TOWARD ADAPTIVE EXPLAINABLE ARTIFICIAL INTELLIGENCE: A SYSTEMATIC LITERATURE REVIEW ACROSS HUMAN, TASK CONTEXT, AND MODEL DIMENSIONS

Lappeenranta–Lahti University of Technology LUT
Master's programme in Digital System and Service Development
2025
Fatemeh Eshghi
Examiner: Associate Professor Annika Wolff

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

LUT School of Engineering Sciences

Software Engineering

Fatemeh Eshghi

Toward adaptive explainable artificial intelligence: a systematic literature review across human, task context, and model dimensions

Master's thesis

2025

126 pages, 9 figures, 5 tables

Examiner: Associate Professor Annika Wolff

Keywords: Adaptive Explainable Artificial Intelligence; Adaptive XAI; Adaptation; Human XAI; Human factors; Task Context; Model Type; Systematic Literature Review; SLR.

This thesis presents the new and growing area of adaptive Explainable Artificial Intelligence (XAI) through a Systematic Literature Review (SLR) carried out under the PRISMA 2020 guidelines. The research investigates the possibility of explanations being adjusted dynamically in three core aspects: Human, Task, and Model. The review of the 43 studies (2015-2025) points out that although the three dimensions have been extensively studied taken separately, very little research has been done to unify them in an adaptive framework. Most of the existing methods are not yet focused on three dimensions at the same time. The results expose critical research areas that need to be investigated, among them, the absence of three-dimensional adaptation mechanisms and the standardization of evaluation criteria. To tackle these challenges, the thesis proposes a conceptual model for the Adaptive XAI, which underlines the necessity for dynamic and balanced explanations that can vary according to the human, the text context, and the model focus.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my husband Hamed Moghni, whose patience, understanding, and continuous support guided me throughout this thesis. I am also grateful to Associate Professor Annika Wolff for being my thesis supervisor. Finally, I would like to thank LUT University for the chance to do this research and the supportive academic environment.

DECLARATION OF AI USAGE

I, the undersigned, in line with LUT AI usage recommendations and policy (https://elut.lut.fi/en/completing-studies/rules-and-regulations/ai-based-tools-policies) hereby declare the following regarding the usage of AI tools in the development of the contents of my thesis work:

1- *ChatGPT-5 was used for the following supportive purposes:*

- *Summarizing and explaining complex studies.*
- *Narrowing down the topic.*
- *Formulating text, checking language, and improving clarity.*
- *Assisting in preparing the abstract and presentation materials based on my own draft.*
- *suggest name for figures and tables and subject this thesis.*

AI tools were not used to generate new research content, analysis, or results

2- *Grammarly was used for Checking grammar and spelling, as well as assisting with AI language editing.*

I declare that no other AI assistance or tools were used in the development of the contents of my thesis work than the above mentioned.

SYMBOLS AND ABBREVIATIONS

Roman characters

h	Human
t	Task
m	Model
A	adaptive
w	weight (Focus)

Abbreviations

AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
SLR	Systematic Literature Review
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ	Research Question
LLM	Large Language Model
KG	Knowledge Graph
NLP	Natural Language Processing
HCI	Human–Computer Interaction
SHAP	SHapley Additive exPlanationsf
LIME	Local Interpretable Model-Agnostic Explanations
CF	Counterfactual Explanation
GAM	Generalized Additive Model
ART	Adaptive Resonance Theory
SENN	Self-Explaining Neural Networks
IDS	Intrusion Detection System
DLT	Distributed Ledger Technology
IoT	Internet of Things
ITS	Intelligent Tutoring System
GLM	Generalized Linear Model
DTD	Deep Taylor Decomposition
RF	Random Forest
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory Network
GNN	Graph Neural Network
GLMs	Generalized Linear Models
GAMs	Generalized Additive Models
ID3	Iterative Dichotomiser 3
CART	Classification and Regression Trees
BRL	Bayesian Rule Lists
FIS	Fuzzy Logic Systems

ANFIS	Adaptive Neuro-Fuzzy Inference System
EBM	Explainable Boosting Machines
k-NN	k-Nearest Neighbors
CPTs	Conditional Probability Tables
ART	Adaptive Resonance Theory
CNNs	Convolutional Neural Networks
XNNs	Explainable Neural Networks
SFA	Stochastic Frontier Analysis
xNN-SF	Explainable Neural Network Inspired by Stochastic Frontier Model
SENN	Self-Explaining Neural Networks
NN	Neural Network
ICNs	Interpretable Convolutional Networks
CBMs	Concept Bottleneck Models
NAMs	Neural Additive Models
GAT	Graph Attention Networks
GNNs	Graph Neural Networks
NCMs	Neural Casual Models
LG-GNN	Logic-Guided Graph Networks
XGNN	Explainable Graph Neural Networks
CNMs	Cognitive Neural Models
PDPs	Partial Dependence Plots
ICE	Individual Conditional Expectation
LRP	Layer-wise Relevance Propagation
PDA	Prediction Difference Analysis
Human × Task × Model	Three-dimensional Adaptive XAI framework

Table of contents

Figures

Tables

1. Introduction

1.1. Background

In recent years, artificial intelligence has undergone rapid growth, ushering into a transformation of industries and everyday livelihood across various domains such as health, finance, autonomous transport, teaching, cybersecurity, transportation, and government services. The advent of rapid machine-learning techniques, particularly those of deep learning and reinforcement learning, could help AI systems gain accomplishments and accuracy in decision- making processes [1] [2]. These, however, pose huge challenges. Many AI systems -the deep neural networks are one classic instance- are considered black-boxes: human beings can see a prediction from the system but cannot understand the reasoning stored in the interiors of these unfathomable systems [3].

Several critical questions arise from this lack of transparency. The first issue concerns trust and reliability: users cannot very well verify or question the AI output. Second is the ethical and legal question: what happens to human well-being when AI decisions affect medical diagnoses, financial risk assessments, or judicial rulings? Then there is the question of misuse or over-reliance: either the user blindly follows whatever the AI says, or they completely disregard its recommendations due to incomprehensibility. All these challenges stress the importance of designing AI systems that can have their behavior and decision-making processes interpreted and understood [4].

These topics have given rise to **Explainable Artificial Intelligence (XAI)**. XAI systems focus on ensuring that AI systems perform well so that they can be fed with interpretable explanations of their predictions, decisions, or behavior. By granting transparency over the system, XAI seeks to facilitate trust, responsibility, and human understanding so that AI may be responsible and integrated into various domains. Hence, within the past decade, research in XAI has gone from building simple interpretable models such as decision trees or linear models to advanced model-agnostic explanation methods such as LIME, SHAP, Grad-CAM, or counterfactual reasoning [4] [2].

Additionally, XAI is increasingly acknowledged as when humans set critical decisions aided by AI systems. It is during these events that concrete and meaningful explanations allow the

users to appropriately rely on the AI recommendation and adjust their trust basis from understanding the rationale of the system [5].

1.2. Importance of Explainability for Human Users

Explainability is of utmost importance, especially when AI systems are involved in supporting decision-making processes impacting human lives. The users must be able to weigh the reliability, considerations of fairness, and aspects of relevance of the system's recommendations. In addition, stakeholders such as a domain expert, the end-user, and a regulator require different levels and types of explanation according to their distinct roles, expertise, and cognitive styles. If explanations are not so tailored, users may misinterpret the outputs of AIs, which in turn further impedes poor decision-making and casts distrust [6].

1.3. Motivation for Adaptation in XAI

To make AI more explainable, most research trends have been to treat users as a single, undifferentiated group. Users, however, differ tremendously from one another in their cognitive styles, domain knowledge, goals, and preferences. Hence, a "one-size-fits-all" explanation may not go a long way to satisfying the needs of most users [6]. This void has created a greater interest in adaptive XAI, in which the explanation is adapted to the characteristics or preferences of individual users. Named explanations can increase satisfaction, help users to make better decisions, and foster deeper trust in AI systems [7] [8]. The landscape of research remains fragmented despite the promise of developing XAI. Different authors focus on different human factors; different personalization is used; different task contexts are targeted; and different XAI models are being used. An attempt at a systematic review must be undertaken to bring this knowledge together, identify trends, and show gaps.

The adaptability of explanation thus must consider its underlying model, complexity, and the sort of abstraction level conducive for the user.

1.4. Problem Statement

Even with the fast growth of XAI techniques, a gap still exists in artificial fields. Most current approaches take a researcher-side view, with little consideration for the end-user who

may have a completely different set of needs. For example, different stakeholders such as domain experts, developers, or lay human users may differ in their levels of domain knowledge and AI literacy. Hence, an explanation that differs from one user type to the other would be valuable.

For example, highlighting certain feature-importance scores can well serve the purpose of helping the data scientist understand the logic of a model, but that solution will often be meaningless to non-technical users. On the other hand, studies argue that current methods of XAI may fail on certain occasions with respect to their intended purpose or lead to unintended consequences such as over-trusting or confusion.

Most research tends to concentrate only on a single dimension—for example, the XAI model, the task domain, or the human user — rarely considering how all three influences interact in one way or another. Hence, the explanations fall victim to being either overly simplistic or overly complex for consideration, which decreases their success down the line and limits their real-world use.

To design effective adaptive explanations, it is not sufficient to consider explanation methods in isolation. Instead, I propose a conceptual model that accounts for the adaptive interaction of three core dimensions: human factors, task context, and XAI model characteristics. Each of these dimensions introduces unique requirements and constraints that shape how explanations should be constructed and delivered. Ignoring any one of them risks producing explanations that are either ineffective, misleading, or even counterproductive.

1.5. The Three Dimensions of Adaptive XAI

Adaptive XAI not only requires identifying the dimensions of the human, task, and model, but also understanding how explanations should dynamically change depending on these factors. Unlike traditional XAI—which often produces fixed, one-size-fits-all explanations—adaptive XAI treats each dimension as a variable that modifies the explanation strategy. The following subsections clarify how each dimension contributes to adaptation.

1.5.1. Human Factors

Human factors are one of the crucial forces in adaptive XAI. Among factors that influence the interpretation of explanations are users' mental models, expertise level, cognitive load, and even cultural background [10]. For example, a novice user might require a simple, narrative-style explanation, while a domain expert might ask for technical details, causal reasoning, or even counterfactual analyses [11]. Therefore, adaptive explanation systems might aid in trust calibration, decision accuracy, and overall satisfaction with AI systems [12]. Hence, to form adaptable explanation strategies, knowing and classifying human factors is key.

1.5.2. Task Context

Explanations are rather efficient depending upon how the AI systems are deployed. The task context includes the domain (healthcare, education, finance, etc.), the risk the decisions carry, and the intended user goal [13]. For instance, in high stakes such as medicine, users seek truthful, transparent, and strictly defensible explanations, whereas in low stakes such as a music recommendation, explanations can be lightweight and preference oriented.
Likewise, adaptive task context also determines which evaluation criteria will be applied: for instance, considering an environment of accuracy, fairness, safety, or even usability may appear appropriate [14].

1.5.3. Model Type

Different XAI models naturally furnish different explanations for their workings. Intrinsically interpretable models (e.g., decision trees, logistic regression) are by nature transparent, whereas Model-agnostic models (e.g., deep neural networks, ensemble methods) are generally studied after fitting, when applying post-hoc interpretation techniques such as SHAP, LIME, or counterfactual explanations [9]. The adaptability of explanation thus must consider its underlying model, complexity, and the sort of abstraction level conducive for the user.

1.6. Research Questions

To explore this growing field, this study is guided by the following four research questions:

RQ1: What human factors have been considered in XAI research, and what ways do they affect the personalization of explanation?
RQ2: What are the task contexts within which adaptive XAI has been applied, and what trends can be observed?
RQ3: What types of XAI models have been looked at in literature, and how does the type of model affect the form and content of the explanations?
RQ4: How have these three dimensions combined in past work, and what gaps remain open for inquiry?

1.7. Objective of the Study

The primary objective of this thesis is to conduct a Systematic Literature Review (SLR) on the state of adaptive XAI, focusing on the interplay between human factors, task contexts, and XAI model types. This study will follow the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, rigor, and reproducibility in the review process [15].

The goals are:

- To identify which human-related aspects are considered in the personalization of XAI.
- To analyze the domains and tasks where personalized and adaptive explanations have been studied.
- To investigate how the type of XAI model influences the design and presentation of explanations.
- To synthesize how these three dimensions integrate and what opportunities exist for future research.

1.8. Structure of the Thesis

This thesis is structured as follows:
Chapter 2- Search Strategy for Identifying Relevant Studies
Chapter 3- Human Factors Considered in XAI
Chapter 4- Domains and Task Context in XAI
Chapter 5- Model Type Methods in XAI
Chapter 6- Three-Dimensional Adaptive Integrations

2. Search Strategy for Identifying Relevant Studies

To be transparent and reproducible, the search procedure follows the PRISMA 2020 guidelines. The PRISMA 2020 flow diagram (Figure 1 - Search Strategy for Identifying Relevant Studies) illustrates the identification, screening, eligibility, and inclusion stages for this systematic review in the fields of human factors, task context, and model features in Explainable AI. The process is explained below.

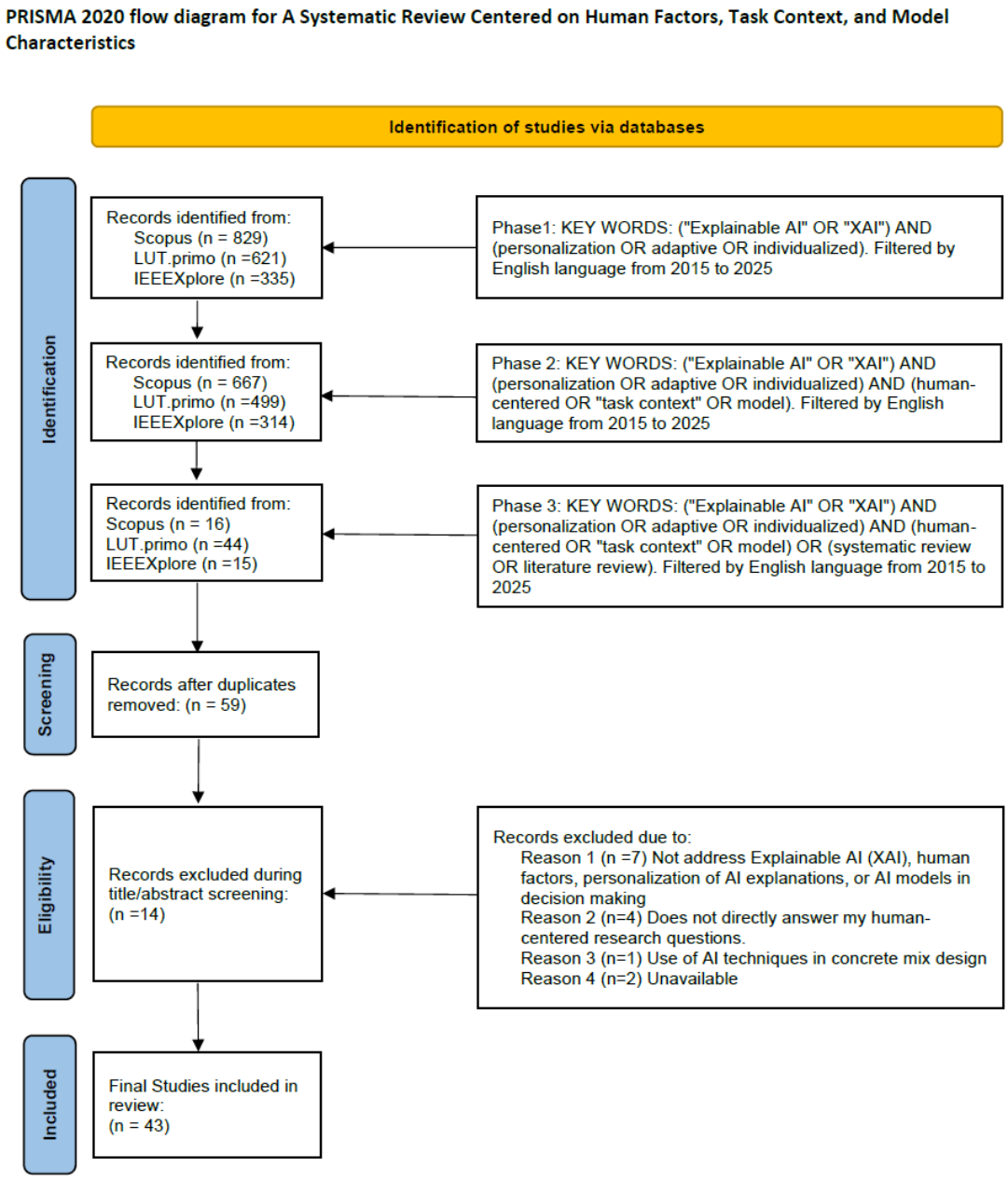

Figure 1 - Search Strategy for Identifying Relevant Studies

2.1. Identification

2.1.1. Phase 1 – Core Topic Identification

KEY WORDS: ("Explainable AI" OR "XAI") AND (personalization OR adaptive OR individualized).
Filters applied: English Language, 2015-2025

Reasoning:
This first search concentrates on a generally broader literature related to Explainable AI (XAI) but covers also searches with the term's personalization, adaptation, and individualization.
The range of years 2015-2025 is selected since most breakthroughs in XAI came after 2015, concurrent with the rise of deep learning and explainability research.
Ensuring English language publication restricts the review to consistent works that would otherwise be complicated by translations.
This step is aimed at capturing the broadest possible set of relevant works before applying more specific constraints.

2.1.2. Phase 2 – Contextual and Application Relevance

KEY WORDS: ("Explainable AI" OR "XAI") AND (personalization OR adaptive OR individualized) AND (human-centered OR "task context" OR model)
Filters applied: English Language, 2015-2025

Reasoning:
In this second stage, the focus is on narrowing down the results toward papers that treat personalized and adaptive XAI from a human factors perspective, or from specific views about tasks or the XAI model type.
In this way, the focus corresponds directly to RQ1-RQ3 in the thesis, meaning that the papers collected are meant to discuss not just personalization theoretically but also how and where it is applied, including which kinds of models are involved.
The introduction of such terms thus screens out Phase 1 results in personalization, but in unrelated contexts.

2.1.3. Phase 2 – Methodological Relevance

KEY WORDS: ("Explainable AI" OR "XAI") AND (personalization OR adaptive OR individualized) AND (human-centered OR "task context" OR model) AND (systematic review OR literature review)
Filters applied: English Language, 2015-2025

Reasoning:
The third phase is focused on reviewing papers to complement the primary studies with an overview of the existing work.
Trying to retrieve systematic and literature review keywords would ensure the retrieval of papers in which evidence is synthesized rather than presented in isolation through one or a few experiments.
These types of sources augment the bird's eye view, focus research gaps, and thereby assist in the validation of the PRISMA methodology followed by you.
Hence, the present step fortifies your research on a macro-level by able mapping of the field and justifying any chosen research direction.

2.2. Screening

This first set of records is exported and saved in Zotero for reference management. Because the same article is frequently indexed in multiple databases, the duplicate detection feature of Zotero is used to weed out the redundant records. A total of sixteen duplicate records is removed at this stage, leaving 59 unique records to be screened for relevance. Accordingly, the studies are taken into the subsequent screening and eligibility assessment phases only once.

2.3. Eligibility

Title and abstract screening are conducted to assess the relevance of the records to the research objectives among the 59 different records saved in Zotero. Eleven of these are removed for several reasons:

> **Reason 1** (n = 7): Studies on Explainable AI (XAI), human factors, personalization or adaptation of AI explanations, or AI systems in decision-making are not considered.

Reason 2 (n = 4): The Study does not answer the human-centered research questions guiding this review.

Reason 3 (n = 1): It dealt with AI methodology for concrete mix design outside the scope of the research.

Reason 4 (n = 2): Full-text retrieval is not possible for 2 records, and they are consequently excluded.

The screen, therefore, filters studies that are strictly personalized human-centric XAI within relevant contexts from proceeding to the eligibility assessment.

2.4. Included

After full-text screening, 43 Publications met all inclusion criteria and are selected for the final synthesis. These publications directly address different facets of personalized, human-centered Explainable AI, including human factors, aspects of task context, and XAI model types relevant to the research questions.
The list of all the included studies found in Table 1 below.

Table 1 - Publications studied

Num	Year	Author	Title
P1	2025	Ouifak, Hafsaa; Idri, Ali	A comprehensive review of fuzzy logic based interpretability and explainability of machine learning techniques across domains
P2	2025	Arnob, Arjun Kumar Bose; Chowdhury, Rajarshi Roy; Chaiti, Nusrat Alam; Saha, Sudipta; Roy, Ajoy	A comprehensive systematic review of intrusion detection systems: emerging techniques, challenges, and future research directions
P3	2024	Wankhede, Vidhi; Verma, Prateek; Gahane, Shailesh	A Review on the Developments in the Field of AI-Based Gait Analysis
P4	2024	Wang, Ziming; Huang, Changwu; Yao, Xin	A Roadmap of Explainable Artificial Intelligence: Explain to Whom, When, What and How
P5	2020	Hatwell, Julian; Gaber, Mohamed Medhat; Atif Azad, R. Muhammad	Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences
P6	2025	Gong, Mingming; Song, Yiliao; Koh, Yun Sing; Xiang, Wei; Wang, Derui	AI 2024: Advances in Artificial Intelligence: 37th Australasian Joint Conference on Artificial Intelligence, AI 2024, Melbourne, VIC, Australia, November 25–29, 2024, Proceedings, Part I

P7	2022	Degen, Helmut; Ntoa, Stavroula	Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings
P8	2025	Pagliaro, Antonio	Artificial Intelligence vs. Efficient Markets: A Critical Reassessment of Predictive Models in the Big Data Era
P9	2025	Moustati, Imane; Gherabi, Noreddine	Bridging Behavioral Insights and Automated Trading: An Internet of Behaviors Approach for Enhanced Financial Decision-Making
P10	2025	Benitez, Ian B.; Teodoro, Mark Anthony G.; Pahati, Melodia D.	Challenges and Opportunities in AI Integration in Power System Protection
P11	2022	Séroussi, B.; Weber, P.; Dhombres, F.	Challenges of Trustable AI and Added-Value on Health: Proceedings of MIE 2022.
P12	2024	Subramanian, Harishankar V.; Canfield, Casey; Shank, Daniel B.	Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review
P13	2025	Wiratsin, In-on; Ragkhitwetsagul, Chaiyong	Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review
P14	2025	Ibrahim, Muhamme; Mahmoud, Moamin A; Islam, Noor; Gunasekara, Saraswathy Shamini	Enhancing Smart Grid Stability Using AI Techniques: A Systematic Literature Review
P15	2024	Aljawawdeh, Hamzeh; Fawareh, Hamed; Al Refai, Mohammed; Alshraideh, Fuad; Salman, Randa; Dabaa't, Sara Abu; Alazzeh, Heba; Al-Shdaifat, Hazim; Khouj, Mohammed	Ethical Considerations in AI-Driven HCI Systems in Software Engineering
P16	2025	Nag, Prashant Kumar; Bhagat, Amit; Vishnu Priya, R.	Expanding AI's Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques
P17	2019	Anjomshoae, Sule; Najjar, Amro; Calvaresi, Davide; Främling, Kary	Explainable Agents and Robots: Results from a Systematic Literature Review

P18	2021	Rayz, Julia	Explainable AI and other applications of fuzzy techniques: proceedings of the 2021 Annual Conference of the North American Fuzzy Information Processing Society, NAFIPS 2021
P19	2024	Tejasvi, Koti; Vinya, V. Lokeshwari; Padmaja, Jagini Naga; Begum, Ruqqaiaya; Jabbar, M. A.; Abraham, Ajith; Gaur, Loveleen	Explainable Artificial Intelligence (XAI) for Managing Customer Needs in E-Commerce: A Systematic Review
P20	2025	Ozay, Dervis; Jahanbakth, Mohammad; Wang, Shouyi	Exploring the Intersection of Big Data and AI With CRM Through Descriptive, Network, and Contextual Methods
P21	2025	Deng, Zhe; Torim, Ants; Ben Yahia, Sadok; Bahsi, Hayretdin	Generative AI in Intrusion Detection Systems for Internet of Things: A Systematic Literature Review
P22	2025	García López, Iván Miguel; Ramírez Montoya, María Soledad; Molina Espinosa, José Martín	Harnessing Open Language Models: A Systematic Literature Review Unleashing AI's Potential for a Smarter Future
P23	2024	Nah, Fiona Fui-Hoon; Siau, Keng Leng	HCI in Business, Government and Organizations: 11th International Conference, HCIBGO 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, Proceedings, Part I
P24	2025	Coman, Adela; Vasilache, Simona; Fui-Hoon Nah, Fiona; Siau, Keng Leng; Wei, June; Margetis, George	HCI International 2024 – Late Breaking Papers: 26th International Conference on Human-Computer Interaction, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, Proceedings, Part II
P25	2020	Kumar, Vivekanandan; Troussas, Christos; Troussas, Christos; Kumar, Vivekanandan	Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8-12, 2020, Proceedings
P26	2022	Farhaoui, Yousef; Rocha, Alvaro; Brahmia, Zouhaier; Bhushab, Bharat	International Conference on Artificial Intelligence and Smart Environment, ICAISE 2024 - Lappeenranta-Lahti University of Technology LUT
P27	2024	Qi, Jun; Yang, Bo	Internet of Things of Big Data for Healthcare: 5th International Workshop, IoTBDH 2023, Birmingham, UK, October 21–25, 2023, Proceedings

P28	2025	Dehal, Ramandeep Singh; Sharma, Mehak; Rajabi, Enayat	Knowledge Graphs and Their Reciprocal Relationship with Large Language Models
P29	2024	Zaphiris, Panayiotis; Ioannou, Andri	Learning and Collaboration Technologies: 11th International Conference, LCT 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part III
P30	2023	Oyebode, Oladapo; Fowles, Jonathon; Steeves, Darren; Orji, Rita	Machine Learning Techniques in Adaptive and Personalized Systems for Health and Wellness
P31	2025	Tekli, Joe; Gamper, Johann; Chbeir, Richard; Manolopoulos, Yannis; Sassi, Salma; Ivanovic, Mirjana; Vargas-Solar, Genoveva; Zumpano, Ester	New Trends in Database and Information Systems: ADBIS 2024 Short Papers, Workshops, Doctoral Consortium and Tutorials, Bayonne, France, August 28–31, 2024, Proceedings
P32	2020	Pandl, Konstantin D.; Thiebes, Scott; Schmidt-Kraepelin, Manuel; Sunyaev, Ali	On the Convergence of Artificial Intelligence and Distributed Ledger Technology: A Scoping Review and Future Research Agenda
P33	2022	Djeddi, Chawki	Pattern recognition and artificial intelligence: 5th Mediterranean Conference, MedPRAI 2021, Istanbul, Turkey, December 17-18, 2021, Proceedings
P34	2023	Lazaar, Mohamed	Proceedings of the 6th international conference on big data and internet of things
P35	2022	Watson, Richard T.; Greenwood, Brad; Wright, Ryan	Towards Personalized Explanations for AI Systems: Designing a Role Model for Explainable AI in Auditing
P36	2020	Grossberg, Stephen	A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action
P37	2024	Durango, Iván; Gallud, José A.; Penichet, Victor M. R.	The data dance: choreographing seamless partnerships between humans, data, and GenAI
P38	2025	Aravazhi, Prasanna Sakthi; Gunasekaran, Praveen; Benjamin, Neo Zhong Yi, et al.	The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions
P39	2025	Ratul, Qudrat E Alahy; Zaman, Eshita; Das, Saikat; Memari, Majid	The Role of AI in Transforming Education: A Systematic Review of Trends

P40	2025	Cappuccio, Eleonora; Esposito, Andrea; Greco, Francesco; Desolda, Giuseppe; Lanzilotti, Rosa; Rinzivillo, Salvatore	Explanation User Interfaces: A Systematic Literature Review
P41	2021	Chromik, Michael; Butz, Andreas	Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces
P42	2025	Budhkar, Aishwarya; Song, Qianqian; Su, Jing; Zhang, Xuhong	Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics
P43	2025	He, Gaole; Aishwarya, Nilay; Gadiraju, Ujwal	Is Conversational XAI All You Need? Human-AI Decision Making with a Conversational XAI Assistant

3. Human Factors in XAI

The effectiveness of explainable artificial intelligence (XAI) relies not just to the extent to which the systems disclose their internal logic, but also on humans' capability of understanding, trusting, and responding to the explanations provided. The human factor is the central core of this communication, and it determines whether AI systems are perceived as being transparent, trustworthy, usable, and additionally supportive of the user's goals [10]. In the quest to reveal the inner workings of these machines, XAI should consider the psychological, emotional, cognitive, and social aspects of the human–machine interaction. Human-centered XAI is developed not just by the developers or experts in the field of human–machine interaction [16].

It recognizes that users differ widely in their expertise, reasoning styles, motivations, and expectations. What is counted as an “explanation” for a data scientist may be entirely different from the needs of a medical patient, financial analyst, or autonomous vehicle driver; therefore, the explanations must be designed according to the human cognitive capacities, trust dynamics, interpretability needs, ethical expectations, and mental models. Providing that these issues are resolved, the XAI systems will not only make it possible for users to understand the AI behavior but will also make them feel strong, keep the willpower, and take sensible, responsible decisions when working with the intelligent systems side by side [17] [10]. The following subsections outline the key human factors that influence Explainable AI (XAI).

Table 2 - Human Factors in XAI

Main Category	Subcategories
Cognitive Factors	Comprehension Ability, Cognitive Load, Attention & Memory, Reasoning Style, Cognitive Biases, Learning Curve, Sensemaking & Mental Simulation
Trust & Reliance	Calibration of Trust, Overtrust/Undertrust Risks, Transparency vs. Confidence, Trust Repair, Temporal Trust, Overconfidence from Attractive Explanations
Interpretability & Usability	Level of Detail, Representation Format, Counterfactuals, Instance-based Explanations, Actionability, Interactivity, Accessibility
Ethical & Emotional Factors	Fairness Perceptions, Privacy Concerns, Emotional Reassurance, Cultural Sensitivity, Inclusivity
User Goals & Motivation	Debugging, Learning, Justifying, Decision Support
User Mental Models	—
Control & Agency	—

3.1. Cognitive Factors

Cognitive factors in Explainable AI concern how people perceive, process, and understand AI explanations. Essentially, XAI systems must take human cognition into account to generate explanations that are understandable, useful, and trustworthy. Different users have differing cognitive abilities, reasoning styles, and prior knowledge that shape how they may interpret outputs from AI. Important cognitive factors include:

3.1.1. Comprehension ability

Users range from some with extensive knowledge of AI to novices who have never worked with one, and the explanations must, thus, fit their respective mental models [18]. XAI must translate complexity into forms that are appropriate for audiences [19] [20] [21].

For example, a data scientist can understand model weights; a patient needs something in plain language, or for a medical AI, a doctor might want "the model weighted infection markers more than age," whereas a patient would rather hear “your wound size and temperature suggest infection” [22].

3.1.2. Cognitive load

Cognitive load refers to the extent of mental effort involved in information processing. An overly detailed explanation would overwhelm and confuse users, while limited information may raise suspicion of a “black box” system. To balance cognitive load, XAI can employ layered explanations with high-level summaries and more detailed layers that users can drill into if they want. This approach wastes less of the users’ effort unless they desire to explore depth. [23] [24] [25] [26] .

3.1.3. Attention and memory

The phrase "attention in working memory: attention is needed but it yearns" is the title of Rhodes and Cowan's 2018 paper. It implies that attention is required to hold information in working memory, but at the same time, attention wants to disengage and be "free" to form new memories or shift attention to some other tasks.

With this, there lies a tension between maintaining information and the need to offload or update it. The "Needed" Aspect Maintenance: Attention is needed for holding information within working memory with activation and accessibility [27].

Refreshing: The use of attention in refreshing items in memory helps in preventing decay from so that they hardly remain in attentional focus [27].

The "Yearn to Be Free" Aspect Offloading: To ease the cognitive burden on their attention, people tend to offload information either by forming new memories or relying upon stored concepts in long-term memory. Reducing Load: When information is successfully offloaded, attention is free to be focused elsewhere-on new tasks or other information rather than attending to the present maintenance. [27] .

3.1.4. Reasoning style

Among human cognitive factors, reasoning style is a major one which describes the way users think and solve problems by their natural mentality. Some of the users depend on analytical, step-by-step logic, while others might use intuitive, lump-sum, or example-based reasoning. Therefore, XAI must give its explanation to fit these individual contrasts so that the users can interpret AI outputs in their respective cognitive approach of choice more conveniently and effectively. Through reasoning style, it happens that explanations even get more intuitive, misunderstanding is lessened, trust is built, understanding, and deciding are enhanced [10] .

3.1.5. Cognitive biases

Users do not always process explanations rationally. They may suffer from biases such as confirmation bias (accepting only those explanations that align with prior beliefs), automation bias (relying too heavily on the AI), or anchoring (focusing on the first explanation) [28] [29]. For example, a doctor might ignore AI advice if the diagnosis does not align with their initial one (confirmation bias).

3.1.6. Learning curve

Users grow in their needs as they become familiar with the AI system: those needs evolve; at first, simple explanations in plain language may suffice, but then more advanced users may want in-depth analysis or trend insight. A patient using a health app may initially need basic feedback ("your heart rate is high"), but after several months of use, they want detailed trend analysis so that they can analyze patterns or risks. Designing for the learning curve will ensure explanations remain relevant and useful down the line [30].

3.1.7. Sensemaking and Mental Simulation

Sensemaking and mental simulation are the things that the users can do to be actively engaged in exploring and experimenting with the AI's reasoning, like thinking of other scenarios, changing the inputs, or predicting 'what-if' outcomes. The user and AI's gradual interaction leads to the respective building of extensive user mental representations thereby, allowing the user to be able to predict possible outcomes and to understand even better the causal relationships in the AI decisions, in the long run, already enabling a deeper level of understanding, trusting and acting because of more information [31] [32].

3.2. Trust & Reliance

Trust and reliance are core elements of human–AI interaction that affect the very success of using AI appropriately and accepting it as well as acting upon the delivered outputs.

3.2.1. Calibration of trust

Calibration of trust means a balanced level of confidence in the AI system. User must not rely on it blindly (overtrust), nor must one disregard it altogether (undertrust). Trust is well-calibrated when explanations assist users in developing a mental model that correctly reflects how the AI operates and how reliable it is under varied circumstances. An accurately developed mental model allows the user to make informed decisions about relying on outputs from the AI by knowing exactly how the AI operates and under which circumstances the AI can be trusted [33] [34] [35].

3.2.2. Overtrust/Undertrust risks

Ensuring overtrust or undertrust stands as a risk to the safety, reliability, and efficacy of human–AI interaction within explainable AI (XAI). When users put excessive trust in AI decisions and follow their recommendations without question, they may be led to serious consequences, particularly in high-stakes domains such as healthcare, finance, or autonomous driving, where users might not analyze or may even have to override inaccurate AI outputs. For example, total reliance by a physician on AI diagnostic tools without verification or consideration of important clinical context could be detrimental. Inferior trust exists when users do not see the AI as a valid or advantageous option to consider. They might have already seen errors or found the explanations bewildering or unintelligible. This casts underutilization of appropriate AI capabilities and may even end up reducing system efficiency because of skepticism from users who disagree with suitable or beneficial recommendations. Consequently, XAI should aim to alleviate the extremes, providing explanations that are balanced, transparent, and context-aware to properly calibrate trust. [33].

3.2.3. Transparency vs. confidence

Transparency is paramount in instilling user confidence. Users must feel that the AI system honestly discloses its certainty and limitations. Yet, presenting raw mathematical uncertainty may overwhelm users, which is why intuitive representation methods should be preferred: probability bars, for example, or even a traffic light system. When confidence is conveyed in a transparent and comprehensible manner, users can evaluate the AI's certainty against contextual factors themselves. [33] [18].

3.2.4. Trust repair

When AI systems make errors, repairing trust is pivotal. Trust repair concerns the practices and explanations that restore confidence in AI after an operational failure. Merely communicating that an error occurred is insufficient. The explanation must also describe why the mistake occurred, what corrective actions have been taken, and measures put in place to prevent similar errors in the future. An effective trust repair mechanism helps users restore confidence in the system's reliability and fairness. To restore trust, after an erroneous

prediction, the system might explain the cause (e.g., "limited training data for this case") and how it adjusted its model or decision boundaries. By fostering transparent communication and corrective feedback, XAI systems retain user trust in the long run, despite occasional errors. [33].

3.2.5. Temporal Trust

The matter of trust in AI systems is very dynamic; it changes with time, depending on the experience of the users and the welfare of AIs. Users gain trust slowly as they observe consistent, accurate outputs. If an AI system makes even one gravely serious mistake, however, it can completely erode the slightest shred of that trust [36]. For example, a driver of an autonomous car can slowly gain confidence through multiple safe trips, yet a near-miss or a sudden malfunction could instantly shake all faith [37]. This shows that AI systems must be reliable and communicate uncertainty and risk so users can decide whether to rely on them less or more as time passes. So-called temporal trust implies that the art of maintaining trust should be viewed as an ongoing process, rather than a one-time achievement.

3.2.6. Overconfidence from “good-looking” explanations

An explanation designed for ease of visual comprehension or intuitiveness can lead the user into overconfidence to the detriment of the actual correctness of the AI. Beautiful visualizations or seemingly logical explanations put the users in a trance, so they do not even think to correctly evaluate an output [38] [39].

For instance, a radiologist might begin to overtrust the AI's classification of tumors just because the colorful heatmaps highlight some regions in the brain, notwithstanding that the model has made an error! What this means is that the aesthetic design and clarity of an interface, while paramount to understanding, must be partnered well with a clear recognition of what the limits of the AI are, or else the trust will break down.

3.3. Interpretability & Usability

Interpretability and usability lie at the core of Explainable AI (XAI); together, they ensure that AI outputs are understandable, meaningful, and actionable for the different categories of users. Interpretability allows the users to understand how decisions were made, and

usability makes sure this knowledge is practically applicable concerning the real-life decision being made [40] [21]. Considerable factors need attention:

3.3.1. Level of detail

Explanations should be commensurate with the level of knowledge or expertise of the user. Beginners or lay users need no more than simple, short explanations in plain English, emphasizing very clear cause-and-effect relations. For instance, a beginner would be told that a bank lender has rejected a bank loan application due to a high debt-to-income ratio. Experts need a lot of rich technical information to feature attributes, model reasoning, or counterfactual analysis that allow them to directly question the AI's outputs. Tailoring explanations to a user's level of expertise facilitates comprehension and acceptance [41].

Lay users → want simple cause-and-effect explanations.

Experts → want something deeper in e.g., feature attribution or counterfactual analysis.

3.3.2. Representation Format

Explanations are offered in myriad forms: text, spoken word, video presentations, static charts, visual highlights, or even counterfactual scenarios. The question one asks must outweigh the actual choice of method-utilizing a heatmap [42]or saliency map [43] in medical imaging, to illustrate the doctor's explanation visually, shows spatial locations that the region contributes considerably to a diagnosis. The right format selection guarantees that the explanation is intuitive and on par with users' cognitive preferences. [44] [24] [45].

3.3.3. Counterfactuals

Explain how changes in inputs would have altered the outcome, helping users to understand cause-and-effect relationships [44].

For example, an explanation by an AI might read: "If your income were €5,000 higher, the loan would have been approved," helping the user to focus on what counts and how it affects the determination.

3.3.4. Instance-based explanations

Instance-based explanations offer all kinds of similar examples to the current input that could make sense of why a decision is being made [44]. For example, the explanation of the negative classification for a review might say: "This review is negative because it resembles other reviews that contain the words 'boring' and 'waste'". This method utilizes analogies or prior cases to both concretize and render familiar the reasoning.

3.3.5. Actionability

Explanations should be actionable, in the sense that they spur users to do something in response or exert some influence in improving the outcome under question [46]. For example: "Bank rejection message: Debt-to-income ratio too high → reduce debt by €2,000 to improve chances". Actionable explanations provide power to users to make informed choices that improve their experiences with the system.

3.3.6. Interactivity

Contemporary users increasingly desire interactive explanations in which they pose questions of "why" and "what if" instead of merely receiving answers. This interactivity fosters the users' ability to explore the model's reasoning in detail and enhance understanding, engagement, and trust. [47].

3.3.7. Accessibility

Explanations underline accessibility for all users, including those with less capacity, both in mind and body [48].
For example, voice explanations assist blind users, while icons or visual marks support users who suffer from cognitive impairments. Accessibility guarantees that a bunch of XAI methods provide prospects to the corresponding population. [48].

3.4. Ethical & Emotional Factors

Ethical and emotional explanations in the field of explainable AI ensure that while a system is technically transparent and reliable, it should also act responsibly and should have empathy and be culturally appropriate. Those factors affect its acceptance by the users, the formation of trust, and the general feeling of well-being, mainly in sensitive domains such

as healthcare, finance, or public services. Some instances of such factors are perceived fairness, privacy concerns, emotional reassurance, cultural sensitivity, or inclusiveness.

3.4.1. Fairness perceptions

Users expect systems of AI to be free of bias or discrimination toward groups, incurring classification based on race, gender, age, or other legally protected attributes. These explanations would have ideally been rather concerned mostly or entirely with rendering the process of decision-making in AI transparent in ways that are comprehensible to the user, regardless of their possibly high degree of mathematical sophistication. It must be felt by users that the outcomes of AI are fair in the sense that no particular groups are unfairly disadvantaged by them [28]

[49] [50] [51] [52].

3.4.2. Privacy concerns

Any explanations are subject to protecting sensitive data. Interpretability should not unnecessarily reveal intimate private information, nor should it not shed light on private things concerning the lives of people or entities that have been included in the AI-making decisions. Empowered through a selective process governing what level of disclosure occurs at any instance, XAI systems maintain users' trust while respecting privacy laws and ethical principles [53] [54] [55] [51] [56] [57].

3.4.3. Emotional Reassurance

It looks at how an AI assistant using empathetic, supportive language helped with emotional well-being in someone managing hypertension, and how a reassuring tone, encouragement, and behavioral reinforcement can help reduce distress and support better outcomes. Especially for healthcare, explanations should be provided in a way that alleviates anxiety rather than creating it [58] [59] [60].

Tone matters → “Your risk is high” vs. “We detected risk factors, but here are steps you can take.” [61]

3.4.4. Cultural sensitivity

Explanations must remain in agreement with culturally accepted norms and communicative modes, which significantly differ across regions and communities. Some cultures might expect the medium to give them direct application of guidance, while others lean towards further reasoning and self-application [62] [51] [21].

For example, a medical AI operating under a collectivist culture might emphasize the well-being of the family along with individual health risks as a way of aligning the explanations with local values and expectations.

3.4.5. Inclusivity

Explanations must consider those representing minorities in society, individuals without technical background, including those with limited capacity to understand or some degree of alienation from the culture, as valid users of the system; the higher this inclusiveness, the less risk of having particular groups of users feel marginalized or put at a disadvantage while providing that AI systems are to be used by a wide array of users with diverse linguistic, educational, and cognitive backgrounds [63].

3.5. User Goals and Motivation

Knowing users' expectations for explanations is a must-know for those building systems of Explainable AI. Different people--be they engineers, researchers, professionals, or end users--bring AI to bear for different purposes. The user's motivation for requesting an explanation determines what kinds of explanations are useful, of what level of detail, and in what format. Such motivations in general are debugging, learning, justification, and decision support [64] [65].

3.5.1. Debugging

In such cases, explanations help engineers and developers confirm whether the AI model operates as intended. They rely on XAI-based tools to verify if the model reasoned in alignment with expected logic or identify where errors were entered, biases crept in, or features were weighed improperly. To illustrate: if the image classifier was misclassifying some objects, developers might consult the explanation to figure out which features the

model was focusing on. This helps in refining the model, data preprocessing, or retraining thereof to ultimately attain higher accuracy and reliability [66] [67] [25].

3.5.2. Learning

An explanation can be used for training and learning, particularly from the viewpoint of students, researchers, or domain learners. An XAI can reveal the internal logic of an AI model, allowing a user to gain insights into domain patterns or decision-making rules. For instance, medical learners at the University might assess which symptoms were weighed more heavily for the diagnosis, and on the other hand, environmental researchers might get to know which variables were more critical for the climate predictions. Hence, XAI supports conceptual understanding and very deep knowledge discovery rather than just the output of the system. [68].

3.5.3. Justifying

Traditionally, AI-assisted decisions must be justified to others—such as supervisors, clients, or courts-in professional and regulatory environments [69]. For example, a doctor may have to provide an explanation to a patient's board of directors or hospital, justifying why a particular AI-based diagnosis recommended treatment for that patient [70]. Similarly, financial analysts or judges may have to justify AI outcomes to their regulators or some other stakeholders [71]. Under such circumstances, reasons should be transparent, traceable, and defendable to common accountability standards and compliance with ethical and legal standards [69].

3.5.4. Decision support

Decision-making should be supported by explanations, especially for end-users. That is, depending on how the AI arrived at a recommendation or why, the user should be able to trust and act faster than before. For instance, if the AI suggests routes to the driver, it must immediately explain the rationale behind the recommendation in simple terms. The same goes for an AI medical diagnosis instrument in a hospital. Explanations help users understand AI outputs quickly, allowing them to act confidently [23] [72] [26].

Also, different types of explanations are needed depending on the user's needs: For example, engineers want causal or feature-importance explanations, learners might want example-based explanations, and decision-makers mainly need contrastive explanations ("Why this, not that?"). Therefore, XAI systems should be sufficiently flexible and adaptable to customize the explanation itself to meet the users' motives, such that they become relevant and useful in their respective contexts [73] [25].

3.6. User Mental Models

A user's mental model refers to the internalized description or view that users develop about how an AI system works, what knowledge it draws upon, and how its outputs should be interpreted. In this sense, mental models are more often cognitive approximations that guide user expectations, decisions, and trust rather than being necessarily actual representations of the technical system [74] [25]. In XAI, developing accurate and helpful mental models is a main explanatory goal [75] [18].

Effective mental model development arises when the explanations and interactions offered by a system allow its users to consider, test, reject, and revise their understanding of system capabilities, limitations, and methods of reasoning. In one concrete example, a clinician may use a medical decision-support system in making diagnoses based on the assumption that AI considers all patient data. However, through transparent explanations (e.g., indicating which features influenced the diagnosis), the clinician can begin aligning their mental model with the system's scope and methods of reasoning [76].

Mental models have a very strong relation to other human factors in XAI. Transparency affords the information necessary to let a user construct and update a model. Trust gets calibrated based on whether the user’s mental model matches the actual performance of the system; a mismatch of mental models is considered to give rise to either over-trust or under-trust. Cognitive alignment is the state in which the explanations of the AI and their reasoning structures fit the user's mode of thinking, so their mental model becomes coherent and actionable [25]. Correct mental models create perceptions of control and agency, which strengthen the user's confidence to convincingly direct, override, or modify the AI outputs [75].

In a nutshell, mental models represent the interface between the transparency of the system and human decision-making. They enter as to whether users can genuinely interpret, trust, and act meaningfully on AI recommendations. Therefore, XAI should be designed not only to explain but to interactively forge proper and robust mental models through collaboration, feedback, and stepwise information disclosure [74].

3.7. Control and Agency

Control and agency relate to the extent to which the users can question, direct, or override AI systems and feel empowered in doing so. Explaining must not be solely about teaching users about how the AI works but must also reinforce the user's autonomy in their interactions. If users feel that they can actively influence the decisions made by the AI, either by choosing the parameters of the decisions, requesting alternative explanations, or rejecting the outputs, then the users actively develop trust in the AI [77].

For instance, I would suggest that commercial aircraft pilots who are utilizing automated flight control systems should be able to appreciate and understand why the AI or autopilot has recommended a certain maneuver, and being at ease with their own judgment, should have the confidence to take, change, or reject it. Simply put, explanations are converted into control-enabling tools that users can use to set specific trust and intervention thresholds or limits. The user must feel in control to keep them from the so-called ‘automation complacency’ or ‘learned helplessness’, the point at which the users unconditionally welcome automated suggestions. As opposed to this, if there is too much resistance to the user or the system is totally opaque, there is less user interaction, and confidence is seriously affected. Consequently, the design of XAI should aim at placing the AI and its human operators in a setting of shared agency—juxtaposing each other’s watchdog and accountable roles during collaborative decision-making [78] [79].

3.8. Conclusion

Human factors are the primary influence in drawing the boundaries of Explainable Artificial Intelligence's (XAI) effectiveness, credibility, and real-world adoption. As demonstrated throughout this chapter, explanations are not just performing a technical artefact but also a process of social-cognitive interactions. They must consider humans'

reasoning, cognitive limitations, emotional responses, and ethical standards. An understanding of the human aspect is critical for even the most sophisticated explainability techniques not to be of any use to the user in terms of their comprehension or building of trust.

An explanation from a cognitive standpoint should reduce mental workload, enable the creation of a correct mental model, and correspond to the user's level of expertise. From the psychological viewpoint, XAI progressively influences the calibration of trust, perceived reliability, and the user's propensity to rely on AI over time. Emotional and social aspects also determine the users' attitudes; the explanations must reassure users during the uncertainty, reduce fear or confusion, and demonstrate fairness and accountability—this is particularly true in cases involving significant decisions.

The usability and interpretability of explanations are equally important. Explanations should be actionable, concise, and a part of the users' workflow. Ethical considerations —such as transparency, privacy, fairness, and user autonomy — define not only what can be explained but also how explanations should be given, considering the communication responsibility. In addition, human factors determine how far trust can be restored in case of a system failure, they underscore how much cultural and contextual adaptability is needed, and they highlight leaving the users empowered by explanations and not taking them into a state of being overwhelmed.

To conclude, the inclusion of human factors in the development of XAI is the only way to finally bridge the gap between algorithmic transparency and meaningful understanding. XAI through human cognition, psychology, and situational context can not only enhance the quality of decisions but also establish the right trust and make AI acceptance easy. With this vision of XAI being human-centered, there is already a great concern that the power of explanation will be misused: to allow the public to engage critically with AI, to be in the position of control, and to receive information that is well-informed and trustworthy as AI in the daily lives of people.

4. Task Context Factors in XAI

Explainable Artificial Intelligence (XAI) is a broad concept that operates on different levels and for various purposes across different domains and in different environmental conditions. It cannot be a one-size-fits-all approach for designing or interpreting explanations. The task context is the major factor affecting an explanation's effectiveness. Thus, it refers to the specific environment, objectives, and limitations of the AI system's operation. The task context determines not only the type of explanation that is to be provided, but also how it is structured, how much detail it is to contain, and the way it is to be communicated to the user. In safety-critical situations such as healthcare, finance, or autonomous systems, explanations must be precise, clear, and justifiable, which in turn fosters trust and accountability. On the other hand, in contexts where the risk is lower, such as entertainment or e-commerce, explanations are mainly for supporting the personalization, adaptation, engagement, and usability of the system rather than justifying the system formally.

In this section, we present the core, extended, and emerging factors that together constitute the task context. These factors are the ones that determine the explainability of AI systems. Domains, risk levels, user goals, evaluation criteria, constraints, and interaction modes are among the decisive characteristics that determine the appropriate and profound form of explanation for a particular case, and how they play their respective roles.

Table 3 - Task Context Factors in XAI

Main Category	Subcategories
Domain / Application Area	Healthcare, Finance, Education, Entertainment, Legal/Regulatory, Autonomous Systems
Decision Stakes / Risk Level	High, Medium, Low
Intended User Goal	Decision support, Learning/Training, Justification/Accountability, Entertainment/Engagement
Evaluation Criteria for Explanations	Accuracy/Correctness, Fairness/Bias, Safety/Risk, Usability/Understandability, Trust/Reliance
Contextual Constraints	Time Pressure, Level of Expertise, Regulatory/Organizational Requirements
Interaction Mode	Passive Explanations, Interactive Explanations
Extended Factors	Data Characteristics/Availability, Temporal Constraints, Collaborative Context, Physical/Device Constraints
Possible Emerging Contexts	Environmental/Situational Context, Human–AI Teaming Context

4.1. Domain / Application Area

The field or application area refers to the specific sector in which an AI solution is applied. Each domain has its own requirements for knowledge, users' expectations, and consequences of AI-assisted decisions. The development of such systems that are understandable to the users, meet their needs as well as follow ethical standards, is very much dependent on identifying such differences.

4.1.1. Healthcare

One of the major AI applications in healthcare is the automated **medical diagnosis** and **treatment recommendation** systems, where trust, accuracy, and transparency are the three most important values. The provided explanations have to be of a medical nature in order to assist the clinical decision-making process and ensure patient safety [80] [81] [28] [82] [40] [83] [26] [80].

4.1.2. Finance

The AI financial domain has various application areas such as **loan approvals, fraud detection, trading, and running e-companies systems**. To prevent unfair deals and ensure compliance with regulations, explanations are requested, which will consequently develop the user's trust in the financial decisions made [84] [39] [40] [51] [26].

4.1.3. Education

As a main concern, the educational sector strives to achieve personalization in learning and grading support, which then leads to the adoption of appropriate teaching methods depending on each student's individual needs. The sector of explainability assures the transparency of the performance evaluation process and thus maintains trust among students, teachers, and institutions [64] [40] [85] [86] [26].

4.1.4. Entertainment

In the entertainment industry, AI is the driving force behind the **recommendation systems for music, movies, games, and social networks**. By explaining the reasons for suggesting certain content to users, the tech further increases user satisfaction and engagement [26].

4.1.5. Legal / Regulatory

In the legal area, AI tools help with the support of judicial decision-making and regulatory analysis. The explainable models play a crucial role here to ensure accountability, fairness, and compliance with legal norms [28].

4.1.6. Autonomous systems

The term "autonomous systems" incorporates a wide variety of technologies, such as **self-driving cars, drones, robotic** applications in various industries, **cybersecurity**, **IoT security,** and **intrusion detection**, and even **autonomous spacecraft**. These systems are all devised to operate through a very low level of human involvement, in the often changing and unpredictable environments, thus the demand for explainability and transparency arises as a crucial means for providing the safety, accountability, and trust of humans. When the system provides an explanation, the users and operators are informed about the behavior of the system, they can also identify errors and check whether the autonomous decisions are in accordance with the ethical and operational standards [40] [87] [88].

4.2. Decision Stakes / Risk Level

The decision stakes or risk level corresponds to the extent of criticality and potential impact an AI system's decision has on the different parties involved or users. In case of high-risk situations, the need for transparent, meticulous, and justifiable explanations becomes disproportionately large. On the other hand, low-risk conditions might be served with more simplified explanations without impairing user comprehension or contentment. Therefore, the depth, detail, and rigor of explanations should be in accordance with the risk level attached to the decision-making task [89] [90].

4.2.1. High stakes

High-stakes applications include life-critical, financial-critical, or legal-critical decisions, and they are found in autonomous driving, medical diagnosis, auditing, or high-impact recommender systems. These fields necessitate rigorous, evidence-based explanations that not only can be challenged but also support accountability, safety, and ethical integrity [20] [91].

4.2.2. Medium stakes

Medium-stakes decisions are usually operational or business-related, in which the results affect efficiency, productivity, or organizational performance. In such situations, the explanation should convey justification of outcomes, clarity of reasoning, and support for informed human oversight. [92]

4.2.3. Low stakes

Low-stakes decisions are those related to personal preferences or leisure activities, for instance, entertainment or casual recommendations. In this context, the use of simple and user-friendly explanations centered on customer satisfaction and personalization is adequate to keep the user's engagement and trust without overburdening them [90] .

4.3. Intended User Goal

The intended user goal is the purpose that the AI system's explanation serves within a certain task or domain. It specifies the target of the explanation - such as decision support, learning facilitation, or accountability provision. This is a system-centered approach, which emphasizes the role of explanations in the context of AI's function or use case. On the other hand, the human factors viewpoint emphasizes the need for an explanation by the user, thus mirroring their inherent motivations, moods, and intellectual demands - such as trust, understanding, or curiosity. Accordingly, the intended user goal describes the operational target of the explanation in the system's context while human factors depict the mental triggers that cause users to desire or appreciate that explanation [10] [23].

4.3.1. Decision support

In decision support situations, the explanations given by AI are meant to make it possible for the users to act on the AI output with full confidence. For instance, in medical or financial fields, explanations unlock the logic behind predictions or recommendations so that the users can make their decisions informed and reliable, thus lessening the uncertainty and risk that they are facing [23] [4].

4.3.2. Learning / training

On the other hand, in the case of learning or training, the explanations are concentrated on showing the domain rules, patterns, or reasoning strategies that the AI is using. This enables

users, such as students or trainees, to comprehend how the AI processes information, leading to the decisions, thus supporting knowledge acquisition and skill development [10] [79]

4.3.3. Justification / accountability

In the contexts of justification or accountability, the explanations are to be tendered to the external stakeholders, like auditors or regulators, to report, defend, or document the decisions. In this scenario, the explanations have to be clear, easily followed, and laid out in a way that supports compliance, ethical standards, and accountability [93] [94].

4.3.4. Entertainment / engagement

The explanations, depending on the case, are there to inform, to satisfy one's curiosity, or to improve the overall user experience. In recommendation systems used in music, movies, or games, the explanations improve the user's experience by clarifying the reasons for the suggestions and thus making the interactions feel more important [95].

4.4. Evaluation Criteria for Explanations

The metrics for deciding whether an explanation is " good enough". Different criteria are relevant depending on the context; for example, safety and correctness are crucial in autonomous cars, while usability may dominate consumer apps [96]. These factors accompany the designers in preparing explanations that are effective and at the same time proper for the context in which they are applied.

4.4.1. Accuracy / correctness

The explanations need to be a correct representation of the AI system's thinking and conclusions. In fact, it would be impossible for the users to rely on the explanations and thus to make informed decisions if the AI's reasoning weren't sound, particularly in high-stakes areas such as finance, healthcare, or autonomous systems. [97]

4.4.2. Fairness / bias

The expose should be known as well as help alleviate the possible biases that might be inherent in AI's decision-making. Fairness assessment would ensure that the system does not activate discriminatory or unjust outcomes, which is particularly important in areas like finance, hiring, or law [98] [51] [57] [96] [99] [100] [101].

4.4.3. Safety / risk

The requirement of safety-critical applications, on the other hand, is that the explanations must help in the risk assessment and hazard mitigation process. Just to mention a couple of instances, autonomous vehicles and medical devices demand explanations that will enable the users to detect possible errors and, thus, to prevent unsafe actions [102] [103] [96] [99].

4.4.4. Usability / understandability

The explanations should always be user-centered, meaning that they should be easy to understand and practical for the targeted user. High usability guarantees that the users will be able to interpret the AI's outputs correctly and easily, no matter if it is for educational tools, consumer apps, or workplace systems [64] [104] [52] [96] [99] [101].

4.4.5. Trust / reliance

The explanations should be such as to generate a reasonable amount of trust in and reliance on the AI system. The users need to determine when to place their trust in the output produced by the system and when to be cautious; thus, they must find a balance between confidence and critical oversight. This is a necessity in every application, be it finance, healthcare, or daily consumer ones. [84] [105] [103] [40] [51] [99] [56] [57] [52] [101] [26] [106].

4.5. Contextual Constraints

Contextual constraints are the factors that include the environment, the situation, and the people involved. They underline every step of the explanation process, from generation to reception and understanding. Depending on the technical context of the system, these factors set the limits for the explanations in their shape, timing, and depth. For instance, in the case of an autonomous car, a prompt, clear explanation is required to back up its real-time decision-making, while a research or auditing system might author a long, elaborate explanation. Thus, these limitations will dictate the mode, complexity, and channel of the AI systems' reasoning communication [107].

4.5.1. Time pressure

The time given to an explanation gives a very strong indication of what kind of explanation it is. In the case of real-time settings, like driverless vehicles and unmanned aerial vehicles,

the nature of the explanation accompanying the quick response must be immediate, brief, and very easy to interpret. This is not the case in the offline or analytical contexts, for instance, research or auditing, where the explanations are allowed to be more detailed and full of information and can be reviewed without any pressure [24] [108].

4.5.2. Level of expertise

The target user's level of expertise is what decides the complexity or simplicity of the explanation. In the working environment, the emphasis is put on being able to operate the AI system when it comes to customizing the structure and technical level of the explanations according to the competence expected, such as providing high-level overviews for beginners and detailed insights for advanced users. The human factors angle, however, extends the whole notion beyond just the user aspect, taking into consideration his or her background cognitive style and interpretive capacity as well [13].

4.5.3. Regulatory / organizational requirements

In some disciplines, the restrictions imposed by legal, ethical, or organizational matters limit the content and structure of the explanations. A legal obligation might, for instance, require that certain details be given or that audit trials be maintained in financial or healthcare systems for the sake of accountability and transparency. These requirements not only keep the user’s needs but also quality standards as per the governing bodies’ regulations under consideration when it comes to providing explanations [109].

4.6. Interaction Mode

The user experience mode pertains to the way users comprehend and interact with a system's explanations. It decides, from the task context point of view, the delivery format, the level of interactivity, and the communication channel through which explanations are delivered. The interaction mode is a major factor in determining the extent of understanding, trust, and use of AI outputs by users [110]. Different modes fit into different environments; for example, static dashboards are great for monitoring, while interactive interfaces are supportive for exploring and learning.

4.6.1. Passive explanations

Users in passive explanation modes get the information without any alterations or questions asked; they only see predefined and static information, like reports, dashboards, or feature summaries. These explanations are appropriate for situations where users need quick insights without any further questioning or manipulation, such as performance monitoring or compliance auditing, since they do not allow for over-the-top interrogation [110].

4.6.2. Interactive explanation

Users in interactive explanation modes can become active participants in the system's reasoning process through a variety of methods. Among these may be "why" or "what-if" questioning, counterfactual reasoning, or the application of layered drill-down methods whereby users gradually expose deeper levels of understanding. Interactive explanations are the ones that are very hard to lose, even in the case of complex or dynamic decision-making processes, and they have the power to engage the users, educate them, and win their trust [110] [111].

4.7. Extended Factors

While the six core Task Context Factors in XAI (Domain, Risk Level, User Goal, Evaluation Criteria, Constraints, and Interaction Mode) are commonly seen as the main basis, some research adds extra situational or environmental dimensions to them, which further specify how explanations should be made. The following factors are mentioned as additional considerations extending the core Task Context dimensions in XAI.

4.7.1. Data characteristics / Availability

The AI system's data quality, diversity, and quantity have a significant impact on the explanation's character. For example, models that are trained with limited or biased datasets need to provide explanations that clearly state uncertainty or limitations, along with communicating uncertainty. On the other hand, rich multimodal data-trained systems are entitled to give more superior and evidence-based justifications. Therefore, the data context plays the role of deciding the depth and trustworthiness of the explanations [112] [113].

4.7.2. Temporal constraints

Aside from immediate time pressure, there are cases when AI systems run over long periods, such as in continuous monitoring or adaptive learning environments. In these situations, explanations must change with time, and thus, reflection of changing conditions or updated knowledge will be the case. An instance can be given of a healthcare AI that monitors the patient's progress; the AI will have to keep changing its explanations as new data is fed in or as treatment objectives shift [91].

4.7.3. Collaborative context

Explanations might have to cover several stakeholders at once, for example, in the case of an AI decision support tool used by doctors, nurses, and administrators. Explanations in such cases must try to find a middle ground among the different perspectives, the different communication styles, and the different responsibilities. Collaborative contexts not only emphasize the need for a shared understanding but also for role-sensitive explanations instead of one-size-fits-all outputs [114].

4.7.4. Physical / Device constraints

Explanation design is also defined by the hardware or user interface environment. Mobile, wearable, or embedded systems (like IoT devices) require concise, visual, or voice-based explanations due to limited screen space or bandwidth. On the other hand, desktop or web-based systems can deliver detailed, layered, or interactive explanations without any restrictions [115] [116]. Device constraints relate to the technical medium of explanation delivery, while environmental context refers to the outside situations that affect the understanding of the explanation.

4.8. Possible Emerging

A few new factors have come to the forefront and are being discussed by recent studies. These factors do not yet have the status of standard ones, but they add value to the analysis:

4.8.1. Environmental / Situational Context

Context is a broad term that refers to the external conditions under which users interact with AI systems and the resulting explanations. Ambient noise, lighting, the availability of devices, multitasking demands, and even stress levels are among the environmental factors

that can strongly affect the user's understanding of explanations. For example, auditory explanations may work best in situations where users are hands-free or using their mobiles, while visual explanations are appropriate in very quiet and highly focused ways, like control rooms or offices. In addition, when the circumstances are very urgent and the level of stress is high, the explanations must be very brief, actionable, and very easy to understand to support quick decision-making. The consideration of environmental and situational factors ensures that explanations remain practical and usable in real-world contexts, even if they are very complex from a technical point of view. Physical/device constraints are the technical possession of how explanations can be delivered, while environmental/situational context is the state of how explanations should be adapted to human and real-world conditions [117].

4.8.2. Human–AI Teaming Context

The Human–AI Teaming Context, which the AI systems, having been to the frontiers of the modern world with humans, is a new contextual dimension to be considered. The Human–AI Teaming Context is a task-context factor in Explainable AI (XAI) that is particularly significant in collaborative scenarios where human and AI systems are dependent on each other. In such environments, explanations are used not only to make the AI's decisions understandable but also to support the building of mutual understanding, coordination, and trust between human team members and AI agents.

Good explanations in this case are necessary for the synchronization of human and AI mental models so that both can correctly interpret and respond to each other's actions. Health care teams or industrial robotics, for example, providing explanations wherever needed that are clear and relevant to the context, helps keep the situation awareness shared, decision-making processes supported, and performance of the team overall enhanced. Thus, the AI explanations must be designed in such a way that they consider the interplay of human-AI collaboration that demands transparency, versatility, and the ability to facilitate interactive dialogue between the members of the team and the AI systems [118].

4.9. Conclusion

The task context is the foundation upon which significant and context-aware XAI can be constructed. The six factors—domain, risk level, user goal, evaluation criteria, contextual constraints, and interaction mode - work together to determine the relevance, depth, and style

of AI's explanations. Further, extended and emerging dimensions such as data availability, temporal continuity, collaborative settings, physical or device limitations, environmental conditions, and human–AI teaming, among others, contribute to the understanding of these factors. Thus, the dimensions of explainability highlight that it is neither a static nor a universal attribute, but rather a flexible one that must conform to the specific requirements of the field, user expectations, and situational demands. Recognition and incorporation of these multifarious contextual factors ultimately ensure that AI justifications are always practical, ethical, and efficient in the various real-world situations—thus, closing the gap between technical clarity and human comprehension.

5. Model Type Methods in XAI

Explainable The methods of Explainable Artificial Intelligence (XAI) can be categorized not just by their approach, i.e., model type (ante-hoc or intrinsic) vs. model-agnostic (post-hoc), but by explanation level as well, which defines the area and the depth of interpretability.

The three main levels of explanation are:

- Feature-level: identification of individual input features' significance.
- Instance-level: justifications for specific predictions or samples.
- Model-level: overall characterization of the model's operations and behaviors.

The study emphasizes the approach-based classification of XAI into Model-Based and Model-Agnostic methods. However, it is further stated that the levels of explanation referred to above are relevant for both groups and that there is a high probability of their overlapping in practical systems.

The intrinsic or ante-hoc Model-based XAI methods add interpretability to the model's structure or learning process. Their transparency is inherent, as it is a direct result of the way the model represents and processes the information. Therefore, there is no need for external explanatory tools to complement the processing. Some of the examples are linear and logistic regression, generalized linear and additive models, decision trees, rule-based systems, fuzzy logic models, and even more advanced interpretable neural architectures such as Explainable Neural Networks (XNNs), Concept Bottleneck Models (CBMs), and Neural Additive Models (NAMs).

The model-agnostic XAI methods (post-hoc approaches), on the other hand, are designed to offer explanations for already-trained models, regardless of their complexity or architecture. These methods function as separate interpretation layers that analyze inputs and outputs for approximating the reasoning of a “black-box” model. LIME, SHAP, Anchors, counterfactuals, and perturbation-, decomposition-, and gradient-based methods are among the most prominent examples.

Both approaches share the same basic objectives—enhancing the clarity, accountability, and trustworthiness of the model—but they differ in timing and the manner of incorporating explainability into the modeling procedure.

Table 4 - Model Type Methods in XAI

Main Category	Subcategories	
Model-based (Intrinsic or Ante-hoc)	Intrinsically Interpretable Models	Linear Regression, Logistic Regression, Generalized Linear Models, Generalized Additive Models, Decision Trees, Rule-Based Models, Fuzzy Logic Systems, Explainable Boosting Machines, k-Nearest Neighbors, Naïve Bayes, Bayesian Networks, Adaptive Resonance Theory, Prototype-Based Models
	Explainable Complex Models	Attention Mechanisms, Explainable Neural Networks, Self-Explaining Neural Networks, Interpretable Convolutional Networks, Concept Bottleneck Models, Neural Additive Models, Graph Attention Networks, Graph Neural Networks, Neural Casual Models, Causal Discovery Models, Interpretable Graph Models with Symbolic, Cognitive Neural Models
Model-agnostic (Post-hoc)	Local Explanation	LIME, SHAP, Anchors, Counterfactual Explanations
	Global Explanation	Partial Dependence Plots, Individual Conditional Explanation, Feature Importance, Permutation Importance, Surrogate Models, Global Surrogate Models
	Gradient-based and Relevance-based Explanations	Saliency Maps, Grad-CAM, Integrated Gradients, Decomposition-based Methods, Layer-wise Relevance Propagation, DeepLIFT,
	Perturbation-based Methods	

5.1. Model-based (Intrinsic or Ante-hoc)

These ways of explaining them are inherent to the models since the models are themselves designed to be open and interpretable. One does not need any further assistance to get the logic behind predictions.

5.1.1. Intrinsically Interpretable Models

They can make reasoning very clear. Their whole thing is to apply basic arithmetic to input features versus outputs.

5.1.1.1. Linear Regression

It sets up a direct relationship between the input variables (features) and the continuous output (target) [119] [120].
The simplest way of writing it mathematically is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

where:

- $\hat{y}$= predicted value (output)
- x_i= input feature
- β_i= coefficient (shows importance and direction of influence)
- β_0= intercept (baseline value when all $x_i = 0$)
- ϵ = error term or residual

Each coefficient (β_i) indicates the effect and the direction of that effect of the feature on the prediction, hence making the model's reasoning quite clear. The epsilon (ϵ) is the error we still make, i.e., the difference between the prediction and the actual outcome. These errors are assumed to follow a Gaussian distribution, which means that we make errors in both negative and positive directions and make many small errors and a few large errors [120] [120].

This implies that there will be no problem in deriving explanations both in a global and local way: globally via the whole model structure and coefficients, and locally by breaking down individual predictions into feature-level contributions. Hence, Linear Regression can give a

very clear, step-by-step explanation that is congruent with human reasoning and supports decision-making in fields that require accountability and trust like healthcare or finance [119] [120].

From the point of view of XAI, Linear Regression is regarded as a benchmark for model interpretability on account of its being simple and feature weighting being very explicit. However, its linear character restricts its ability to model complex, non-linear interactions among variables, which, on the other hand, may result in the loss of predictive power in these challenging real-world situations. Nonetheless, it still plays an important role in the XAI area as a transparent single model as well as a surrogate model to reveal the workings of non-transparent systems such as deep neural networks [119].

For example, in healthcare:

$$\text{Risk Score} = 0.8 \times \text{Blood Pressure} + 1.2 \times \text{Cholesterol} - 0.5 \times \text{Exercise Level}$$

Doctors can see undeniably what each factor adds up to the patient's predicted risk, satisfying the requirements of transparency and trusting in the clinical decision support systems.

Pros: Simple; transparent; interpretable; fast; baseline-friendly
Cons: Linear assumption; sensitive to outliers; poor for complex patterns; affected by multicollinearity

5.1.1.2. Logistic Regression

Logistic Regression is typically associated with binary classification. It assigns a probability of membership in a class to an input sample through a logistic (sigmoid) function applied to a linear combination of parameters [119]:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

maps any linear combination of the input features to a value between 0 and 1, using the sigmoid (logistic) function.

Here's what each part means:

- β_0: the intercept, representing the baseline log-odds when all features are zero.

- β_i : the weight (coefficient) of feature x_i , showing how strongly that feature influences the prediction.
- $e^{-(\cdot)}$: the exponential transformation that converts the linear output into a probability.

The coefficient (β_i) of the feature indicates the direction (positive or negative) and the strength of the influence of that feature on the predicted probability. This direct and interpretable mapping allows the users to really have a good idea as to what individual features contribute to the model's decision-making process. This transparency is what makes Logistic Regression a preferred choice for high-stakes applications—like diagnoses in healthcare, determining credit risk, or making policy decisions—where interpretability and trustworthiness are vital [119].

Logistic Regression from an XAI point of view offers both global and local explanations without post-hoc methods being involved. In the global, the model indicates the way the total outcome is affected by each feature in the dinosaur population of all samples, while in the local it will identify the reason why a particular instance was put in a specific category. Coefficient plots or feature contribution graphs are visualization techniques that can further improve the interpretation of the model. However, like Linear Regression, Logistic Regression is also restricted by its linear boundary of decision making and cannot very well identify complex non-linear patterns. Nevertheless, it still is one of the most transparent and hence explainable models and very often used as a benchmark for assessing the interpretability of more complicated machine earning methods [119] [120].

Pros: Simple; transparent; interpretable; backed by mathematics, probabilistic.
Cons: Can only represent linear relationships and perform poorly when the number of dimensions is very high or the problem is nonlinear in nature [121].

5.1.1.3. Generalized Linear Models (GLMs)

The extension is made to linear regression to allow dependent variables to have distributions (Gaussian, binomial, Poisson, etc.). GLMs use a link function (such as logit, log, or identity) that relates linear predictors, which are a weighted sum of features, to the expected value of the response.

Example:
Logistic regression is a GLM with a logit link and a binomial distribution.
Poisson regression is a GLM for count data with a log link.
The interpretability stems from the coefficients, just as it does in linear/logistic regression, but it is in the context of the link function that has been chosen [122].

Mathematically, a GAM can be represented as:

$$g(E[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + B_p(X_p)$$

Where:

- $f_i(X_i)$*= smooth function (e.g., splines) of predictor* X_i
- $g(\cdot)$*= link function*
- *The contributions are* ***additive****, making the model interpretable in XAI.*

where $f_i(X_i)$*is a smooth function estimated from data, and* g*is a link function (e.g., logit, identity, log).*

Pros: Flexible; interpretable; versatile.
Cons: Complex; link-dependent; Linearity assumption.

5.1.1.4. Generalized Additive Models (GAMs)

Generalized Additive Models (GAMs) belong to the category of XAI methods that are based on the model and that build upon GLM by representing the connection between the input variables and the output as a smooth, possibly non-linear function instead of simply a linear coefficient or coefficients [123].
The prediction is calculated in a GAM by adding up all the contributions of the smooth functions of the respective features, thus making the model additive and interpretable. This implies that the effect of each variable can be depicted separately, thus indicating the manner and extent to which it influences the prediction, all while maintaining the model's structure as transparent and understandable [124].

Mathematically, a GAM can be represented as:

$$g(E[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

Where:

- $f_i(X_i)$= *smooth function (e.g., splines) of predictor* X_i
- $g(\cdot)$= *link function*
- *The contributions are additive, making the model interpretable in XAI.*

Due to the additive property, GAMs can create visual explanations, where the contribution of each feature can be displayed, letting the users know not only if it is increasing or decreasing the prediction but also by how much. The characteristics turn out to be a great advantage of GAMs in healthcare, risk assessment, and policy modeling because these applications require not only predictive accuracy but also transparency and interpretability [123] [125] [126].
On the other hand, the situation changes when the feature interactions are introduced (for instance, via pairwise or higher-order terms), as then the interpretability becomes more difficult, because the additive assumption is no longer strictly applicable [119].

Pros: They are more flexible than GLMs; but still interpretable, more expressive than linear models; visually explainable.
Cons: When interaction with features is added; the interpretation becomes more difficult.

5.1.1.5. Decision Trees

Linear regression and logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other. Time to shine for the decision tree! Tree-based models split the data multiple times according to certain cutoff values in the features. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called terminal or leaf nodes, and the intermediate subsets are called internal nodes or split nodes. To predict the outcome in each leaf node, the average outcome of the training data in this node is used. Trees can be used for classification and regression [119].

Decision Trees are very easy to understand because they use a set of human-readable rules (if - then - else conditions) to do their job. A way from root to leaf is a decision rule that makes it easy to understand the reason for the prediction. They are particularly helpful in areas like medicine and banking, where one needs to understand the reasoning behind the predictions. In other words, A Decision Tree is a model that is structured like a tree. It has a

hierarchy made up of branches and nodes. Every internal node reflects a decision condition that is based on a feature (e.g., age < 35), and every branch portrays the result of that condition (true or false). This process goes on until a leaf node is reached, which stands for the last prediction or decision of the model [127].Consequently, it can be said that every path from the root to a leaf corresponds to a logical "if–then" rule, hence the model being inherently interpretable.

Decision trees can be constructed using algorithms like ID3 (Iterative Dichotomiser 3) [128], C4.5 [129], or CART (Classification and Regression Trees) [119]. These algorithms recursively partition data into subsets that are as like each other regarding the target variable as possible. Usually, the divisions depend on criteria such as information gain, Gini impurity, or variance reduction, which vary with the type of problem [130].

From the XAI perspective, the interpretability of decision trees is supported by their explicit structure and rule-based reasoning. A tree's path represents a decision rule that can easily be expressed in everyday language. Thus, it is possible for the user to understand the reasons and way the model has planned without needing any additional post-hoc explanation techniques.
Random Forests and Gradient Boosted Trees are among the ensembles that have been developed to address the limitations of scalability and performance. Even though these ensemble methods are more precise, they still lose a considerable portion of the interpretability associated with a single tree. Consequently, researchers are attempting to lessen the impact of this trade-off by utilizing surrogate trees—simplified tree forms that replicate the predictions of complex black-box models while preserving interpretability [119].

Pros: Transparency; understandability; ease of visualization.
Cons: They might overfit; in the case of very large trees, they might become complicated and lose their interpretability.

5.1.1.6. Rule-Based Models

Rule-Based Models constitute another variant of AI that is explainable by default, where the reasoning process is represented in the form of explicit sets of if–then rules. In contrast to

Decision Trees, models based on rules are flat and non-hierarchical, which means that every rule works separately. These models are created to imitate human-like logical reasoning and often can be characterized by the natural language they are.
Since every rule clearly shows part of the model's logic, Rule-Based Models are, by their very nature, transparent. They are particularly useful in extremely sensitive fields like healthcare or law, where interpretability, trust, and expert validation are crucial [123].
Rule-based XAI models can generally be classified into two types:

Rule Lists (Ordered)

Models that make predictions using explicit IF–THEN rules.
Rule Lists are rules ordered in sequence for the model to check in accordance with the order of appearance [119] [131] [132].
Sample Rule List:
If male and adult, then the survival probability is 21% (19%–23%)
else if 3rd class then survival probability 44% (38%–51%)
else if 1st class then survival probability 96% (92%–99%)
else survival probability 88% (82%–94%).

For example:
Bayesian Rule Lists (BRL): The method employs Bayesian inference to produce sorted lists of probabilistic rules, giving the confidence level of each rule as a measure [119] [131] [123].

Rule Sets (Unordered)

Several rules may apply, and therefore, they are commonly combined with voting methods or probabilistic methods.
Sample Rule Set: Several rules would be applied simultaneously, and the system is designed to aggregate them [131] [132] [123].

Examples of Rule Sets XAI Models:
Anchors: This approach develops very accurate if–then rules ("anchors") that are enough to give a local explanation for a prediction [123].
For instance: IF income > 50k AND credit history = good → loan approved

RuleFit: This method extracts rules from Tree ensembles (e.g., Random Forest) and combines them into a linear model. Thus, the model can achieve high accuracy while intelligibly presenting the reasoning behind its decisions [119] [133] [123].

RIPPER: This is a rule-based XAI algorithm that provides a set of if–then rules to classify data. It gradually builds rules, cuts them down to avoid overfitting, and finally puts them together in a short, easy-to-read rule list. Unlike hierarchical decision trees, the structure is flat and linear. Explicit rules of RIPPER not only help in predicting but also in validating the predictions, which are especially helpful in transparency-requiring fields like healthcare and finance [123].

Pros: High interpretability; transparency and accountability; ease of validation; faithful explanations; flexibility (ordered or unordered)
Cons: Can grow complex with many rules; harder to capture interactions; may trade accuracy for interpretability.

5.1.1.7. Fuzzy Logic Systems (FIS, Neuro-Fuzzy Models)

Fuzzy Logic Systems (FIS) simulate human reasoning. They use linguistic variables, fuzzy sets, and if–then rules. Instead of applying the conventional binary logic of true/false, the fuzzy logic systems deal with the truth degrees between 0 and 1, which is a way of reasoning under uncertainty [80]. The 3 basic components form the structure of the model:

- Fuzzification: translates exact numerical inputs into fuzzy values using membership functions.
- Inference engine takes human-readable if–then rules (for instance, “If temperature is high and pressure is low, then risk is medium”).
- Defuzzification: transforms fuzzy outputs into a clear decision or prediction.

Because each rule and membership function is transparent and interpretable, Fuzzy Logic Systems are considered intrinsically interpretable models within Model-based XAI. Their rule-based and linguistic nature makes them especially effective for applications requiring human understanding and trust — such as healthcare diagnostics, energy management, and decision-support systems [80].

Neuro-Fuzzy Models, for example, the Adaptive Neuro-Fuzzy Inference System (ANFIS), take FIS to a higher level by using neural networks to automatically learn the parameters of membership functions and rules from data while keeping the interpretability. This mixing of the two allows for an adaptive learning process without sacrificing the explainable, rule-based framework of fuzzy systems [80].

Pros: Interpretable by design with the help of linguistic rules; uncertainty and imprecision are dealt with effectively; learning mechanisms in neuro-fuzzy systems make them adaptable.
Cons: Requires rules set by experts or careful tuning; interpretability can decrease as rule complexity grows.

5.1.1.8. Explainable Boosting Machines (EBM)

Explainable Boosting Machines (EBMs) form a modern set of interpretable, model-based XAI methods that are based on the Generalized Additive Models (GAMs) principles. They are an advanced version of GAMs, as through boosted decision trees, each feature's contribution is learned rather than using simple smooth functions. This is what allows EBMs to model complex, nonlinear interactions between features and outcomes while still being additive and interpretable by humans [126].
In EBMs, the prediction is computed as the sum of individual feature effects (and sometimes low-order interactions)

$$\hat{y} = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_n(X_n) + f_{ij}(X_i, X_j)$$

Here, $f_i(X_i)$ explains the learned shape function for each feature and fij(Xi,Xj) denotes a small number of pairwise interactions that are still interpretable.

The effect of each feature can be visualized directly, so the user can tell how a certain variable change will affect the prediction, regardless of the other variables. The clarity that comes with EBMs makes them very effective in high-stakes domains like healthcare, finance, and policymaking, where the need for accountability and fairness is critical [126].

Pros: Highly interpretable, visualizable feature effects, nonlinear relationships capturing, and often competing with black-box models regarding performance.

Cons: Large datasets cost more in terms of computation, interpretable interactions are limited to low-order ones, and longer training compared to linear models.

5.1.1.9. k-Nearest Neighbors (k-NN)

The k-nearest neighbor method can be used for regression and classification and uses the nearest neighbors of a data point for prediction. For classification, the k-nearest neighbor method assigns the most common class of the nearest neighbors of an instance. For regression, it takes the average of the outcomes of the neighbors. The tricky parts are finding the right k and deciding how to measure the distance between instances, which ultimately defines the neighborhood [119].

It estimates the outcome of a new input by finding the k closest training samples in the feature space and combining their labels (for example, majority voting for classification or averaging for regression) [119] [134] [135].

Working process of k-NN an XAI perspective:

1- Calculating the distance between the new sample and all training points using common metrics such as Euclidean or Manhattan
2- Choosing the nearest neighbors
3- Guessing depending on the neighbors:
    a. Classification: majority labels' vote
    b. Regression: mean (weighted mean) of neighbors' values

Interpretability in XAI:

Local explainability: The outcome of each prediction can be interpreted via the k-earest neighbors, which mainly affects its result.

For instance, "This patient is classified as high-risk because 7 out of the 10 most similar patients had high-risk outcomes."

Pros: Locally easy to understand; intuitive; no training model complexity.

Cons: It is harder to interpret global behavior; the performance is greatly affected by the choice of k and distance metric; large datasets make it computationally demanding.

5.1.1.10. Naïve Bayes

The Naive Bayes classifier uses Bayes' theorem of conditional probabilities. For each feature, it calculates the probability for a class depending on the value of the feature. The Naive Bayes classifier calculates the class probability for each feature independently, which is equivalent to a strong (= naive) assumption of independence of the features [119]. Naive Bayes is a conditional probability model and models the probability of a class *Ck* as follows:

$$P(C_k \mid X) = \frac{1}{Z} P(C_k) \cdot \prod_{i=1}^{n} P(x_i \mid C_k)$$

The term Z is a scaling parameter that ensures that the sum of probabilities for all classes is 1 (otherwise they would not be probabilities). The conditional probability of a class is the class probability times the probability of each feature given the class, normalized by Z. This formula can be derived by using Bayes' theorem [119].

Naive Bayes is an interpretable model because of the independence assumption. It can be interpreted on the modular level. It is very clear for each feature how much it contributes towards a certain class prediction, since we can interpret the conditional probability [119].

An explanation can be given for each prediction based on the contribution of every feature towards the posterior probability.

For instance: "The classification of the email is spam due to the presence of words such as 'free' and 'winner' which significantly raise the spam probability."

Probabilities give a clear-cut and straightforward explanation by their very nature.

Pros: Easy; quick; understandable; suitable for high-dimensional data.

Cons: Relies on the independence of features (often not the case); might not perform well if features have a strong correlation.

5.1.1.11. Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models (PGMs) that represent and reason relationships among variables using Bayesian inference within a graphical structure [136] [137].

A Bayesian Network is largely formed of:

1- **Structure (Graph):**

- It is a Directed Acyclic Graph (DAG).
- Every node in the graph symbolizes a random variable (for instance, “Fever,” “Infection,” “Temperature”).
- Every directed edge indicates a probabilistic dependency between the variables (for instance, an “Infection” leads to “Fever”).
- The arrows on the edges signify the cause or effect of one variable on the other.

2- **Parameters (Conditional Probability Tables – CPTs):**

- For measuring the degree of the relations between a node and its parent nodes, every node has a Conditional Probability Table (CPT) attached to it.
- CPT represents the conditional probabilities that determine how the state of a node depends on the state of its parent variables.

Mathematically, a Bayesian Network defines joint probability distribution over a set of variables $X = (X_1, X_2, \dots, X_n)$as:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P\big(X_i \mid \text{Parents}(X_i)\big)$$

where $\text{Parents}(X_i)$ denotes the immediate predecessors (causal variables) of X_i in the network.

Bayesian Networks are model-based (intrinsically interpretable) in the sense of XAI as they display the way predictions are made through their graph structure. The users may not only visually follow the path of reasoning through the network but also inspect the impact of the input data on the probabilities of the outcomes through probabilistic inference. A healthcare BN, for instance, may reveal the relation "Fever → Infection → Prescription of Antibiotics," which means the model may have prompted the user unintentionally to see the causal chain of the diagnosis. In addition, the capabilities of sensitivity analysis and updating the posterior probabilities make Bayesian Networks very fitting for providing explanations, and in fact, for being transparent and communicating uncertainty, which are the main goals of XAI [136] [138] [137].

Pros: Clear and understandable structure; supporting both causality and probability reasoning; and excellent uncertainty management.

Cons: Slower for large or dense networks; structural learning requires domain expertise for correct structure learning.

5.1.1.12. Adaptive Resonance Theory (ART) and ARTMAP

Adaptive Resonance Theory (ART) is a family of neural architectures influenced by biological and cognitive factors, thereby enabling self-organizing, interpretable category learning. ART networks form durable recognition categories by contrasting the bottom-up sensory input with the top-down expectations. The system has found a consistent interpretation of the input when these two signals communicate with the resonance at the same time—this is what resonance means in this context. The scenario of code instability exemplifies the situation where top-down expectations are not present, and the system gets deprived of this resonance mechanism, making it unstable and resulting in either unbounded or inconsistent formation of categories [60].

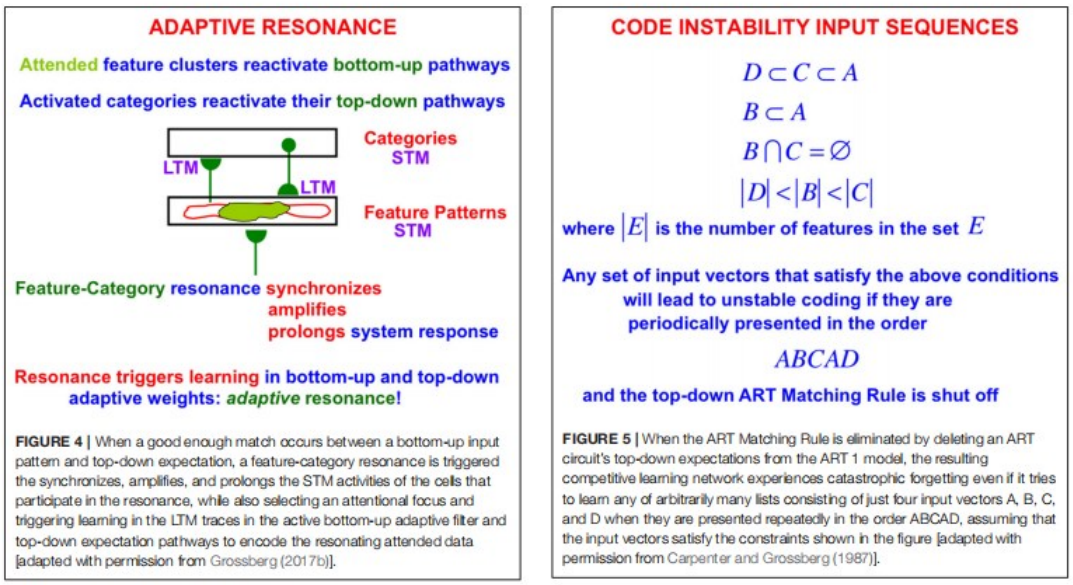

Figure 2 - Adaptive Resonance Mechanism and Code Instability in ART Networks

The process of resonance in the ART is inherently interpretable by design; that is, every category is represented by a human-understandable prototype that describes the necessary and sufficient features for a specific class.

The ART and its derivatives, namely Fuzzy ART, ARTMAP, and Fusion ART, are among the most popular solutions for pattern recognition, robotics, and continual learning, as they can learn new patterns incrementally and persistently at the same time, thus overcoming the "catastrophic forgetting" problem [60].

Pros: Prototypes that portray categories are understandable; compatible with biological processes; enable stable incremental learning.
Cons: Requires careful tuning of the vigilance determining parameter; the scalability of the model to those datasets that are high-dimensional is restricted; often not the best choice for non-linear domains that are complex.

5.1.1.13. Prototype-Based Models

Models based on prototypes represent a category of model-based explainable AI (XAI) paradigms that try to reason and make the deep models more interpretable and clearer by using representative examples (prototypes) through their similarity with the abstract numerical features [139] [140].
In prototype-based models, a class is represented by a set of prototypes—trained samples that embody the vital features or visual traits of that class. The model, during the training process, detects these representative prototypes from the data and the network stores. With the new input given, the model compares it with the prototypes and predicts according to the similarity. This process of reasoning of the model can be expressed as "the new input resembles this prototype," which is easy for a human to follow and visualize how the decision was made [140].

The Prototype Propagation Network (ProtoPNet) is one of the most notable models in this area. ProtoPNet combines prototype learning with convolutional neural networks (CNNs) by linking latent-space patches of training images to class prototypes. At the time of inference, the network points out the parts of the input image that are most like these prototypes, thus offering a case-based and interpretable for humans reasoning of its categorization [141].

As an example, in the task of image classification where birds are to be identified, ProtoPNet could come up with the prototypes that represent the features such as "bird wing" or "bird beak." Once the model classifies a new image as that of a bird, it can highlight the portions

of the image (e.g., a wing patch) that were like the learned prototypes and hence explain the reason for the given decision [141].

Pros: Example-based and visual explanations; interpretable by design; faithful reasoning.
Cons: May slightly reduce accuracy compared to black-box models; explanation quality depends on choice; computationally intensive for large datasets.

5.1.2. Explainable Complex Models

They are like deep neural networks but intend to enhance interpretability through built-in mechanisms. They usually make use of a high-performance plus a partial transparency approach using different methods.

5.1.2.1. Attention Mechanisms

The attention mechanisms signify novel techniques that are incorporated in neural networks and allow the model to selectively focus on the most enlightening parts of the input data when predicting the results. Instead of merging all input features, the model assigns different attention weights to its input components according to their importance. For example, sometimes even the same word in a sentence or a certain area in an image could be significant respectively, and this would be revealed by the allocated weight in the decision process [142] [143] [144].

In the Transformer model, attention mechanisms allow the model to concentrate differentially on various sections of the input in accordance with each step in the output sequence, thus emulating the human-like attention to certain parts of a sentence [144]. By visualizing the attention weights, one can get a glimpse into the model's decision-making process, which is, in fact, the input parts that the model judges to be important. This kind of visualization helps in comprehending the model's decision-making process, in addition to the input segments that are given significant attention; thus, it aids in making black-box models more interpretable and transparent. AttentionViz [145], which is a method for visualizing self-attention, brings forth the query and key embeddings that make it possible to analyze global patterns throughout sequences and to get even further insights into Transformers' pattern recognition and connectivity than the previous visualization methods. In another research, a technique is proposed for providing explanations for the predictions

made by Transformer-based systems, with a specific focus on the use of multi-modal data and co-attention in such cases [146]. This method suggests generic explainability solutions for the three predominant components of Transformer architecture: pure self-attention, a mixture of self-attention and co-attention, and encoder-decoder attention [147].

In the field of Natural Language Processing (NLP), attention permits the machine to indicate which words are mainly responsible for a given output, for example, as far as polarity detection of a sentence is concerned (“not good” vs. “very good”). Likewise, in the realm of Computer Vision, attention mechanisms are used to spotlight the areas of an image that are important for its classification just like encircling an imaginary circle around the dog’s face rather than its tail and the background. The way attention weights are applied is such that it facilitates the visualization of the process, and hence it can be assumed that the model's inputs and outputs are linked more clearly, and the reasoning is more interpretable to humans [148].

Attention mechanisms have turned out to be a must-have in the opposite of utilizing complicated pipelines in the deep learning models of Transformers, where they not only enhance the performance but also guide the understandability of the model. This is because they take the internal focus of the model and highlight it, thus giving the prediction a reason beyond just stating what it is.

Pros: Providing input-level explanations for complex models; improving performance by focusing on informative features; and enhancing transparency through attention visualizations.
Cons: Attention weights may not reflect true casual importance and can be difficult to interpret in deep or complex models.

5.1.2.2. Explainable Neural Networks (XNNs)

Explainable Neural Networks (XNNs) represent a category of neural networks that try to combine the advantages of the deepest learning models in terms of their transparency and interpretability with the disadvantage of a loss in their predictive power. They are designed to be the opposite of traditional "black box" neural networks. XNNs come with a unique composition that is built upon Additive Index Models. The output relies on the combination

of simple and interpretable functions. Each of these functions denotes the impact of a particular configuration of input features on the prediction [149]. There are three main components embedded in the architecture of this network:

Firstly, the projection layer makes linear combinations of the inputs.

Secondly, the subnetworks discover the one-dimensional nonlinear relationships (called ridge functions).

Finally, the combination layer that adds them together with determined weights.

The whole structure gives people the ability to see and comprehend not only what prediction and how much they do, but also which features have influenced it and in what manner. Thus, XNNs give an implicit interpretability to the users by demonstrating the input-output relationships in a straightforward manner and venturing into applications where trust and transparency are necessary, like finance, healthcare, and regulation [149].

xNN-SF (An Extended Vision)

xNN-SF (Explainable Neural Network Inspired by Stochastic Frontier Model), one of the breakthroughs brought by Zhao et al., 2024 (Zhejiang University), combines the XNN framework and the concepts of Stochastic Frontier Analysis (SFA) together [150].

In the Stochastic Frontier Model, the output is split into two segments:

- The Efficient (frontier) component that signifies the best practical performance under the given inputs.
- The Inefficiency component which indicates the divergence from the frontier due to loss, error, or inefficiency.

The xNN-SF framework adopts this notion by creating subnetworks that antagonistically learn:

- The frontier function (the highest possible output), and
- The inefficiency term (the difference that remains).

Elucidation is one of the major features in which xNN-SF and the likes of XAI differ. Additionally, not just predictions but also gaps in performance are the parts that analysts

particularly want to interpret in domains like finance, energy, and economics; hence, this characteristic makes the model very interpretable throughout such domains [150].

In such a manner, xNN-SF gives a wider scope of explainability by not only letting the user know how the outputs are predicted but also allowing them to see why the outputs differ from the best-achievable outcome [150].

Pros: Interpretable by design; transparent decisions; builds user trust; easier debugging; regulation-friendly; good accuracy-interpretability balance; no post-hoc explanation is needed (inherently interpretable); reliable predictions (good performance); both global and local interpretability; acting as a surrogate model to explain other black-box models.
Cons: Hard design, limited scalability; risk of oversimplification; fewer mature tools; and lower performance on complex tasks.

5.1.2.3. Self-Explaining Neural Networks (SENN)

Self-Explaining Neural Networks (SENN) are basically a turtle covering itself using its shell. They are an example of a new-generation technology that is the neural network (NN) yet making NN completely transparent decision-wise. The conception of three parts characterizes the paper from Alvarez-Melis & Jaakkola (2018) [151]: there is a concertizer that is converting input into human-understandable concepts; a parametrizer that is granting relevance scores to these concepts, and finally an aggregator that is uniting them for the output-generating process. This arrangement lets the users learn what features are responsible for getting the output of the model influenced in a certain way; therefore, it is particularly assuring, robust, and responsible in critical areas like healthcare or finance, where it is a must [151].

Moreover, the research, for example, the paper by Elbaghdadi (2020) [152], consider the SENN extensions, such as configuring them for different types of data like tabular, image, and text data, and at the same time making interpretability more powerful without significantly losing the performance. On the one hand, SENNs deliver unambiguous explanations of the predictions and are adaptable to the specific applications of domains; on the other hand, they still run the risk of being a tad less powerful than the complex black-box neural networks in case of handling the most sophisticated tasks [152].

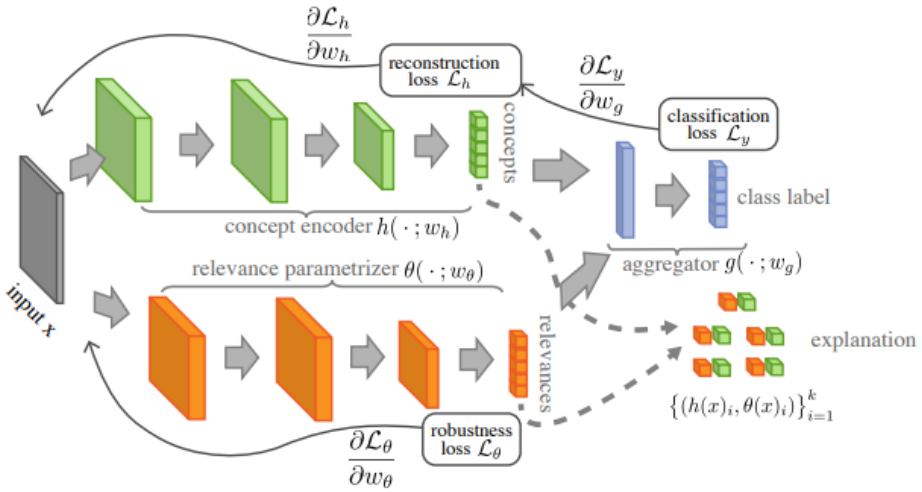

Figure 3 - SENN architecture (Alvarez Melis and Jaakkola, 2018)

Pros: Transparent and locally explainable; flexible design; robust explanations; works across data types.

Cons: Depends on concept quality; complex to train, hard to measure the quality of explanations; few tools and benchmarks.

5.1.2.4. Interpretable Convolutional Networks

Interpretable Convolutional Networks (ICNs) are deep learning architectures based on Convolutional Neural Networks (CNNs) that enable researchers to interpret the model's behavior internally. This is achieved by mapping every filter in the high-level convolutional layers to a specific part of the object that can be understood by humans [153]. In traditional CNNs, filters usually react to mixtures of features, but ICNs impose further regularization to ensure that:

1. Inter-category activation is sparse: A filter should be active mainly for one object category only.
2. Spatial activation is sparse: The activation of a filter should cover only one part of the object (i.e., one object part).

These restrictions are applied as entropy-driven losses in addition to the standard classification loss, during training. The outcome is a situation where filters are semantically

clear, thereby making it possible to identify directly which object parts affect predictions without the need for part annotations.

ICNs are thus representative of a model-based XAI approach, as the interpretability feature is embedded in the architecture rather than being applied afterwards [153].

Pros: Filters correspond to object parts; transparent decisions; no extra annotations needed; built-in interpretability.
Cons: Slight accuracy drop; hard to scale to deep networks; filter meaning not always intuitive; mostly for object classification.

5.1.2.5. Concept Bottleneck Models (CBMs)

Concept Bottleneck Models (CBMs) are one of the classes of model-based XAI methods, which make neural networks interpretable due to their inherent nature. Rather than mapping raw inputs directly to predictions, CBMs create a bottleneck layer of human-understandable concepts [154].

- Concept prediction where the model first predicts the interpretable concepts (e.g., "has wings," "is red," "irregular border").
- Label prediction where the final decision regarding the label (e.g., "bird," "malignant tumor") is made using the predicted concepts.

Every model decision is thus explained through the intermediate concepts made available to the users for inspection, verification, or even modification of the reasoning process.
Later works, such as Probabilistic Concept Bottleneck Models [155], expand on this idea by introducing the notion of uncertainty in concept prediction, while Causally Reliable CBMs [156] have focused on the aspect of spurious correlations between concepts and labels, improving robustness.

Pros: High interpretability; easy to debug via concept inspection; intervene human-in-the-loop edits; ethical and transparent.
Cons: Required concept annotations may limit model expressiveness; risk of biased and incomplete concepts; less accurate than black-box models.

5.1.2.6. Neural Additive Models (NAMs)

Neural Additive Models (NAMs) extended the traditional GAMs by replacing each simple function in GAMs with a neural network sub-model [157]. In a typical generalized additive model (GAM), each of the features contributes through a certain function to the prediction:

$$y = \sum_{i=1}^{p} f_i(x_i)$$

In neural additive models (NAMs), each (f_i) is represented by a small neural network instead of a simple linear or spline function. The summation of all the feature-specific networks' outputs yields the final prediction. The predictions from all the networks dedicated to a particular feature are added together to get the final prediction [157].
This design makes it possible for NAMs to capture complex, non-linear relationships of every single feature and, at the same time, be completely transparent - the effect of each feature on the prediction can be visualized separately, which is a must for interpretability in sensitive areas like medicine, banking, or policymaking [157].

Bayesian Neural Additive Models [158] extend NAMs by the introduction of probabilistic priors, Bayesian inference, and uncertainty quantification. The advantages of this method are robustness, generalization, and reliability; users can not only get to know the contributions of features but also the confidence intervals for predictions.

Pros: Human-understandable predictions; flexibility (neural subnetworks capture nonlinear relationships); transparent.
Cons: Limited feature interactions; higher computational cost; requires careful regularization to avoid overfitting; less flexible than black-box networks.

5.1.2.7. Graph Attention Networks (GAT)

Graph Attention Networks (GATs) are an approach that brings the concept of attention into the neighborhood of Graph Neural Networks (GNNs). By applying attention mechanisms, GAT allows each node to assess the importance of its neighbors and weigh the information accordingly [159].

Instead, GATs take a more flexible and non-fixed approach by dynamically calculating attention coefficients that label the extent to which each adjacent node will be contributing to the new representation of the target node [159].

Mathematically, each node i aggregates information from its neighbors j using learned attention weights α_{ij}:

$$h_i' = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j\right)$$

Where:

- h_j*is the neighbor's feature vector,*
- W*is a learnable weight matrix,*
- α_{ij}*is the normalized attention coefficient,*
- σ*is a nonlinear activation (e.g., ELU).*

In XAI, the GATs methodology becomes partially interpretable as the attention weights thus assigned can be visualized and examined to disclose which neighbors have had the greatest impact on the prediction of that node. This opens a way of providing a transparent signal of the model's internal reasoning through the graph structure—a feature that is not present in traditional GNNs [159].

Nonetheless, Brody et al. (2021) [160] performed subsequent research that pointed out that the attention weights of GATs do not necessarily reflect the ground truth regarding feature importance or causal influence. There is a work that brings out the learned attention mechanism as having unstable and at times even misleading characteristics, implying that GATs are probably not as "attentive" or explainable as initially thought.

Pros: Providing node-level interpretability; flexible and adaptive work for both transudative **(**fixed graphs) and inductive (new nodes/graphs) tasks.

Cons: Limiting reliability in XAI; complex; computationally expensive; partial interpretability.

5.1.2.8. Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) are a type of deep learning that processes graph-like data, where nodes are entities and the edges are the relationships. The GNNs are iteratively aggregating and transforming node information from neighbors to learn node embeddings, thereby examining local interactions and global structural patterns in a graph. They generalize the concept of traditional neural networks to structured non-Euclidean data like social networks, molecules or knowledge graphs [161].

GNNs are the fundamental of Explainable AI (XAI), as they can be made interpretable by integrating methods such as attention, masking, or concept-based reasoning to interpret model predictions [162].
For instance, Graph Attention Networks (GATs) bring in an attention mechanism that pinpoints which neighbors are most impactful in a node's forecast—yielding inherent interpretability [159].

Likewise, GNNExplainer boosts understanding by revealing essential subgraphs, nodes, and attributes that lead to a particular prediction, thereby making it easier to see how the model works with the graph structures [162].

Yuan et al. (2022) [163] furnished a detailed classification of GNNs' explainability techniques by organizing them into model-based (intrinsic) and post-hoc (model-agnostic) approaches.

Model-based approaches that are like GAT and PGExplainer, alter the architecture to improve its interpretability, thereby characterizing GNNs as ones that are compatible with model-based XAI principles where an explanation is a part of the model itself [163].

Pros: Complex structural dependencies; integrating attention or masking mechanisms for interpretable predictions; offering node, edge, or subgraph-level explanations.
Cons: High complexity; unstable explanations; requiring post-hoc validation to reliable explanations.

5.1.2.9. Neural Casual Models (NCMs)

Neural Causal Models (NCMs) have combined the inherent capabilities of neural networks with the formal abstractions of causal inference. They take one step further than Structural Causal Models (SCMs) by representing the connections between variables using neural networks as the agents of causation [164] [165].

It is worth mentioning that, while at traditional neural networks the only aspect being considered is correlation, NCMs are different since they want to depict the cause–and–effect relationships besides granting the reasoning support, answering "what-if" and "why" questions. A given NCM, for instance, could mimic the case where the specific input variable was interfered with (do-calculus), thereby making them relevant for interventions and counterfactuals in the understanding of real-world systems [166].

What NCMs essentially do is link causal modeling and deep learning: They keep the interpretability and the theoretical foundation of causal models while still using the expressiveness and scalability of neural networks. Nevertheless, their performance is reliant on the accuracy of the proposed causal graph and whether causal effects can be identified. In the absence of proper causal assumptions or data diversity (for instance, interventional data), even the most powerful neural models will end up being wrong in their deduction of causal relationships [166].

CNMs have been called by other names, and the reasons for that are as follows:

- Causal Neural Models (CNMs): Same as NCMs, just a different arrangement of words [166].
- Deep Causal Models / Deep Causal Learning Models: When the learning of causal relationships involves deep neural networks [167].
- SCM-NN (SCM-NN): This is when neural networks simulate the functions of a Structural Causal Model [166].
- Causal Representation Learning Models: The main goal is to learn causal representations, and CNMs are viewed as one of the categories in this context.
- Differentiable Causal Models / Neural Causal Graphs: If the neural model can be differentiated and is working on learning or predicting the structures of causal graphs.

The reason for name variation:
The causality and machine learning community is not yet in possession of a common naming standard. Some names underline the graph structure, others the neural networks, and still others the interpretability and representation learning aspect. Depending on the use case (prediction, intervention, or counterfactual reasoning), various papers apply different terms, but in essence, they all cover the same ground—the pairing of neural networks with causality in research.

Pros: Combine neural flexibility with causal interpretability; enable interventional and counterfactual reasoning; model nonlinear, high-dimensional causal mechanisms.
Cons: Require prior causal structure or graph knowledge; difficult to guarantee identifiability from observational data alone; computationally complex to train and validate.

5.1.2.10. Causal Discovery Models

Causal discovery models are created to reveal cause–and–effect relationships from observational or experimental data, so that the explanations provided will be more than just correlations. Causal models, as opposed to conventional machine learning models, can respond to the questions "why did this happen?" and "what would happen if we intervened?" which are the primary needs of high-quality XAI [168].

Causal discovery methods (e.g., PC, FCI, GES, LiNGAM, and deep causal learning models) produce causal graphs that visualize the direct and indirect impacts of the variables on each other. Such graphs allow for counterfactual explanations like "if this feature changed, the decision would be different, which are among the most compelling types of explanations [169] [168].

The latest developments involve combining deep learning with causal discovery, thereby making it possible to create a hybrid causal–neural XAI that can provide explanations that are consistent, unbiased, and applicable to various tasks. Not only that, but these approaches also increase flexibility: user groups with different levels of granularity can be presented with causal stories, or their specific goals can be supported with task-relevant causal pathways [169].

Neural Causal Models and Causal Discovery Models both deal with the same issue, but each has a different approach.
Causal Discovery Models are first to identify the unknown causal structure through data, which means the outcomes are determined by the variables and vice versa.
To put it in a nutshell, NCMs clarify by causal reasoning and Causal Discovery Models do it by revealing the causal relationships.

Pros: Casual explanation beyond; support counterfactual and interventional reasoning; structural transparency; bias and confounder detection; improved generalization and robustness, integration with deep learning.
Cons: Dependence on strong assumptions; sensitivity to data quality and confounders; limited identifiability from observational data alone; high computational complexity; challenge of validation; integration complexity in deep frameworks.

5.1.2.11. Interpretable Graph Models with Symbolic

Graph-based interpretable models utilize graph-structured representations and symbolic reasoning to clarify the decision-making process in a manner that is comprehensible to humans. By encoding the entities as nodes and the relationships as edges, these models visualize and logically trace the decision pathway. The integration of symbolic rules or constraints—such as logical clauses, rule-based systems, and graph traversal patterns—enables the model to provide explanations that can be scrutinized, confirmed, and even modified by human experts [170].

The use of graphs in explainability is a very potent tool in areas where relationships, dependencies, and structures are important: knowledge graphs, medical diagnosis graphs, molecular structures, social networks, and recommender systems. In these systems, often the explanations are made through path reasoning, subgraph extraction, contrastive rules, or causal structural dependencies. Due to their explicit structure, graph-based models are naturally suited to adaptive explanations, in which different users (e.g., expert vs. novice) may receive different levels of graph detail or symbolic abstraction [170].

Graph reasoning also fits nicely with hybrid neuro-symbolic XAI methods, where deep learning models provide embeddings, and the symbolic layers convert those embeddings

into interpretable rules. This combination has been found to be beneficial in building trust, consistency, and correctness in critical areas [171] [172].
Generally, such models work in two successive and complementary stages:

1- Graph representation learning reveals the nodes' characteristics and structural dependencies.
2- Symbolic reasoning applies interpretable rules or constraints over these learned relations to derive conclusions or verify consistency.

The reasoning steps are made explicit and are human-readable, so the users can easily trace the link, label, or outcome that was produced and the reason for that.
The combination of neural relational modeling and symbolic logic thus brings about both interpretability and high performance.
The models are frequently applied in tasks like knowledge-graph completion, biomedical relation extraction, social-network analysis, and causal inference, where the comprehension of relational logic is as critical as predictive accuracy.

Examples:

- Explainable Graph Neural Networks (XGNN) with rule extraction [163]
- Logic-Guided Graph Networks (LG-GNN) [173]

Pros: High interpretability, transparent decision logic, improve trust and accountability, better generalization, with constraints, facilitates knowledge integration.
Cons: High computational complexity, limited scalability, difficult knowledge of engineering, reduced flexibility, implementation complexity.

5.1.2.12. Cognitive Neural Models (CNMs)

Cognitive Neural Models (CNMs) are a kind of neural architecture that takes inspiration from human cognitive processes such as perception, attention, memory, reasoning, and learning. They are supposed to be not only interpretable, but they also let each user find the analogies with the corresponding cognitive functions that are believed to be human-understandable [174].

Instead of just being statistical models, CNMs are looking at the human way of processing and integrating the information, and thus they allow us to map the internal mechanisms of the model into known cognitive operations.

These models can reveal the process of input perception, the list of features that attracted attention, how information was stored or combined, and the reason for the decision made at the end [175].
The close relationship of computational modules with cognitive functions makes CNMs particularly advantageous for AI-human interaction, decision making, and educational systems, where the explanations must be meaningful for the users.

Some of the models that fall under this category include the ACT-R neural implementations, Deep ACT-R [176], CLARION, and hybrid connectionist–cognitive architectures [177] [175].

Pros: Human-friendly modular structure, contact with cognitive reasoning, enhanced transparency and trust.
Cons: Heavy on computing; not adaptable to big data; mostly tied to a specific task rather than a broad application.

5.2. Model-agnostic (Post-hoc)

They give explanations for any trained model without any alteration. They decipher the inputs as well as the outputs and produce the explanations which make them flexible and applicable to a broad range of cases. This category can be used for any black-box model. Explanation is added to the model from outside.

5.2.1. Local Explanations

Concentrate on the justification of a single prediction, that is, to unveil the reason behind the model's decision-making in a particular instance.

5.2.1.1. LIME (Local Interpretable Model-agnostic Explanations)

LIME is a very commonly used model-neutral technique that interprets the output of any non-transparent model by approximating it with a simple, understandable model (like linear

regression or decision trees) to the first order. The gist is that while a complicated model (like a neural network) may be globally difficult to understand, its action can be approximated in a local context (around a single prediction) using a simpler, transparent model [119] [178].

How It Works

- **Select an instance to explain:** Pick one sample (like one image or one text) for which you want to get an explanation.
- **Generate perturbations**: The input data will be slightly modified to derive new synthetic data points, such as altering pixels, removing words, or varying feature values, around the instance.
- **Predict with the black-box model**: Get the model's output for these altered samples.
- **Fit an interpretable local model**: A simple model that is easy to interpret (like a linear regression) will be trained with the modified data and the respective predictions given.
- **Computer feature importance**: The coefficients of the interpretable model indicate which features had the strongest influence on the black-box model's prediction for that instance.

Mathematically, LIME minimizes the following objective:

$$\xi(x) = \arg\min_{g \in G} \; L(f, g, \pi_x) + \Omega(g)$$

where:

- f *= black-box model,*
- g *= interpretable local model,*
- L *= loss function measuring the fidelity between* f *and* g*,*
- π_x *= proximity measure (weights samples closer to the instance more),*
- $\Omega(g)$ *= complexity penalty for* g *(simpler models preferred).*

This ensures that g *is:*

Locally faithful (accurate near x*)*

Globally simple (easy to interpret)

Pros: Flexible, providing human-readable explanations (produces feature-level explanations), easy to implement and widely supported (e.g., via lime python library), model-agnostic.

Cons: instability (change if input slightly changes), local only (does not provide a global view of model behavior), sampling cost (generating and evaluating perturbations can be slow).

5.2.1.2. SHAP (SHapley Additive exPlanations)

The SHapley Additive exPlanations (SHAP) method has a unified, game-theoretic approach for interpreting machine learning model predictions. It uses the concept of Shapley values from cooperative game theory to determine a feature's importance score based on its contribution to the model output. The key properties of local accuracy, consistency, and missingness are satisfied by SHAP, thereby ensuring a fair and theoretically sound distinction of feature importance. It can be localized or globally spread, thus providing knowledge about individual predictions as well as the overall behavior of the model [179] [180].

For a model $f(x)$and a feature set $F = \{x_1, x_2, \ldots, x_n\}$, the Shapley value for feature $i$is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

Where:

- S *is a subset of all features excluding* i.
- $f(S)$ *is the model prediction using only the features in* S.
- ϕ_i *quantifies the average marginal contribution of feature* i *across all possible feature combinations.*

This is the reason why SHAP satisfies three desirable axioms:

Local Accuracy (Additivity): The total of all SHAP values equals the model's prediction difference from the baseline.

Consistency: If a model is modified in such a way that a feature becomes more contributing, its SHAP value must not decrease.

Missingness: Features that are not part of the model input at all will have a SHAP value of zero [179].

Lundberg, Erion, and Lee (2018) extended a new method called a tree-based model based on the original work done on SHAP that allowed SHAP to become more efficient when applied to tree-based models. This can compute Shapley values with the help of polynomial time. This made it feasible to use SHAP for large ensemble models such as random forests and gradient-boosted trees while still being able to offer theorems about the methods being consistent and locally accurate [180].

Pros: Fair; explain local and global predictions; consistent; supports trust and transparency.
Cons: Computationally expensive; complexity in interpretations; may assume feature independence.

5.2.1.3. Anchors

Anchors is an approach to local explanation that is not bound to any specific model and is based on rules. This method gives local explanations in the form of very precise "if-then" rules [181].
The main concept is to single out a collection of input conditions (a "rule" or anchor) such that the model's prediction keeps being the same with a quite high probability whenever those conditions are true. More precisely, for instance x, an anchor A is defined so that $\mathrm{A}(\mathrm{x})=1$ (the conditions of the rule apply), and the output of the model for perturbed samples z satisfying $\mathrm{A}(\mathrm{z})=1$ is the same as for x, with at least a threshold τ probability.

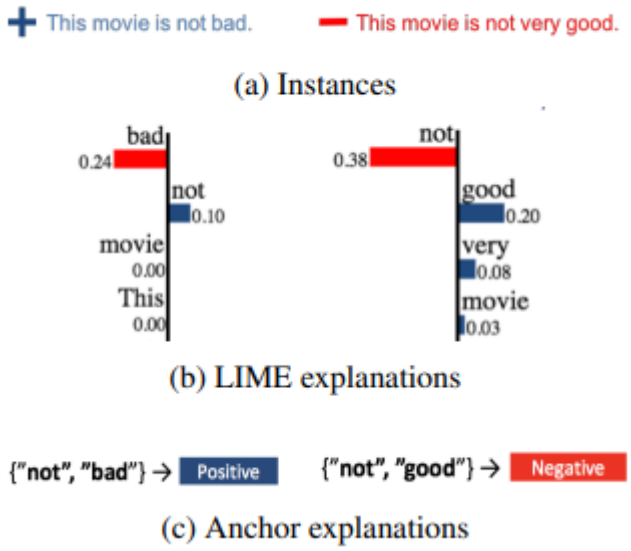

Figure 4 - Sentiment predictions, LSTM

For example, the anchors in Figure 4 state that the presence of the words "not bad" virtually guarantees a prediction of positive sentiment (and "not good" of negative sentiment). Anchors are intuitive, easy to comprehend, and have extremely clear coverage – they only apply when all the conditions in the rule are met, and if they apply, the precision is high (by design) [181].

As stated in the Interpretable Machine Learning guide by Molnar [119], what makes anchors stand out is their transparency and coverage: users can easily grasp them and they are only applicable when the conditions laid down by the rule are off, but Molnar also brings the fact that the anchor method may become very slow when there are a lot of features since the search for the rules tends to grow with the number of features and also that the creation of a good perturbation space is the foundation for meaningful explanations [119].

Pros: Simple; high precision; model-agnostic; provides clear conditions when prediction is stable.

Cons: Computationally expensive; limited coverage; harder to generalize beyond local neighborhoods; Sensitive to how data perturbations are generated.

5.2.1.4. Counterfactual Explanations

Counterfactual explanations demonstrate how a prediction would change if certain input features were modified. They do not expose the black box or uncover the inner workings; instead, they present the least amount of change that a person needs to make to have a different (typically more favorable) result. Thus, this is consistent with a practical concept of interpretability: showing people what inputs they could modify to shift the future decision in their favor. For example: “If I hadn’t taken a sip of this hot coffee, I wouldn't have burned my tongue". Event Y is that I burned my tongue; cause X is that I had a hot coffee [182]. This ensures that the model's decision gets flipped through the least, thus making it more manageable and keeping it realistic and coherent with possible real-world values.

Counterfactuals are of great importance in sectors such as finance, healthcare, and recruitment, to name a few, because they not only explain why the current prediction was made but also show the actionable steps needed to obtain a desired prediction [183].
What makes counterfactuals important is that they give practical feedback: they tell which features to modify (and by how much) to steer the prediction, hence their critical importance in high-risk or supervised areas like lending or hiring.

On the other hand, there is a trade-off in the process of producing good counterfactuals: they must be realistic (i.e., possible for a person to change), they must be sparse (only a few features are altered), and they must be cohesive (in harmony with the data distribution) [119].

In terms of law and ethics, the three authors Wachter, Mittelstadt, and Russell argued that counterfactual explanations are a good fit for the objectives set by GDPR (General Data Protection Regulation). The authors posit that such explanations would enable individuals to realize the decision rationale, question unfavorable decisions, and even prepare a strategy to get different results—all these, without revealing the proprietary know-how of the model [94].

Nonetheless, according to Verma, Dickerson, and Hines, the practical uptake of counterfactual reasoning has been slower than anticipated. One of the challenges they mention is the difficulty in making the suggested changes across different people actionable, valid, and fair [183].

Pros: Actionable; model-agnostic, encourages fairness and accountability in decision systems; helpful for decision support in sensitive applications.

Cons: Realism challenge; complexity in generation, may produce multiple or ambiguous explanations; adoption barriers; sensitive feature dependencies and data constraints.

5.2.2. Global Explanations

Global explanation methods aim to describe the overall behavior and logic of a trained model rather than explaining individual predictions. Because they are model-agnostic, they can be applied to any black-box algorithm without modifying its internal structure. These techniques analyze the relationships between input features and predicted outcomes across the entire dataset, revealing how the model learns patterns, identifies trends, and distributes importance among variables. In this way, global explanations provide a holistic understanding of the model's functioning, complementing local interpretability methods such as LIME or SHAP.

5.2.2.1. Partial Dependence Plots (PDPs)

Partial Dependence Plots (PDPs) are globally applicable and model-independent interpretation methods that depict the marginal effect of either one or two input features on the predicted outcome of a machine learning model. The technique attempts to indicate the extent of the prediction as a function of the selected feature while simultaneously eliminating the impact of the rest of the features [119] [184]. The partial dependence function is defined mathematically as:

$$\hat{f}_{x_S}(x_S) = \mathbb{E}_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, x_C)\, d\mathbb{p}(x_C)$$

Where:

$\hat{f}(x_S, x_C)$*: the prediction of the trained model using all features.*

X_S*: the subset of features you are analyzing (e.g., "age").*

X_C*: the complement set — all other features except* X_S*.*

$P(X_C)$*: the marginal probability distribution of those other features.*

$\mathbb{E}_{X_C}[\cdot]$*: expectation over that distribution (i.e., average model output).*

Basically, PDPs demonstrate the change in the model's average prediction for a given feature value change while other features' contributions are assumed to be constant (averaged). PDPs have been developed into interactive visual tools wherein users can witness the impact of certain variables on predictions. The visualization framework of the method provides the analyst facility to examine the nonlinear effects of features and spot the possible interactions among features [185].

The idea of Individual Conditional Expectation (ICE) plots was one of the key features that came along with the PDPs. ICE plots give insights into the individual prediction change of an instance with respect to a feature, thereby disclosing the degree of heterogeneity that was masked by the average in PDPs [186].

Simplicity, the model-independent nature, and the visualization ease are the main reasons for the broad acceptance of PDPs. They are cited as one of the major tools for global interpretability, assisting in the recognition of both monotonic and nonlinear relationships between features and predictions [119] [187].

Pros: Model-agnostic; simple and intuitive visualization; helps identify feature relationships; effective for global interpretation.
Cons: Assumes feature independence; averages out individual differences; computationally expensive; limited to low-dimensional visualization.

5.2.2.2. Individual Conditional Expectation (ICE) plots

Individual Conditional Expectation (ICE) plots are one of the most effective visualization techniques used to interpret complex machine learning models on a per-instance basis. They show the predicted outcome of a model as one feature changes while the other features for a specific observation are kept constant [188].

In ICE plots, for each instance $(x_S^{(i)}, x_C^{(i)})$ where $i = 1, \dots, N$, a curve $\hat{f}_S^{(i)}$ is plotted as a function of $x_S^{(i)}$, while the complementary features $x_C^{(i)}$ are kept fixed [119].

While Partial Dependence Plots (PDPs) provide the average impact of a feature on forecasts over the whole dataset, ICE plots disaggregate this impact, revealing heterogeneity that

global averages may mask. ICE Plots permit the analysts to observe the non-linear and interaction effects by varying the response of each instance unconditionally [186].

ICE plots not only localize but also globalize the interpretation. Having multiple ICE curves, one can easily tell whether all individuals are responding the same way to a feature (parallel curves) or they are responding differently (diverging curves), which indicates the presence of interaction or non-linearity [119].

Pros: Captures individual-level feature effects hidden by averages; reveals feature interactions and non-linear relationships; provides both local (instance) and global (aggregate) interpretability when combined with PDPs.
Cons: Becomes cluttered for large datasets (too many curves); harder to interpret without smoothing or clustering; computationally intensive for large models.

5.2.2.3. Feature Importance

The importance of a feature is the increase in the prediction error of the model after the feature's values are permuted, which breaks the relationship between the feature and the true outcome. Feature Importance stands out among the global explanation techniques in XAI as the one that best fits human intuition. It enables one to quantify the predictive performance of the model with respect to the different features, i.e., how much each feature was responsible for it.

The main idea is that if we look at how much the model's accuracy or its loss changes by either altering or taking out a certain feature, we would be able to determine the importance of that feature. Among various methods, Molnar (2022) [119] points out permutation importance as being very efficient and applicable for all models - it quantifies the reduction in model performance after a particular feature's values have been shuffled while the others remain unchanged [119].

Feature importance was first introduced by Breiman (2001) [189] in relation to Random Forest; here, it is determined by the reduction in impurity (e.g., Gini impurity) or improved accuracy of prediction through the decision trees over the feature. This method provides insight into the relative importance of various features in the prediction of the model and is still extensively adopted in ensemble techniques.

Fisher, Rudin, and Dominici (2019) [190], on the other hand, expand this concept by presenting MCR Model Class Reliance, which is one of the statistically robust approaches that capture a feature's importance across a whole predictive model class rather than for a single one. Their findings underline that importance must be put in context in a wider sense so as not to extract misleading conclusions from one model's internal functioning.

Pros: Simple and visual; identifies key feature; model-agnostic.
Cons: Based on correlated features; ignoring features interactions; depend on the chosen metric

5.2.2.4. Permutation Importance

Permutation Importance evaluates the contribution of each feature to the model's predictive accuracy by measuring the change in performance [189] [119].

The main point is straightforward:

A feature's importance is evaluated by permuting its values randomly (shuffling) while leaving all other features intact. This breaks the association between that feature and the outcome variable. The decrease in model performance (e.g., accuracy, F1-score, or R^2) after the permutation is an indication of how much the model relies on that feature. A more substantial drop indicates a more significant feature [190] [189].

As per Molnar (2022) [119] in Interpretable Machine Learning, permutation importance is a beneficial method since it is a natural way of interpretation, easy to apply, and compatible with all kinds of black-box models. It provides a clear picture of the overall feature importance and does not depend on the internal model-specific structures (unlike the case with tree-based models, where internal structures are exploited).

Nonetheless, it has some drawbacks: under situations when features are correlated, importance might be misrepresented, and it is also possible for it to depend on the specific test data distribution or performance metric chosen.
The article "Permutation importance: A corrected feature importance measure" [191] says that the original permutation importance in Random Forests may sometimes be misleading, especially in situations where the features are correlated or have different scales or the

number of categories is different. In such cases, the measure may overemphasize the importance of continuous or high-cardinality variables while at the same time underestimating the importance of correlated or categorical ones.

To overcome this, the paper advances a corrected permutation importance measure that removes these biases by applying conditional inference trees and conditional permutations. In other words, instead of permuting the feature values completely at random, the corrected procedure permutes them conditional on correlated features thus maintaining their statistical relationships.
This correction makes available a more trustworthy and unbiased estimate of variable importance, particularly in datasets where feature correlation exists.

Pros: Shows real feature impact; simple and model-agnostic.
Cons: Biased with correlated features; slow on large data; depends on metric used.

5.2.2.5. Surrogate Models

Surrogate models can be seen as interpretable approximations of complex black-box models. They are designed to replicate the original model's forecasts using simpler, more exposable methods —such as a decision tree, linear regression, or rule-based model. The concept is that by getting the surrogates, we can investigate the black-box model and its ways of operating [119].

Surrogates can be either global or local ones.

- Global surrogate gives an account of the general logic of the model.
- Local surrogate explains individual predictions.

The whole thing is done by first getting predictions from the black-box model for numerous input samples and then fitting the surrogate to these input–output pairs [119]. The upside to this technique, however, is that the surrogate models are approximations and not exact representations. Therefore, it is important to trust the replication quality—the fidelity, as it is called the surrogate models when interpreting the black-box model. To this end, Molnar himself points out that the surrogate model accuracy must be confirmed before any conclusions are made [119].

One way to measure how well the surrogate replicates the black box model is the R-squared measure:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}^{*(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n} (\hat{y}^{(i)} - \bar{\hat{y}})^2}$$

where $\hat{y}^{*(i)}$ is the prediction for the i-th instance of the surrogate model, $\hat{y}^{(i)}$ the prediction of the black box model and $\bar{\hat{y}}$ the meaning of the black box model predictions. SSE stands for sum of squares error, and SST for sum of squares total. The R-squared measure can be interpreted as the percentage of variance that is captured by the surrogate model. If R-squared is close to 1 (= low SSE), then the interpretable model approximates the behavior of the black box model very well. If the interpretable model is very close, you might want to replace the complex model with the interpretable model. If the R-squared is close to 0 (= high SSE), then the interpretable model fails to explain the black box model [119].

How it works:

- Firstly, a dataset is prepared containing input samples and the corresponding predictions made by the black-box model.
- Next, the dataset is used to train an interpretable model (e.g., a linear model or a small decision tree).
- Finally, the surrogate can be employed to carry out feature importance, decision paths, or variable relationships analyses.

Pros: Useful for both global and local explanations; model-agnostic; simple, interpretable.
Cons: May lose fidelity; interpretations depend on surrogate accuracy; risk of oversimplification in complex models; no clear R^2 to ensure closeness to the black-box model; conclusions apply to the model, not the data.

5.2.2.6. Global Surrogate Models

The concept of global surrogate models is also known as model distillation models, which are simple, interpretable models, such as decision trees or linear regressors, that are trained to imitate the performance of more complex, black-box models like ensemble methods or deep neural networks. The authors put forth a tailored methodology for developing these

surrogates, with an objective to enhance both the fidelity (the closeness of the surrogate to the original model's predictions) and interpretability [192].

Their approach is model agnostic, which implies that it is applicable to any machine learning algorithm, and it focuses on deriving global insights regarding feature importance and decision logic. The article reports on trials carried out on assorted datasets through which it proves that its tailored surrogate models have not only significantly increased approximation accuracy but also provided a better understanding of model behavior. This, in turn, leads to the development of trust in AI systems, particularly in delicate fields such as medicine and finance, where comprehending model decisions is vital.

In contrast, Global Surrogate Models aim to explain the model's behavior across the entire input space; some surrogate-based methods can also focus on local interpretability. For example, LIME are particularly focusing on explaining single predictions with the help of local surrogates, whereas others like SHAP are drawing wider understanding through accumulating local insights.

Global surrogates could offer an alternative approach that provides both high predictive power and explainability [193].Global surrogates are an option for insurers that have developed a black-box model that is difficult to implement, either due to regulatory constraints or other logistical issues [193].

Pros: Flexibility to choose any statistical framework; balances interpretability and predictive fidelity; simple; provides global insights into model behavior.
Cons: May lose accuracy compared to the original model; explanations depend on surrogate fidelity; oversimplification may hide subtle relationships.

5.2.3. Gradient-based and Relevance-based Explanations

The mentioned techniques are primarily developed for deep learning models, where mostly vision or NLP are the applications. They investigate the internal workings of the model (like gradients or activations) to figure out which part of the input had the largest influence on the prediction.

5.2.3.1. Saliency Maps

Saliency Maps are visualization gradient-based techniques for getting an explanation in XAI of the Convolutional Neural Networks (CNNs) that are used for computer vision tasks. The saliency maps aim at visually showing which parts (i.e., pixels or areas) of the input image had the most impact on the model's prediction.

In this process, the gradient of the model's output concerning the input image is obtained. The formula for the saliency map from Simonyan et al., 2014 [43] is:

$$w = \frac{\partial S_c}{\partial I} |_{I_0}$$

Where:

- S_c *= the score (output before softmax) for class* c*.*
- I *= the input image (all pixel values).*
- I_0 *= the specific input image for which we compute the saliency.*
- w *= the saliency map (gradient of the class score with respect to each pixel).*

The size of these gradients denotes the output's dependency on the corresponding input pixel—huge gradients suggest that even a small variation in those pixels would have a considerable influence on the model's decision. So, the saliency map thus created represents the importance of the pixels for the given prediction as a visual.

Saliency maps provide detailed, fine-grained, pixel-level explanations, which assist researchers and practitioners in recognizing the image features that lead the model to classify in a certain way. They, however, might not always be clear or easy to interpret, and anything from a slight change in model parameters to a change in inputs may result in different saliency patterns [43].

Pros: Simple and intuitive visual explanations; model-agnostic; offers pixel-level feature importance.

Cons: May be noisy and unstable; does not show true casual importance; sensitive to gradient saturation or network architecture.

5.2.3.2. Grad-CAM

Grad-CAM is a method that has been specially created for visualizing the predictions made by convolutional neural networks (CNNs) by showing, through color maps, which parts of the input image were the most influential in the prediction. The method works by calculating the gradients of the score assigned to the predicted class with respect to the feature maps of a specific convolutional layer and then taking the average of these gradients to form a weight map that points out the importance of each feature map. Superimposing the weights on their respective feature maps creates a heatmap of the regions that have the highest contribution to the model's output [42].

The technique offers a network's attention visually and hence is a very effective way of understanding the underlying reasons of a specific CNN prediction. It supports all CNN architecture without the need for any modifications and can help in the detection of model errors or biases through the analysis of the regions affecting the predictions. On the downside, the explanations given are mostly rough, and their quality is related to the selection of the convolutional layer. Grad-CAM remains a provider of qualitative insights as opposed to giving precise numerical feature importances [42].

Mathematical explanations

Let y^c be the score for class c before SoftMax and let A^k be the k-th feature map in a convolutional layer. Grad-CAM computes the importance weight α_k^c for each feature map using the gradients of y^c with respect to A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \frac{\sum_j \quad (\partial y^c)}{\partial A_{ij}^k}$$

where Z is the number of pixels in the feature map, and i, j index the spatial locations.
The Grad-CAM heatmap $L^c_{\text{Grad-CAM}}$ is then computed as:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}(\sum_k \quad \alpha_k^c A^k)$$

The ReLU ensures that only positive contributions are considered, highlighting regions that positively influence the class prediction. This heatmap can be unsampled and overlaid on the input image for visualization.

Pros: Intuitive visual explanations; model-agnostic for CNNs; helps find model errors or biases.
Cons: Produces coarse heatmaps, explanations depend on layer, qualitative only.

5.2.3.3. Integrated Gradients

Integrated Gradients is a method for attributing credits for inputs through gradients which was developed especially for making predictions of differentiable models elucidating, and more specifically, neural networks. It offers a solution to the issue of input feature credit assignment for the output of a model and, at the same time, adheres to the criteria of sensitivity and implementation invariance [195].

The process involves a comparison of the input x with some baseline input x' (for instance, a zero vector or a neutral reference). The attribution of each feature i is estimated as the integral of the gradients along a straight line from the input to the baseline:

$$\mathrm{IG}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

In this formula, F represents the model function while α signifies the position on the straight line that connects the baseline and the input. As a result, a score reflecting the importance of the feature is produced, indicating how much each feature has contributed to the prediction.

Pros: Provides faithful attributes; applicable to any different model; meets theoretical criteria.
Cons: Needs a baseline; computationally expensive; specific to a single input (local).

5.2.3.4. Decomposition-Based Methods

Decomposition-based procedures strive to elucidate difficult machine learning algorithms by making it possible to see which input features contributed to the model's output [195].

On the other hand, perturbation methods vary the input features to measure the sensitivity of the output. Decomposition approaches, however, turn the prediction back through the layers of the model, and according to specific propagation rules, they give the output score back to the input features in the defined ways [195].

A prominent example is Deep Taylor Decomposition (DTD), where the relevance of each neuron is computed by performing a Taylor series expansion of the function f around a root point. This not only makes the relevance propagation theoretically justified but also guarantees thar the total relevance across layers is equal. By this property, the explanation is assured to be the same, no matter if it is given by the model from output to input [195].
Hence, one of the main advantages of decomposition-based methods is that they can reveal very small details about neural networks predictions and, consequently, they have been the basis for later methods like Layer-wise Relevance Propagation (LRP) and DeepLIFT.

Pros: Layer-wise, mathematically grounded explanations; ensures relevance conservation; interpretation consistency; pixels or features insights in deep models.

Cons: Not model agnostic; implementation can be complex for modern architecture; sensitive to choice of root points and propagation rules.

5.2.3.5. Layer-wise Relevance Propagation (LRP)

Layer-wise Relevance Propagation (LRP), deriving from Deep Taylor Decomposition (DTD), is among the most prominent gradient-based methods in Explainable Artificial Intelligence (XAI). The origin of LRP is in the explanation of the decisions made by non-linear classifiers, especially in deep learning, through the retracing of the model's prediction to the input features.

It was Bach et al. (2015) [196] who presented LRP as a pixel-wise explanation approach for deep classifiers. Their paper, which concerned image classification, provided evidence that LRP can generate heatmaps that indicate the extent to which each pixel had an impact on the prediction. The whole process of explaining convolutional neural networks (CNNs) and, in fact, every complicated, nonlinear classifier, works by score redistribution in the reverse direction through the layers.

In advance of CNN architecture featuring local renormalization layers, Binder et al. (2016) [197] came up with an LRP extension that greatly enhanced its applicability to the more sophisticated continuous functions and networks. Strictly following the propagation rules, this work made the operations of normalization very gentle and still interpretable.

The LRP method was used by Sundararajan et al. (2017) [194] in Axiomatic Attribution for Deep Networks to provide a theoretical basis for gradient-based interpretability methods. They marked that LRP was one of the few that possessed the critical axioms (like sensitivity and implementation invariance) necessary for fair and consistent feature attribution.

Sturm et al (2017) [198] illustrated the benefit of LRP in uncovering the inner workings of neural networks in EEG signal classification tasks. Accordingly, their research portrayed that LRP could reveal areas of the brain or patterns of signals that are critical for the performance of certain cognitive tasks, hence proving its potential beyond the field of computer vision.

Mudabbir Ali et al. (2022) [199] portrayed LRP as a technique belonging to the category of surrogate explanation models, stressing that LRP explanations could and even should be employed to guide the creation of custom surrogate models for the purpose of better interpreting the model and gaining its trust.

Basic LRP Formula [196]:

$$R_d^{(1)} = \begin{cases} f(x)\dfrac{|\,\alpha_d\phi_d(x_d)\,|}{\sum_d |\,\alpha_d\phi_d(x_d)\,|}, & \text{if } \sum_d |\,\alpha_d\phi_d(x_d)\,| \neq 0 \\ \dfrac{b}{V}, & \text{if } \sum_d |\,\alpha_d\phi_d(x_d)\,| = 0 \end{cases}$$

This version distributes the model output $f(x)$among all input features $d$in proportion to the absolute weighted feature contributions $|\,\alpha_d\phi_d(x_d)\,|$.

Improved LRP Formula [196]:

$$R_d^{(1)} = \frac{b}{V} + \alpha_d\phi_d(x_d)$$

This version provides more meaningful positive and negative relevance values, allowing us to see which features support ($R_d > 0$) or oppose ($R_d < 0$) the model's output.

Pros: Delivers detailed, elementwise, or attribute-specific explanations; works well for deep nonlinear models; fulfills theoretical principles for fair and consistent attritions; creates transparent, simple visual maps that help interpretation.

Cons: computationally expensive; highly dependent on the model design; Different propagation strategies (ε-rule, αβ-rule) can lead to variable explanations; interpretation accuracy may decline if the network architecture is suboptimal.

5.2.3.6. DeepLIFT

DeepLIFT is one of the model-dependent explainable AI techniques that ascribe the input features of a neural network to its prediction by measuring the activations of the neurons with a reference (baseline) input. DeepLIFT goes further than traditional methods by computing the differences in activations instead of applying gradients to show the input's contribution to the gap between the actual output and the reference output.

DeepLIFT assigns back "contribution scores" starting from the output of the network, measuring the difference between the actual activation of the neuron and the activation for the baseline input. For a neuron $t = f(x)$, the contribution $C_{\Delta x \to \Delta t}$satisfies:

$$\sum_{i=1}^{n} C_{\Delta x_i \to \Delta t} = \Delta t$$

where difference-from-reference as $\Delta t = t - t_0$

This guarantees that the total contributions are equal to the changed output, thus preserving the completeness, which is an important axiomatic property that Integrated Gradients shares with DeepLIFT [200].

Propagation Rules

DeepLIFT prescribes rules for the backward propagation of relevance through layers:

- **Rescale Rule:** Allocates contribution proportionally to the variations in activation and weights.
- **RevealCancel Rule:** Treats positive and negative contributions separately to prevent mutual cancellation, thus making the interpretation of nonlinear activations clearer [201].

As a result, these rules permit the DeepLIFT technique to work with saturated gradients (the point where the gradient-based methods fail) by looking at differences and not derivatives [200].

Sundararajan et al. (2017) [194] recommended axioms such as Sensitivity and Implementation Invariance so that consistent attributes could be derived. DeepLIFT obliges these axioms by its very nature, thus providing attributions that are simultaneously true and stable.

Ancona et al. (2018) [201] thereafter put forward DeepLIFT together with Integrated Gradients and Layer-wise Relevance Propagation, revealing that when using the Rescale rule, DeepLIFT may be interpreted as a discrete approximation of Integrated Gradients.

Pors: Handles vanishing gradient issue; produces consistent, complete attributions; efficient for deep nonlinear models, provides meaningful and intuitive feature importance maps.
Cons: Depends on chosen reference baseline; may produce unstable non-differentiable activations; less generalizable for complex architectures (e.g., attention-based)

5.2.4. Perturbation-based Methods

Making alterations to inputs is one of the most straightforward methods of getting the explanations of complex models' predictions. These are the methods that Samek and Müller (2019) [204] and Montavon et al. (2018) [195] have explained, which evaluate the effect of each input feature by gradually and systematically making controlled changes (perturbations) to input data and then observing the variation of the model's output as the change happens.
The whole concept is predicated on the assumption that the features that matter the most are those whose exclusion or modification brings about a drastic change in the predicted output of the model.

Common examples include:
Occlusion Sensitivity Maps: which cover certain parts of an image to find out their contribution to the score of the prediction [205].
Feature Masking: where importance is assessed by replacing or mixing up certain features.
Prediction Difference Analysis (PDA): which does conditional sampling to completely strip information from given input areas but still retain realistic data distributions [204].

These methods are not attached to a specific model in that they are determined only by the input-output behavior of the model and not by its internal parameters, such as gradients or activations. Nonetheless, they are computationally expensive, because the model must be computed several times for different perturbations. In addition to that, sometimes perturbation methods can alter the input distribution, resulting in a biased explanation if the perturbation is too unrealistic [202].

Pros: Simple and intuitive; model-agnostic; provides direct insight into model sensitivity.
Cons: Computationally expensive; sensitive to perturbation design; can produce biased results if perturbations are unrealistic.

5.3. Conclusion

Model-based XAI techniques integrate interpretability in the very core of their construction, thus providing reasoning that is inherently transparent and can sometimes be visualized, inspected, or mathematically traced. Such models, which include not only linear and rule-based ones but also prototype-driven and interpretable neural networks, are greatly suitable for those high-stake domains that require explainability by design, for instance, healthcare, finance, and policy making.

On the other hand, model-agnostic XAI techniques act like external exploratory tools that are implemented after the model is trained. They can explain the behavior of any model through approximation, visualization, or attribute feature contributions, which makes them more adaptable yet sometimes less true to the actual internal reasoning of the model.
The understanding of both perspectives constitutes the conceptual background of this thesis and supports the subsequent analysis of personalized, adaptive, and human-centered explainability frameworks.

6. Combining Adaptive Three Dimensions and Gaps

In this chapter, the results of the systematic review are synthesized in terms of how the selected publications address the three core research questions (RQ1–RQ4). The evaluation table, which summarizes human factors, task context, and model-specific considerations for the 43 studies, provides the analytical base for this chapter. This table is particularly important, as it not only shows the extent of studies in each aspect but also highlights the disparities between these aspects in terms of previous research.

The gaps that were seen, mainly the few cross-dimensional combinations and the lack of adaptive explanation frameworks, are signs of the fragmentation of the present XAI research. The major concern is that most of the studies are based on only one aspect of explainability, which in turn opens massive areas for deep integration across the three perspectives of human, task, and model.

This chapter not only builds on but also links the limitations that appear with the three-dimensional Adaptive XAI model proposed by the thesis. By putting the strengths and weaknesses of literature onto this model, the chapter elaborates on why a flexible, multi-dimensional approach is a prerequisite for the provision of properly individualized explanations.

The chapter finally concludes by writing down the future research directions. Among these are the development of adaptive mechanisms that can adjust to the changes in user profiles, task requirements, and model behaviors; more extensive empirical studies; and the formulation of conceptual frameworks that can realize true adaptability in applications.

6.1. Published papers' Evaluation

After reviewing all the publications, some of them got eliminated because they didn't talk about explainability or XAI at all. How papers excluded are represented in the PRISMA (Figure 1). The rest of the papers that mentioned XAI were subjected to analysis, and the results are presented in Table 5.

The evaluation in this table was done considering the Adaptive XAI concept, which lays down different levels of adaptivity. The evaluation of each paper in the Human, Task, and Model dimensions, as well as the integration of these dimensions, was done.

The levels of adaptive consideration in every paper were classified as Yes, Partial, or No, as detailed below. These categories demonstrate the level of adaptive explainability each study integrates. Each paper was also evaluated individually, and the results can be found in the following text.

Yes: the paper explicitly mentions adaptive mechanisms — meaning the explanation or model changes dynamically according to the user, task, or context. Shows actual personalization or dynamic adjustment.

Partial: The paper discusses Human, Task, or Model aspects but lacks true dynamic adaptation — it only conceptually mentions them or uses fixed explanations. Indicates awareness of personalization factors but does not implement adaptive behavior.

No: The paper only provides general or technical explanations of XAI or interpretability, with no adaptive or personalized aspects. Descriptive only — not adaptive or context-aware.

Table 5 – Human, Task, and Model Adaptive Evaluation

	Year	Author	Human factors	Task context	Model type	Combined Dimensions
P1	2025	Ouifak, Hafsaa, Idri, Ali	No	Partial	Yes	No
P2	2025	Arnob, Arjun Kumar Bose, et al.	No	Partial	No	No
P3	2024	Wankhede, Vidhi, et al	No	Partial	Partial	No
P4	2024	Wang, Ziming, et al.	Partial	Partial	Partial	Partial
P5	2020	Hatwell, Julian, et al.	No	Partial	Yes	No
P6	2025	Gong, Mingming, et al.	No	No	No	No
P7	2022	Degen, Helmut; Ntoa, Stavroula	Partial	No	No	No
P8	2025	Pagliaro, Antonio	No	Partial	No	No
P9	2025	Moustati, Imane; Gherabi, Noreddine	Partial	Yes	No	Partial
P10	2025	Benitez, Ian B., et al.	No	Partial	No	No
P11	2022	Séroussi, B., et al.	Partial	Yes	No	No
P12	2024	Subramanian, Harishankar V., et al.	Partial	Yes	No	No
P13	2025	Wiratsin, In-on, et al.	Yes	Partial	Partial	Partial
P14	2025	Ibrahim, Muhamme, et al.	No	Yes	Partial	No
P15	2024	Aljawawdeh, Hamzeh, et al.	Partial	Partial	No	No
P16	2025	Nag, Prashant Kumar, et al.	Partial	Yes	No	No
P17	2019	Anjomshoae, Sule, et al.	Partial	Partial	Yes	Partial
P18	2021	Rayz, Julia	No	Partial	Yes	No

P19	2024	Tejasvi, Koti, et al.	Partial	Yes	Partial	Partial
P20	2025	Ozay, Dervis, et al.	No	Partial	No	No
P21	2025	Deng, Zhe; Torim, Ants, et al.	No	Yes	Partial	No
P22	2025	García López, Iván Miguel, et al.	No	Partial	Yes	No
P23	2024	Nah, Fiona Fui-Hoon,et al.	Partial	Partial	No	No
P24	2025	Coman, Adela, et al.	Partial	Partial	No	No
P25	2020	Kumar, Vivekanandan, et al.	Partial	Partial	No	No
P26	2022	Farhaoui, Yousef, et al.	No	Partial	No	No
P27	2024	Qi, Jun; Yang, Bo	No	Yes	No	No
P28	2025	Dehal, Ramandeep Singh, et al.	No	No	Partial	No
P29	2024	Zaphiris, Panayiotis; Ioannou, Andri	Partial	Yes	No	No
P30	2023	Oyebode, Oladapo, et al.	Partial	Yes	No	No
P31	2025	Tekli, Joe; Gamper, Johann, et al.	No	No	No	No
P32	2020	Pandl, Konstantin D., et al.	No	No	No	No
P33	2022	Djeddi, Chawki	No	No	No	No
P34	2023	Lazaar, Mohamed	No	Partial	No	No
P35	2022	Watson, Richard T., et al.	Yes	Yes	Partial	Partial
P36	2020	Grossberg, Stephen	Yes	Partial	Yes	Partial
P37	2024	Durango, Iván, et al.	Partial	No	No	No
P38	2025	Aravazhi, Prasanna Sakthi, et al.	Partial	Yes	No	No
P39	2025	Ratul, Qudrat E Alahy, et al.	Partial	Yes	No	No
P40	2025	Cappuccio, Eleonora, et al.	Partial	Partial	No	No
P41	2021	Chromik, Michael; Butz, Andreas	Yes	Partial	Partial	Partial
P42	2025	Budhkar, Aishwarya, et al.	No	Yes	Partial	No
P43	2025	He, Gaole, et al.	Yes	Partial	Partial	Partial

P1: This research provides a comprehensive overview of fuzzy logic-based explainable models and emphasizes the interpretability benefits of rule-based systems. Nevertheless, the study does not include adaptive personalization or user-specific explanation strategies. Even though it covers different application domains, the explanations are not restricted to the task contexts or user profiles. The review leans towards model-centric interpretability instead of advocating joint Human–Task–Model adaptation. Hence, while it adds to the model side of XAI, it does not provide adaptive explainability mechanisms and consequently leaves large gaps in human-aware and task-adaptive explanation design.

P2: The SLR that is concerned with IDS presents an overview of different security environments, such as IoT and cloud systems, but it does not yet provide an explanation

mechanism for such diverse adaptive environments to different user skills or roles. The focus is on detection precision, system scalability, and toughness rather than on any form of collaborative or adaptive explainability. There is no indication of Human–Task–Model integration, nor is there any user-tailored personalization. Thus, the contribution remains outside the adaptive XAI category, thereby signaling a lack of ability to make adaptive explanations available for security analysts, or to provide context-sensitive intrusion responses.

P3: The paper provides a detailed review of the status of gait analysis systems used in two main sectors, i.e., health-related and biometric authentication applications, highlighting the use of AI techniques like CNNs and hybrid deep architecture. It has a clear context of a given task (health & biometrics movement analysis), but does not allow for adaptive or personalized explanations, nor does it mention any of the XAI methods beyond the standard interpretability issues. It laid more stress on detection accuracy in the technical aspects than on user-centered explanation design; that is, the task dimension was discussed, but the human and adaptive explanation layers were missing, and model interpretability was not closely linked to the explanation method. There is a critical gap in the lack of adaptive explainability strategies specifically designed for medical professionals, physiotherapists, or different rehabilitation tasks.

P4: By using the Whom-When-What-How framework, this paper implicitly illustrates the necessity of the human, task, and model dimensions in XAI. It depicts the user's expertise, the timing of the decision, the aim of the task, and the structure of the model as factors that affect in different ways the resulting explanations. Nevertheless, it aids the theorist rather than the practitioner with its conceptual guidance instead of providing operational adaptive XAI mechanisms. The dimensions are covered at a theoretical level, and partial integration exists, but no empirical or practical adaptive systems have been implemented. The gap is in the need to realize this conceptual model into actual adaptive XAI pipelines and assess their performance across various user types and real-world tasks.

P5: This paper introduces the Ada-WHIPS, a suitable enhancement of AdaBoost for healthcare classification under the premise of being interpretable. It makes model-level explainability gain through rule-based and feature-importance-style explanations. However, the proposed method does not customize explanations according to user expertise, the

complexity of the task, or the clinical role. The strength of the project is in model-centric explainability, while it is weak in human-adaptive and task-specific adaptation. The critical factor is a priori the mismatch of explanation modes that are designed for the different clinicians (e.g., specialists versus interns) or clinical situations. Thus, although the model aspect has been fully considered, no interaction between the factors of human and task has appeared, indicating a deficiency in adaptive and personalized clinical XAI.

P6: The present volume of conference proceedings is mainly concerned with improvements of fundamental AI algorithms and does not consider adaptive explainability at all. There is no discussion of any of the three aspects of adaptive XAI: human personalization, task-aware explanation, or model-driven explanation tailoring. Consequently, the work is limited to one dimension only, signifying a major gap in the understanding of how adaptive and user-centric mechanisms of explainability can be built into the broader AI innovation research.

P7: The present volume claims affiliation with HCI, but it is not clear from the text that it concerns adaptive explainability. The authors refer to principles of human–AI interaction but do not specify how those would lead to the development of personalized or cognitively-adaptive explanation mechanisms. The concept of concrete task-context adaptation or model-specific explanation tailoring is absent as well. Thus, the three adaptive XAI dimensions remain unintegrated, revealing a gap in linking HCI insights with adaptive XAI theory and implementation.

P8: The review driven by economic theory compares the performance of AI with that of market efficiency, but the matter of explainable or adaptive AI is not discussed. It does place AI amidst the financial task environment but lacks to presenting the task-specific, user-personalized, or model-adaptive explanation mechanisms. Subsequently, it does not play a part in the integration of the three dimensions of adaptive XAI and highlights the lack of adaptive explainability research on financial system evaluation.

P9: The paper connects behavioral psychology and automated trading, thereby recording human decision biases in financial tasks. Nevertheless, the research does not use adaptive explainability; the human factors are treated as input signals and not as personalization goals. Though the task context is clear-cut, the model dimension and adaptive explanation

mechanisms are missing. Thus, the study resonates with adaptive XAI via human-task insights but doesn't extend to the model layer or produce adaptive explanations.

P10: In this systematic review, the AI applications in power system protection are investigated, but the issue of explainable or adaptive AI strategies is not touched upon. Human factors and model dimensions are not there, task relevance is specific to the domain but not personalized, and although model types are referred to, they are not examined through an XAI lens. Therefore, the integration of human, task, and model dimensions does not take place, thus revealing a gap in the application of adaptive explainability to AI systems in safety-critical infrastructures.

P11: The paper intends to discuss the issues of trust, transparency, and ethical considerations regarding AI in healthcare, demonstrating the necessity of humans as the main decision-makers and ensuring by the AI systems trustworthiness, as the main supporters of others' decisions. It does admit human factors like doctors' trust and the security of a decision taken but it does not provide or assess personalized explanations. The paper sets the context of its work very clearly as clinical decision-making, but it does not mention any specific model-based XAI techniques or even ways of adapting them. Thus, it does not include the human, task, and model dimensions together, nor does it suggest adaptive XAI mechanisms. The other side of the coin is that it lacks user-tailored explanations, and through the design of adaptive interactions, it has not served the needs of the medical stakeholders.

P12: The present scoping review deals with the necessity of explainability for human-AI cooperation in the healthcare domain and puts forth the claims of human participation, trust, cognitive load, and communication needs as key factors. Though it nicely depicts the specific environments in medical teams, it still does not reveal the way explanation strategies are adjusted to different user roles or levels of expert knowledge. The model aspect is not thoroughly investigated, and the adaptive XAI strategy does not materialize in the context of the paper. The primary gap is the non-existence of linkage between user roles, model behavior, and explanation adaptivity, therefore raising the issue of future possibilities of clinical teams receiving tailored collaborative XAI through the passage of time.

P13: This systematic literature review (SLR) assesses the performance of different XAI methodologies with respect to the dimensions of human trust, interpretability, and user

perception. It outlines the relationship between a person's confidence and the different explanation techniques, which makes it strong in the human dimension. At the same time, it fails to perform a systematic and detailed analysis of task-specific or adaptive explanation demands, only treating the domains generally. Meanwhile, a model analysis is done, but it remains at a descriptive level only. The research has not integrated the integration of Human × Task × Model into an adaptive explanation framework. The gap is the absence of adaptive personalization and task-aware explanation tailored to user needs.

P14: The current review on AI technologies applied to smart-grid stability presents the aspects of system performance, optimization, and reliability as the main factors its concern. The paper does not cover topics like human interpretability or personalized explanation mechanisms. The task domain is well-defined, but there is almost no mention of explainability or adaptive methods. So, one can say that the paper has technical merits, but it does not help in integrating or developing adaptive XAI research fields. A significant gap remains in applying human-assisted and adaptive explainability techniques to high-risk energy systems.

P15: In this paper, the authors discuss the ethical and trust issues related to human-computer interaction systems in connection with software engineering. They point out the human sides of the issue, such as fairness and openness, but do not go further to present the case for personalized or adaptive explanations. Besides, the task context still being very broad, no model-specific or adaptive explanation strategies were suggested. Therefore, the integration of Human × Task × Model is not there, and the principal adaptive XAI gaps still prevail because of linking the ethical HCI design issue with the area of personalized explainability.

P16: The work discusses AI methods for assessing patients' emotions and cognition in health care environments, emphasizing the potential of recognizing and using these aspects to create more intelligent systems. Although this is a human-centered approach, the human factors are still considered merely as data inputs rather than making the explanations adaptable through their being the reason for the explanations; thus, the paper does not touch on adaptive explanation mechanisms. The paper firmly establishes its work in the medical field that has an impact on the task factor; however, it overlooks the need for XAI-specific model interpretability and does not connect the dimensions of human, task, and model

together. As a result, the paper has an indirect contribution to human-centered AI and no contribution to Adaptive XAI.

P17: This paper presents a review of XAI methods in the fields of autonomous agents and robotics, which are mostly related to trust, transparency, and interaction between humans and intelligent systems. The issue of human factors is recognized; however, there are no proposals for any adaptive personalization of explanations. The robotics tasks are taken into consideration, but the task context does not dynamically shape the explanations. Strong insights into model-based agent explanation strategies are provided by the study, but the human and task dimensions are not fully integrated. Thus, this research has merely contributed to Adaptive XAI, showing a certain level of interaction awareness without the adaptive coordination across dimensions.

P18: The paper delves into fuzzy logic and neuro-fuzzy methods for explaining algorithms, and it continually draws attention to the transparency and to the reasoning through rules that are the inherent characteristics of fuzzy systems. Nevertheless, this interpretability is not customized according to user types or cognitive needs, which implies that there is no adaptive human dimension. The authors claim applications in various domains, which are more illustrative than task specific. Their main contribution is model-driven interpretability, which does not consider human and task dimensions. This is why their contribution is aligned with traditional interpretability rather than Adaptive XAI.

P19: This paper provides a comprehensive survey of XAI applications in the field of customer-centric e-commerce systems. Customer experience and understanding are present in the paper, but at the same time, no user modeling by means of adaptive mechanisms is suggested to fit the explanations. The context of the task (recommendation, buying behavior analysis) is very clear, but the style of explanation does not change depending on the task or the type of user. A few different XAI techniques are discussed, yet the connection between the type of model and the form of explanation is still under-explained. Thus, the paper reveals a domain-focused XAI but falls short in achieving multidimensional adaptive integration.

P20: This study employs network and contextual techniques to unveil customer perceptions at the intersection of AI, big data, and CRM. On the contrary, the study emphasizes neither

explainability nor the delivery of human-centric and adaptive explanation elements. The paper merely cites CRM tasks but does not investigate the possibility of varying explanations according to user roles or decision contexts. Therefore, this study is not within the XAI sphere and does not entail any contribution to the Adaptive XAI frameworks.

P21: The review analyzes the generative AI's role in IoT-based intrusion detection systems, and at the same time, discusses the task-specific challenges and security contexts. It covers very well model development and domain-specific use cases, but it does not introduce personalization or human-adaptive explanation methods. The research is more focused on system performance and automation than on user understanding or adaptive interpretability. So, it contributes to the task and model dimensions but lacks human-centered adaptation, which results in no integration across the three adaptive XAI dimensions.

P22: This review is systematic and deals with the capabilities and applications of open-language models, putting a strong emphasis on model architecture and interpretability limitations. Multiple domains are covered, but the analysis remains general and does not include task-adaptive explanation tailoring. The same goes for the study, which does not point out any human-adaptive or personalized explanation strategies. Consequently, it provides the model and somewhat the task dimension with information, but it still lacks an integrated adaptive XAI perspective, thus leaving a big gap in user-centric and context-aware explainability.

P23: The editors of this conference volume have gathered a wide range of contributions coming from different disciplines but having one common interest in AI, among them, the human-computer interaction problem. The book is organized into human-centered issues, such as trust and usability, but does not include adaptive explanation strategies that are tuned to the needs of different user groups. The discussion of task contexts is broad rather than task-specific, and model explainability is left out of the picture. As a result, the research is only partially human and task factor driven and does not link them with model-based adaptation, thus missing the alignment with Adaptive XAI.

P24: This proceedings compilation showcases the Near to the State-of-the-art Research in HCI, and several papers deal with the Ease of Understanding topics in Human-Computer Interaction (HCI), one way or another. While users' unmet needs, trust, and transparency are

mentioned as the "must" issues, the authors do not imply that the said situation can lead to personalized explanations through Adaptive Personalization of Explanations. The same is true for the consideration of the task; it is present but not tailored to variations across tasks explicitly. The same dimension of the model is not given separate attention. Overall, this source documents human and task issues without the depth of integrated 3D adaptive XAI alignment.

P25: This work deals with Intelligent Tutoring Systems and concentrates on the aspect of learner modeling and the giving of personalized instruction as the key to the system's success. It considers the learner's knowledge level (human factor) and operates within a well-defined task context, i.e., education, but it does not use the explication of adaptive explainability. It basically the system uses personalization of learning content instead of personalization of explanations as it adapts in the learning process. No XAI models or explanation frameworks are systematically examined, and the integration of Human × Task × Model perspectives extrapolated from an adaptive XAI viewpoint is also absent. Thus, even though it is relevant to adaptive systems, it does not facilitate the field of adaptive explainable AI. Thus, while relevant to adaptive systems, it does not contribute to adaptive explainable AI research and reveals a gap where educational ITS systems lack transparent and tailored explanations.

P26: The proceedings of the ICAISE 2024 mainly deal with technical breakthroughs in artificial intelligence for smart homes and IoT. The collection of papers introduces background scenarios of intelligent systems but no mention of explainability or personalization as an adaptation based on the user. An integration of the three dimensions (human, task, and model) was not found, proving that the work is outside the main scope of Adaptive XAI. Thus, reference only helps in understanding the domain and does not participate in the development of adaptive explainability frameworks.

P27: This is a review that details the applications of IoT and big data in the healthcare industry; it discusses scenarios like remote monitoring of patients, telehealth, as well as medical diagnostics. However, the paper does not go beyond establishing a well-defined task domain to the human-centered factors or adaptive explanation generation mechanisms. User expertise, cognitive demands, and trust calibration are not discussed at all, nor is there any analysis of XAI models or how the architectural characteristics influence the various forms

of explanation. As a result, the research, although contributing to the understanding of the domain, does not involve the Human × Task × Model dimensions that are essential for Adaptive XAI and, thus, leaves behind quite a lot of ground in terms of personalized and context-aware explanation mechanisms.

P28: The present research examines the ways in which Knowledge Graphs (KGs) and Large Language Models (LLMs) interact with each other, mainly looking at how combining symbolic structured knowledge can lead to more reliable and better reasoned LLM outputs. The paper works towards model-oriented explainability by stressing the role of structured knowledge in grounding but does not bring in human-centered personalization or adaptive explanatory systems. The work does not define the user's expertise, cognitive needs, or trust issues, nor does it provide explanations aligned with specific users' tasks. So, although the model angle is to some extent covered via KG-supported transparency, the human, task, and model perspectives that are required for adaptive XAI are not integrated; thus, the research highlights the gap between knowledge-enhanced language models and their personalized explanatory systems.

P29: This research paper brings in learning and collaboration technologies in educational settings, putting emphasis on user experience, engagement, and collaborative behavior, thus only touching the human aspect of XAI at a general cognitive-interaction level. It extensively describes AI use cases in education that cover the task dimension by examining the ways in which AI supports collaborative learning scenarios. However, it does not deal with model-level explainability nor the generation of adaptive explanations. More critically, the research does not unite the human, task, and model perspectives in an adaptive XAI framework. It is lacking the explanation-tailoring mechanisms that consider learner's expertise, behavior, or changing learning needs. As a result, while the work contributes to understanding human interaction and task design in educational AI, it leaves an open gap for future research on personalized, context-aware, and model-driven adaptive explanations in learning systems.

P30: The research presented in this article explores machine learning systems that are both adaptive and personalized within the health and wellness domain and focus heavily on user needs, behavior personalization, and context adaptation. The authors meaningfully discuss the factors of human personalization and task-specific contexts, but their adaptive mechanisms are confined to the ML system's behavior and output, rather than the

explanations produced. The research does not investigate XAI techniques, nor does it consider how the properties of the model affect the process of explanation. Thus, while the project is in line with user-adaptive AI, it does not incorporate the Human × Task × Model triad for Adaptive XAI, hence leaving a major gap between personalized decision support and personalized explainability for end-users in the healthcare systems to be filled by the latter.

P31: This paper is a collection of short communications from the area of data processing and information systems, chiefly comprising new topics in the aforementioned areas, such as machine learning and data-driven computation. Parts of the document occasionally point to machine learning and data analytics methods, but the conversation still must go primarily in the direction of data architecture, system engineering, and computational frameworks as the ones posited by Adaptive XAI. The three main XAI dimensions (Human – Task – Model) are not explicitly discussed. Thus, even though the conference proceedings dealt with AI and data-related topics, there was no attempt made to bring together the Human - Task - Model dimensions in developing an adaptive explainability framework. The work in question stays outside the boundaries of Adaptive XAI research.

P32: focuses on the intersection of Artificial Intelligence (AI) and Distributed Ledger Technologies (DLT), presenting a scoping review that highlights the architecting of the two technologies together, security, transparency, and decentralized learning as the main areas. The paper covering these topics in a technical and system-design manner is very detailed, but still, the issue of user-friendly AI is not approached, and solutions like human-adaptive explainability, user-specific personalization, or cognitive-aligned explanation strategies are not provided. The context of the task is only briefly mentioned at the level of the domain, like, secure data exchange, decentralized decision-making, without considering adaptive explanation. Similarly, although the different AI components are cited, there is no mention of any model-specific explanation mechanisms or XAI techniques being analyzed. Thus, the authors do not blend the human, task, and model components that are necessary for Adaptive XAI, and they point out the lack of alignment of interpretability views with AI-DLT research.

P33: The conference series' main theme was pattern recognition and AI, with only a minor part dedicated to the discussion of explainable machine learning. The Model aspect was

looked at through the opening of classifiers and recognition models but did not include a structural analysis of their explainability. The Task aspect was represented in different PR sectors, but without any alteration. The Human aspect was almost completely overlooked, as there were no interactions or personalization's considered. There did not seem to be any attempt to bring the three aspects together or to create adaptive XAI frameworks.

P34: The papers mainly focus on big data and IoT, and some briefly mention explainability, but neither systematically nor adaptively. The Model dimension has limited references to big data/IoT ML models. The Task dimension appears in the context of IoT and smart systems, but there is no personalization of explanations. The Human dimension is missing, as the focus remains on system-level performance. Therefore, no three-dimensional adaptive alignment is present.

P35: Although the authors Watson et al. (2022) propose a conceptual role model for personalized explanation in auditing, their method only considers the dimensions of XAI combined two at a time—connecting user roles with either tasks or model transparency in one hand. On one other hand, my thesis advocates that the Human, Task, and Model dimensions be integrated together through a unified Adaptive-XAI framework. This framework is not limited to static or role-based personalization but seeks dynamic adaptivity where user characteristics, task context, and model behavior interaction cause the explanations to adjust continuously. Thus, after Watson et al. have provided a domain-restricted and two-dimensional view of personalization, my research has proposed a generalizable and fully integrated three-dimensional model of adaptive explainability across domains.

P36: Grossberg (2020) presents an example of a system that is adaptive in learning but not in explaining its decisions. This is where biologically inspired cognitive models, notably the Adaptive Resonance Theory (ART), come into play. By the ART, human cognitive and emotional mechanisms are explicitly modeled—perception, attention, learning, motivation, and reasoning. These mechanisms establish a direct link between human factors and the processes of explainability and adaptive intelligence. The task contexts are delineated in cognitive and perceptual domains but not specified for explanations. The paper does offer excellent model-based interpretability through networks of ART; however, it does not couple the human-task-model triad into an adaptive explanation framework. Thus, it

theorizes and contributes to an understanding of cognition-aligned explainability but does not operationalize Adaptive XAI. Therefore, the work is foundationally adaptive but not an Adaptive-XAI system in the sense defined by this thesis.

P37: The research conducted, human-data-AI collaboration, has as its main purpose the facilitation of the human-GenAI co-adaptation process. It brings to the fore human factors like trust, engagement, cognitive alignment, and interaction fluency, and recognizes the importance of task-oriented collaboration scenarios. It does not, however, present any domain-specific adaptive explanation mechanisms. The paper refrains from discussing the model-level explainability structures or the frameworks that adjust the explanations based on the user's knowledge or the complexity of the task, even though it uncovers certain adaptive interaction principles. Thus, the unifying of the adaptive XAI dimensions is only conceptual and not systematic, with a major lack in the areas of personalization and model-adaptive explanation techniques' operationalization remaining.

P38: The paper reviews AI adoption trends in the field of clinical medicine and in the areas of decision support, diagnostics, and workflow optimization. It points out specific healthcare situations where AI can be used and mentions trust, reliability, and ethics at a conceptual level, while not giving any user-personalized or adaptive explanation. The focus is on clinical scenarios and system challenges rather than on interactive explanation models or on model-driven adaptation. Thus, although it indirectly refers to human and task aspects, there is no structured assessment of how explanations should vary according to roles (e.g., doctor vs. patient) or clinical contexts, which leaves a considerable gap for adaptive XAI.

P39: The paper gives an overview of AI applications in the educational sector, with the emphasis being on adaptive learning systems and student support systems. It goes into detail to analyze the task dimension within the educational domain, which is the learning personalization and student-modeling. Nevertheless, it refers to learner characteristics but does not deploy or investigate personalized explanations. It focuses on content-adaptation rather than explanation-adaptation. The model dimension is very slightly dealt with in respect of explainability. There is no unification of the human-task-model perspectives for adaptive XAI. Thus, the paper points to a gap in the process of translation from adaptive learning to adaptive explanatory systems.

P40: A review paper on Explanation User Interfaces is presented, and human factors were discussed, such as usability, trust, interpretability, user understanding, and cognitive load; however, it did not incorporate the dynamic explaining mechanisms that are user expertise, role, or goals-based. These factors were pointed out as the evaluation dimensions rather than the adaptive processes. The research looks at the different methods of explanation presentation throughout various fields, but it does not consider the task-adaptive mechanisms in detail, nor does it look at how the specific semantics of the task variably affect the users' needs for explanations. Moreover, the characteristics of model-level explainability are not systematically observed. It emphasizes the integration of the Human, Task, and Model dimensions as a very important step in the development of effective explanation interfaces and indicates this gap explicitly as future research. Nevertheless, it does not implement this integration within a single adaptive framework.

P41: The author of this paper focuses a lot on human-AI interaction, giving principles for the effective explanation of user interfaces and stressing such factors as trust, cognitive effort, and mental models. Different task scenarios are referred to, but no direct methods for task-adaptive personalization are given. The paper acknowledges that model behavior has an impact on the explanation interfaces but does not offer any frameworks that integrate model characteristics with human and task dimensions. It partially backs up adaptive XAI, thus lacking the full integration across all three dimensions.

P42: This research paper surveys the applications of XAI in bioinformatics and the biomedical fields giving a deep insight into the tasks in the respective domains and the models used in omics studies. Besides discussing interpretability problems, the paper overlooks user-adaptive factors and does not analyze how different biomedical user roles or expertise levels affect the explanations. Therefore, despite skimming through task and model parameters, it falls short of human-centered customization and does not combine all three Adaptive-XAI dimensions.

P43: This study proposes conversational XAI as a means of enhancing human-AI collaborative decision-making. It extensively deals with human factors by means of adaptive dialogue-driven explanations. Nonetheless, the evaluation is conducted on only one decision-support scenario, which limits the task-adaptive generalization. Model-level considerations are lightly touched upon and not deeply analyzed concerning explanation

structure. Thus, it partially moves forward with adaptive XAI through human-centric dialogue mechanisms but does not fully integrate task-dependent or model-aware adaptation.

6.2. Coverage of Adaptive 3D in the Reviewed Papers

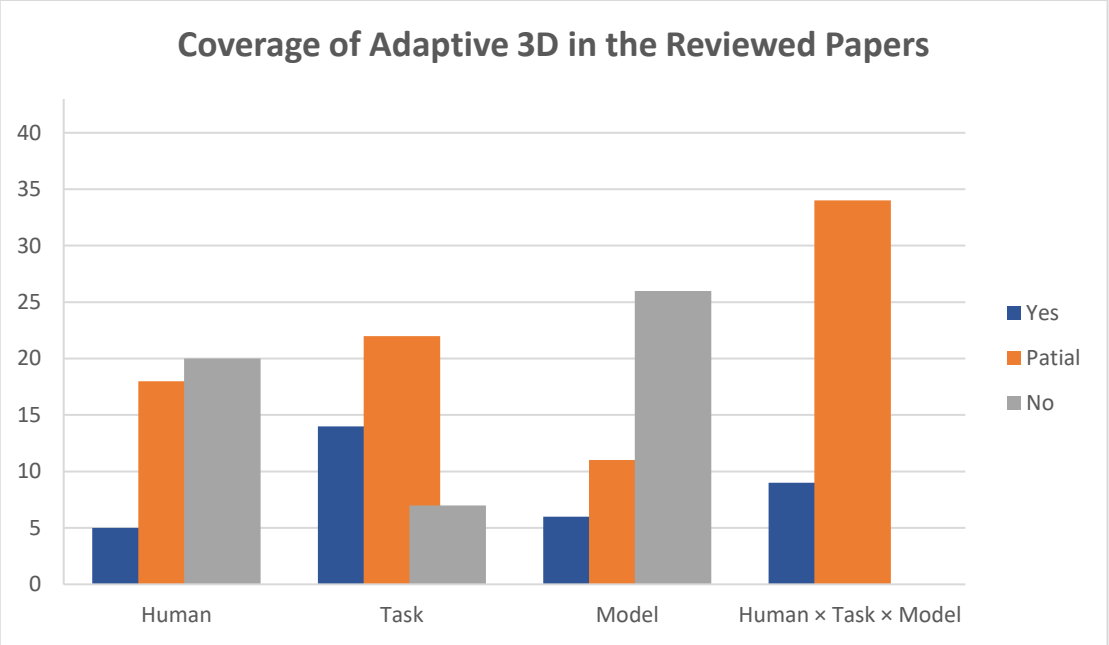

Figure 5 - Coverage of Adaptive 3Dimensions in the Reviewed Papers

Figure 4 illustrates the coverage of the three Adaptive XAI dimensions—Human, Task, and Model—and their combined integration (Human × Task × Model) over the 43 reviewed publications. The number of studies is indicated in three categories for each bar group: Yes (adaptive), Partial (conceptually mentioned but not dynamically adaptive), and No (no adaptive consideration).

In the chart, the Human dimension is represented by 5 papers (11.6%) that were assigned to the fully adaptive category, while the Task dimension is represented by 14 papers (32%) that discussed adaptive or task-specific contexts. The Model dimension was covered in 6 papers (13.95%), mainly dealing with the model-based explainability and interpretability aspects.

The Human × Task × Model category comprises 9 papers (20.9%) that consider the Human, Task, and Model dimensions together, indicating partial integration rather than full unification.

The findings represented in the chart reveal that the Task dimension got the most overall attention, as over 30% of the studies looked at adaptive considerations in fields like healthcare, education, and cybersecurity. The Human dimension was only rarely addressed; just a few publications referred to human-adaptive factors like user expertise, cognitive effort, or trust calibration. The Model dimension showed the least adaptability, since most of the work concentrated on static interpretability (e.g., transparency or feature attribution) that did not connect model behavior with human or task adaptation.

The last group (Human × Task × Model) shows that even though several papers tried to merge the three dimensions, none of them were fully integrated. Most of these efforts were still at the theoretical stage — recognizing the requirement for Human × Task × Model alignment — but not applying adaptive systems in reality. Thus, Figure 5 gives a visual summary of how often each adaptive dimension, or their combination, was found in the literature review.

6.3. Findings

Most studies usually focus on one dimension only, either the technical interpretability of models (e.g., transparency, feature attribution), the human-centered aspects (such as usability, trust, and comprehension), or the task context (such as domain characteristics or risk level). Only a small number of works are trying to put all three—Human, Task, and Model—together within a single adaptive framework.

Although personalization has been a prevalent topic in recent XAI debates, most of the applications are still confined to fixed solutions that assume one explanation method fits all users and all contexts uniformly. Just a handful of studies have investigated how explanations should automatically change depending on user skills, cognitive load, domain needs, or model behavior.

Additionally, current XAI evaluation frameworks mainly focus on measuring trust, interpretability, and satisfaction while rarely evaluating the adaptiveness of explanations in terms of their effectiveness in differing users or contexts. This indicates a methodological shortcoming in the field's ability to comprehensively assess adaptive systems.

Moreover, among the reviews, four articles that were specifically linked to explainable AI did not refer to the concept of Adaptive XAI at all. Even though, to a certain extent, their research can be seen as contributing towards interpretability and transparency, they entirely ignore the concept of adaptiveness or context-sensitive explanation. The total lack of mentioning Adaptive XAI in these studies is indicative of fragmentation in the current situation-picture of the area, where at times the issue of explainability is pursued without connecting it to the issue of adaptivity.

The literature reviewed in total indicates a shift towards the need for such frameworks of explanation that are aware of the context, are personalized, and are model-adaptive—these are systems that can automatically adjust the explanation according to the user's needs, task complexity, and model characteristics. The developments are making the future of Adaptive and Human-Centered XAI brighter, as it is expected to deliver better transparency, usability, and generalization in the real-world application of AI.

6.4. Limitations of Considering Each Dimension in Isolation

The three dimensions—human factors, task context, and model characteristics—each contribute significantly to the development of explanations, but none of them alone is capable to providing the desired level of effective and trustworthy explainability.

Human Factors are not enough by themselves. If the design of explanations is based entirely on user needs and cognitive preferences, it might disregard the limits of the task or the nature of the underlying model. For example, a simplified explanation might be clear to a patient, but if the model in question is a complex neural network in the context of a high-stakes medical diagnosis, neglecting the model and task context could result in misleading or even harmful interpretations.

Task Context cannot stand alone. When the design of explanation is shaped only by the task environment, explanations might meet the standards of professionalism or regulation but be unable to adjust their level to the user’s cognitive ability or to the model’s technical aspects. For instance, in judicial decision-making, transparency might be required by law, but if the explanations are not adapted to the judge's comprehension or the complexity of the XAI model is not considered, the explanations may well become ineffective.

Model Characteristics cannot suffice. If the explanations focus solely on the model's characteristics, they can be technically accurate but practically insignificant. For instance, saliency maps or gradient-based visualizations can be methods that accurately illustrate the working of the neural network in terms of data processing, yet they are of little help to the teacher, the doctor, or the layperson when it comes to the needed actionable insights.

6.5. The Need for Integration

The fusion of the Human, Task, and Model aspects is a must for attaining to be adaptive and meaningful explanation in AI systems. When the three views are taken together, the explanations become not only technically correct but also practically relevant and individually comprehensible.

The Human aspect makes sure that the explanation is consistent with the user's traits like the level of expertise, cognitive capacity, goals, and trust expectations. Explanations made with this viewpoint are human-centered and thus user's mental model and communication preferences are taken into account.

The Task aspect ensures that the explanation can recognize the context, taking into consideration the nature, complexity, and criticality of the task being executed. For instance, in safety-critical areas like healthcare or self-driving cars, the explanations must point out the reliability and risks while in educational systems they should concentrate on clarity and learning support.

The Model aspect provides model-truthfulness — the assurance that the explanation elucidating the AI system's reasoning and does not distort or oversimplify the decision-making process. It makes the explanation linked to the model's internal logic, parameters, and prediction behavior.

Merging these three dimensions in a single adaptive framework makes the explanation process cover the whole picture and be dynamic in such a way that it constantly adjusts to the user's, the task's, and the model's interaction. This wide view not only gets rid of the limitations of treating each dimension separately but also lays down the groundwork for the creating of personalized, context-aware, and model-aligned adaptive explanations.

The combination of these methods allows for the development of state-of-the-art Adaptive XAI systems, where at each interaction the user gets to know things that are correctly, appropriately, and specifically adjusted to their needs during the whole process of communication—thanks to the fact that the explanations are changing along with the context.

6.6. Adaptive Explanation Model

Based on analysis of the available literature, I propose a conceptual model of adaptive explanations (Figure 6) that take into consideration Human (h), Task (t), and Model (m) contexts and treats explainability as a dynamic weighting problem rather than a static relation.

Adaptive explainability can emerge only when all three dimensions are jointly considered, because each contributes a non-replaceable property:

- The Human dimension provides interpretability, trust, usability, and cognitive.
- The Task dimension provides relevance, usefulness, and task-specific appropriateness.
- The Model dimension provides the technical accuracy, fidelity, and completeness.

Missing one dimension leads to making an explanation that is either misleading, irrelevant, or incomprehensible. That is why adaptive XAI should incorporate all three viewpoints and constantly adjust them according to the situation to deliver explanations that are both truthful and significant.

An Adaptive XAI system constantly rebalances the three dimensions according to the context. For example:

- In **educational or user-training situations**, the Human dimension is given more weight.
- In **critical or high-risk areas (such as healthcare**), the Task dimension is considered above all others.
- In **auditing or regulatory situations**, the Model dimension is predominant.

Yet, in none of these cases are the other two dimensions completely sidelined — their impact (w_i) is just made significant relative to the context. Consequently, the merging and dynamic

balancing of all three dimensions is what gives XAI its Adaptive Explainability. The Dynamic Model depicts the gradual movement of the focus of explainability in XAI among the three principal axes: Human (h), Task (t), and Model (m) through the dynamic triangle's vertices.

The explanation is considered adaptive only if:

$$\mathrm{A} = h + t + m = 3$$

Which indicates that all three *perspectives are represented, where* $h, t, m \in \{0,1\}$ *once values of* $\mathrm{A} < 3$ *donate incomplete or unbalanced explanations.*

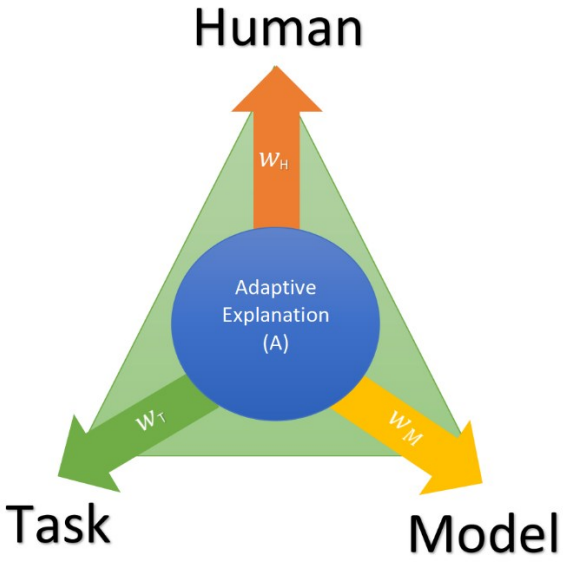

Figure 6 - Dynamic Model of the Human, Task, and Model Dimensions in Adaptive XAI

Each one of the three main axes is corresponded by a corner of the triangle. The apex that goes up the highest shows which axis is the strongest now.

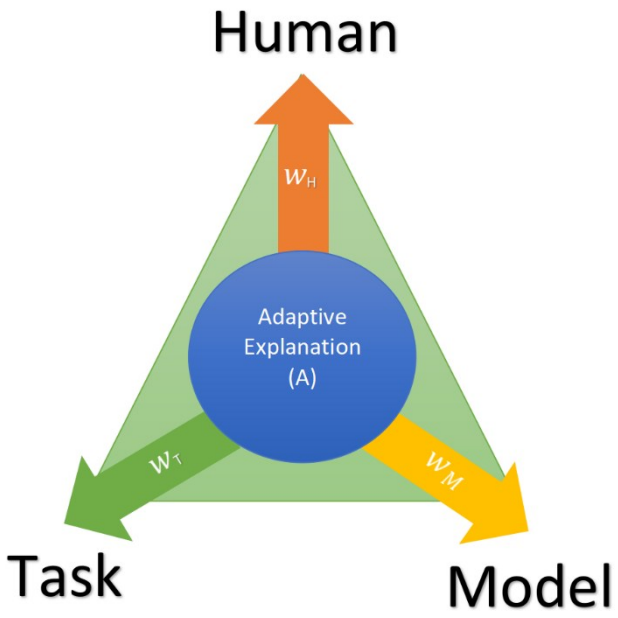

Figure 7 - Adaptive Explainability – Human-Dominant Configuration (w_H highest)

When the Human aspect is the most important one (w_H) (e.g., the case of a user who is a non-expert or when interpretability is an absolute requirement), the triangle will rotate in such a manner that Human becomes the top vertex while Task and Model become the base (Figure 7).

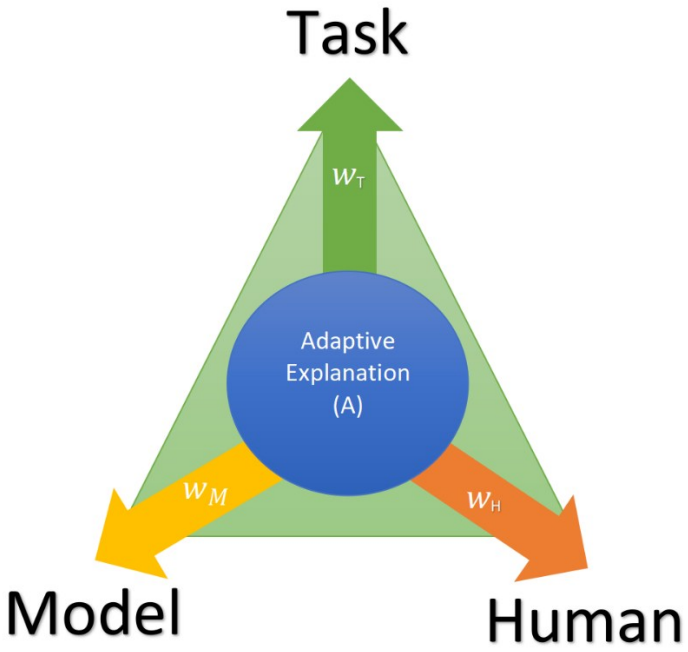

Figure 8 - Adaptive Explainability – Task-Dominant Configuration (w_T highest)

If the Task aspect is the most important one (w_T) (e.g. in a time-critical situation or high-stakes decision-making), then it will be at the top, and the other two will be at the same level below it (Figure 8).

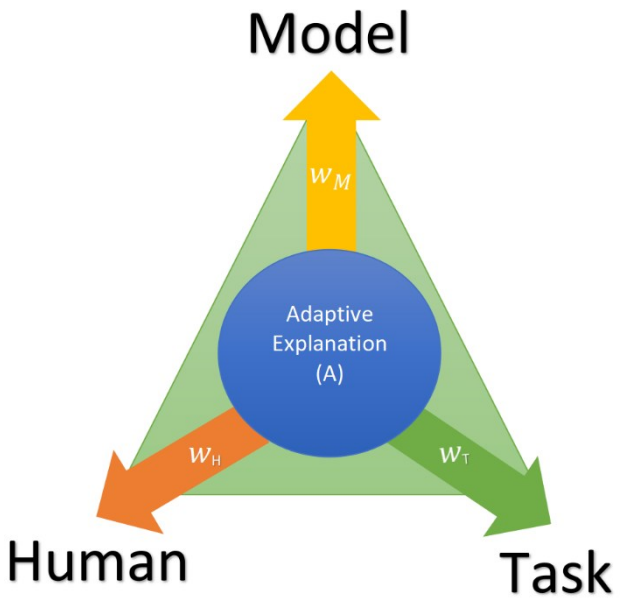

Figure 9 - Adaptive Explainability – Model-Dominant Configuration (w_M highest)

In case the Model aspect is given priority (w_M) (e.g., in auditing or technical validation), then it will be at the top, representing the highest faithfulness and stability (Figure 9).
This rotation illustrates the concept of the adaptive balance between the three dimensions quite effectively. Even though one dimension takes over, the other two remain active and supportive, thus ensuring a comprehensive view while shifting the focus.
The dynamics of the triangle can thus be simply stated as:

When w_i increases, vertex i becomes the apex of the triangle.
If $w_H > w_T, w_M \Rightarrow$ Human-focused explanation.
If $w_T > w_H, w_M \Rightarrow$ Task-focused explanation.
If $w_M > w_H, w_T \Rightarrow$ Model-focused explanation.

Hence, the dynamic triangle is a visual representation of the adaptive quality of XAI — it always embraces all three dimensions, but the emphasis and the direction change according to the situation.

6.7. Future Work & Gap

Based on the findings of this study, several key research gaps exist that could be investigated in future studies.

1. This research has presented a classification of XAI methods that is based on the three dimensions: Human, Task, and Model. Nevertheless, only a partial completion of such categorization took place because of the limited review of publications as well as time limitations. Therefore, researchers in the future should develop more comprehensive taxonomies to continue this process by including the pros and cons along with the real-world applications of each approach.
2. Future research should be based on such organized foundations and focus on developing a dynamic adaptive model that would be capable of intelligently adjusting the explanations according to the various human roles, task contexts, and model types.
3. Future research should concentrate on the development of computational mechanisms that would be able to dynamically adjust the importance of the three dimensions—Human, Task, and Model—depending on the conditions. Lastly, these conceptual foundations should be transformed into practical implementations as well as cross-domain validation studies to assess the performance of adaptive explainability across different real-world environments.

6.8. Conclusion

This article performed a Systematic Literature Review (SLR) on Explainable Artificial Intelligence (XAI) personalized and adaptive technologies, spotlighting human factors, task context, and model characteristics interplay among these three main dimensions. Following the PRISMA 2020 guidelines, a thorough analysis of 43 publications from 2015 to 2025 uncovered that the evolution of AI explainability is passing from static, one-size-fits-all, to dynamic, user-centered, and context-aware adaptive explanations.

The research results indicate that the area of human factors, such as cognitive load, trust calibration, and interpretability, has attracted a substantial amount of research, but still, most of the studies have considered them independently. Only a tiny fraction (around 9%) of the

reviewed literature presented frameworks that accommodate partially three dimensions — Human × Task × Model — in an adaptive way. This observation implies that present-day XAI research lacks comprehensive frameworks to dynamically adjust explanations based on user expertise, task criticality, and model complexity; therefore, still unable to cater to the specific user needs.

In the case of human dimensions, personalization techniques are increasingly acknowledged as the key to the development of trust that is calibrated and reliance that is sustainable. Explanations based on user knowledge, goals, and reasoning styles not only improve understanding but also foster ethical and emotional alignment. With the task dimension, contextual variables such as domain, risk level, and intended user goals determine how explanations should be structured, evaluated, and communicated. Lastly, the model dimension points out that the nature of the AI model — be it interpretable, hybrid, or black-box — imposes limitations on the types of explanations that can be produced and the extent of their adaptability.

The combination of these dimensions makes it clear that adaptive XAI is not just a technological feature but a socio-technical necessity. The real adaptivity comes with the constant rebalancing of human interpretability, task relevance, and model fidelity according to the situation. If any of these aspects is overlooked, the resulting explanations will be of one of these three types: either misleading, irrelevant, or incomprehensible.

Future research needs to develop the integrated adaptive frameworks that dynamically mix all three dimensions.

The present work provides a complete view of the personalized and adaptive XAI landscape. It lays the theoretical foundation for the AI systems of the future that are capable of not only showing what they do but also explaining how and why they change their explanations — thus, promoting more open, reliable, and human-aligned AI.

7. References

[1] K. de Fine Licht and J. de Fine Licht, "Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy," pp. 919-926, 2020.

[2] A. Berrada and M. Adadi, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai),'' IEEE access, vol. 6," pp. 52-138, 2018.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado and et al, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58," pp. 82-115, 2020.

[4] W. Samek, A. Vedaldi, L. Hansen, K. Müller and G. Montavan, "Explainable AI: Interpreting, explaining and visualizing deep learning. Springer," 2019.

[5] Z. Atf and P. Lewis, "Is Trust Correlated With Explainability in AI? A Meta-Analysis," 2025.

[6] O. M. Mastromichalakis, J. Liartis and G. Stamou, "Beyond One-Size-Fits-All: Adapting Counterfactual Explanations to User Objectives," 2024.

[7] K. Porayska-Pomsta, C. Conati and M. Mavrikis, "AI in education needs interpretable machine learning: lessons from open learner modelling, in: ICML," 2020.

[8] S. Kraus, A. Azaria, J. Fiosina, M. Greve and a. et, "AI for Explaining Decisions in Multi-Agent Environments," pp. 13534-13537, 2020.

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," pp. 1-38, 2019.

[10] Q. Vera Liao, D. Gruen and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," 2020.

[11] Y. Zhang, Q. Vera Liao and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," 2020.

[12] V. Beaudouin, I. Bloch and et al, "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach," 2020.

[13] R. R. Hoffman, S. T. Mueller, G. Klein and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," 2018.

[14] S. Ali, T. AbuHmed, S. H. El–Sappagh and et al, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," 2023.

[15] M. Ppage, J. McKenzie, P. Bossuyt, I. Boutron, T. Hoffmann, C. Mulrow, L. Shamseer, J. Tetzlaff, E. Akl, S. Brennan, R. Chou, J. Glanville, J. Grimshaw, A. Hróbjartsson, M. Lalu, T. Li, E. Loder, E. Mayo-Wilson, S. McDonald, L. McGuinness, L. Stewart, J. Thomas, A. Tricco, V. Welch, P. Whiting and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71," 2020.

[16] A. Ashraf, J. Vermeulen, D. Wang and B. Lim, "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda," 2018.

[17] U. Ehsan and M. O. Riedl, "Human-Centered Explainable AI: Toward a Reflective Sociotechnical Approach," 2020.

[18] R. S. Dehal, M. Sharma and E. Rajabi, "Knowledge Graphs and Their Reciprocal Relationship with Large Language Models," 2025.

[19] F. M. Cau and L. D. Spano, "Exploring the Impact of Explainable AI and Cognitive Capabilities on Users' Decisions," 2025.

[20] Z. Wang, C. Huang and X. Yao, "A Roadmap of Explainable Artificial Intelligence: Explain to Whom, When, What and How," 2024.

[21] A. Coman, S. Vasilache, F. Fui-Hoon Nah and et al, "HCI International 2024 – Late Breaking Papers: 26th International Conference on Human-Computer Interaction, HCII 2024 , Proceedings, Part II," Washington, DC, USA, June 29 – July 4, 2024.

[22] V. Borst, A. Riedmann, T. Dege and et al, "WoundAIssist: A Patient-Centered Mobile App for AI-Assisted Wound Care With Physicians in the Loop," 2025.

[23] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," 2017.

[24] V. Kumar and C. Troussas, "Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Proceedings," Athens, Greece, June 8-12, 2020.

[25] E. Cappuccio, A. Esposito, F. Greco, G. Desolda and et al, "Explanation User Interfaces: A Systematic Literature Review," 2022.

[26] M. Chromik and A. Butz, "Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces," 2021.

[27] S. Rhodes and N. Cowan, "Attention in working memory: attention is needed but it yearns," 2018.

[28] M. Gong, Y. Song, Y. S. Koh, W. Xiang and D. Wang, "AI 2024: Advances in Artificial Intelligence : 37th Australasian Joint Conference on Artificial Intelligence, AI 2024, Melbourne, VIC, Australia, November 25–29, 2024, Proceedings, Part I," 2025.

[29] A. Bussone, S. Stumpf and D. O'Sullivan, "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems," 2015.

[30] T. Kulesza, M. Burnett, W.-K. Wong and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," 2015.

[31] T. Kulesza, M. Burnett, W.-K. Wong and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," 2015.

[32] T. Kulesza, S. Stumpf and et al., "Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models," 2013.

[33] J. Zerilli, U. Bhatt and A. Weller, "How transparency modulates trust in artificial intelligence," 2022.

[34] J. Hatwell, M. M. Gaber and R. M. Atif Azad, "Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences," 2020.

[35] I. Durango, J. A. Gallud and V. M. R. Penichet, "The data dance: choreographing seamless partnerships between humans, data, and GenAI," 2024.

[36] K. Anthony Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," 2014.

[37] L. John D and S. Katrina A, "Trust in Automation: Designing for Appropriate Reliance," 2004.

[38] A. Jacovi and Y. Goldberg, "Towards Faithfully Interpretable NLP Systems:," 2020.

[39] I. Moustati and N. Gherabi, "Bridging Behavioral Insights and Automated Trading: An Internet of Behaviors Approach for Enhanced Financial Decision-Making," 2025.

[40] I.-o. Wiratsin and C. Ragkhitwetsagul, "Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review," 2025.

[41] C. Ford and M. T. Keane, "Explaining Classifications to Non-Experts: An XAI User Study of Post-Hoc Explanations for a Classifier When People Lack Expertise," 2023.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," no. https://arxiv.org/abs/1610.02391, 2019.

[43] K. Simonyan, A. Vedaldi and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2014.

[44] A. B. Haque, N. Islam and P. Mikalef, "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research," 2023.

[45] M. Lazaar, "Proceedings of the 6th international conference on big data and internet of things," 2023.

[46] R. Singh, P. Dourish, P. Howe and et al, "Directive Explanations for Actionable Explainability in Machine Learning Applications," 2021.

[47] D. Weld and G. Bansal, "The Challenge of Crafting Intelligible Intelligence," 2018.

[48] H. C. Nwokoye, M. J. P. Peixoto, A. Pandey, L. Pardy and M. Sukhai, "A SURVEY OF ACCESSIBLE EXPLAINABLE ARTIFICIAL INTELLIGENCE RESEARCH," 2024.

[49] J. Schoeffer, M. De-Arteaga and N. Kühl, "Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making," 2024.

[50] J. M. Alvarez, A. B. Colmenarejo, A. Elobaid and et al, "Policy advice and best practices on bias and fairness in AI," 2024.

[51] P. K. Nag, A. Bhagat and R. Vishnu Priya, "Expanding AI's Role in Healthcare Applications: A Systematic Review of Emotional and Cognitive Analysis Techniques," 2025.

[52] F. F.-H. Nah and K. L. Siau, "HCI in Business, Government and Organizations: 11th International Conference, HCIBGO 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Proceedings, Part I," Washington, DC, USA, June 29 – July 4, 2024.

[53] F. Ezzeddine, "Privacy Implications of Explainable AI in Data-Driven Systems," 2024.

[54] T. T. NGUYEN, T. T. HUYNH, Z. REN and et al, "Privacy-preserving explainable AI: a survey," 2025.

[55] R. Aydo˘gan, N. Criado, J. Lang and et al, "PRIMA 2022: Principles and Practice of Multi-Agent Systems 24th International Conference," Valencia, November 16–18, 2022.

[56] K. Tejasvi, V. L. Vinya, J. N. Padmaja, R. Begum, M. A. Jabbar, A. Abraham and L. Loveleen, "Explainable Artificial Intelligence (XAI) for Managing Customer Needs in E-Commerce: A Systematic Review," 2024.

[57] v. M. García López, M. S. Ramírez Montoya and J. M. Molina Espinosa, "Harnessing Open Language Models: A Systematic Literature Review Unleashing AI's Potential for a Smarter Future," 2025.

[58] A. Al Fraidan, "AI language and emotional support as a physician assistant in hypertension management: an N-of-1 case study on virtual encouragement and blood pressure control," 2025.

[59] C. Woodcock, B. Mittelstadt and D. Busb, "The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study," 2021.

[60] S. Grossberg, "A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action," 2020.

[61] M. R. Turchioe, A. Hermann and N. C. Benda, "Recentering responsible and explainable artificial intelligence research on patients: implications in perinatal psychiatry," 2024.

[62] U. Peters and M. Carman, "Cultural Bias in Explainable AI Research: A Systematic Analysis," 2023.

[63] U. Ehsan, P. Tambwekar and L. Chan, "Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions," 2019.

[64] H. Degen and S. Ntoa, "Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings," 2022.

[65] O. Oyebode, J. Fowles, D. Steeves and R. Orji, "Machine Learning Techniques in Adaptive and Personalized Systems for Health and Wellness," 2023.

[66] E. Yan, S. Burattini and J. F. Hübner, "A multi-level explainability framework for engineering and understanding BDI agents," 2025.

[67] L. Arora, A. Raj and A. Shetgaonkar, "Explainable Artificial Intelligence Techniques for Software Development Lifecycle: A Phase-specific Survey," 2025.

[68] M. Suffian, J. M. Alonso-Moral and A. Bogliolo, "Toward enriched Cognitive Learning with XAI," 2023.

[69] W. Maxwell and B. Dumas, "MEANINGFUL XAI BASED ON ON USER-CENTRIC DESIGN METHODOLOGY," 2023.

[70] J. Amann, A. Blasimme, E. Vayen and et al, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," 2020.

[71] E. Martinez, L. Robinson and H. Clark, "The Role of Explainable AI (XAI) in Financial Decision-Making and Regulatory Audits," p. 2024.

[72] S. Chae, S. Lee, H. Hauptmann and et al, "The Role of Explanation Styles and Perceived Accuracy on Decision Making in Predictive Process Monitoring," 2025.

[73] F. M. Cau, L. D. Spano and et al, "Exploring the Impact of Explainable AI and Cognitive Capabilities on Users' Decisions," 2025.

[74] R. R. Hoffman, S. T. Muelle, G. Klein and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," 2023.

[75] Q. V. Liao, D. Gruen and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," 2020 .

[76] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," 2019.

[77] M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human–Automation Research," 2016.

[78] M. R. Endsley, "Automation and Situation Awareness," 1991.

[79] A. Holzinger, "From Machine Learning to Explainable AI," 2018.

[80] H. Ouifak and A. Idri, "A comprehensive review of fuzzy logic based interpretability and explainability of machine learning techniques across domains," 2025.

[81] Wankhede, Vidhi; Verma, Prateek; Gahane, Shailesh, "A Review on the Developments in the Field of AI-Based Gait Analysis," 2024.

[82] B. Séroussi, P. Weber and F. Dhombres, "Challenges of Trustable AI and Added-Value on Health: Proceedings of MIE 2022.," 2022.

[83] Aljawawdeh, Hamzeh; Fawareh, Hamed; Al Refai, Moha, "Ethical Considerations in AI-Driven HCI Systems in Software Engineering," 2024.

[84] A. Pagliaro, "Artificial Intelligence vs. Efficient Markets: A Critical Reassessment of Predictive Models in the Big Data Era," 2025.

[85] K. Vivekanandan, C. Troussas and C. Troussas, "Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8-12, 2020, Proceedings," 2020.

[86] A. Sallah, E. A. Abdellaoui Alaoui and S. Agoujil, "Explainable Machine Learning for Identifying Malicious Profiles in Online Social Networks," 2023.

[87] M. Akwu Adaji, E. Etuh, B. Isa Sambo, S. B. Francis, C. Markus, M. E. Ezema and S. Owoicho Olofu, "Effectiveness of Machine Learning Algorithms in Threat Detection and Mitigation in Cyberspace: A Systematic Review - Lappeenranta-Lahti University of Technology LUT," 2024.

[88] A. K. B. Arnob, R. R. Chowdhury, N. A. Chaiti, S. Saha and A. Roy, "A comprehensive systematic review of intrusion detection systems: emerging techniques, challenges, and future research directions," 2025.

[89] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," 2022.

[90] S. Baron, A. J. Latham and S. Varga, "Explainable AI and stakes in medicine: A user study," 2025.

[91] N. Fernando, B. Nakisa, A. Ahmad and M. N. Rastgoo, "Adaptive XAI in High Stakes Environments: Modeling Swift Trust with Multimodal Feedback in Human AI Teams," 2025.

[92] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," 2021.

[93] R. Binns, M. Van Kleek, M. Veale and et al, "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," 2018.

[94] S. Wachter, B. Mittelstadt and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," 2018.

[95] N. Tintarev and J. Masthoff, "Explaining Recommendations: Design and Evaluation," 2015.

[96] D. Kowald , S. Scher, V. Pammer-Schindler and et al, "Establishing and evaluating trustworthy AI: overview and research challenges," 2024.

[97] S. Mhammedi, H. El Massari, N. Gherabi, M. El Hajji, M. Sedki and S. Agoujil, "CF Recommender System Based on Ontology and Nonnegative Matrix Factorization (NMF)," 2023.

[98] I. Moustati and N. Gherabi, "Bridging Behavioral Insights and Automated Trading: An Internet of Behaviors Approach for Enhanced Financial Decision-Making," 2025.

[99] A. Nastoska, B. Jancheska, M. Rizinski and D. Trajanov, "Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries," 2025.

[100] P. Zaphiris and A. Ioannou, Learning and Collaboration Technologies: 11th International Conference, LCT 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Proceedings, Part III, Washington, DC, USA, June 29–July 4, 2024.

[101] P. Zaphiris and A. Ioannou, "Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Proceedings," Athens, Greece, June 8-12, 2020.

[102] A. Pagliaro, "Artificial Intelligence vs. Efficient Markets: A Critical Reassessment of Predictive Models in the Big Data Era," 2025.

[103] H. V. Subramanian, C. Canfield and D. B. Shank, "Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review," 2024.

[104] S. Anjomshoae, A. Najjar, D. Calvaresi and K. Främling, "Explainable Agents and Robots: Results from a Systematic Literature Review," 2019.

[105] I. B. Benitez, M. A. G. Teodoro and M. D. Pahati, "Challenges and Opportunities in AI Integration in Power System Protection," 2025.

[106] G. He, N. Aishwarya and U. Gadiraju, "Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant," 2025.

[107] H. Schuff, A. Jacovi and H. Adel, "Human Interpretation of Saliency-based Explanation Over Text," 2022.

[108] D. Göndöcs, S. Horváth and V. Dörfler, "Uncovering the Dynamics of Human-AI Hybrid Performance: A Qualitative Meta-Analysis of Empirical Studies," 2025.

[109] S. Wachter, B. Mittelstadt and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," 2017.

[110] K. Sokol and P. Flach, "One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency," 2020.

[111] Y.-L. Chou, C. Moreira and et al, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," pp. 59-83, 2022.

[112] D. Folgado, M. Barandas, L. Famiglini, R. Santos, F. Cabitza and H. Gamboa, "Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series," 2023.

[113] A. Pahud de Mortanges, H. Luo, S. Z. Shu, A. Kamath, Y. Suter and et al., "Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging," 2024.

[114] M. Kim, S. Kim, J. Kim, T.-j. Song and Y. Kim, "Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces," 2024.

[115] A. Dobrovolskis, E. Kazanavičius and L. Kižauskienė, "Building XAI-Based Agents for IoT Systems," 2023.

[116] Z. Deng, A. Torim, S. Ben Yahia and H. Bahsi, "Generative AI in Intrusion Detection Systems for Internet of Things: A Systematic Literature Review," 2025.

[117] M. Ridley, "Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paperscience," 2024.

[118] J. B. Schmutz, N. Outland, S. Kerstan, E. Georganta and A.-S. Ulfert, "AI-teaming: Redefining collaboration in the digital era," 2023.

[119] C. Molnar, "Interpretable Machine Learning A Guide for Making Black Box Models Explainable," 2019.

[120] G. James, D. Witten, T. Hastie and R. Tibshirani, "An Introduction to Statistical Learning," 2021.

[121] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, 2009.

[122] p. McCullagh and J. Nelder, "Generalized Linear Models (2nd ed.)," 1989.

[123] U. Kamath and J. Liu, Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, Nashville, USA, 2021.

[124] Y. Lou, R. Caruana and J. Gehrke, "Intelligible Models for Classification and Regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, p.150-158*, 2012.

[125] T. Hastie and R. Tibshirani, "Generalized Additive Models," 1986.

[126] M. Krúpová, N. Rachdi and Q. Guibert, "Explainable Boosting Machine for Predicting Claim Severity and Frequency in Car Insurance," 2025.

[127] B. Mahbooba, M. Timilsina, R. Sahal and M. Serrano, "Explainable Artificial Intelligence (XAI) to Enhance TrustManagement in Intrusion Detection Systems Using DecisionTree Model," 2021.

[128] A. E. Khedr, A. M. Idrees and A. I. El Seddawy, "Enhancing Iterative Dichotomiser3 algorithm for classificationdecision tree," 2016.

[129] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," 1996.

[130] L. Longo, S. Lapuschkin and C. Seifert, "Explainable Artificial Intellihence: Second World Conference, xAI 2024, Part II," in *Communications in Computer and Information Science 2154*, Valletta, Malta, 2024.

[131] B. Letham, C. Rudin, T. McCormick and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis," 2015.

[132] F. Wang and C. Rudin, "Causal Falling Rule Lists," 2017.

[133] . J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," 2008.

[134] D. w. Aha, D. Kibler and M. K. Albert, "Instance-Based Learning Algorithms," 1991.

[135] J. E. Miranda-Vega, M. Rivas-López and W. F. Fuentes, "k-Nearest Neighbor Classification for Pattern Recognition of a Reference Source Light for Machine Vision System," 2021.

[136] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, London, England, 2009.

[137] C. Charon, P.-H. Wuillemin and J. Belmin, "Learning Bayesian Networks for the Prediction of Unfavorable Health Events in Nursing Homes," 2022.

[138] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, 1988.

[139] C. Chen, O. Li, C. Tao and et al., "This Looks Like That: Deep Learning for Interpretable Image Recognition," 2019.

[140] O. Li, H. Liu, C. Chen and C. Rudin, "Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions on Artificial Intelligence (AAAI-18)," in *The Thirty-Second AAAI Conference*, 2018.

[141] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su and C. Rudin, "This Looks Like That: Deep Learning for Interpretable Image Recognition," 2019.

[142] N. E. H. Dehimi and Z. Tolba, "Attention Mechanisms in Deep Learning : Towards Explainable Artificial Intelligence," 2024.

[143] A. Brown, A. Tuor, B. Hutchinson and N. Nichols, "Recurrent Neural Network Attention Mechanisms for Interpretable System Log Anomaly Detection," pp. 1-8, 2018.

[144] A. Vaswani, N. Shazeer, N. Parmar and et al., "Attention Is All You Need," 2023.

[145] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas and M. Wattenberg, "AttentionViz: A Global View of Transformer Attention," no. IEEE Transactions on Visualization and Computer Graphics, 2023.

[146] H. Chefer, S. Gur and L. Wolf, "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers," pp. 397-406, 2021.

[147] M. Mersha, K. Lam, J. Wood, A. AlShami and J. Kalita, "Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction," 2025.

[148] A. M. P. Brasoveanu and R. Andonie, "Visualizing Transformers for NLP: A Brief Survey," in *In Proceedings of the 2020 24th International Conference Information Visualisation (IV)*, Melbourne, VIC, Australia, 7–11 September 2020.

[149] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen and V. N. Nair, "Explainable Neural Networks based on Additive Index Models," 2018.

[150] S. Zhao, "xNN-SF: An Explainable Neural Network Inspired by Stochastic Frontier Model," 2024.

[151] D. Alvarez-Melis and T. S. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks," 2018.

[152] O. Elbaghdadi, A. Hussain, C. Hoenes and I. Bardarov, "Self Explaining Neural Networks: A Review with Extensions," 2020.

[153] Q. Zhang, Y. N. Wu and S.-C. Zhu, "Interpretable Convolutional Neural Networks," USA, 2018.

[154] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim and P. Liang, "Concept Bottleneck Models," 2020.

[155] E. Kim, D. Jung, S. Park, S. Kim and S. Yoon, "Probabilistic Concept Bottleneck Models," 2023.

[156] G. D. Felice, A. C. Flores, F. D. Santis and et al., "Causally Reliable Concept Bottleneck Models," 2025.

[157] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana and G. E. Hinton, "Neural Additive Models: Interpretable Machine Learning with Neural Nets," 2021.

[158] K. Bouchiat, A. Immer, H. Yeche, G. Ratsch and V. Fortuin, "Improving Neural Additive Models with Bayesian Principles," 2024.

[159] P. Veličković, G. Cucurull, A. Casanova, A. Romero and et al., "Graph Attention Networks," 2018.

[160] S. Brody, U. Alon and E. Yahav, "How Attentive Are Graph Attention Networks?," 2022.

[161] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," ICLR, 2017.

[162] R. Ying, D. Bourgeois, J. You and et al., "GNNExplainer: Generating Explanations," 2019.

[163] H. Yuan, H. Yu, S. Gui and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," 2022.

[164] F. P. Busch, J. Seng, M. Willig and M. Zecevic, "Continually Updating Neural Causal Models," *In Proceedings of The First AAAI Bridge Program on Continual Causality, PMLR 208,* pp. 30-37, 2023.

[165] B. Sick and O. Durr, "Interpretable Neural Causal Models with TRAM-DAGs," in *4th Conference on Causal Learning and Reasoning https://arxiv.org/abs/2503.16206*, 2025.

[166] K. Xia, K.-Z. Lee, Y. Bengio and E. Bareinboim, "The Causal-Neural Connection: Expressiveness, Learnability, and Inference," no. https://arxiv.org/abs/2107.00793, 2021.

[167] Z. Li, X. Guo and S. Qiang, "A Survey of Deep Causal Models and Their Industrial Applications," no. https://arxiv.org/abs/2209.08860, 2022 .

[168] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi and J. Gama, "Methods and tools for causal discovery and causal inference," no. https://doi-org.ezproxy.cc.lut.fi/10.1002/widm.1449, 2022.

[169] W. Niu , Z. Gao and et al., "Comprehensive Review and Empirical Evaluation of Causal Discovery Algorithms for Numerical Data," no. arXiv:2407.13054v1, 2024.

[170] L. N. DeLong, R. F. Mir and J. D. Fleuriot, "Neuro-Symbolic AI for Reasoning over Knowledge Graphs: A Survey," no. arXiv:2302.07200, 2024.

[171] U. Nawaz, M. Anees-ur-Rahaman and Z. Saeed, "A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems," vol. 26, 2025.

[172] Z. Wan, L. Che-Kai, H. Yang, C. Li, H. You and e. al., "Towards Cognitive AI Systems: a Survey and Prospective on Neuro-Symbolic AI," 2024.

[173] Y. Liu, X. Wang, T. Meng, W. Ai and K. Li, "LG-GNN: Local and Global Information-aware Graph Neural Network for default detection," vol. 169, 2024.

[174] K. Sandbrink and C. Summerfield , "Modelling cognitive flexibility with deep neural networks," vol. 57, 2024.

[175] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: core cognitive abilities and practical applications," 2018.

[176] F. E. Ritter, F. Tehranchi and J. D. Oury, "ACT-R: A cognitive architecture for modeling cognition," 2018.

[177] R. Sun, "The Importance of Cognitive Architectures: An Analysis Based on CLARION," 2006.

[178] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[179] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017.

[180] S. M. Lundberg, G. G. Erion and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," 2019.

[181] M. T. Ribeiro, S. Singh and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," 2018.

[182] S. Wachter, B. Mittelstadt and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology,* vol. 31, 2018.

[183] S. Verma, J. P. Dickerson and K. Hines, "Counterfactual Explanations for Machine Learning: Challenges Revisited," 2021.

[184] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics,* vol. 29, no. n0.5, p. 1189–1232, 2001.

[185] J. Krause, A. Perer and K. Ng, "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models," 2016.

[186] D. W. Apley and J. Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *Journal of the Royal Statistical Society Series B: Statistical Methodology,* vol. 82, no. 4, p. 1059–1086, 2020.

[187] R. Guidotti, A. Monreale, S. Ruggieri and et al., "A Survey of Methods for Explaining Black Box Models," *A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93,* 2018.

[188] A. Goldstein, A. Kapelner, J. Bleich and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation," no. https://doi-org.ezproxy.cc.lut.fi/10.1080/10618600.2014.907095, pp. 44-65, 2015.

[189] L. Breiman, "Random Forests. Machine Learning," vol. 45, no. https://doi.org/10.1023/A:1010933404324, p. 5–32, 2001.

[190] A. Fisher, C. Rudin and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," vol. 2o (177), pp. 1-81, 2019.

[191] A. Altmann, L. Toloşi, O. Sander and T. Lengauer, "Permutation importance: A corrected feature importance measure. Bioinformatics," *https://doi.org/10.1093/bioinformatics/btq134,* vol. 26, no. 10, pp. 1340-1347, 2010.

[192] J. Herbinger, S. Dandl, F. K. Ewald, S. Loibl and G. Casalicchio, "Leveraging Model-based Trees as Interpretable Surrogate Models for Model Distillation," no. https://arxiv.org/abs/2310.03112, 2023.

[193] H. Jones and S. Chen, "Surrogate Models: A Comfortable Middle Ground?," 2021.

[194] M. Sundararajan, A. Taly and Q. Yan, "Axiomatic Attribution for Deep Networks," no. https://arxiv.org/abs/1703.01365, 2017.

[195] G. Montavona, W. Samek and . K.-R. Müller, "Methods for Interpreting and Understanding Deep Neural Networks," no. https://arxiv.org/abs/1706.07979, 2017.

[196] S. Bach, A. Binder, G. Montavon and et al., "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," no. https://doi.org/10.1371/journal.pone.0130140, 2015.

[197] A. Binder, G. Montavon, S. Bach and et al., "Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers," 2016.

[198] I. Sturm, S. Lapuschkin, W. Samek and K. R. Müller, "Interpretable deep neural networks for single-trial EEG classification," vol. 274, pp. 141-145, 2017.

[199] M. Ali, A. M. Khattak, Z. Ali and et al., "Estimation and Interpretation of Machine Learning Models with Customized Surrogate Model," 2022.

[200] A. Shrikumar, P. Greenside and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," no. https://arxiv.org/abs/1704.02685, 2019.

[201] M. Ancona, E. Ceolini, C. Öztireli and M. Gross, "Towards better understanding of gradient-based attribution methods for Deep Neural Networks," in *ICLR 2018*.

[202] W. Samek and K.-R. Müller, "Towards Explainable Artificial Intelligence," pp. 5-22, 2019.

[203] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," 2013.

[204] L. M. Zintgraf, T. S. Cohen, T. Adel and M. Welling, "Visualizing Deep Neural Network Decisions: Prediction Difference Analysis," 2017.