explainable artificial intelligence (XAI)

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

explainable machine learning, explainable ML

Abstract

Explainable artificial intelligence (XAI) is the subfield of artificial intelligence (AI) concerned with making the predictions of machine learning (ML) methods understandable to humans. XAI can be framed as function approximation: a learned hypothesis is explained by a simpler function that a human can comprehend, either locally around a data point (local interpretable model-agnostic explanations (LIME), counterfactuals) or via additive feature contributions (SHapley Additive exPlanations (SHAP)). Such post hoc explanations treat the learned hypothesis as fixed; alternatively, explainable empirical risk minimization (EERM) builds explainability into training. If the hypothesis space is sufficiently simple, a learned hypothesis is its own explanation (see interpretable machine learning (interpretable ML)).

Definition

Explainable artificial intelligence (XAI) is the subfield of artificial intelligence (AI) concerned with making the predictions of machine learning (ML) methods understandable to humans. Each prediction is complemented by an explanation of how it has been obtained. If the ML method uses a sufficiently simple hypothesis space, the learned hypothesis may be inherently interpretable and no separate explanation is needed (Rudin, 2019). The term XAI was popularized by a program of the US Defense Advanced Research Projects Agency (Gunning and Aha, 2019); in the context of ML, the synonymous term explainable machine learning (explainable ML) is also used. The underlying property is explainability, which the international terminology standard for AI defines for artificial intelligence systems (AI systems) in general (International Organization for Standardization and International Electrotechnical Commission, 2022, Sect. 3.5.7).

XAI can be framed as function approximation. A trained hypothesis $\hat{h} : \mathcal{X} \rightarrow \mathcal{Y}$, e.g., delivered by an opaque deep net, is typically a highly non-linear function on a potentially high-dimensional feature space $\mathcal{X}$. Explaining $\hat{h}$ amounts to describing its behavior locally, around a given data point, in a form that a

human can comprehend (see Fig. 1). One form of explanation is a simple local approximation of $\hat{h}$, such as the linear function fitted by local interpretable model-agnostic explanations (LIME), which makes the contribution of each feature to the prediction explicit (Ribeiro et al., 2016). SHapley Additive exPlanations (SHAP) decomposes the prediction into additive feature contributions based on Shapley values (Lundberg and Lee, 2017; Molnar, 2025). Another form of explanation is a counterfactual: instead of approximating $\hat{h}$, it points out changes of the features that change the value of $\hat{h}$, i.e., the prediction. A counterfactual identifies the smallest such change, measured by a chosen metric such as the Euclidean norm (Wachter et al., 2018). Concept activation vectors (CAVs) probe internal representations of a trained hypothesis with linear classifiers to test sensitivity to human-defined concepts (Kim et al., 2018).

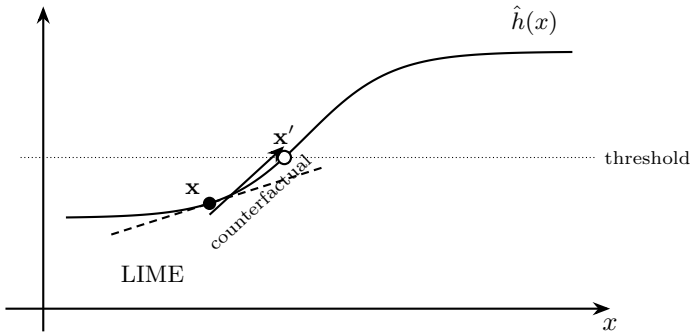

Figure 1: XAI as function approximation. The trained hypothesis $\hat{h}$ (solid curve) maps features to predictions. At a given data point $\mathbf{x}$ (filled dot), LIME (dashed) constructs a local linear approximation. A counterfactual identifies the closest point $\mathbf{x}'$ (open circle), with respect to a chosen metric, at which the prediction crosses a decision threshold

In a loan-approval setting, for example, LIME can indicate which applicant features (income, credit history) contributed most to a rejection, while a counterfactual can state the smallest change in those features that would lead to approval.

These are post hoc methods: they treat the learned hypothesis as fixed and construct the explanation afterwards, from a simpler hypothesis out of a user-parseable hypothesis space (see Fig. 2).

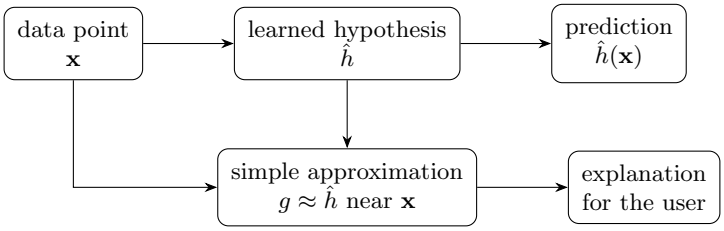

Figure 2: XAI by post hoc approximation. The learned hypothesis $\hat{h}$ delivers the prediction $\hat{h}(\mathbf{x})$ for a data point with feature vector $\mathbf{x}$. An explanation is constructed from a simple hypothesis g, e.g., a linear map, that approximates $\hat{h}$ near $\mathbf{x}$ (cf. LIME)

Explanations are useful only if they match the background of the user: explainability is measured relative to a specific user (Colin et al., 2022; Jung and Nardelli, 2020). Explanations can be local, concerning a single prediction, or global, characterizing the learned hypothesis as a whole (Molnar, 2025). Instead of constructing explanations post hoc, explainable empirical risk minimization (EERM) builds the requirement of explainability into training itself, via a regularization term that favors hypotheses which are intrinsically explainable for a specific user (Zhang et al., 2024).

A goal of XAI is to quantify the influence of features on the prediction of a learned hypothesis, whether the feature is an input attribute of a data point or, in the sense of mechanistic interpretability, a concept encoded internally, e.g., by the activations of an artificial neural network (ANN).

Cross References

explainability, interpretability, explanation, interpretable ML, EERM, LIME, SHAP, counterfactual, feature, mechanistic interpretability

References

1. Colin et al. (2022). What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. In: Adv. Neural Inf. Process. Syst., pp. 2832–2845.

2. Gunning and Aha (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), pp. 44–58.

3. International Organization for Standardization and International Electrotechnical Commission (2022). ISO/IEC 22989:2022 — Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. International Standard, 1st edition. URL: https://www.iso.org/standard/74296.html.

4. Jung and Nardelli (2020). An Information-Theoretic Approach to Personalized Explainable Machine Learning. IEEE Signal Process. Lett., 27, pp. 825–829.

5. Lundberg and Lee (2017). A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774.

6. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.

7. Ribeiro et al. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135–1144.

8. Wachter et al. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), pp. 841–887.

9. Zhang et al. (2024). Explainable empirical risk minimization. Neural Comput. Appl., 36(8), pp. 3983–3996.

10. Kim et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Mach. Learn., pp. 2668–2677.

11. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.