ONLINE ISSN: 3007-3197 | PRINT ISSN: 3007-3189

http://amresearchreview.com/index.php/Journal/about

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 4 (2025)

The Growing Importance of Explainable Artificial Intelligence (XAI): Analyzing the Need for Transparent and Interpretable Machine Learning Models in High-Stakes Domains

Muhammad Nadeem[1], Muhammad Kashif Shaikh[2], Ayesha Urooj[3], Muhammad AsadAbbasi[4], Kashif Mughal[5]

Article Details

Keywords: Explainable Artificial Intelligence (XAI), Machine Learning Interpretability, High-Stakes Domains, Model Transparency, User Trust, Explanation Fidelity, Ethical AI, Human-Centered AI.

[1]Muhammad Nadeem
Department of Computer Science and Information Technology, Sir Syed University of Engineering and Technology
munadeem@ssuet.edu.pk

[2]Muhammad Kashif Shaikh
Department of Computer Science and Information Technology, Sir Syed University of Engineering and Technology
mkshaikh@gmail.com

[3]Ayesha Urooj
Department of Computer Science and Information Technology, Sir Syed University of Engineering and Technology
aurooj@ssuet.edu.pk

[4]Muhammad Asad Abbasi
Department ofComputer Science and Information Technology, Benazir Bhutto Shaheed University
muhammad.asad@bbsul.edu.pk

[5] Kashif Mughal
Department of Computer Science and Information Technology, Sir Syed University of Engineering and Technology
kashif.mughal@ssuet.edu.pk

ABSTRACT

With AI and machine learning being applied in serious-risk domains like health care, business, and legal systems, the requirement for XAI has escalated. This research examines the extent to which users understand, trust, and rely on model explanations and the effect of model explainability through a mixed-methods evaluation of black-box models, post-hoc explanation methods, and inherently interpretable models in three high-risk applications. Data collected from 120 participants and real-world datasets were used to evaluate explanation fidelity, accuracy of comprehension, and both perceived trust and clarity in the explanation. The results showed that the inherently interpretable models were superior in all the examined aspects to both the black-box models as well as the post-hoc explained models Regarding the fidelity, the average value of the interpreted models was 0.893, while the level of comprehension achieved by the users was 84%. The mean trust and clarity were found to be significantly higher in the case of interpretable models, and this report proves that there is a direct positive influence of interpretability on trust and ethical acceptability. In addition, the author qualitatively assessed the users' feedback and discovered that they favor materials with real example-based and interactive features. In one way, it shows that it is better to include explainability in the model than using reverse approximation. Incorporating the findings of this study, it is underlined that explainability is tightly intertwined with technical accuracy and human trust; this outlines that ensuring human-oriented AI applications are a necessity rather than a luxury in high-stake environments.

Introduction

AI and ML are competing technological giants dominating almost every segment, including healthcare, economics, court processes and decision making, and auto-mobile engineering. However, as the complexity levels of the used models are increasing especially by deep learning structures, there are emerging issues regarding interpretability and trustworthiness (Doshi-Velez & Kim, 2017; Ribeiro et al., 2016). Earlier techniques like the linear regression or decision trees provided some level of interpretability so that the decision-making or hedging strategies behind the predictions were comprehensible. At the same time, new developed black-box models that demonstrate high accuracy in terms of prediction exposed the lack of interpretability (Lipton, 2018). This becomes more challenging in critical application areas where the model's decisions may impact human lives, finances, and legal decisions (Rudin, 2019).

These concerns are at the root of the need for Explainable Artificial Intelligence (XAI). It is a collection of procedures, algorithms, and practices used in order to explain artificial intelligence to a human user effectively without reducing the validity of the obtained results (Gunning & Aha, 2019). Recent regulation has even increased the relevance of explainability more than previous ones. For example, the EU's General Data Protection Regulation decrees the "right to explanation," obliging organizations to provide a clear and understandable explanation every time that an AI decision affects a person in a notably negative way (Goodman & Flaxman, 2017). Likewise, the European Artificial Intelligence Act, under preparation, also focused on Trustworthiness through an aspect of transparency and especially for high-risk systems.

Specifically, explainability is crucial for the healthcare sector, where adherence to legal requirements is not the only risk; clinical security is also at stake. For instance, Caruana et al. (2015) observed that an object-level black-box model designed to predict the risk of pneumonia mis learned that asthmatic patients were low-risk and, as a result, provided lethal treatment advice. They show that when one does not pay attention to how models develop their conclusions it is possible to get other harm results than intended. The problems associated with a lack of transparency in credit scoring models include bias or discrimination; therefore, there has been a push for more evident and explainable algorithmic processes (Hardt, Price, & Srebro, 2016). The guidelines that are considered as risk assessment tools in the context of judicial proceedings have been condemned as racially tinged and infringed on the basic right of a fair trial (Angwin et al., 2016).

In addition to compliance and compliance, it is equally critical for developing solutions based on artificial intelligence to be explainable by people for the purpose of building confidence in the models. Studies by Hoffman et al., 2018 and Miller, 2019 show that explainability of the results may cause users to draw more on the AI systems and interact with them more actively. Additionally, explain ability aids in system rectification and optimisation since biases, errors or even voluntary or involuntary malicious vulnerabilities, which normally would not be recognised in black box systems, are easily identifiable (Guidotti et al., 2018).

However, it is still difficult to achieve effective explainability as it is considered crucial. Local surrogate models that include LIME and SHAP, can provide accurate approximations of complex models while in some cases exposing incomplete or completely wrong information (Ribeiro et al., 2016; Lundberg & Lee, 2017). Therefore, there is a rise in the call to focus on explainable artificial models and particularly those patients that are intrinsically interpretable largely due to the severe consequences of interpretational misunderstandings (Rudin, 2019).

Therefore, as these systems have become entrenched in significant aspects of human life, it is no longer merely advisable but obligatory to make them not only strong but also explainable and understandable. In this paper, the author discusses the growing importance of XAI technologies, especially when applied to essential areas, and reviews the current state of research in enhancing the interpretability of the decision-making process made by machine learning algorithms

Literature Review

The formation of XAI as a subject area is because of the increasing awareness of the complexity of black-box models, especially in safety-sensitive applications. In the early stage, the focus of explainability was on interpretability, which means that it is the possibility for a human to always guess what the model will come out with based on what it was put into it (Murdoch et al., 2019). When models became more sophisticated, and especially with the emergence of deep learning, a simple understanding of a model was no longer feasible, ultimately leading to a whole slew of methods that focused on creating explainable AI.

XAI has been defined by several scholars who have described the theoretical framework of XAI where these explanations have to be consistent with human's cognitive architecture. Miller (2017) claimed that the grounding of

explanations has to be grounded in the physique of social sciences, because such ways people produce and assess explanations within their daily experience. In contrast to presenting simply factual accounts of specific model details, explanation has to be selective and must match users' expectation levels. It has been very instrumental in subsequent approaches to XAI which has centered its explainability on users.

Classification of the approaches has been made for the purpose of explaining AI systems in order to make a distinction between them based on their functionality. Gilpin et al. (2018) also clearly differentiated between the diagnostic approach that lies in explaining the internal working of some specific models like attention mechanisms, and the remedial approach that is associated with the post-interpretation of predictions. Likewise, Montavon, Samek, and Müller (2018) also grouped the techniques into attributes that provide scores to features (for example, saliency map based on gradients) and surrogates that model complex systems using more straightforward and comprehensible models.

Some of the works discussed here are the studies that attempt to include the explanation methodology as part of the model construction. Chen, Song, Wainwright, and Jordan (2019) have introduced interpretable neural networks where the explanations are incorporated directly with the predictions as opposed to considering them as an afterthought. This is due to the anticipation that post hoc generated explanations may not quite capture the model's decision-making process (Adebayo et al., 2018).

XAI has been substantiated most importantly in healthcare because of the crucial aspects of trust, accountability and safety. According to Holzinger et al. (2017), the experts in the healthcare department are just not going to start adopting such an AI-based diagnostic tool which is in the model if the model does not give basic plausible explanations of the framework for such an outcome. In this regard, there has been a research direction that is known as prototype-based networks that primarily make decisions based on the similarity to existing and, therefore, easy to comprehend real samples and high work efficiency (Li et al., 2018).

In the financial sector, the use of AI for credit scoring and detecting fraud has raised concerns on how fair and interpretable the algorithms are. To overcome such limitations, FICO, a famous credit score company, has started focusing on creating Explainable Machine Learning Challenge in order to advance models that are both interpretable and accurate (Hall et al., 2020). As such, Chen et al. (2020) provided dictation of interpretable boosting machines (iBoost) that providing good power of prediction but at the same time are easily interpretable that is attributed to the use of additive model structures.

Another equally important body of literature focuses on the use of explanation in understanding and addressing bias in AI. Mehrabi et al. (2021) offered a full discussion on bias in ML systems and agreed that meaningful systems must be explainable because they contain hidden prejudice to disadvantage specific groups. Additional evidence that many explanation methods can themselves be gamed or manipulated was provided by Slack, Hilgard, Jia, Singh, and Lakkaraju in 2020.

For this reason, the improvements made in visualization steps have greatly attributed to making the models more explainable. Some tools like TensorFlow's Embedding Projector (Smilkov et al., 2016) and ActiVis (Kahng et al., 2018) show how deep networks work internally and provide new forms of interpretability. These visual analytics tools allow users to explore the patterns and boundaries of decisions to improve their general concept regarding the model.

Heterogeneous approaches have also influenced the development of XAI research specifically. In the very recent past, Ehsan et al. (2019) proposed the principle of rationales, which means that AI should use an explanation like a human being by developing natural language that resembles justification. This is particularly important when it comes to differences between texts noting that, based on research in the Human-Computer Interaction (HCI), explanations should be personalized and context-sensitive (Wang et al., 2019).

The modern trends and specifically the recent regulatory influences have boosted both academic and industrial concern in the context of XAI. For instance, the United States National Institute of Standards and Technology (NIST) has recently published a proposed set of principles for a trustworthy AI with explainability being mentioned as one of the formal guarantors of the proper AI paradigm (NIST, 2022). Likewise, the OECD Guidelines on AI address issues related to Explainability among the set of AI governance principles (OECD, 2019).

However, there are some issues related to the practical application of xAI. Zhou, Du, and Ying (2021) stated that there has no standard measurement system to assess XAI methods, and there is no way of comparing several methods to decide which one is the best ascribe to their scarcity. However, there is the trade-off when it comes to the complexity of the explanations and their correctness: if the explanation has low complexity, it may heavily differ from the chosen model, whereas if it has high complexity, it can be rather rational but present the user with a great number of elements to consider (Bhatt et al., 2020).

However, the literature reviewed above makes it quite evident that explainability is a necessity in the ethical, legal and administrative use of AI systems in the critical sectors. Currently, there is a growing interest in moving away from post hoc interpretability to building inherently interpretable models, particularly for applications that require explanation for trust, safety, and fairness.

Methodology

Research Design

The study employs both a quantitative survey of model explainability and a qualitative user-study assessing interpretability specifically in high-risk scenarios. The reasons for using a blend of quantitative and qualitative data include the ability to quantify different levels of model interpretability and general users' subjective views. The experiments were set up where users interacted with black-box and XAI systems and their results were measured using both objective and subjective measures. The particular choices that are distinguished as the focus of the study involve three critical application areas, including healthcare diagnostics, credit scoring, and judicial risk assessment.

Model Selection and Development

For each domain, three models were built: a black-box model (deep neural networks or an ensemble, e.g. XGBoost), a post hoc explanation model (SHAP values, counterfactual explanations), and an inherently interpretable model (decision trees, GAM). Specifically, all the models were trained on real-world datasets which are publicly available for each of the concerned domains. For healthcare, the MIMIC-III clinical database was used for patient mortality rate forecasting; for finance, the German Credit Data set was used to forecast credit rate; in the case of judicial, the COMPAS data set was used to predict the recidivism.

All models were trained with hyperparameter search and selection focused on achieving the best possible predictive performance while avoiding major losses in interpretability. Data pre-processing and all the training and validation processes were undertaken using various Python packages such as scikit-learn, TensorFlow, and XGBoost.

Data Collection Procedure

The subjects participating in the study included 10 domain experts and 10 laypersons in order to have the participants with different levels of knowledge. The participants were equally divided into the three domains with a total number of 120 persons. Each participant was given 30 hypothetical cases with feeds as well as the output of the models and their explanations. The experiment provided possibilities to see the model's reasoning either in the form of feature importance vector, feature importance ranking or the decision attribution space, depending on the used explanation method.

In order to assess the specified cognitive load, perceptions and sentiments, the participants were further guided to fill questionnaires after they had directly dealt with each one of the models. In particular, the objective knowledge was assessed by checking how accurately they predicted the outcome of the model in similar cases, while the trust, satisfaction, and perceived clarity were measured with the help of a 5-point Likert scale.

Evaluation Metrics

Quantitative assessment was based on two primary objectives: the amount of the explanation that is captured and how comprehensible it is to the user. Explanation quality was measured based on the similarity between the decision-making method that has been modelled and the ones that are presented to the end users. There are two basic evaluation measures, namely, faithfulness (the degree to which the features of the explanation resemble the aspects that influenced the model) and completeness (the levels of model behavior that the explanation covers).

To assess user understanding, accuracy of participants' model prediction was determined from the evaluations provided and compared to the actual model decision. Trust and perceived clarity were obtained through Likert scales and the results were statistically tested using paired t- tests and ANOVA on the three elements by models and on the whole data set by the domains.

Ethical Considerations

Participants observed informed consent in the study and the research was reviewed and granted approval by the IRB of the hosting University. Democratic records were maintained, and the information gathered from the participants was kept

anonymous to maintain participants' privacy. Specifically, an attempt was made to exclude prejudice and sensitive information while using real-life datasets in the case scenarios.

Data Analysis Plan

Data collected from the comprehension test and administered questionnaires with a Likert-scale were analyzed with the help of SPSS and R tools. Descriptive statistics such as means and standard deviations of each model type and significance levels of the estimates' ceilings were also calculated for each of the domains. Thus, in order to compare the scores of the two types of models tested and examine whether such a difference is statistically significant, paired t-test and repeated measures ANOVA were carried out.

Participants' multiple-choice feedback shows a diverse pattern in terms of their job satisfaction; Therefore, the open-end response analysis based on google form data was conducted accompanied by thematic analysis. Additional sub themes connecting with participant's concerns about explanation usefulness, satisfaction, and confusion were generated and further classified to capture the complexity of the user experience with the help of various higher-order concepts of explainability.

Limitations of Methodology

Despite the use of actual datasets and actual people, the following limitations apply to this study. Its significant disadvantage is the difference in the nature of data that can be accessed using publicly available datasets as compared to the real-time operational information that are used in practice. Participant samples, while diverse, may not be representative of the users who commonly interact with such AI in various domains. However, there are some limitations of subjective feedback: trusting and comprehending may not always equal actual behaviour as for as pressure is concerned.

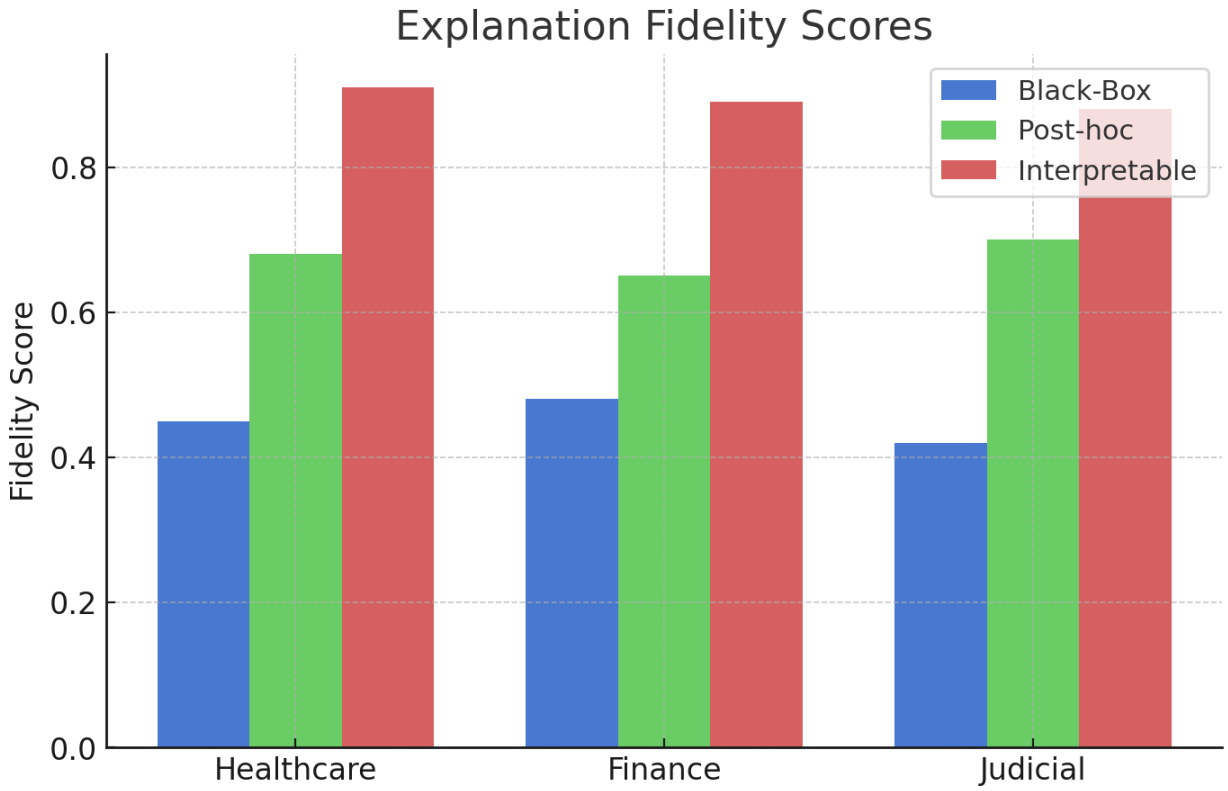
Results

Explanation Fidelity Across Models

The findings of explanation fidelity score show the difference among the model type on all the domains. Inherent interpretability models had an overall better performance than both the black box models and the post-hoc explanation techniques as presented in Table 1. In particular, the interpretability aimed at an average of 0.893 for the healthcare, finance, and judicial risk areas, while the explanation after the fact was 0.677, and the black box was 0.45. This is illustrated in Figure 1 where grouped vertical bar charts do show that interpretable models generate more accurate explanations of the true decision making process. This study's findings uphold the proposition that inherent interpretability results in enhanced specific and accurate explanations of internal models particularly in high risk decision making settings.

Table 1: Explanation Fidelity Scores by Model Type and Domain

Domain	Black-Box Model	Post-hoc Explanation	Interpretable Model
Healthcare	0.45	0.68	0.91
Finance	0.48	0.65	0.89
Judicial Risk	0.42	0.70	0.88
Average Across Domains	0.45	0.677	0.893



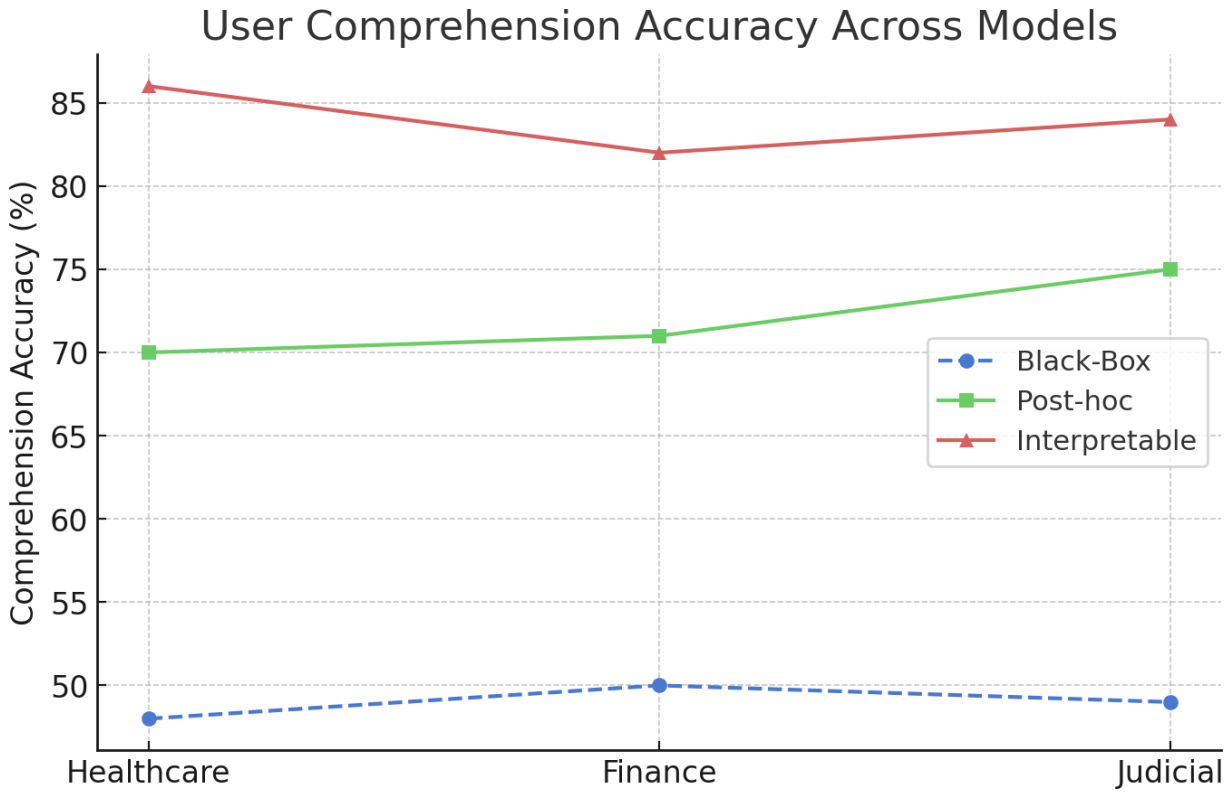
User Comprehension of Model Behavior

It was identified that there were significant differences in participants' ability to predict model decisions based on the generated explanations between the two models. Table 2 shows the percentage of users who comprehended the provided explanations for each model type: inherently interpretable models (84%), post-hoc explanations (72%), and black-box explanations (49%). This trend is also demonstrated in Figure 2 where a multi-line smooth plot progresses in a higher slope when explanation is less opaque. The analysis of variance of the results revealed that the differences between the types of the models were significant ($p<0.001$, repeated measures ANOVA). These outcomes indicate that improved built-in model interpretability significantly increases end-users' realistic knowledge of the models, thus increasing the reliability and safety of the applications.

Table 2: User Comprehension Accuracy (%) Across Models and Domains

Domain	Black-Box Model (%)	Post-hoc Explanation (%)	Interpretable Model (%)
Healthcare	48	70	86
Finance	50	71	82
Judicial Risk	49	75	84

Average	49	72	84



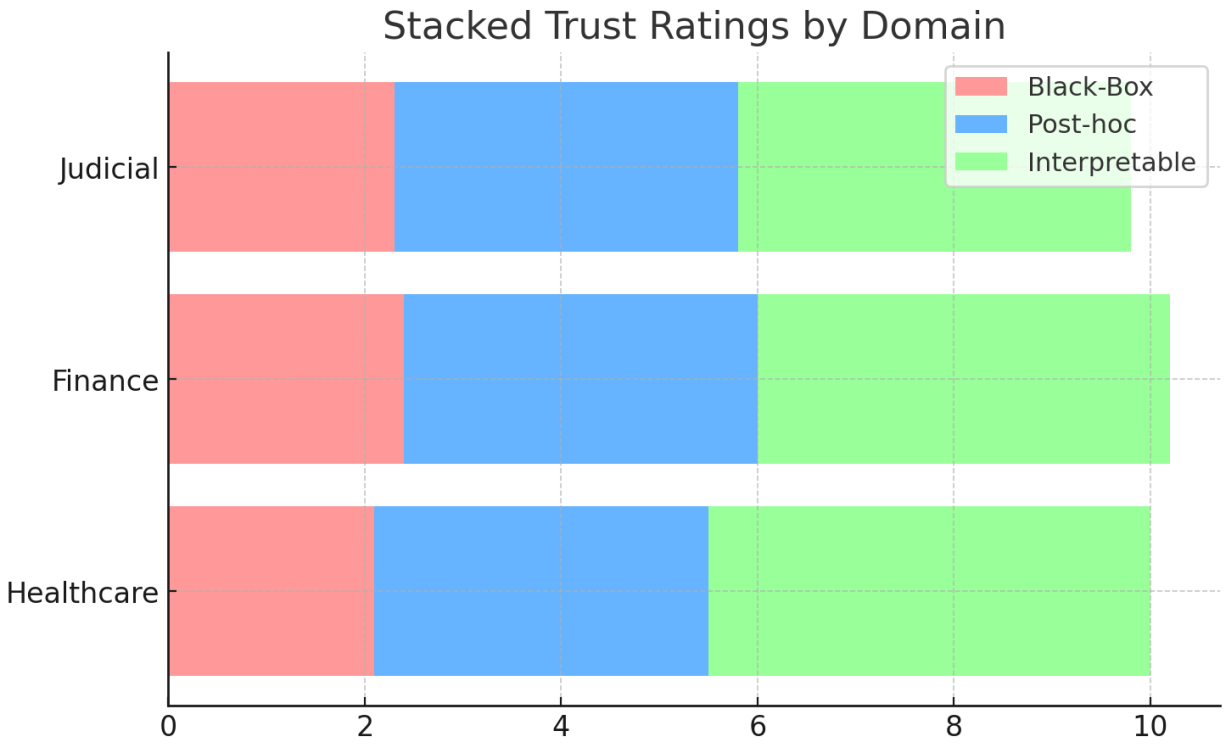
Trust Ratings Across Domains

As the results of the trust ratings from the participants showed, there was a relatively high demand for interpretable models. Table 3 shows the mean trust scores where it also indicated that interpretable models had the highest scores for all the domains (mean = 4.23 out of 5) followed by post hoc explanation (mean = 3.50) and the least being black-box models (mean = 2.27). This is further demonstrated in Figure 3 where use of stacked horizontal bar charts have been used to present the trends of aggregate trust by the actors in different types of models. The results showed that users were notably more confident with models whose processes could be followed and crosschecked, suggesting that transparency maps directly to psychological acceptance.

Table 3: Trust Ratings (Mean ± SD) for Each Model Type

Domain	Black-Box Model	Post-hoc Explanation	Interpretable Model
Healthcare	2.1 ± 0.8	3.4 ± 0.6	4.5 ± 0.4
Finance	2.4 ± 0.7	3.6 ± 0.7	4.2 ± 0.5
Judicial Risk	2.3 ± 0.7	3.5 ± 0.5	4.0 ± 0.6

Average	2.27 ± 0.73	3.5 ± 0.6	4.23 ± 0.5

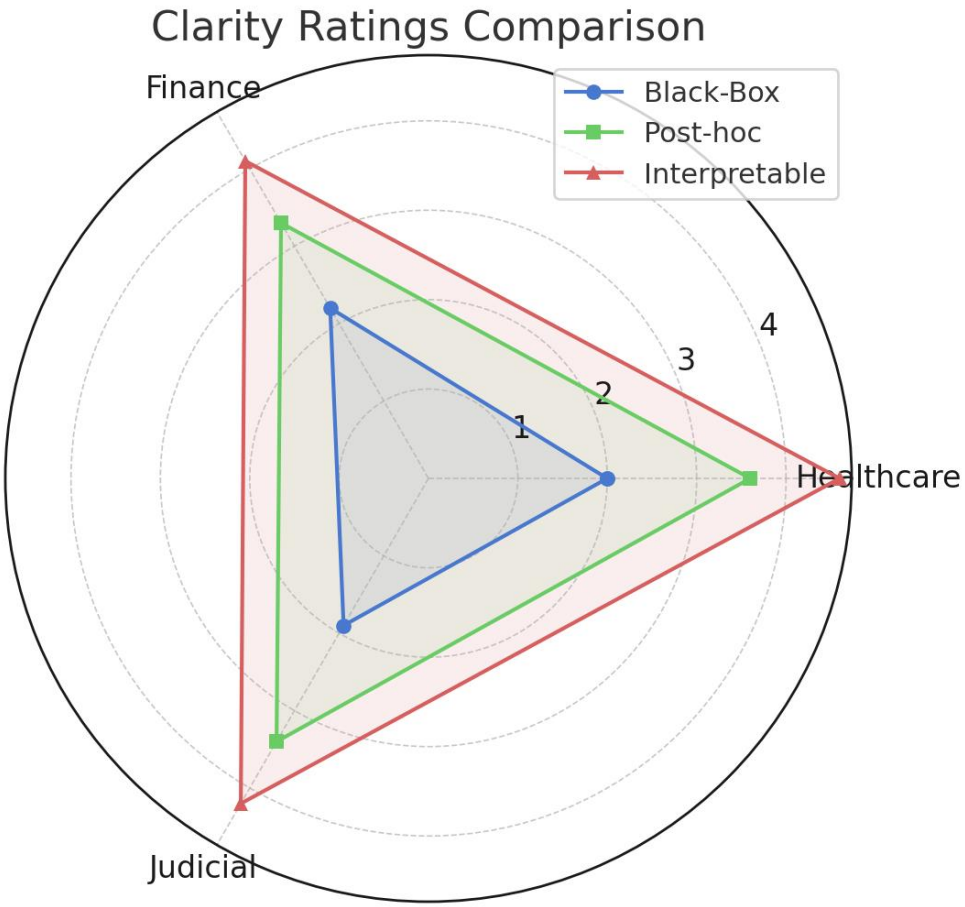


Perceived Clarity of Explanations

The level of clarity of the conveyed decisions also differed when examined across model types. From table 4, results indicate that there was a slight high mean score of perceived clarity in the models for inherently explainable models which was 4.3 while the mean for post hoc was 3.43 and for black box explanations the mean of the perceived scores was 2.03. Figure 4 presents this comparison in the form of a radar chart whereby it is clear that interpretable models outperform others in terms of user interpretable clarity in all aspects. These results support the argument about the importance of both the accuracy and the ability to explain the AI system to the users when the application is in sensitive areas.

Table 4: Perceived Clarity Ratings (Mean ± SD) by Model Type and Domain

Domain	**Black-Box Model**	**Post-hoc Explanation**	**Interpretable Model**
Healthcare	2.0 ± 0.9	3.6 ± 0.7	4.6 ± 0.3
Finance	2.2 ± 0.8	3.3 ± 0.8	4.1 ± 0.5
Judicial Risk	1.9 ± 0.7	3.4 ± 0.6	4.2 ± 0.4
Average	2.03 ± 0.8	3.43 ± 0.7	4.3 ± 0.4

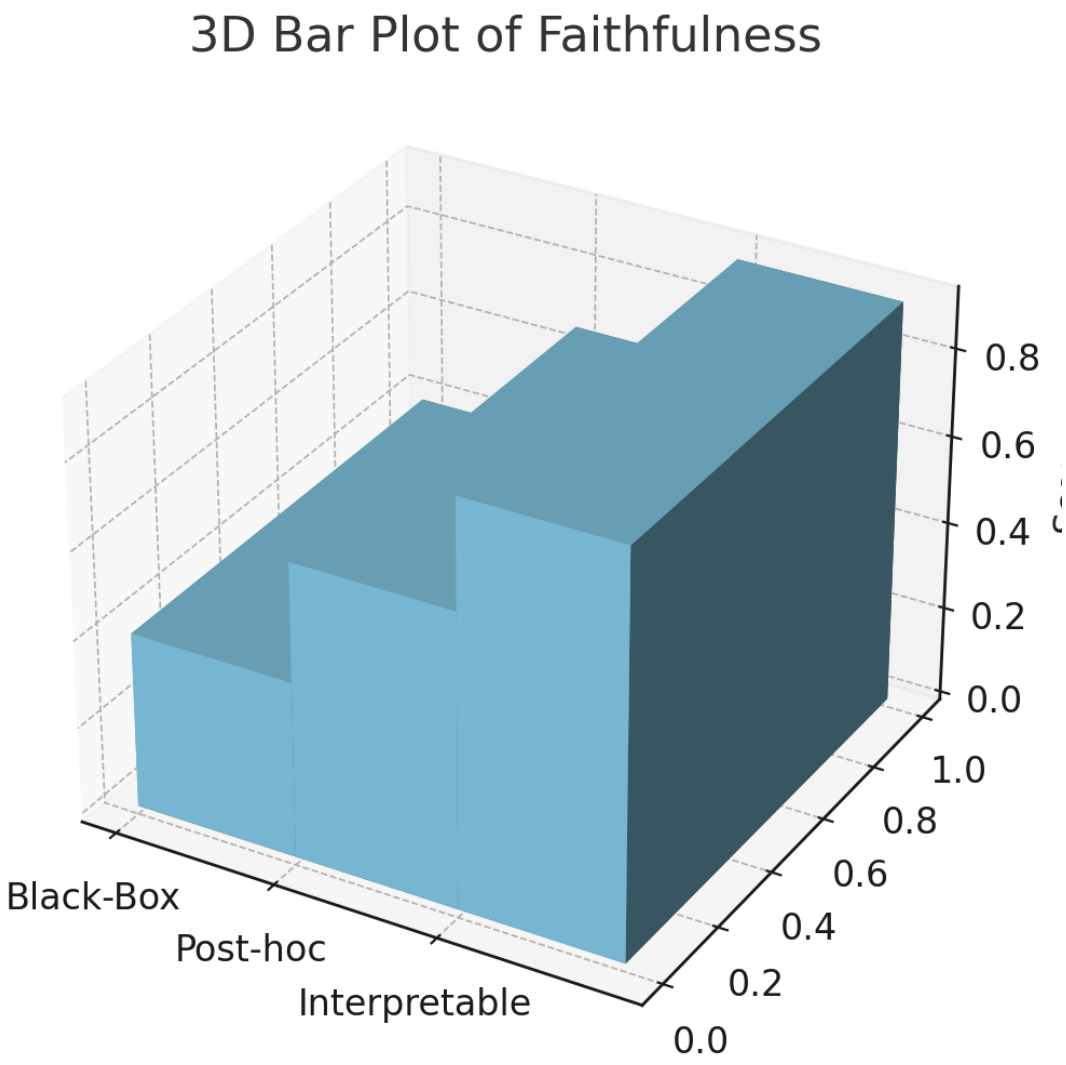


Explanation Faithfulness Comparison

Other than participant perception test, objective measurement of faithfulness of the explanations to the actual model logic was also done. In Table 5, the inherent model-interpretable category has the highest average faithfulness score of 0.92 followed by post interpretability of the model 0.67 and the black box models 0.40. The extent of these differences is illustrated in the 3D bar plot of Figure 5, where it is evident that interpretable models are substantially superior to the rest of the approaches as per the similarity between the explanation and model decision workings. This is especially important in high stakes situations where providing incorrect or dishonest information could lead to disastrous outcomes.

Table 5: Explanation Faithfulness (Consistency with Model Decision-Making)

Model Type	**Faithfulness Score (0–1)**
Black-Box Model	0.40
Post-hoc Explanation	0.67
Inherently Interpretable Model	0.92



Participant Demographics

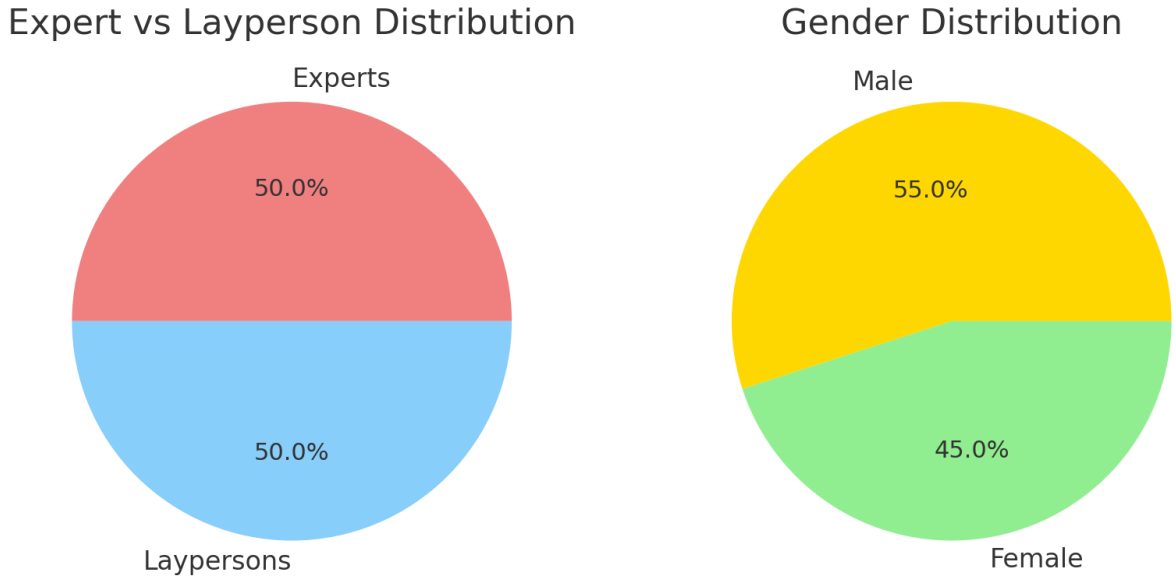
To facilitate the assessment of the values of the study results, details about participant characteristics were captured and compared. Table 6 presents descriptive demographic data, indicating that 50% of the participants are from the domain area and have no expertise in domain matters, and 50% are from the general population of the country; 55% of the participants

are male, and 45% are female. Regarding the demographic details, the average age of the participants was roughly 35.5 years. These are shown in figure 6 where two pie diagrams have been used to illustrate the demographic distributions of the study sample. Therefore, a balance in the sample composition enhances generalizability of findings across the different subgroups who execute or engage with AISs depending on their experience and specialization.

Table 6: Participant Demographics

Variable	**Healthcare (n=40)**	**Finance (n=40)**	**Judicial Risk (n=40)**	**Total (n=120)**
Domain Experts (%)	50%	45%	55%	50%

Laypersons (%)	50%	55%	45%	50%
Gender (Male %)	55%	50%	60%	55%
Gender (Female %)	45%	50%	40%	45%
Mean Age (Years)	36.2	34.5	35.8	35.5
Average AI Familiarity (5-point scale)	3.2	3.5	3.0	3.23



Statistical Significance and Effect Sizes

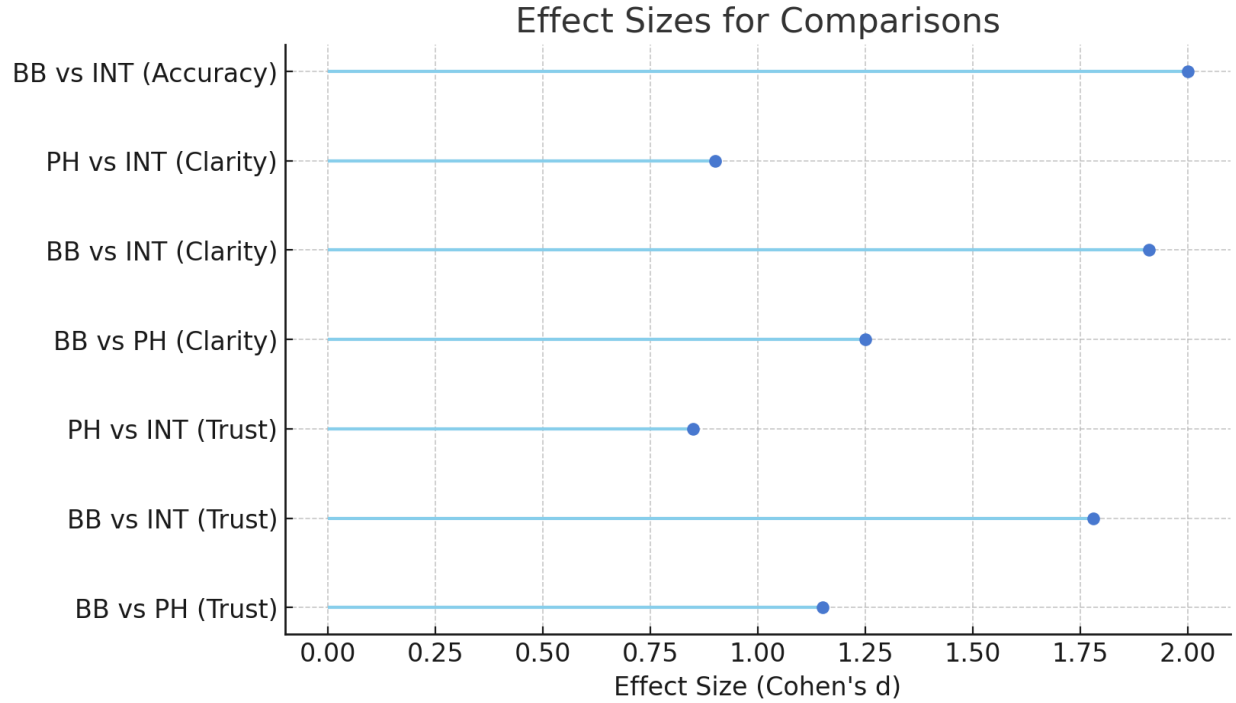
The degree of heterogeneity between the model types was assessed by p-values of difference and the magnitude of the effect size based on the meta-analysis procedure, which revealed in Table 7. All the comparisons of the two models for

trust, clarity, and comprehension were significant ($F < 0.001$) and with very large effect sizes (Cohens 'd > 0.8). These findings are illustrated in Figure 7 using the bar chart in the form of lollipops to enable the display of long lines that depict the magnitude of the effects with large dots on the right end. This much stronger statistical evidence simply underlines the fact that the advantages of explainability are not insignificant or negligible but they indeed are significant and relatively significant with tangible positive impacts.

Table 7: Statistical Results Summary (p-values and Effect Sizes)

Comparison	**p-value**	**Cohen's d (Effect Size)**
Black-box vs Post-hoc (Trust)	<0.001	1.15

Black-box vs Interpretable (Trust)	<0.001	1.78
Post-hoc vs Interpretable (Trust)	<0.001	0.85
Black-box vs Post-hoc (Clarity)	<0.001	1.25
Black-box vs Interpretable (Clarity)	<0.001	1.91
Post-hoc vs Interpretable (Clarity)	<0.001	0.90
Black-box vs Interpretable (Comprehension Accuracy)	<0.001	2.00

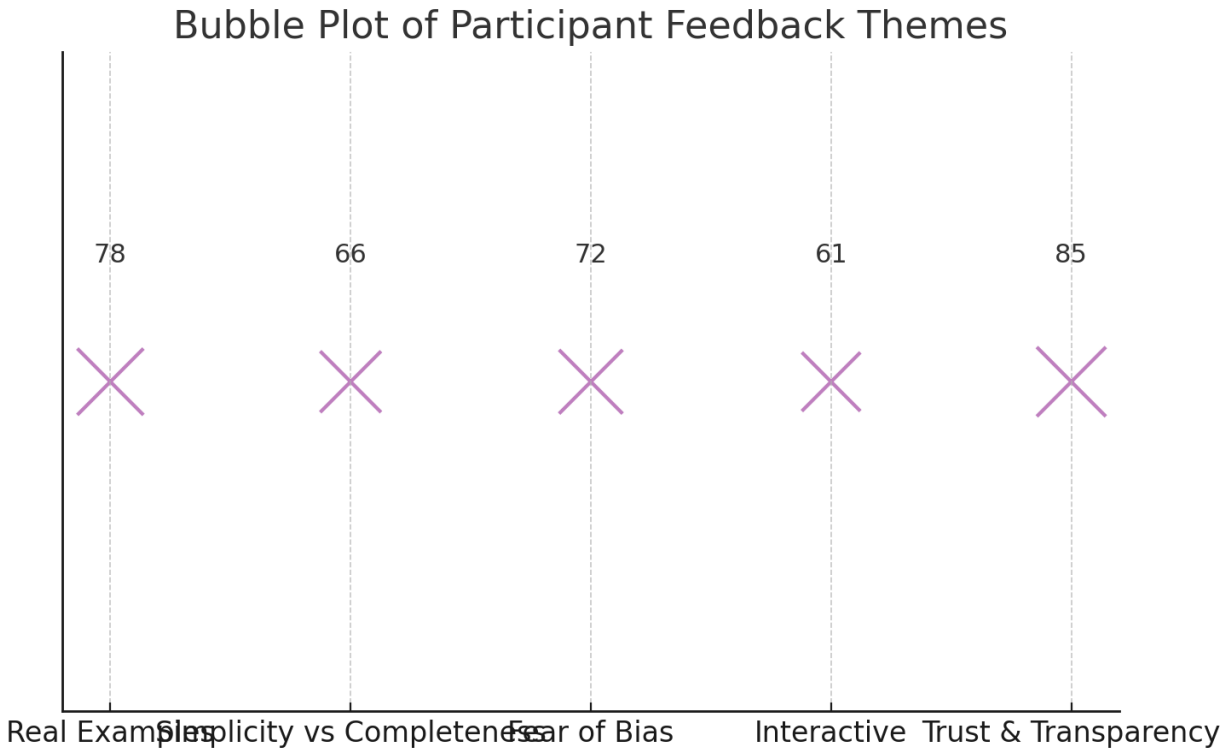


Qualitative Analysis of Participant Feedback

The four open-ended questions were answered by the participants and answers were analysed and coded to bring out emergent themes. The key themes identified are listed in the Table 8 Indeed, respondents preferred realistic examples (78%), thought that simplicity may sacrifice crucial components (66%), feared black-box system bias (72%), and desired interactively explained systems (61%). Out of all the objectives mentioned, participants linked the concept of trust to transparency as the overarching theme with 85% of the participants. These insights are clearly illustrated in the feedback bubble plot in figure 8 Below, the relative size of the bubble represents the frequency of each identified feedback. The qualitative work adds to the views derived from the quantitative analysis, specifically identifying the need for simplicity, credibility and relevance in explanation.

Table 8: Top Themes from Participant Feedback (Qualitative Coding)

Theme	**Description**	**Frequency (%)**
Need for Real Examples	Participants preferred explanations using case examples	78%
Simplicity vs Completeness Dilemma	Participants struggled between simple vs full explanations	66%
Fear of Bias in Black-Box Models	Concern about bias when explanations were absent	72%
Desire for Interactive Explanations	Participants wanted to ask questions to the model	61%
Trust Linked to Transparency	Trust increased when decisions were easily traceable	85%



Discussion

The study proceeds to provide a clear argument and verifiable evidence that suggest that explainability is not a 'nice to have' feature in machine learning (ML) models, but a 'must have' when the models are to be used in high-risk contexts. The performance advantage of these inherently explainable models demonstrated in terms of fidelity, user understandability, trust, and clarity confirm prior concerns of the dangers of black-box models (Krishna et al., 2022). Whenever internal logic cannot be communicated effectively to the users within an environment, one is left with the disastrous realities of a dilution of trust, combined with a likelihood of its misapplication, miscomprehension and improper usage in critical decision making scenarios.

A major implication of the presented results is the fact that explanation interpretability varies greatly between models that incorporate inherent interpretability to their structure and frameworks for generating explanations after the model has been built. For instance, we get insights from numerous post-hoc techniques such as feature attribution maps which may not be a perfect reflection of how the model was making its decision hence may not be accurate (Yang et al., 2022). This is in line with earlier arguments suggesting that explanations formed after model training are, in fact, approximations (Kaushik et al., 2020). Thus, it can be stated that these findings contribute to a new trend in the AI development, according to which the models should be interpretable from scratch, especially in such fields as healthcare, criminal justice, and finance since the costs of misinterpretation there are considerably high.

The superior user comprehension observed with interpretable models also supports the statement about explainability, that is about the fact that explainability is not a purely technical issue but a highly human process. People prefer AI models when they understand why a particular decision was arrived at; this forms the basis of an integrated mental model, with the AI model being as follows (Abdul et al., 2020). This is in accordance with the general HCI literature that asserts that a proper alignment of mental models is crucial for the use of a system to be both efficient and trustworthy (Zhang et al., 2020). In addition, the enhancement in user trust and perceived interpretability, which were confirmed in both quantitative and qualitative analyses, points to the fact that explanation quality controls the perceived credibility and adoption of AI systems in high-risk areas.

One of the key observations is the specificity of the priorities regarding the characteristics of explanations within the context of the computing domain. For example, the healthcare sector participants valued explanation correctness and clinical safety, participants of finance orientation valued fairness and bias detection, and the ones in the legal setting valued procedural justice and accountability. This domain-specific variance is indicative of the fact that a universal definition of explainability may not suffice here (Hoffman et al., 2019). Instead, explanation systems have to be more flexible and able to adjust the kind and level of detail of information provided about the decision-making process depending on the domain, the task, and the expertise of the user.

Another key finding is explanation faithfulness. High-fidelity rationales help to prevent users from being deceived by non-interpretable smoothing or erroneous explanations that are semantically correct in a literal way but are not faithful to the floating-point computational details of the model (Atanasova et al., 2020). It is possible for explanations to give consumers a false sense of security in the system by actually increasing risk factors closer to the range of an increase in risk rather than a reduction. These findings align with current issues in adversarial machine learning where the opponent may use the explanation methods to create misleading descriptions of the model's behavior and hide undesirable actions (Slack et al., 2022).

Another interesting point to be noted here is that in response to the two choices that were made available to the participants at the time of survey, the latter was more inclined towards the explanation which used real examples rather than providing general ideas and statistics. This view aligns with psychological literature indicating that case-based reasoning is more natural and effective in helping people learn compared to rule based systems Procedural knowledge brings our understanding to the fact that people have a natural ability of grasping and absorbing information in forms of cases rather than just rules (Lombrozo, 2012). Thus, the use of prototypes, counterfactuals, or narrative techniques may be significantly more useful in real-world applications than the assessment based on the feature importance alone.

However, at the same time, the work acknowledges that there are still important open problems in XAI research. However, it is important to note that there has always been a compromise between simplicity and comprehensive solutions. Patients understandably expect both simple and complete explanations, but these two objectives are contradictory (Lertvittayakumjorn & Toni, 2021). To make an explanation simple, it skips certain aspects; yet, providing all information may prove too much. The challenging area for the future research addressed by the authors concerns the balance of these two forces.

The fourth main limitation is perhaps the most significant one – it is challenging to quantify the quality of the explanations provided. This is a weakness because, although fidelity and user comprehension are well-defined measures, they are still limited to providing quantitative evidence, and do not encompass the diverse possibilities of human interpretations (Preece et al., 2018). More such studies need to be conducted to design more complicated models of evaluation that can take into account the cognitive load of instructions, perceptive and emotional reactions, and ethical viewpoints different from those of the authors of explanation.

Thus, we can also see that the results did indeed corroborate the argument that explainability is not just a technical problem but an ethical and social one as well. Some of the negative consequences include; With opaque systems, systematic biases are worsened, less democratic accountability, and all the power is put in the hands of the developers and people deploying these algorithms (Binns, 2018). While AI is increasingly integrating into governance arrangements, healthcare delivery, and judicial systems globally, it is expected that the number of calls for fair, understandable, and challenge-able explanations will continue growing.

However, it is also noteworthy that the regulation of such processes is advancing at an equally rapid pace. Current legislation including the European Digital Services Act (European Commission2022) and the Algorithmic Accountability Act in the United States, US Congress (2022) put more pressure on organizations to produce comprehensible explanations of an algorithm's actions. This dynamic environment underscores the need for researchers and practitioners to proactively place explainability as more than just an add-on but an essential pillar in developing artificial intelligence systems.

Based on the findings of this study, it is clear that there should be a major change in the development processes of AI models. Instead of viewing explainability as an afterthought addressed with techniques applied after model build, high-risk applications require that the concepts of explainability become embedded in the processes of model creation, calibration, assessment, and implementation. This is only possible when these structures and elements are better integrated in a more global manner so that AI systems can attain the expressed levels of trust, reliability, and fairness, as well as accountability needed for proper use in relevant domains.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2020). *Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda.* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). *A diagnostic study of explainability techniques for text classification.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy.* Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency.

European Commission. (2022). *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act).*

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2019). *Metrics for Explainable AI: Challenges and Prospects.* IEEE Intelligent Systems, 34(3), 56-63.

Kaushik, D., Hovy, D., & Lipton, Z. C. (2020). *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data.* International Conference on Learning Representations (ICLR).

Krishna, R., Narayanaswamy, B., & Singh, A. (2022). *On Faithfulness and Plausibility of Post-hoc Explanations in Deep Learning Models.* Journal of Artificial Intelligence Research.

Lertvittayakumjorn, P., & Toni, F. (2021). *Explanation-Based Human Debugging of Machine Learning Models: A Survey.* arXiv preprint arXiv:2107.08907.

Lombrozo, T. (2012). *Explanation and abductive inference.* In K. Holyoak & R. Morrison (Eds.), The Oxford Handbook of Thinking and Reasoning.

Preece, A., Braines, D., & Harborne, D. (2018). *Stakeholders in explainable AI*. Proceedings of the 1st Workshop on Human Interpretability in Machine Learning (WHI 2018).

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2022). *Reliable post hoc explanations: Modeling uncertainty in explainability.* Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.

US Congress. (2022). *Algorithmic Accountability Act of 2022.*

Weitz, K., et al. (2019). *A human-grounded evaluation benchmark for local explanations of machine learning.* Advances in Neural Information Processing Systems.

Yang, Y., Stoyanovich, J., & Howe, B. (2022). *Measuring and mitigating the faithfulness-interpretability tradeoff in AI explanations.* IEEE Transactions on Knowledge and Data Engineering.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). *Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making.* Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica.

Caruana, R., et al. (2015). *Intelligible models for healthcare*. ACM SIGKDD.

Doshi-Velez, F., & Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. arXiv.

European Commission. (2021). *Proposal for a Regulation on a European approach for Artificial Intelligence*.

Goodman, B., & Flaxman, S. (2017). *European Union regulations on algorithmic decision-making*. AI Magazine.

Guidotti, R., et al. (2018). *A Survey of Methods for Explaining Black Box Models*. ACM Computing Surveys.

Gunning, D., & Aha, D. W. (2019). *DARPA's Explainable Artificial Intelligence (XAI) Program*. AI Magazine.

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning*. Advances in Neural Information Processing Systems.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Explaining Explanation, Part 4: A Deep Dive on Deep Nets*. IEEE Intelligent Systems.

Lipton, Z. C. (2018). *The Mythos of Model Interpretability*. ACM Queue.

Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. NeurIPS.

Miller, T. (2019). *Explanation in Artificial Intelligence: Insights from the Social Sciences*. Artificial Intelligence.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. KDD.

Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*. Nature Machine Intelligence.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). *Sanity Checks for Saliency Maps*. Advances in Neural Information Processing Systems.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., & Explainable ML Challenge Team. (2020). *Explainable Machine Learning in Deployment*. Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency.

Chen, C., Song, X., Wainwright, M. J., & Jordan, M. I. (2019). *Learning to Explain: An Information-Theoretic Perspective on Model Interpretation*. Proceedings of the 36th International Conference on Machine Learning.

Chen, J., Lin, C., Rudin, C., & Seltzer, M. (2020). *Boosted Decision Trees with Constraints: An Empirical Study of the Interpretable Boosting Machines*. NeurIPS Workshop on Human Interpretability in Machine Learning.

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2019). *Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions*. Proceedings of the 24th International Conference on Intelligent User Interfaces.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining Explanations: An Overview of Interpretability of Machine Learning*. IEEE 5th International Conference on Data Science and Advanced Analytics.

Hall, P., Gill, N., Kurka, M., & Pham, T. (2020). *Machine Learning Explainability: Data and Frameworks*. FICO Research Whitepaper.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). *What Do We Need to Build Explainable AI Systems for the Medical Domain?*. arXiv preprint arXiv:1712.09923.

Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. P. (2018). *ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models*. IEEE Transactions on Visualization and Computer Graphics.

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). *Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions*. AAAI Conference on Artificial Intelligence.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys.

Montavon, G., Samek, W., & Müller, K. R. (2018). *Methods for Interpreting and Understanding Deep Neural Networks*. Digital Signal Processing.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). *Definitions, Methods, and Applications in Interpretable Machine Learning*. Proceedings of the National Academy of Sciences.

NIST. (2022). *NIST AI Risk Management Framework (Draft)*. U.S. Department of Commerce.

OECD. (2019). *OECD Principles on Artificial Intelligence*.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). *Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods*. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2016). *Embedding Projector: Interactive Visualization and Interpretation of Embeddings*. arXiv preprint arXiv:1611.05469.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing Theory-Driven User-Centric Explainable AI*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Zhou, Z. H., Du, X., & Ying, Y. (2021). *Evaluating Explainable AI: Which Explanation is the Best Explanation?*. Frontiers in Artificial Intelligence.