Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29218

Human-Centered Explainable Artificial Intelligence: Concepts, Techniques, and Applications

Lingbo Feng

College of Science, China Jiliang University, Hangzhou, China
2200802211@cjlu.edu.cn

Abstract. Artificial intelligence (AI) drives the development of various fields in society,and the wide application of complex “black box” models, such as neural networks and random forests, has opaque decision-making, causing users to distrust. Explainable Artificial Intelligence (XAI) mitigates this by providing certain explanations, yet it balances fidelity and usability, lacks unified evaluation, and fails at multimodal problems. Human-Centered Explainable Artificial Intelligence (HCXAI) emerges as a solution by incorporating human-computer interaction, psychology, sociology, and other concepts into the explanation of the occasion and pays more attention to the user's needs, situational features, and social context. This study systematically sorts out the relevant concepts and core features of HCXAI, including user-orientedness, situationalization, and operability. By combining the three typical domains of healthcare, education, and industry, it explores the application value of HCXAI in scenarios such as clinical decision support systems. In addition, this paper analyzes the challenges faced by HCXAI in terms of trust management and assessment methodology, i.e., challenges in interdisciplinary integration, and proposes future development directions. The study shows that HCXAI enhances user trust, facilitates human-computer collaboration, and promotes AI in critical areas, boosting sustainable development of AI.

Keywords: Explainable Artificial Intelligence, Human-Centered Explainable Artificial Intelligence, Human-Computer Interaction

1. Introduction

In recent years, Artificial Intelligence (AI), with its excellent learning paradigms and algorithms, has become an important force in promoting social development and disciplinary progress. With the continuous improvement of machine learning algorithms, the use of AI in society has gradually increased, from search engines and image recognition to industrial production and medical healthcare, and it has turned into a mainstream method in many research fields. AI has demonstrated strong capabilities and values in natural language processing, computer vision, financial analysis, and other fields. AI technologies strengthen the intelligent development of related industries and trigger a high degree of concern in all sectors of society.

Early machine learning methods, such as linear regression and decision trees, are transparent white-box models. Their internal logic and the contribution of each feature to the final decision can

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29218

be easily understood and traced by humans. However, most contemporary AI and machine learning models, including neural networks, random forests, and large language models are typically black-box models. This means that while their inputs and outputs can be observed, their internal decision-making processes remain opaque and difficult to explain. Although black-box models are highly expressive and able to capture high-dimensional, nonlinear, and complex patterns, their complexity often makes it difficult for human beings to comprehend the underlying logic. This is primarily due to the large number of parameters, which leads to human beings questioning the credibility of their decision-making. This concern is especially prevalent in the medical field, where doctors are wary of machine learning. The opaque nature of black-box models makes it difficult to use such powerful models in industries where human safety is at stake.

In order to solve this problem, human trust needs to be strengthened, which requires transparency of models and understanding of model decisions through approximate explanations. This need led to the emergence of explainable AI (XAI). According to a large number of published comprehensive studies on XAI, doctors can provide global and local interpretations of random forest models through shallow decision trees or utilize LIME for specific predictions of heart disease. These visualization methods provide global interpretations of black-box models, which largely address concerns of doctors about black-box models. In other domains, especially in safety-critical areas such as finance and autonomous driving, XAI has likewise contributed greatly, removing the barrier of human trust in models. As a result, XAI has attracted the attention of more researchers and is being used more widely.

In order to better observe the decision logic and processes of black-box models, researchers have been working on how to enhance the interpretability and explicability of AI. While the development of XAI has been confronted with opportunities, it has also faced many challenges. For example, it is difficult to maintain a balance between model fidelity and usability; there is a lack of uniform evaluation criteria for different tasks with different XAI goals; it is difficult to explain model fusion and relationships in multimodal (image, text, audio, etc.) models; and there are large differences in interpretation needs in different languages and cultures, which may lead to misunderstanding or bias [1]. However, these challenges are not insurmountable, and Liao, Gruen, & Miller first proposed a human-centered approach, shifting from computer-centered to human-centered, i.e., the interdisciplinary fusion of human-computer interaction and XAI (HCXAI) [2]. HCXAI emphasizes human-computer interaction, pays more attention to the model logic and needs of users, and at the same time enriches the XAI toolkit. This approach provides new ways to address the shortcomings of traditional XAI and offers a new direction.

The main contributions of this paper can be categorized into the following three aspects. First, this paper reviews the research development of HCXAI, which helps to understand the development of this field by providing an overview of their research paradigms and related technologies. Second, this paper also analyzes the application of HCXAI in typical domains by combining the theoretical and practical parts, and thus illustrates the advantages of HCXAI in related tasks, as well as the challenges that still need to be overcome. Finally, this paper summarizes the existing results of HCXAI and proposes possible future research directions for subsequent research.

2. HCXAI-related concepts

Before delving into the applications and benefits of HCXAI, this section provides an introduction to its related technologies and concepts. In 2.1, the XAI foundation is introduced and then extended to 2.2, the definition and characteristics of HCXAI, as well as the implementation process and

DOI: 10.54254/2755-2721/2026.TJ29218

architecture of HCXAI. By organizing the concepts, subsequent applied research can be better understood.

2.1. Foundations of Explainable Artificial Intelligence (XAI)

David Gunning first introduced the concept of explainable AI in 2016, which aims to provide explanations of individual decisions that enable one to understand overall strengths and weaknesses, as well as convey an understanding of the future behavior of the system [3]. Its core goals are to enable machine learning models with high predictive power and to provide users with understandable and trustworthy explanations. It has also been suggested that the goal of interpretability is to make certain aspects of a system understandable to humans, including developers, designers, regulators, or users from the community [4]. XAI can be categorized into local and global interpretability, and another distinction can be made between ex-ante and ex-post interpretability. These techniques will be explained in more detail in Section III.

2.2. Definition and characteristics of HCXAI

According to Ridley's overview, HCXAI is a research paradigm that shifts XAI from computer-centered to human-centered and emphasizes the role of the user, the situation, and the social context in the explanation more than traditional XAI models [5]. The goal of HCXAI is to explain the basic logic of the model but also to fit the user's needs and context, thus enhancing the user's trust in the AI system. Summarizing the relevant reviews of HCXAI, the core features of HCXAI can be classified into the following aspects:

User-oriented: HCXAI focuses on the needs of different types of users, including experts and non-experts, and is more attuned to the users' perceptions and requirements.

Situationalization: Explainable models are closely related to specific social contexts and have different explanatory dimensions in different domains.

Actionability: HCXAI ensures actionability, as it supports user actions, enables users to interact with the system, helps them make better decisions, and enhances their trust in the system.

2.3. HCXAI implementation process and architecture

HCXAI needs a systematic process structure for achieving interaction with users and making decisions. The specific process structure can be divided into the following parts:

Data Input and Preprocessing: cleaning, standardization, and feature selection of multimodal data. This operation helps to improve the predictive performance of the model and makes the interpretation more credible.

Interpretable model construction: this stage will select appropriate interpretable algorithms to approximate the black box model so as to generate decision logic that can be understood by users.

Interpretation Generation: This phase makes the core part of HCXAI, making decisions based on the model and providing interpretations that meet the user's needs.

Human-computer interaction and feedback: The user provides feedback to the system through the interactive interface so as to optimize the model and achieve the effect of human-computer collaboration.

Model Optimization and Iteration: After collecting user feedback, the system will optimize the explanation to enhance its understandability and credibility.

DOI: 10.54254/2755-2721/2026.TJ29218

3. HCXAI-related technologies

In this section, the researcher will provide a systematic overview of the existing technologies of HCXAI by introducing several types of technical tools, namely interpretable algorithms, visualization methods, and human-computer interaction mechanisms.

3.1. Model interpretability techniques

Model interpretability is the foundation of HCXAI. By approximating the interpretation so that the user can understand the decision logic of the model, interpretability techniques are subdivided into the following categories:

Endogenous Interpretable Models: These models are specific and transparent in structure and logic, typically represented by Linear Regression, Decision Tree, and Plain Bayes, whose simple structure can provide a link between features and predictions, enabling users to intuitively understand the decision logic. Such models perform well in handling low-dimensional linear tasks, but are difficult to cope with high-dimensional nonlinear complex tasks. Therefore, Agarwal et al. have proposed the Neural Additive Models (NAMs) [6], which visualize features while maintaining the expressive power of the deep network to help users understand complex nonlinear relationships.

Post-hoc interpretation techniques: These techniques are commonly used to interpret black-box models, such as deep neural networks, random forests, and reinforcement learning models, by approximating the model to the features. Typical post-hoc interpretation techniques are local interpretation (e.g., LIME, SHAP) and global interpretation (e.g., approximate decision trees), and recent research has proposed new forms of interpretation, such as counterfactual post-hoc interpretations based on causal graphs and generative models [7]. This approach combines causal and counterfactual samples to provide explanations while explaining the links between variables and clarifying causal pathways.

Multimodal Interpretable Models: These models focus on multimodal scenarios such as images, text, language, etc., and are able to interpret different types of information simultaneously. Most of the traditional interpretable techniques are limited to a single modality, which makes it difficult to cope with the complexity of multimodality. This type of model interprets the model by aligning the feature space of different modalities and establishing the connection between different modalities. Recently, Yang and Wang have proposed the Language Model as Visual Explainer (LVX) [8], which combines LLM and joint visual-linguistic explanations. LVX allows the system to give predictions while providing visual heat maps and natural language explanations, helping users understand the decision logic.

3.2. Human-Computer Interaction (HCI)

Although traditional XAI models can handle a large number of tasks, the user's involvement is not high, leading to a lack of trust in the interpretable model, especially in safety-critical areas such as healthcare. To address these issues, the researcher introduces human-computer interaction techniques, which are crucial in the interpretation process. In this section, human-computer interaction techniques are introduced:

Interactive Interpretation Interface: This type of approach allows users to actively explore the model decision-making process through an interactive visualization interface. For example, adjusting the interpretation granularity and simulating the prediction results in different environments can enhance the credibility and comprehensibility of the model. Recently, Wang et al.

DOI: 10.54254/2755-2721/2026.TJ29218

have proposed the GAM Changer [9], which demonstrated the ability of interactive visualization in detecting bias and debugging models.

User Feedback Mechanism: This mechanism collects feedback from users, and the system incorporates it into the model and completes the optimization and looping while satisfying the user's needs. For example, users can annotate incorrect interpretations, and the system adjusts the model and reformulates decisions and interpretations in subsequent loops. This mechanism increases the degree of human-computer collaboration and plays an important role in improving the fit and operability of the system's explanations.

Scenario-based and personalized explanations: The system provides differentiated explanations according to user roles, task types, and usage scenarios, thus enhancing operational convenience and user trust. Scenario-based explanations emphasize the collection of corresponding environments to generate contextual explanations, e.g., in a medical scenario, the doctor is concerned with the cause of the evolution of the lesion, but the patient is concerned with the generalized explanation of the pathology. In contrast, personalized explanation emphasizes the differences between individuals and gives explanations that conform to the users' way of understanding through their perceptions and preferences. In recent years, researchers have combined user profiling and multimodal modeling techniques to make the explanation system more interactive with targeted effects. For example, Rahimi et al. have proposed the USER-LLM R1 [10], which combines chain-of-thought reasoning models and vision-language models.

4. Application areas

With the rapid development of artificial intelligence, HCXAI has been widely used in many fields and shows unique value and application prospects. This section will focus on three typical application scenarios: healthcare, education, and industry.

4.1. Healthcare

Healthcare is a field where human-computer collaborative XAI is used more frequently, and traditional black-box models show great potential in disease judgment and prediction due to their excellent decision-making ability. However, because of its opaque decision-making process and XAI's after-the-fact explanations, it is often difficult for physicians to be convinced of the results and basis of their judgments, and they worry that the model predictions are in error from the actual situation.

HCXAI effectively alleviates this challenge by proposing an interdisciplinary fusion of HCI and XAI that emphasizes human-centeredness and attention to user experience. In recent years, HCXAI has been widely utilized in Clinical Decision Support Systems (CDSS). Typical HCXAI can be classified into three categories: application-oriented assessment, human-oriented assessment, and agent-based assessment. Application-oriented invites experts to use the system in a real-world environment with high diagnostic correctness and trust, but with high implementation costs; human-oriented assessment evaluates the usability and satisfaction of physicians when using the system by simulating an experimental environment; and agent-based assessment quantifies the interpretability of the system through agent-based tasks to assess the completeness and consistency of the interpretation [11]. There is a trade-off between cost and fidelity between the three types of methods, with application-oriented evaluation being considered the most suitable for the medical domain but also the most expensive and complex.

DOI: 10.54254/2755-2721/2026.TJ29218

4.2. Education

HCXAI also plays a significant role in the field of education, where the most common XAI approaches are post-hoc explainers such as LIME and SHAP. These approaches treat the underlying model as a black box and provide explanations after the training is completed. However, these methods suffer from significant drawbacks, lacking sufficient comprehensibility, and their explanation process often stops at feature importance or correlation analysis and lacks responsiveness to student needs.

In order to enhance the role of AI in education, the researchers constructed a human-centered model and combined it with LLM to improve the human interpretability of XAI. This structure allows LLM to act as a communicator of explanations, effectively increasing students' trust and understanding of explanations. One of the most typical models is iLLuMinaTE [12], which contains three stages of causal association, explanation selection, and explanation presentation, and combines eight mainstream social science theories of explanation, three kinds of interpreters (LIME, Counterfactual, and MC-LIME), three kinds of online courses, and three kinds of LLM (GPT-4o, Gemma2-9B, and Llama3-70B). The experimental results showed that students preferred the explanations provided by iLLuMinaTE in 89.52% of the cases, which reflects that those students have more trust and understanding of the human-centered XAI model.

4.3. Industry

The process industry is another important application scenario for HCXAI, which covers a wide range of industries such as chemical manufacturing, oil and gas refining, food processing, pharmaceuticals, and paper manufacturing. These industries share common characteristics: complex and continuous production processes and strict safety requirements for the system.

The human-centered design philosophy of HCXAI ensures that explanations are more closely aligned with the user's needs by combining time series, visualization, and other methods to make explanations easier to understand. In addition, HCXAI provides feedback for users to determine and modify the output of the system, thus promoting continuous optimization of the system in a human-computer collaboration mode. The model plays a significant role in a number of core aspects, for example, in the anomaly detection and alarm system, HCXAI will find abnormal alarms and also explain which set of variables set off the alarm, and the user can also optimize the model through feedback. In the human-centered MLOps lifecycle management, HCXAI embeds human supervision and interpretation into the whole process of data collection, model development, deployment and supervision, ensuring model transparency and user participation.

Pilot cases from the EXPLAIN program have shown that the application of HCXAI in the process industry reduces false positives while increasing user trust in the system and meeting stringent EU regulatory requirements [13]. These results demonstrate the important role HCXAI plays in enabling intelligence in process industry basins.

5. Challenges and directions for development

HCXAI has demonstrated the value and benefits of its applications in areas such as healthcare, education, and industry, and researchers have found that it faces a number of challenges in practice but also provides direction for future development.

DOI: 10.54254/2755-2721/2026.TJ29218

5.1. Challenges

The management of user trust is still a challenge for AI promotion. HCXAI is proposed to enhance user trust in AI systems, but how to balance trust and dependence becomes a new challenge. Excessive user trust leads to dependence on the system and a lack of independent and critical thinking, while a lack of user trust leads to rejection of the system's decisions, and this challenge has become a key challenge for HCXAI promotion.

Another major challenge facing HCXAI is the limitation of evaluation methods. Current HCXAI models use application-oriented evaluation, human-oriented evaluation, and agent evaluation, but there are significant problems with each of these methods: application-oriented evaluation is too costly and difficult to popularize, human-oriented simulation environments lack realism, and agent tasks deviate from actual needs. These shortcomings make HCXAI difficult to achieve generalization in many domains.

The lack of interdisciplinary integration also greatly limits the popularization of HCXAI, which requires not only the algorithmic support of computer science, but also the theoretical guidance of human-computer interaction, sociology and other fields. However, most of the research now remains under the framework dominated by computer science, and the interdisciplinary cooperation is obviously insufficient, which leads to the lack of theoretical support for the explanatory system in the design of user cognition and interaction mechanisms.

5.2. Development direction

The future development of HCXAI should pursue trust enhancement and trust regulation. This requires the introduction of uncertainty expressions (e.g., confidence intervals, error ranges) in the explanations, combined with the risk reminder mechanism, to guide users to make rational judgments. At the same time, it is also necessary to clarify the boundaries of the responsibility of humans and machines in the decision-making process to avoid the transfer of responsibility caused by over-reliance on AI.

In response to the limitations of evaluation methods, a hybrid evaluation framework can be adopted in the future, which organically combines the authenticity of application-oriented evaluation, the controllability of human-oriented evaluation and the scalability of agent-based evaluation in order to take into account the cost, efficiency and applicability. Meanwhile, a multi-level coupled evaluation system can also be investigated to directly measure the role of interpretation in assisting user decision-making in combination with actual task performance, such as clinical diagnosis accuracy.

The further development of HCXAI also requires the deep integration of computer science with human-computer interaction, psychology, sociology, and other disciplines. Future directions include cognitive science-based explanation modeling methods, drawing on sociological theories of explanation, and introducing social contextualized explanations into the design of HCXAI. These designs can build a more complete interdisciplinary theoretical framework so that HCXAI can truly realize the transition from algorithmic explanation to humanistic explanation.

6. Conclusion

This paper provides a systematic review of the research development of HCXAI, introduces the shift from algorithm-centered to human-centered XAI, and clarifies the important role of HCXAI in dealing with the trust crisis caused by the “black box” model, improving user experience and

DOI: 10.54254/2755-2721/2026.TJ29218

promoting human-computer collaboration. The study shows that HCXAI not only emphasizes the transparency and rationalization of the decision-making process but also pays attention to the personalization, contextualization, and interactivity of the users, thus promoting the development of the AI research paradigm. By analyzing the application of HCXAI in the typical fields of healthcare, education and industry, it can be found that HCXAI has demonstrated significant advantages in improving model credibility, enhancing user participation and understanding, and optimizing the application effect, which also lays the foundation for its promotion and application in other key fields. In summary, HCXAI opens up a new path for the application of AI in high-risk areas, provides a theoretical foundation and practical experience for promoting transparent and trustworthy AI practices, and also provides a development direction for realizing a more humane and sustainable system.

References

[1] Mersha, M., Lam, K., Wood, J., AlShami, A., & Kalita, J. (2024). Explainable Artificial Intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing, 599, 128111. https: //doi.org/10.1016/j.neucom.2024.128111

[2] Liao, Q. V., & Varshney, K. R. (2022). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv preprint. https: //arxiv.org/abs/2110.10790

[3] Gunning, D. (2016). Explainable Artificial Intelligence (XAI). DARPA/I2O Proposers Day. https: //www.darpa.mil/news-events/2016-08-11

[4] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106, 102301. https: //doi.org/10.1016/j.inffus.2024.102301

[5] Ridley, M. (2024). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. Journal of the Association for Information Science & Technology, 76(1), 98–120. https: //doi.org/10.1002/asi.24889

[6] Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. (2021). Neural Additive Models: Interpretable Machine Learning with Neural Nets. In Advances in Neural Information Processing Systems (NeurIPS).

[7] Taylor-Melanson, W., Sadeghi, Z., & Matwin, S. (2024). Causal generative explainers using counterfactual inference: A case study on the Morpho-MNIST dataset. Pattern Analysis and Applications, 27(3), 89. https: //doi.org/10.1007/s10044-024-01306-8

[8] Yang, X., & Wang, X. (2024). Language Model as Visual Explainer (LVX). arXiv preprint. https: //arxiv.org/abs/2412.07802

[9] Wang, Z. J., Kale, A., Nori, H., Stella, P., Nunnally, M., Chau, D. H., Vorvoreanu, M., Vaughan, J. W., & Caruana, R. (2021). GAM Changer: Editing Generalized Additive Models with Interactive Visualization. arXiv preprint. https: //arxiv.org/abs/2102.03024

[10] Rahimi, H., Cattoni, J., Beghili, M., Abrini, M., Khoramshahi, M., Pino, M., & Chetouani, M. (2025). Reasoning LLMs for User-Aware Multimodal Conversational Agents. arXiv preprint. https: //arxiv.org/abs/2504.01700

[11] Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. International Journal of Human-Computer Studies, 154, 102684. https: //doi.org/10.1016/j.ijhcs.2021.102684

[12] Swamy, V., Romano, D., Desikan, B. S., Camburu, O.-M., & Käser, T. (2024, Sept.). iLLuMinaTE: An LLM-XAI Framework Leveraging Social Science Explanation Theories Towards Actionable Student Performance Feedback. arXiv preprint. https: //arxiv.org/abs/2409.08027

[13] EXPLAIN Project. (2024). EXPLAIN guidebook: Human-centered explainable AI for process industries. ITEA4. https: //itea4.org/news/explain-guidebook-human-centered-explainable-ai-for-process-industries.html