



IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence

IEEE Computer Society

Developed by the
Artificial Intelligence Standards Committee

IEEE Std 2894™-2024

STANDARDS

IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence

Developed by the

Artificial Intelligence Standards Committee
of the
IEEE Computer Society

Approved 15 February 2024

IEEE SA Standards Board

Abstract: A new wave of artificial intelligence applications that offer extensive benefits to our daily lives has been led to by dramatic success in machine learning. The loss of explainability during this transition, however, means vulnerability to vicious data, poor model structure design, and suspicion of stakeholders and the general public—all with a range of legal implications. The study of explainable AI (XAI), which is an active research field that aims to make AI systems results more understandable to humans, has been called for by this dilemma. This is a field with great hopes for improving the trust and transparency of AI-based systems and is considered a necessary route for AI to move forward. A technological blueprint for building, deploying, and managing machine learning models, while meeting the requirements of transparent and trustworthy AI by adopting a variety of XAI methodologies, is provided by this guide. It defines the architectural framework and application guidelines for explainable AI, including: description and definition of XAI; the types of XAI methods and the application scenarios to which each type applies; and performance evaluation of XAI.

Keywords: AI, architectural framework, artificial intelligence, explainable AI, explainable artificial intelligence, IEEE 2894™, machine learning, XAI

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2024 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published 30 August 2024. Printed in the United States of America.

IEEE is a registered trademark in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 979-8-8557-0785-4 STD26972
Print: ISBN 979-8-8557-0786-1 STDPD26972

IEEE prohibits discrimination, harassment, and bullying.

For more information, visit <https://www.ieee.org/about/corporate/governance/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system or otherwise, without the prior written permission of the publisher.

Important Notices and Disclaimers Concerning IEEE Standards Documents

IEEE Standards documents are made available for use subject to important notices and legal disclaimers. These notices and disclaimers, or a reference to this page (<https://standards.ieee.org/ipr/disclaimers.html>), appear in all IEEE standards and may be found under the heading “Important Notices and Disclaimers Concerning IEEE Standards Documents.”

Notice and Disclaimer of Liability Concerning the Use of IEEE Standards Documents

IEEE Standards documents are developed within IEEE Societies and subcommittees of IEEE Standards Association (IEEE SA) Board of Governors. IEEE develops its standards through an accredited consensus development process, which brings together volunteers representing varied viewpoints and interests to achieve the final product. IEEE standards are documents developed by volunteers with scientific, academic, and industry-based expertise in technical working groups. Volunteers involved in technical working groups are not necessarily members of IEEE or IEEE SA and participate without compensation from IEEE. While IEEE administers the process and establishes rules to promote fairness in the consensus development process, IEEE does not independently evaluate, test, or verify the accuracy of any of the information or the soundness of any judgments contained in its standards.

IEEE makes no warranties or representations concerning its standards, and expressly disclaims all warranties, express or implied, concerning all standards, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. IEEE Standards documents do not guarantee safety, security, health, or environmental protection, or compliance with law, or guarantee against interference with or from other devices or networks. In addition, IEEE does not warrant or represent that the use of the material contained in its standards is free from patent infringement. IEEE Standards documents are supplied “AS IS” and “WITH ALL FAULTS.”

Use of an IEEE standard is wholly voluntary. The existence of an IEEE standard does not imply that there are no other ways to produce, test, measure, purchase, market, or provide other goods and services related to the scope of the IEEE standard. Furthermore, the viewpoint expressed at the time a standard is approved and issued is subject to change brought about through developments in the state of the art and comments received from users of the standard.

In publishing and making its standards available, IEEE is not suggesting or rendering professional or other services for, or on behalf of, any person or entity, nor is IEEE undertaking to perform any duty owed by any other person or entity to another. Any person utilizing any IEEE Standards document should rely upon their own independent judgment in the exercise of reasonable care in any given circumstances or, as appropriate, seek the advice of a competent professional in determining the appropriateness of a given IEEE standard.

IN NO EVENT SHALL IEEE BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: THE NEED TO PROCURE SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE PUBLICATION, USE OF, OR RELIANCE UPON ANY STANDARD, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Translations

The IEEE consensus balloting process involves the review of documents in English only. In the event that an IEEE standard is translated, only the English language version published by IEEE is the approved IEEE standard.

Use by artificial intelligence systems

In no event shall material in any IEEE Standards documents be used for the purpose of creating, training, enhancing, developing, maintaining, or contributing to any artificial intelligence systems without the express, written consent of IEEE SA in advance. “Artificial intelligence” refers to any software, application, or other system that uses artificial intelligence, machine learning, or similar technologies, to analyze, train, process, or generate content. Requests for consent can be submitted using the Contact Us form.

Official statements

A statement, written or oral, that is not processed in accordance with the IEEE SA Standards Board Operations Manual is not, and shall not be considered or inferred to be, the official position of IEEE or any of its committees and shall not be considered to be, or be relied upon as, a formal position of IEEE or IEEE SA. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that the presenter’s views should be considered the personal views of that individual rather than the formal position of IEEE, IEEE SA, the Standards Committee, or the Working Group. Statements made by volunteers may not represent the formal position of their employer(s) or affiliation(s). News releases about IEEE standards issued by entities other than IEEE SA should be considered the view of the entity issuing the release rather than the formal position of IEEE or IEEE SA.

Comments on standards

Comments for revision of IEEE Standards documents are welcome from any interested party, regardless of membership affiliation with IEEE or IEEE SA. However, **IEEE does not provide interpretations, consulting information, or advice pertaining to IEEE Standards documents.**

Suggestions for changes in documents should be in the form of a proposed change of text, together with appropriate supporting comments. Since IEEE standards represent a consensus of concerned interests, it is important that any responses to comments and questions also receive the concurrence of a balance of interests. For this reason, IEEE and the members of its Societies and subcommittees of the IEEE SA Board of Governors are not able to provide an instant response to comments or questions, except in those cases where the matter has previously been addressed. For the same reason, IEEE does not respond to interpretation requests. Any person who would like to participate in evaluating comments or revisions to an IEEE standard is welcome to join the relevant IEEE SA working group. You can indicate interest in a working group using the Interests tab in the Manage Profile & Interests area of the [IEEE SA myProject system](#).¹ An IEEE Account is needed to access the application.

Comments on standards should be submitted using the [Contact Us](#) form.²

¹ Available at: <https://development.standards.ieee.org/myproject-web/public/view.html#landing>.

² Available at: <https://standards.ieee.org/about/contact/>.

Laws and regulations

Users of IEEE Standards documents should consult all applicable laws and regulations. Compliance with the provisions of any IEEE Standards document does not constitute compliance to any applicable regulatory requirements. Implementers of the standard are responsible for observing or referring to the applicable regulatory requirements. IEEE does not, by the publication of its standards, intend to urge action that is not in compliance with applicable laws, and these documents may not be construed as doing so.

Data privacy

Users of IEEE Standards documents should evaluate the standards for considerations of data privacy and data ownership in the context of assessing and using the standards in compliance with applicable laws and regulations.

Copyrights

IEEE draft and approved standards are copyrighted by IEEE under U.S. and international copyright laws. They are made available by IEEE and are adopted for a wide variety of both public and private uses. These include both use by reference, in laws and regulations, and use in private self-regulation, standardization, and the promotion of engineering practices and methods. By making these documents available for use and adoption by public authorities and private users, neither IEEE nor its licensors waive any rights in copyright to the documents.

Photocopies

Subject to payment of the appropriate licensing fees, IEEE will grant users a limited, non-exclusive license to photocopy portions of any individual standard for company or organizational internal use or individual, non-commercial use only. To arrange for payment of licensing fees, please contact Copyright Clearance Center, Customer Service, 222 Rosewood Drive, Danvers, MA 01923 USA; +1 978 750 8400; <https://www.copyright.com/>. Permission to photocopy portions of any individual standard for educational classroom use can also be obtained through the Copyright Clearance Center.

Updating of IEEE Standards documents

Users of IEEE Standards documents should be aware that these documents may be superseded at any time by the issuance of new editions or may be amended from time to time through the issuance of amendments, corrigenda, or errata. An official IEEE document at any point in time consists of the current edition of the document together with any amendments, corrigenda, or errata then in effect.

Every IEEE standard is subjected to review at least every 10 years. When a document is more than 10 years old and has not undergone a revision process, it is reasonable to conclude that its contents, although still of some value, do not wholly reflect the present state of the art. Users are cautioned to check to determine that they have the latest edition of any IEEE standard.

In order to determine whether a given document is the current edition and whether it has been amended through the issuance of amendments, corrigenda, or errata, visit [IEEE Xplore](#) or [contact IEEE](#).³ For more information about the IEEE SA or IEEE's standards development process, visit the IEEE SA Website.

Errata

Errata, if any, for all IEEE standards can be accessed on the [IEEE SA Website](#).⁴ Search for standard number and year of approval to access the web page of the published standard. Errata links are located under the Additional Resources Details section. Errata are also available in [IEEE Xplore](#). Users are encouraged to periodically check for errata.

Patents

IEEE standards are developed in compliance with the [IEEE SA Patent Policy](#).⁵

Attention is called to the possibility that implementation of this standard may require use of subject matter covered by patent rights. By publication of this standard, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. If a patent holder or patent applicant has filed a statement of assurance via an Accepted Letter of Assurance, then the statement is listed on the IEEE SA Website at <https://standards.ieee.org/about/sasb/patcom/patents.html>. Letters of Assurance may indicate whether the Submitter is willing or unwilling to grant licenses under patent rights without compensation or under reasonable rates, with reasonable terms and conditions that are demonstrably free of any unfair discrimination to applicants desiring to obtain such licenses.

Essential Patent Claims may exist for which a Letter of Assurance has not been received. The IEEE is not responsible for identifying Essential Patent Claims for which a license may be required, for conducting inquiries into the legal validity or scope of Patents Claims, or determining whether any licensing terms or conditions provided in connection with submission of a Letter of Assurance, if any, or in any licensing agreements are reasonable or non-discriminatory. Users of this standard are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. Further information may be obtained from the IEEE Standards Association.

IMPORTANT NOTICE

Technologies, application of technologies, and recommended procedures in various industries evolve over time. The IEEE standards development process allows participants to review developments in industries, technologies, and practices, and to determine what, if any, updates should be made to the IEEE standard. During this evolution, the technologies and recommendations in IEEE standards may be implemented in ways not foreseen during the standard's development. IEEE standards development activities consider research and information presented to the standards development group in developing any safety recommendations. Other information about safety practices, changes in technology or technology implementation, or impact by peripheral systems also may be pertinent to safety considerations during implementation of the standard. Implementers and users of IEEE Standards documents are responsible for determining and complying with all appropriate safety, security, environmental, health, data privacy, and interference protection practices and all applicable laws and regulations.

³ Available at: <https://ieeexplore.ieee.org/browse/standards/collection/ieee>.

⁴ Available at: <https://standards.ieee.org/standard/index.html>.

⁵ Available at: <https://standards.ieee.org/about/sasb/patcom/materials.html>.

Participants

At the time this guide was completed, the Explainable Artificial Intelligence Working Group had the following membership:

Lixin Fan, Chair
Yixin Chen, Vice Chair

<i>Organization Represented</i>	<i>Name of Representative</i>
Beijing Baidu Netcom Science Technology Co., Ltd.	Xuhong Li
CETC Big Data Research Institute Co., Ltd.	Xu Cheng
China Asset Management	Xiaoya Wang
China Mobile Communications Co. Ltd.	Xiaojia Song
China Telecommunications Corporation	Zuping Wu
ENN Group	Xiaoxu Ma
Hisense Co., Ltd.	Xuesong Gao
Huawei Technologies Co., Ltd.	Chi Shu
Institute of Computing Technology, Chinese Academy of Sciences	Yiqiang Chen
JD.com, Inc.	Junbo Zhang
Shanghai Artificial Intelligence Association	Chunhao Zhao
Shanghai Development Center of Computer Software Technology	Mingang Chen
Sinovation Ventures	Ji Feng
YITU Technology	Chuqi Yang

The following members of the entity Standards Association balloting group voted on this guide. Balloters may have voted for approval, disapproval, or abstention.

0xSenses Corporation	CRRC Zhuzhou Institute Co. Ltd.	Shandong Academy of Intelligent Computing Technology
Alipay (China) Technology Co. Ltd.	Institute of Biomedical Engineering	Shandong Computer Science Center (National Supercomputing Center in Jinan)
Beijing Tsingzhi Shuyuan Tech Inc.	Intel	YITU Technology
Boeing Company	Lenovo Group Limited	
China Telecommunications Corporation		

When the IEEE SA Standards Board approved this guide on 15 February 2024, it had the following membership:

David J. Law, Chair
Vacant Position, Vice Chair
Gary Hoffman, Past Chair
Alpesh Shah, Secretary

Sara R. Biyabani	Yousef Kimiagar	Robby Robson
Ted Burse	Joseph L. Koepfinger*	Jon Walter Rosdahl
Stephen Dukes	Howard Li	Mark Siira
Doug Edwards	Xiaohui Liu	Lei Wang
J. Travis Griffith	John Haiying Lu	F. Keith Waters
Guido R. Hiertz	Kevin W. Lu	Sha Wei
Ronald W. Hotchkiss	Hiroshi Mano	Philip B. Winston
Hao Hu	Paul Nikolich	Don Wright

*Member Emeritus

Introduction

This introduction is not part of IEEE Std 2894-2024, IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence.

Dramatic success in machine learning has led to a new wave of artificial intelligence (AI) applications that offer extensive benefits to our daily lives. But the complete transition to an algorithmic society has still not yet come since the insight and trust in the optimization modules decrease as we move from traditional mathematically dependent optimization to data-dependent deep learning models.

We have been lingering in front of the fourth industrial revolution for a long time, while it is limited by the machines' current inability to explain their decisions and actions to human users. As a result, wider and deeper adoption of AI-based systems is hindered by their lack of transparency. In fact, the black-box nature of these systems can provide predictive outputs but cannot be directly explained. This is typically critical in industries like education, manufacturing, law enforcement, healthcare, and finance, where we find that the need to trust these AI-based systems for making predictions and decisions is paramount. This problem gradually emerged with the large-scale success of deep learning in the last decade. This is due to remarkable performance improvement as well as intrinsic opaqueness in complex deep neural network architectures. Although said to imitate the information transmission of the biological nervous system, deep learning is still regarded as constructing superficially continuous functions fitted on highly structured data. The loss of explainability means vulnerability to vicious data, poor model structure design, and suspicion of stakeholders and the public—all with a range of legal implications.

The dilemma has sparked a discussion about explainable AI (XAI). This is a field with great hopes for improving the trust and transparency of AI-based systems and is considered a necessary route for AI to move forward. The term *XAI* tends to refer to the initiatives and efforts made in response to AI transparency and trust concerns. In general, XAI is a research field that aims to make AI systems' results more understandable to humans.

Technically, there is no standard, generally accepted definition or formal concept of XAI. For the purpose of systemizing and standardizing the research community in this area, IEEE has the IEEE Explainable Artificial Intelligence Working Group. The standard will serve as a technological framework that allows the adoption of XAI in a large variety of application systems, which are required to meet governance criteria concerning privacy, transparency, accountability, responsibility, and fairness for multiple application areas such as finance, business, and healthcare.

Contents

1. Overview	10
1.1 Scope	10
1.2 Purpose	10
1.3 Word usage	11
2. Normative references.....	11
3. Definitions, acronyms, and abbreviations	11
3.1 Definitions	11
3.2 Acronyms and abbreviations	12
4. XAI architectures.....	12
4.1 General	12
4.2 Interpretability goals of AI	13
4.3 Taxonomy of explainability.....	14
4.4 Explanations properties	16
4.5 XAI reference architecture	19
5. XAI requirements	21
5.1 Explanation method properties	21
5.2 Levels of detail of explanation	22
6. XAI Explanation assessment.....	23
6.1 General	23
6.2 Transparency explanation.....	23
6.3 Post hoc explanation.....	24
6.4 Qualitative assessment.....	24
6.5 Model-specific explanation methods.....	33
6.6 Data-centric explanation methods	38
Annex A (informative) List of advantages and disadvantages of post-modeling evaluation methods	40
Annex B (informative) A case study of infidelity and sensitivity for gradient explanations.....	41
B.1 Gradient	41
B.2 Finance.....	42
B.3 Health.....	44
B.4 Urban computing	45
B.5 Energy.....	46
B.6 Telecommunication.....	47
Annex C (informative) Bibliography.....	50

IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence

1. Overview

Artificial intelligence (AI) technologies are now used in an increasingly wider domains of applications, including healthcare, finance, telecommunications, and so forth. On one hand, outcomes or decisions made by these AI systems impact billions of people almost every day or even every second. On the other hand, the underlying mechanisms that govern this decision-making are not always explicit and understandable to stakeholders of the AI system. The lack of explainability may cast doubts on the adoptions of these AI systems, which are expected to be trustworthy and adhere to human ethics principles. This challenge has sparked the study of a variety of explainable artificial intelligence (XAI) technologies that aim to provide explanations of AI systems and, subsequently, improve transparency and trustworthiness of the systems.

1.1 Scope

This guide provides a technological framework that aims to increase trustworthiness of AI systems using explainable artificial intelligence (XAI) technologies and methods. The document also provides measurable solutions to evaluate AI systems in terms of explainability. Specifically, the document illustrates the following aspects of XAI systems:

- a) The requirements of providing human-understandable explanations for AI systems in different use-cases, for example, healthcare and financial applications
- b) Approaches to offer a series of available tools for giving an AI model tenable explanations
- c) A set of measurable solutions to evaluate AI systems and corresponding performance, such as the availability, resiliency, accuracy, safety, security, and privacy of the AI system under such status

1.2 Purpose

Artificial intelligence systems are expected to be trustworthy and adhere to human ethics principles including fairness, privacy, transparency, and so forth. Achieving the trustworthiness requires underlying mechanisms of AI systems that are transparent and understandable to all stakeholders of AI systems. This mandate motivates the study of a variety of explainable artificial intelligence (XAI) technologies and methods. Given the urgent need and profound influence of XAI, this guide aims to provide a technological framework that facilitates the adoption of relevant methods, facilitates the evaluation and comparison of

different approaches, and showcases typical scenarios in which XAI can bring great value to AI system stakeholders and our society.

1.3 Word usage

The word *shall* indicates mandatory requirements strictly to be followed in order to conform to the standard and from which no deviation is permitted (*shall* equals *is required to*).^{6,7}

The word *should* indicates that among several possibilities one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required (*should* equals *is recommended that*).

The word *may* is used to indicate a course of action permissible within the limits of the standard (*may* equals *is permitted to*).

The word *can* is used for statements of possibility and capability, whether material, physical, or causal (*can* equals *is able to*).

2. Normative references

The following referenced documents are indispensable for the application of this document (i.e., they must be understood and used, so each referenced document is cited in text and its relationship to this document is explained). For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments or corrigenda) applies.

There are no normative references in this guide.

3. Definitions, acronyms, and abbreviations

3.1 Definitions

For the purposes of this document, the following terms and definitions apply. The *IEEE Standards Dictionary Online* should be consulted for terms not defined in this clause.⁸

artificial intelligence (AI) customer: An organization or entity that uses an artificial intelligence (AI) product or services either directly or by its provision to AI users.

artificial intelligence (AI) developer: An organization or entity that is concerned with the development of artificial intelligence services and products.

artificial intelligence (AI) subject: An organization, entity or individual people that is impacted by an artificial intelligence system, service or product.

⁶ The use of the word *must* is deprecated and cannot be used when stating mandatory requirements; *must* is used only to describe unavoidable situations.

⁷ The use of *will* is deprecated and cannot be used when stating mandatory requirements; *will* is only used in statements of fact.

⁸ *IEEE Standards Dictionary Online* is available at: <http://dictionary.ieee.org>. An IEEE account is required for access to the dictionary, and one can be created at no charge on the dictionary sign-in page.

artificial intelligence (AI) system explainability: The nature of the artificial intelligence (AI) system's ability to express the important factors affecting decision-making in a way that humans can understand.

infidelity: Infidelity is the expected difference between two terms: **(A)** the dot product of the interpreted input disturbance and **(B)** the output disturbance.

instance-based: The training samples are assigned directly according to the stored training samples.

relevant authorities: Organizations or entities that can have an impact on an artificial intelligence (AI) system, service, or product, including policy makers and regulators.

rule-based: With a given threshold, the rules are extracted for classification.

sensitivity: Measures the degree to which the interpretation is affected by the small disturbance from the test point.

Shapley value: By considering the contribution made by each agent, the cooperation benefits can be distributed fairly.

single characteristic-based: Identify the single feature representation shared by the source domain and the target domain and then use this feature to transfer knowledge to the target domain.

3.2 Acronyms and abbreviations

AI	artificial intelligence
ALE	accumulated local effects
FI	feature interaction
GSM	global surrogate model
ICE	individual conditional expectation
LIME	Local Interpretable Model-Agnostic Explanation
PDP	partial dependency plots
PFI	permutation feature importance
SHAP	SHapley Additive exPlanation
XAI	explainable artificial intelligence

4. XAI architectures

4.1 General

Clause 4 first illustrates interpretability goals, taxonomy, and properties for XAI, followed by reference architectures of XAI systems.

4.2 Interpretability goals of AI

4.2.1 General

Interpretability means that people need to obtain enough information that they can understand how an AI system accomplishes something and what factors affect the system's decision-making. When people cannot obtain enough information, it is known to be inexplicable. The interpretability of the deep learning model is often not considered when the model is built. And humans do not know how it makes decisions. It is thus considered to be a black box.

Poor interpretability is one of the biggest flaws of deep learning and thus affects the implementation of AI systems in various fields. To improve the trustworthiness of AI systems for people, it is a priority to make people understand the mechanism of decision-making in AI models. The following subclauses contain aspects relevant to achieving the interpretability goals.

4.2.2 Understanding the principle

The current AI decision-making mechanism is mostly based on data, which is intrinsically incomplete and biased in general. A model that performs perfectly on the training dataset actually does not reflect the causality and, thus, may have much poorer performance on other datasets. It is necessary to find out the true relationship between input and output data and the principle behind the mechanism. Therefore, it is important for interpretable systems to reflect the true relationship in data, causality, and so forth.

4.2.3 Application adoption

There still lacks clear academic explanations for decision-making mechanisms, especially for prevailing training of neural networks. The bias of system, adversarial attacks, and complexity of the model cause the critical application difficult to be accepted by people. People need a clear guidance to evaluate *which part of the model and data has the greatest impact* on the model results. In addition, the guidance also provides descriptions of the way in which the model and data influence the model results.

4.2.4 Required by supervision

Regulation and policies give principles regarding the requirements on interpretability of AI and methods to be taken to help ensure the implementation of trustworthy AI in order to let people understand and accept the result of automated decision-making and performance of the AI system. The goal of this explanation is to make sure that the AI system in question is in compliance with required regulation and policies.

4.2.5 To improve trustworthiness

Explainable AI (XAI) expresses the important factors affecting the AI system in a way that humans can understand and convinces stakeholders that the AI systems in question have the attributes to fulfill requirements for trustworthy AI. Specifically, trustworthy AI systems should satisfy the following principles: human agency and oversight, diversity, nondiscrimination, fairness, privacy and data management, accountability, technological stability and security, transparency, and environmental and social well-being. By adopting different explainable AI methods, XAI systems are expected to convince stakeholders these trustworthy principles are fulfilled.

4.3 Taxonomy of explainability

4.3.1 General

The explainability should be considered from the following three aspects:

- a) Content type (what to explain)
- b) Communication (how to explain)
- c) Stakeholders (to whom the explanation addressed)

4.3.2 Content

The performance of a machine learning model depends on its content; therefore, different models need different type of explanations, which are summarized as follows:

- a) *Local explanation*: Local explanation is to explain the specific decisions made by the model for a single instance or the specific decisions made by the model for a group of instances. It focuses more on a single sample or a group of samples. Looking at a single sample, the predicted value given by the model may have a deterministic relationship with some features such as a linear relationship.
- b) *Counterfactual explanation*: Counterfactual explanation provides facts for stakeholders to verify the explanation (Sundararajan et al. [B37]⁹). It also improves the decision-making mechanism in case that it does not end up as one intended.

Example-based explanations can be categorized as the following three types:

- *Prototype (local) explanation*: Prototype explainability is provided by reporting similar examples to explain the initial decision. Examples are prototypical instances that are similar to the unseen instance. Providing examples helps to equip a model with explainability. Prototype explanation provides examples to achieve explainability when the distribution of data is of low complexity. These examples are similar prototypical instances to the ones not provided. And the nearest neighbor classifiers is a common method to generate prototypes.
- *Criticism (local) explanation*: Relying only on examples to provide the explanation can lead to over-generalization and misunderstanding of the models. Criticism explanation provides explanation for the part of data where prototypical examples do not fit very well. Criticism supports prototypes since it detects what the prototypical instance did not capture better and gives prototypes by focusing on which instances are not captured in prototypes. Examples used in criticism explanation are weighted and then selected. For example, MMD-critic, in addition to prototypes, selects criticism examples using a regularized witness function score (Kim et al. [B23]).
- *Global explanation*: Global explanation refers to how the model makes decisions based on entire feature space, model structure, and parameters. It includes which features are important and what will happen to feature interactions. The global explanation of the model can help people understand the distribution of the target variable for different characteristics. In fact, when the model has a large number of parameters, it is crucial to have a global explanation to help people understand how the features interact and why such prediction results are obtained.

⁹ The numbers in brackets correspond to those of the bibliography in Annex C.

4.3.3 Communication

4.3.3.1 General

The communication type determines how the explanation is provided to the user.

4.3.3.2 Text

Explainability is provided in the form of texts. For example, the explanation can be the caption of an image or language description.

4.3.3.3 Graphics

Explainability is provided in the form of graphs. The flow of parameters in the model can be visualized (see Loh, 2011 [B26]).

It is noted that depending on different types of models in question, different forms of graphs can be used to improve explainability. For example, Figure 1 shows partitions (left) and the decision tree structure (right) for a CART classification model with three classes labeled 1, 2, and 3. At each intermediate node, a case goes to the left child node if—and only if—the condition is satisfied. The predicted class is given beneath each leaf node.

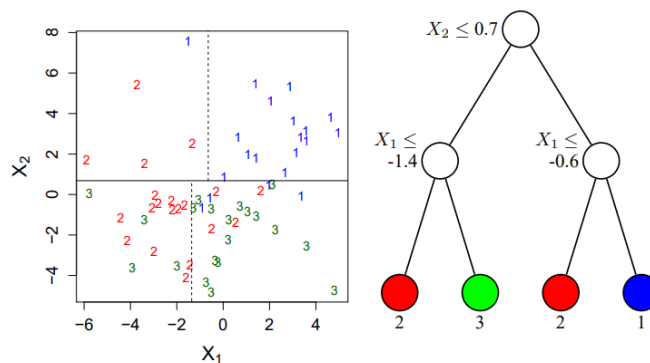


Figure 1—The illustrations of decision rules used, respectively, by partitions (left) and tree (right) models

4.3.3.4 Multimedia

Explainability is provided in the combination of various types of media including texts, video, images, and audio.

4.3.4 Stakeholders

The explainability should have different meaning towards stakeholders of AI. Different stakeholders have different underlying goals and needs for explanation in different stages. The stakeholder can be unfamiliar with AI technologies or he/she/they can be a domain expert. The explainability requirements, depending on the stakeholders, can be considered from the following aspects:

- *AI subject and AI customer:* The AI subject is an organization or person that is impacted by an AI system or product. The AI customer is an organization or person who uses an AI system or product. The AI subject and the AI customer are normally nonexperts who have neither technical nor domain knowledge of the system. His goal is to understand why a certain prediction result is made by the AI system, for example, why his credit loan was denied. Furthermore, he might be interested in understanding whether the explanation domain is composed of simple or complex input features.
- *AI developer:* The AI developer is an organization or person who develops and tests the AI system or trains the model. The organization or person should have technical knowledge in the field of AI. His goal is to understand how the algorithm works and what features it exploits so that he can debug and improve the model from a technical standpoint, for example, the accuracy, robustness, and fairness of the model.
- *Relevant authorities:* The relevant authorities are organizations who have enforcement of regulations and policies on AI system applications. Throughout the life cycle of an AI system, every step should be supervised by relevant authorities in order to ensure the implementation of AI trustworthiness.

4.4 Explanations properties

4.4.1 General

Comparing the interpretability between explanations generated by different interpretations can help stakeholders to evaluate these methods. Some properties of explanations are summarized to assist stakeholders such as accuracy, fidelity, consistency, stability, comprehensibility, certainty, importance, novelty, representativeness, and so forth (see Table 1 at the end of 4.4).¹⁰ Some of these properties are still hard to quantify because of subjective reasons that humans with different levels of background knowledge have a different understanding of the same explanation (see W. Guo et al. [B16]).

4.4.2 Accuracy

Accuracy measures the ability of explanation in an unseen dataset. The accuracy of explanations has an uncertain impact on the interpretability; for example, the accuracy of a model and explanations on an unseen dataset are close and low, explanations can give the reason for the low accuracy of model in this situation. The value of accuracy could not be used alone to assess interpretability of explanations; nevertheless, it could assist other properties to evaluate interpretability. This property is quantified by calculating the number of correct instances judged by explanations.

¹⁰ Notes in text, tables, and figures of a standard are given for information only and do not contain requirements needed to implement this standard.