explainability

Authors

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

Abstract

A machine learning (ML) method is explainable if there is an effective way to explain its predictions. The method delivers an explanation along with every prediction, and that explanation is effective if it lets a human user comprehend how the features of a data point drive the prediction made for it. Explainability is always relative to a specific user (group): an explanation can be very effective for a specific user (group) but useless for another user (group). Explainability can be measured by comparing the predictions of a learned hypothesis to the prediction of a user before and after they are provided with an explanation.

Definition

The explainability of a machine learning (ML) method is the level to which a human user can anticipate the predictions it delivers, based on the explanations it provides (Colin et al., 2022; Jung and Nardelli, 2020). Explainability thus includes the notion of an explanation: each prediction is delivered with an explanation for this specific prediction, such as the feature values that drove it or, for image data, the relevant pixels (see Fig. 1). The definition asks for the existence of an explanation, not for a particular one: an ML method is explainable as soon as some explanation lets the user anticipate its predictions. Explainability is therefore certified by exhibiting one explanation that works, and a single explanation the user cannot follow refutes nothing. Explainability is user-relative: the same learned hypothesis can be explainable for one user and inscrutable for another.

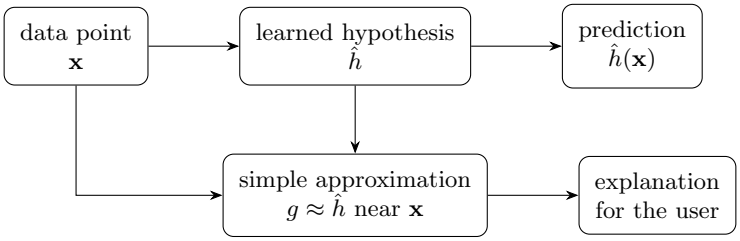

Figure 1: Explainability refers to explanations. The learned hypothesis $\hat{h}$ delivers the prediction $\hat{h}(\mathbf{x})$ for a data point with feature vector $\mathbf{x}$. An explanation is constructed from a simple hypothesis g, e.g., a linear map, that approximates $\hat{h}$ near $\mathbf{x}$ (cf. local interpretable model-agnostic explanations (LIME)). The explanation is delivered to the user along with the prediction

The usefulness of an explanation can be measured by how much it enables the user to anticipate the predictions on a curated test set: the user states a prediction for each data point in the test set. If the user comprehends the method, these should be close to the predictions of the learned hypothesis (Colin et al., 2022; Zhang et al., 2024). A measure of this form is referred to as predictability (International Organization for Standardization and International Electrotechnical Commission, 2022, Sect. 5.15.7).

Using a probabilistic model for data generation allows measuring explainability by the conditional differential entropy of the predictions (Chen et al., 2018; Jung and Nardelli, 2020). The conditional differential entropy quantifies the uncertainty about the predictions that remains once the anticipations are known, so a smaller value means that the anticipations determine the predictions more tightly. In practice it is unknown and must be replaced by an estimator, e.g., a plug-in estimate computed from discretized predictions and anticipations on a test set.

An explanation raises explainability when the user can reason with it. Fig. 2 illustrates this for a user who reasons in terms of linear maps and anticipates the predictions of an opaque hypothesis (a kernel method): without explanations, the anticipations deviate strongly from the predictions. Given a local linear approximation of the hypothesis around each data point (cf. LIME), the anticipations match the predictions almost exactly. The example presupposes that the user knows how to apply a linear approximation: the anticipations improve only because the user can evaluate the explanation for a data point. A user who cannot is left where they started.

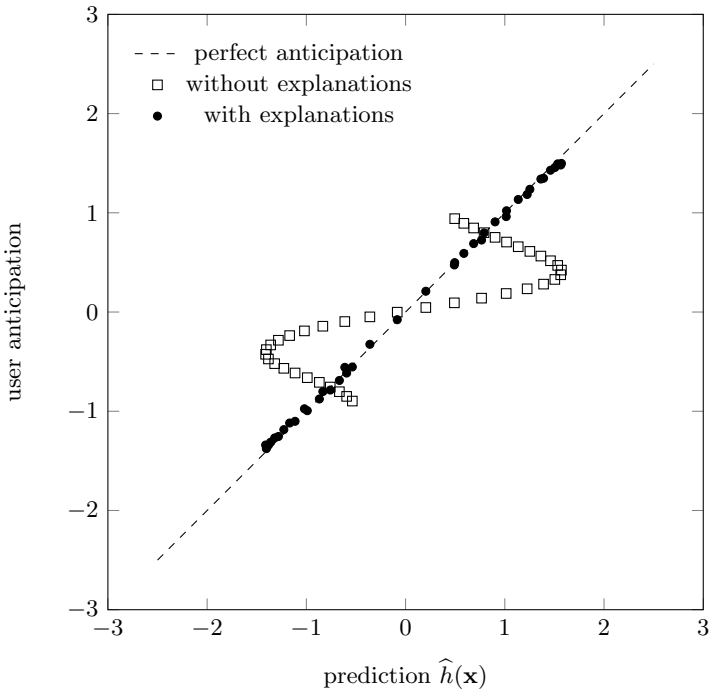

Figure 2: User anticipations against the predictions of an opaque hypothesis (Gaussian kernel ridge regression) on a test set, for a user who reasons in terms of linear maps. Without explanations (open squares), the anticipations deviate strongly from the predictions; with a local linear approximation around each data point as explanation (filled circles), the anticipations track the diagonal of perfect anticipation. The estimated conditional differential entropy (a plug-in estimate on the test set) of the predictions given the anticipations drops from 1.22 to 0.09 bits. Data generated by `pythondemos/explainability.py`

The international terminology standard for artificial intelligence (AI) defines explainability as the property of an artificial intelligence system (AI system) to express important factors influencing its results in a way that humans can understand. An explanation is intended to answer the question of why the AI system produced a result, without arguing that the result was optimal (International Organization for Standardization and International Electrotechnical Commission, 2022, Sect. 3.5.7). The AI risk management framework of the US National Institute of Standards and Technology ties explainability to a representation of the mechanisms underlying the operation of an AI system, and reserves interpretability for the meaning of the output in the context of the designed purpose (National Institute of Standards and Technology, 2023).

Regulation treats explainability as an ingredient of transparency. The EU AI Act, in its Article 13, requires that high-risk artificial intelligence systems (high-risk AI systems) are sufficiently transparent to enable deployers to interpret their

outputs (European Parliament and Council of the European Union, 2024). For individual automated decisions, the right to explanation entitles an affected person to a clear and meaningful account of the role that an AI system played in the decision (European Parliament and Council of the European Union, 2024, Art. 86).

What counts as an automated decision is drawn widely: the Court of Justice of the European Union held that a credit information agency computing a person's ability to meet future payments already makes one, when a third party draws strongly on that value (Court of Justice of the European Union, 2023). The Colorado Automated Decision-Making Technology Act of 2026 requires that the deployer give the affected consumer a plain-language description of the decision and of the role the technology played in it, together with the means to request further information (Colorado General Assembly, 2026).

Cross References

explanation, interpretability, XAI, EERM, LIME, transparency, right to explanation, trustworthy AI, regularization

References

1. Court of Justice of the European Union (2023). Judgment of 7 December 2023, SCHUFA Holding (Scoring), C-634/21, ECLI:EU:C:2023:957. Court of Justice of the European Union, First Chamber. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0634.

2. Colorado General Assembly (2026). Senate Bill 26-189 — Automated Decision-Making Technology. Colorado General Assembly, 2026 Regular Session, signed May 14, 2026. URL: https://leg.colorado.gov/bills/sb26-189.

3. Chen et al. (2018). Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In: Proc. 35th Int. Conf. Mach. Learn., pp. 883–892.

4. 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.

5. 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.

6. 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.

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

8. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. URL: https://doi.org/10.6028/NIST.AI.100-1.

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