2019, p. 60-68. Consulté le: 2 décembre 2022. [En ligne]. Disponible sur: https://hal.sorbonne-universite.fr/hal-02184519

[8] T. Knuuttila, N. Carrillo, et R. Koskinen, *The Routledge handbook of philosophy of scientific modeling*. in Routledge handbooks in philosophy. Abingdon, Oxon New York, NY: Routledge, 2024.

[9] C. Arlotti et K. Pasini, « Combining XAI and semiotics to interpret hallucinations in deep generative models », présenté à 4th International Conference on Human and Artificial Rationalities (HAR), Paris, juin 2025. [En ligne]. Disponible sur: https://hal.science/hal-05170068

[10] F. Pero, « Target systems », in *The Routledge handbook of philosophy of scientific modeling*, T. Knuuttila, N. Carrillo, et R. Koskinen, Éd., in Routledge handbooks in philosophy. , Abingdon, Oxon New York, NY: Routledge, 2024, p. 126-137.

[11] T. Knuuttila, « The artifactual approach to modeling », in *The Routledge handbook of philosophy of scientific modeling*, T. Knuuttila, N. Carrillo, et R. Koskinen, Éd., in Routledge handbooks in philosophy. , Abingdon, Oxon New York, NY: Routledge, 2024, p. 111-125.

[12] J. Sánchez-Dorado, « Representation », in *The Routledge handbook of philosophy of scientific modeling*, T. Knuuttila, N. Carrillo, et R. Koskinen, Éd., in Routledge handbooks in philosophy. , Abingdon, Oxon New York, NY: Routledge, 2024, p. 59-73.

[13] C. Molnar *et al.*, « Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process », in *Explainable Artificial Intelligence*, vol. 1901, L. Longo, Éd., in Communications in Computer and Information Science, vol. 1901. , Cham: Springer Nature Switzerland, 2023, p. 456-479. doi: 10.1007/978-3-031-44064-9_24.

[14] C. A. Scholbeck, C. Molnar, C. Heumann, B. Bischl, et G. Casalicchio, « Sampling, Intervention, Prediction, Aggregation: A Generalized Framework for Model-Agnostic Interpretations », in *Machine Learning and Knowledge Discovery in Databases*, vol. 1167, P. Cellier et K. Driessens, Éd., in Communications in Computer and Information Science, vol. 1167. , Cham: Springer International Publishing, 2020, p. 205-216. doi: 10.1007/978-3-030-43823-4_18.

[15] I. C. Covert, S. Lundberg, et S.-I. Lee, « Explaining by removing: a unified framework for model explanation », *J. Mach. Learn. Res.*, vol. 22, nº 1, p. 209:9477-209:9566, janv. 2021.

[16] M. Pégny et M. I. Ibnouhsein, « Quelle transparence pour les algorithmes d'apprentissage machine ? », 2018, *⟨hal-01791021⟩*.

Biography:

Clément Arlotti is a researcher at IRT SystemX, in the Human-AI interaction team. PhD in physics and Master's degree in philosophy of science, his current interests hinge around developing interdisciplinary research in AI and epistemology. In particular eXplainable Artificial Intelligence (XAI), Uncertainty Quantification (UQ) and hallucination characterization for Deep Generative Models (DGMs).