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. Much of it can be posed as a function approximation problem: a learned hypothesis is explained by a simpler function that a human can comprehend. Explanations differ in what they are made of. One kind reports a score for each feature, read off a linear map fitted near the data point (local interpretable model-agnostic explanations (LIME)) or obtained as additive contributions (SHapley Additive exPlanations (SHAP)). A counterfactual instead names another feature vector, the closest one whose prediction differs. A concept activation vector (CAV) states the explanation in a concept the user names, located in the activations inside the trained method. All three are post hoc and treat the learned hypothesis as fixed. The alternative is a hypothesis that needs no separate explanation, reached by restricting the hypothesis space to hypotheses a human comprehends or by regularization that favors explainable ones (explainable empirical risk minimization (EERM)).

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. It aims to complement each prediction 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); see interpretable machine learning (interpretable ML).

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 posed as a function approximation problem. A trained hypothesis $\hat{h} : \mathcal{X} \to \mathcal{Y}$, e.g., delivered by an opaque deep net, is typically a highly nonlinear function on a potentially high-dimensional feature space $\mathcal{X}$ (Goodfellow et al., 2016, Ch. 6). Explaining $\hat{h}$ at a data point with feature vector $\mathbf{x} \in \mathcal{X}$ means answering one question about that function in a form a human can comprehend. The examples of explanations below are of three kinds, distinguished by what the explanation is made of. The first is a score for each feature of the data point. The scores are read off a simpler hypothesis out of a user-parseable hypothesis space that agrees with $\hat{h}$ near $\mathbf{x}$, for instance a linear map whose weights are those scores, as fitted by local interpretable model-agnostic explanations (LIME). Fig. 1 draws that construction. The second is another feature vector, one whose prediction differs. The third is a concept the user names, located inside the trained method rather than among the features of the data point. All three are post hoc: they treat the learned hypothesis as fixed and construct the explanation after training.

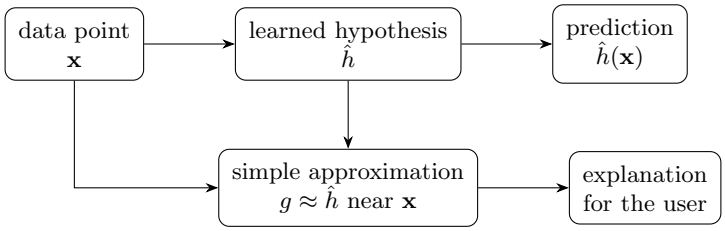

Figure 1: The first kind of explanation. The learned hypothesis $\hat{h}$ delivers the prediction $\hat{h}(\mathbf{x})$ for a data point with feature vector $\mathbf{x}$. The explanation is constructed from a simple hypothesis g, e.g., a linear map, that approximates $\hat{h}$ near $\mathbf{x}$; the score reported for a feature is the weight g gives it (cf. LIME)

The first kind of explanation reports one score per feature, saying how much that feature contributed to the prediction. LIME obtains the scores by fitting a linear map to $\hat{h}$ near $\mathbf{x}$ and reading off its weights (Ribeiro et al., 2016). SHapley Additive exPlanations (SHAP) obtains them by decomposing the prediction into a base value and one contribution per feature, the contributions being Shapley values (Lundberg and Lee, 2017; Molnar, 2025). The two deliver the same kind of explanation and differ in how its scores are computed. When features are properties of image pixels, the feature scores are a second image of the same size: one relevance score per pixel (see Fig. 2). The class activation map (CAM) is one construction of such a per-pixel score.

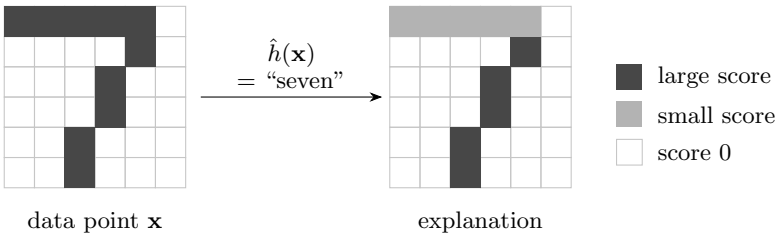

Figure 2: An explanation of the first kind for an image, one score per feature, as a CAM delivers it. The objects are named as in Fig. 1. Left: a data point $\mathbf{x}$, an image on a grid of 6×6 pixels, for which the learned hypothesis delivers the prediction $\hat{h}(\mathbf{x}) =$ "seven". Right: the explanation of that prediction, one relevance score per pixel. The pixels of the diagonal stroke carry the largest scores, the top bar smaller ones, and the remaining pixels score zero

The second kind of explanation is a counterfactual. It reports no scores and approximates nothing. It asks instead where $\hat{h}$ takes a different value, pointing out how the features must be modified for the prediction $\hat{h}(\mathbf{x})$ to come out differently. A counterfactual identifies the smallest such modification, with closeness measured by a chosen metric (Wachter et al., 2018) (see Fig. 3).

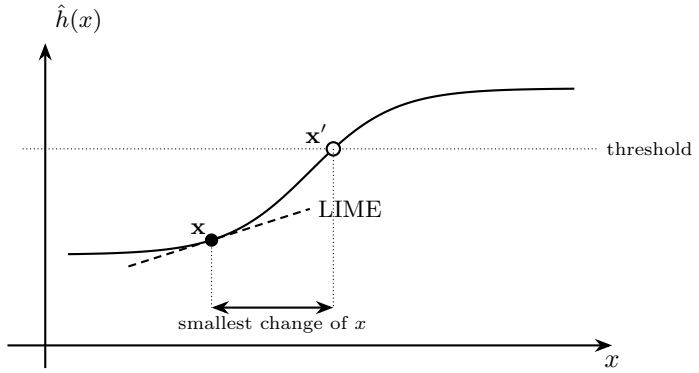

Figure 3: The first two kinds of explanation for one prediction. The trained hypothesis $\hat{h}$ (solid curve) delivers a prediction for each value of the feature. At a given data point $\mathbf{x}$ (filled circle), LIME (dashed line) constructs a local linear approximation. A counterfactual identifies the closest point $\mathbf{x}'$ (open circle), with respect to a chosen metric, at which the prediction crosses a decision threshold. The double arrow marks the change of the feature it reports

In a loan-approval setting, for example, LIME reports which of the applicant features, such as income and credit history, its local linear approximation assigns the largest weight to in a rejection, while a counterfactual states the smallest modification of those features that would turn the prediction into an approval.

The third kind of explanation refers not to the features of the data point but to a concept the user specifies. Such a concept is defined by a dataset of examples the user supplies. A concept activation vector (CAV) locates such a concept as a direction in the space of the activations of one hidden layer of an artificial neural network (ANN): the normal vector of a decision boundary that separates the supplied examples carrying the concept from those that do not (Kim et al., 2018).

In contrast to LIME and counterfactual explanations, a CAV needs access to the internal computations of a machine learning system (ML system). One such internal computation is the activation $\mathbf{z} = f(\mathbf{x})$ of a hidden layer in a deep net. A deep net computes $\hat{h}$ by feeding the activations $\mathbf{z}$ to the final layers, which deliver a score $s(\mathbf{z})$ for each label; the prediction $\hat{h}(\mathbf{x})$ follows from those scores (see CAV). In an image classifier, for example, a user who suspects that photographs are labeled zebra through the concept "stripes" supplies photographs showing stripes and photographs not showing them. Each of these photographs results in an activation $\mathbf{z}$ in the layer of interest. These activations are then used to fit a linear classifier that separates the two sets, and its normal vector is the CAV for "stripes". How much the concept contributed to the prediction of a zebra photograph is then measured by the directional derivative of the score s along that CAV.

The above XAI methods construct an explanation after training, for a learned hypothesis that is already fixed. The alternative is to arrive at one that needs no separate explanation. Interpretable ML does so by admitting only hypotheses a human comprehends directly. This can be achieved by manually choosing a hypothesis space that is simple enough, or by regularization that favors hypotheses with specific properties, such as sparsity or predictability (Tibshirani, 1996; Zhang et al., 2024). Explainable empirical risk minimization (EERM) takes the second route for explainability: the user supplies their own predictions for the data points of a training set, and the penalty term charges the part of the learned hypothesis that those predictions do not already account for (Zhang et al., 2024).

Three aspects of an explanation are studied: whom it serves, how much of the learned hypothesis it covers, and whether it is faithful. How much an explanation achieves depends on who receives it: 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).

An explanation must be faithful, i.e., reflect the computation that the learned hypothesis actually carries out. When the explanation scores the features, this can be tested rather than asserted. A CAM is faithful for a prediction if flipping the pixels it scores highest changes that prediction more often than flipping as many of the pixels it scores low. Perturbing the highest-scoring regions first and recording how quickly the predicted class score falls is the standard form of the test (Samek et al., 2017, Sect. III-C).

For the image of Fig. 2 and a linear classifier fitted to images of that kind,

flipping the three highest-scoring pixels changes the prediction, while flipping the pixels in the opposite order leaves it unchanged through all 36 of them; the map-guided order also changes the prediction sooner on each of 200 noisy variants of the image. Faithfulness also limits what a post hoc explanation can achieve: one that agreed with the learned hypothesis everywhere would be that hypothesis itself, so a simpler explanation deviates from it somewhere (Rudin, 2019).

Two legal instruments explicitly state requirements on XAI. Under the general data protection regulation (GDPR), a person subjected to a decision taken by automated means is entitled to meaningful information about the logic involved in that decision (European Parliament and Council of the European Union, 2016, Art. 15(1)(h)). Communicating the algorithm itself is not a sufficiently concise and intelligible explanation (Court of Justice of the European Union, 2025). Counterfactual information can be appropriate: the extent to which a variation in the personal data would have led to a different result. The second instrument is the EU AI Act, which requires for a high-risk artificial intelligence system (high-risk AI system) that affected persons obtain clear and meaningful explanations of the role of the AI system in the decision procedure (European Parliament and Council of the European Union, 2024, Art. 86) (see right to explanation).

Cross References

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

References

1. Court of Justice of the European Union (2025). Case C-203/22, CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien. Judgment of the Court (First Chamber) of 27 February 2025. URL: https://curia.europa.eu/juris/liste.jsf?num=C-203/22.

2. 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.

3. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

4. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

5. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

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

7. 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.

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

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

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

11. 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.

12. Samek et al. (2017). Evaluating the Visualization of What a Deep Neural Network Has Learned. IEEE Trans. Neural Netw. Learn. Syst., 28(11), pp. 2660–2673.

13. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Statist. Soc.: Ser. B (Methodological), 58(1), pp. 267–288.

14. 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.

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

16. 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.

17. 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.