TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

EXPLAINABLE AI FOR UNDERSTANDING HUMAN DECISION-MAKING PATTERNS

PRIYA DALAL[1], BHARTI SHARMA[2], TRIPTI SHARMA[3], PUNEET GARG[4], KAHKSHA AHMED[5]

[1]ASSISTANT PROFESSOR, MSIT, JANAKPURI, NEW DELHI, INDIA
[2]ASSOCIATE PROFESSOR, MSIT, JANAKPURI, NEW DELHI, INDIA
[3]PROFESSOR, MSIT, JANAKPURI, NEW DELHI, INDIA
[4]ASSOCIATE PROFESSOR, KIET GROUP OF INSTITUTIONS, DELHI NCR, GHAZIABAD, INDIA
[5]ASSISTANT PROFESSOR, SAITM, GURUGRAM, DELHI NCR, INDIA
EMAIL: priya@msit.in[1], bhartisharma@msit.in[2], tripti_sharma@msit.in[3], puneetgarg.er@gmail.com[4], kahkasha.ahmed@gmail.com[5]

CORRESPONDING AUTHOR: PUNEET GARG

Abstract
Explainable artificial intelligence (XAI) has become an essential research area for making complex Machine-Learning models transparent, trustworthy and actionable for human decision makers. As artificial intelligence (AI) systems increasingly influence high-stakes decisions in domains such as finance, healthcare and education, understanding how explanations impact human judgment is critical. This paper presents a comprehensive examination of XAI for understanding human decision-making patterns, synthesising theoretical foundations, recent empirical findings and design considerations. The paper develops conceptual frameworks linking explanation types to cognitive processes, summarises empirical evidence regarding the effects of explanations on task performance, trust and cognitive load, and discusses challenges such as the *white-box paradox*, algorithmic aversion and the risk of overreliance. We further propose guidelines for designing human-centred XAI systems that align with users' mental models, support various stakeholder needs and incorporate mechanisms to recognise when explanations are insufficient. Finally, we highlight open challenges and future directions for research at the intersection of XAI and human decision-making.
Keywords: Explainable artificial intelligence; human decision making; cognitive load; trust and reliance; human–computer interaction; interpretability; evaluation metrics.

1 INTRODUCTION

1.1 Motivation

Modern AI models often operate as black boxes, mapping input features to predictions without revealing the internal reasoning process. While such models achieve high performance, their opacity undermines human trust and raises concerns about fairness, accountability and regulatory compliance [7][8]. Without transparent explanations, stakeholders cannot evaluate whether predictions are reasonable, understand which factors drive outcomes or identify potential biases. The absence of interpretability hampers adoption in high-stakes contexts such as credit scoring, medical diagnosis and automated hiring, where decisions have serious consequences. Explainable AI seeks to address these issues by providing explanations that make models' behaviour understandable to humans, thereby enhancing trust and enabling humans to make informed decisions [9][10][11].

1.2 Scope and objectives

This research aims to provide a detailed overview of XAI for understanding human decision-making patterns. The study does not examine case studies in specific domains; instead, it generalises principles and empirical findings across multiple fields. Our objectives are to:

a) Summarise the taxonomy of explanation methods and relate them to cognitive processes;
b) Analyse empirical evidence on how explanations affect task performance, trust, reliance and cognitive load;
c) Discuss theoretical frameworks such as the *white-box paradox* and the *halo effect* that highlight potential pitfalls of XAI;
d) Present guidelines for designing human-centred explanation systems;
e) Identify open research challenges and propose future directions.

2 THEORETICAL FOUNDATIONS Of EXPLAINABLE AI

2.1 Interpretability and transparency

The terms *interpretability*, *transparency* and *explainability* are often used interchangeably but denote different concepts. *Intrinsic interpretability* refers to models whose internal structure can be directly understood by humans, such as decision trees or generalised additive models (GAMs) [12]. *Post-hoc explanations* are generated after model training and aim to approximate the decision logic of a black-box model through surrogate models, feature

attribution or example-based reasoning. Transparent models provide comprehensive access to all parameters and operations, whereas interpretable models prioritise human comprehensibility over full transparency [13]. The goal of XAI is not to reveal every detail of a complex neural network but to produce explanations that are faithful, comprehensible and actionable for specific users [14].

2.2 Why interpretability matters

Explanations serve multiple purposes. *First*, they enhance trust by enabling users to verify that a model's decision logic aligns with domain knowledge and ethical principles. *Second*, explanations facilitate *Error Detection* and model debugging by revealing spurious correlations and biases. *Third*, regulatory frameworks such as the EU's General Data Protection Regulation (GDPR) and guidelines from US government agencies mandate that algorithmic decisions be explainable to affected individuals. *Lastly*, explanations support *knowledge transfer*, allowing human experts to learn from model insights and integrate them into decision making [15][16][17].

2.3 Taxonomy of explanation methods

Figure 1 presents a high-level pipeline illustrating how data, a black-box model, an explanation module and the human decision maker interact. Explanations are produced as post-hoc summaries of model predictions and may influence the final human decision [18][19].

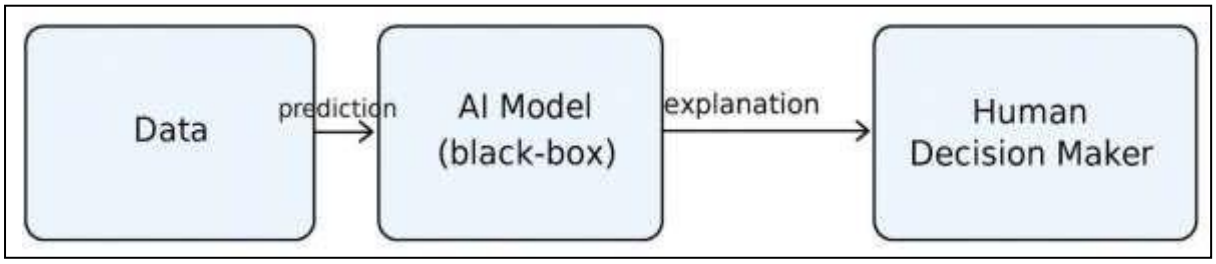

Figure 1: XAI pipeline for human decision making.

Intrinsic methods include decision trees, rule-based systems and interpretable linear models (GAMs). These models are inherently transparent and can be inspected directly [20]. However, they often sacrifice performance relative to complex models and may struggle to capture non-linear relationships [21]. *Post-hoc methods* operate on trained black-box models and derive explanations by analysing the relationship between input features and predictions. They can be further divided into:

a) **Feature attribution methods**, which assign importance scores to input features (e.g., LIME, SHAP, integrated gradients). They are widely used because they are model-agnostic and produce heatmaps or bar plots that highlight influential features.
b) **Example-based methods**, which select representative training examples or prototypes to justify predictions. These include k-nearest neighbours, exemplar-based explanations and case-based reasoning.
c) **Rule-based explanations**, which extract logical rules that approximate the model's decision boundaries.
d) **Counterfactual explanations**, which suggest minimal changes to input features required to alter the prediction. They enable users to understand how decisions could differ under alternative circumstances.
e) **Surrogate models**, which train interpretable models on the inputs and outputs of the original black-box model (e.g., decision tree surrogates).

Figure 2 depicts a taxonomy of XAI methods across two dimensions: intrinsic vs post-hoc and local vs global explanations.

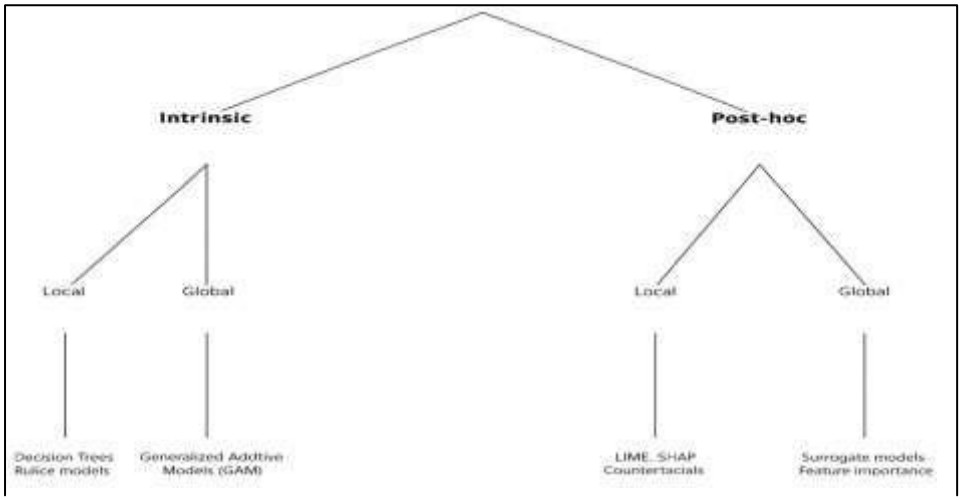

Figure 2: Taxonomy of explanation methods

Beyond the intrinsic/post-hoc distinction, several other dimensions enrich the taxonomy of explanations. *Local vs global* refers to whether an explanation pertains to a single instance or provides a summary across many instances [22][23]. *Model-specific vs model-agnostic* distinguishes explanations that exploit the structure of a specific model

(e.g., saliency maps for convolutional neural networks) from those that treat the model as a black box and probe it by perturbing inputs. *Ante-hoc vs post-hoc* separates methods that build interpretability into the model from those that derive interpretability after training. *Actionable vs descriptive* distinguishes explanations that suggest how to alter an outcome from those that merely describe factors influencing a prediction [24][25].

2.4 Cognitive frameworks and explanation questions

Explanations must align with human cognitive processes to be effective. Human reasoning often proceeds by asking and answering different types of questions: *Why* did an event happen, *why not*, *how*, *what if* and *what else*. Aligning explanation types with such questions supports deeper understanding and helps users integrate model output into their mental models [26][27]. Figure 3 illustrates a conceptual framework linking explanation questions to mental models and cognitive processes. Explanations in categories such as *why* or *why not* feed into users' reasoning processes, update their mental models and ultimately influence decision outcomes [28][29][30].

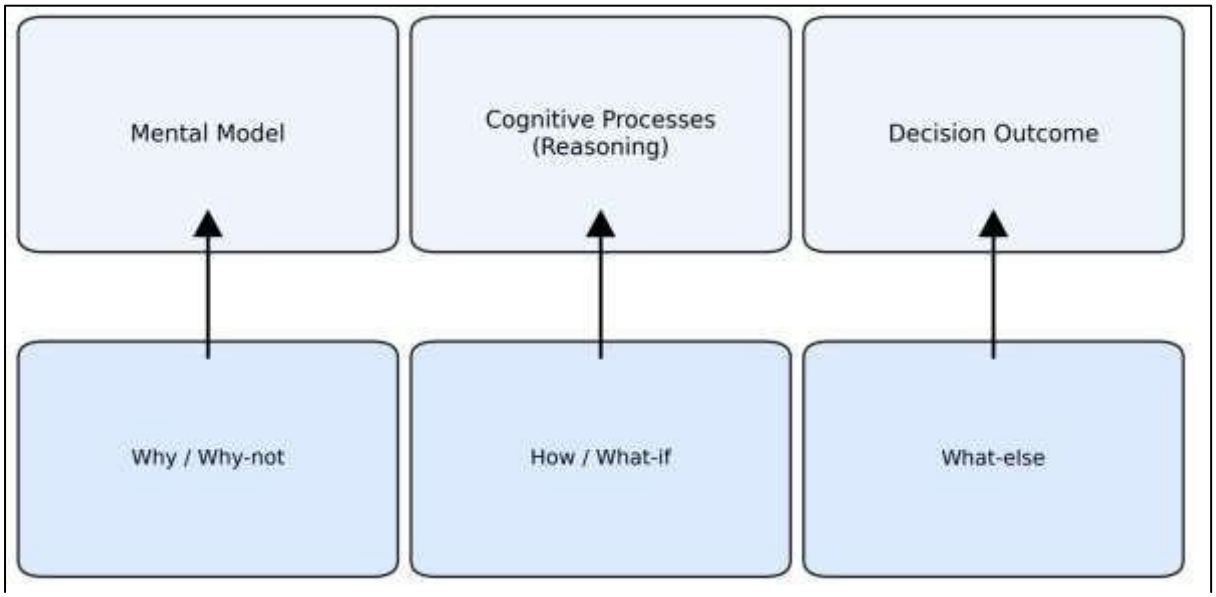

Figure 3: Cognitive framework linking explanation questions to human mental models.

Theories from cognitive psychology provide deeper insights into how explanations influence decision making. *Mental models* theory posits that people construct internal representations of external systems to simulate scenarios and reason about them. Explanations aid in refining these models by elucidating causal relationships, highlighting alternative possibilities and correcting misconceptions. *Dual-process theories* differentiate between fast, intuitive cognition (System 1) and slow, deliberative cognition (System 2). Simple visualisations or narratives may suffice for System 1 processing, while complex tasks require analytical explanations that engage System 2. Cognitive biases—including confirmation bias, anchoring and availability bias—can shape how explanations are interpreted. For instance, users might favour explanations that confirm pre-existing beliefs, thereby reinforcing misconceptions. Understanding these cognitive principles helps designers create explanations that are both persuasive and accurate, while avoiding unintended biases [31][32][33].

2.5 Stakeholder perspectives and human-centred XAI

Traditional XAI methods often target technical stakeholders such as data scientists or developers, who need to debug models and ensure fairness. However, many AI applications involve multiple stakeholders—decision subjects, domain experts, regulators—who require different types of explanations [34][35]. For example, in credit risk assessment the applicant may ask why their loan was denied, a regulator may ask whether the model complies with fairness regulations, while a data scientist seeks to know which features drive predictions. A one-size-fits-all approach is inadequate, underscoring the need for human-centred XAI that allows interactive exploration, supports diverse questions and acknowledges the limitations of the model. Explanations must be tailored to the user's role, knowledge and cognitive style to be effective [36].

2.6 Historical evolution of XAI

The quest for interpretability predates the advent of modern deep learning. Early expert systems in the 1970s, such as MYCIN, were built upon symbolic rules encoded by human experts and included modules that explained the reasoning behind diagnoses. These systems highlighted the importance of trust and interpretability but were limited by their inability to learn from data [37]. In the 1990s and 2000s, machine learning models such as decision trees and logistic regression offered a compromise between accuracy and interpretability. As big data and deep neural networks emerged, black-box models dramatically improved predictive performance but at the cost of transparency [38]. This trade-off spurred a resurgence of interest in explainability. The development of model-agnostic explanation techniques like LIME and SHAP in the mid-2010s facilitated the examination of complex models, marking a turning point in XAI research. Subsequent years saw the rise of counterfactual explanations, example-based reasoning, surrogate models and hybrid approaches that combine multiple

explanation modalities. Recent work emphasises holistic frameworks that integrate user feedback, causal reasoning and domain knowledge to create explanations that are more meaningful for diverse stakeholders [39][40].

3 Linking Explanations to Human Decision-making Patterns

3.1 Cognitive load and mental workload

Human decision makers operate under varying levels of cognitive load. Cognitive load refers to the amount of mental resources required to process information, and it is influenced by task complexity, time pressure, fatigue and individual abilities [41]. High workload can impair the ability to process explanations, leading to overreliance or underreliance on AI. Researchers use both subjective measures, such as the NASA Task Load Index (TLX), and objective measures, including EEG or eye-tracking, to assess cognitive load. Senoner et. al. [1] measured mental workload and its effect on appropriate reliance on AI. Participants were asked to determine the trustworthiness of a source with or without explanations while their brain activity was recorded. Cau et. al [2] found that under low workload participants appropriately calibrated their reliance on AI, while high workload led to overreliance regardless of whether explanations were provided. Explanations alone did not mitigate the negative effects of high mental workload. These results align with cognitive theory: when working memory is overloaded, individuals may resort to heuristics or defer to the AI without properly evaluating the explanation [42]. To address this, XAI systems can monitor indicators of mental workload (e.g., through physiological sensors or performance data) and adapt the complexity of explanations accordingly. Figure 4 visualises how workload influences reliance on AI. Data are adapted from the EEG study. Under low workload, participants showed near-optimal reliance on AI, and explanations had negligible effect. Under high workload, reliance decreased markedly, again with little difference between explanation conditions [43][44].

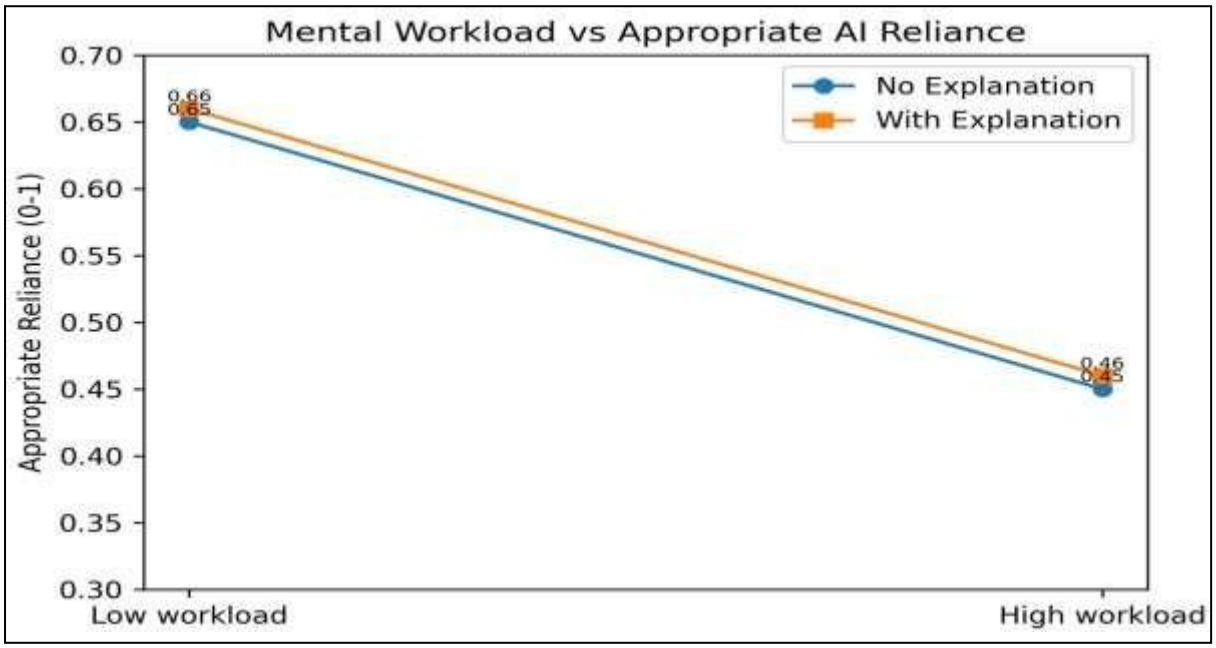

Figure 4: Relationship between mental workload and appropriate reliance on AI.

3.2 Effects of explanation type on task performance and trust

Multiple studies have examined whether different explanation styles improve task performance and trust. An experiment involving loan approvals compared feature-based, example-based, rule-based and counterfactual explanations. High AI confidence increased user reliance and reduced cognitive load, but feature-based explanations did not improve accuracy and counterfactuals improved accuracy yet were harder to understand. Another research work of Shajalal et. al. [3] with 742 participants found that guided explanations (which provided tailored suggestions) increased reliance more than transparent strategies, while no explanation sometimes also led to high reliance [45][46]. These findings suggest that simply adding an explanation is insufficient; the explanation must be designed to suit the task and user. Explanations may serve different functions: descriptive explanations help users understand what the model did, while prescriptive or guided explanations tell users how to act. In high-risk domains, users may prefer guided advice that highlights critical factors and suggests actions (e.g., "review this loan's debt-to-income ratio"). Conversely, in educational settings, descriptive explanations that encourage exploration may foster learning [47][48][49].

Synthesising results across experiments is challenging because tasks differ widely—from medical diagnoses to credit scoring to image classification—and participants vary in expertise. Nonetheless, a meta-analysis of classification tasks concluded that XAI improves task performance overall, although the choice of explanation type plays only a minor role. The meta-analysis emphasises that other factors, such as the complexity of the task,

user expertise, and the risk associated with decisions, moderate the effect of explanations. For example, novices may benefit more from example-based explanations that ground predictions in familiar instances, whereas experts may prefer feature attribution that highlights domain-specific cues [50][51]. In high-stakes contexts, such as medical diagnosis, even a small improvement in accuracy may justify the cost of implementing XAI, but the explanation must align with clinical reasoning to be accepted. Researchers should thus consider context, user goals and domain conventions when selecting explanation types.

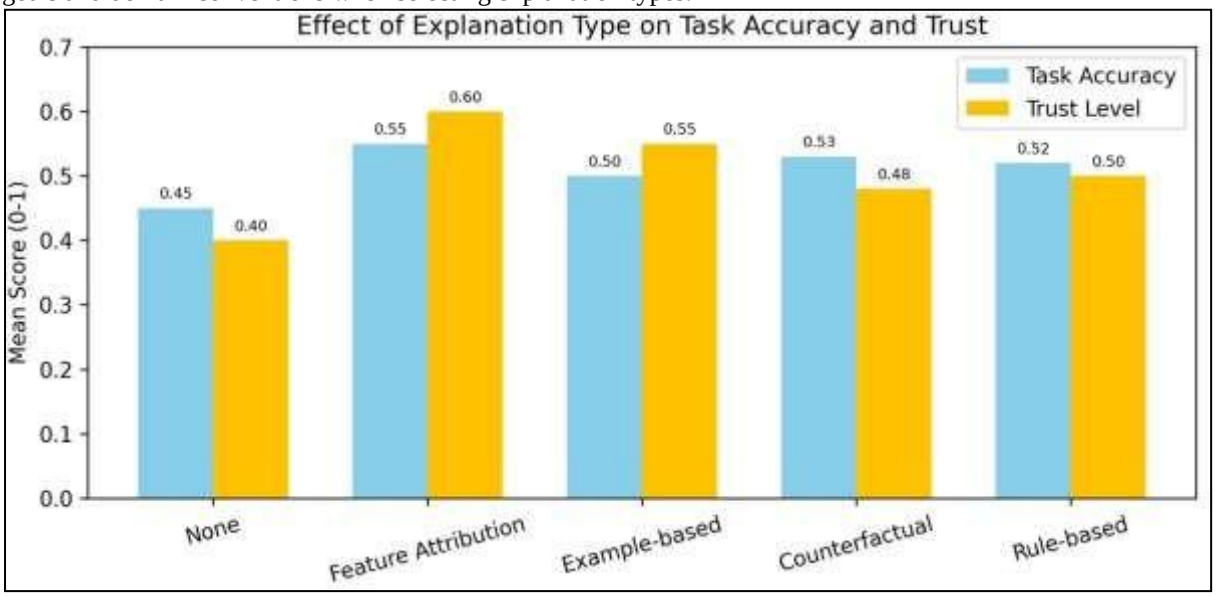

Figure 5. Comparison of task accuracy and trust across different explanation types.

The Figure 5 summarises general patterns observed in empirical studies. Although explanations improve trust, the impact on accuracy is often marginal.

3.3 Individual differences and decision-making patterns

People differ in their propensity to seek or defer responsibility. Kim et. al. [4] explored how decision-making styles—vigilance, hypervigilance and *buckpassing*—affect reliance on AI suggestions [52]. Buckpassing describes the tendency to relinquish responsibility to others (in this case AI) and was found to correlate with greater reliance on AI and less time spent reading explanations. Hypervigilant individuals, who experience anxiety and indecision, also showed elevated reliance on AI but were more prone to reject explanations that conflicted with their own judgement. In contrast, vigilant decision makers scrutinised explanations carefully and relied less on AI, using explanations as an aid rather than a crutch. Another study examined the *Need for Cognition*, a personality trait reflecting enjoyment of effortful thinking; participants with high need for cognition benefited more from counterfactual explanations, whereas those with low need for cognition preferred simpler explanations. These findings suggest that explanation interfaces should adapt to individual differences, offering more detailed explanations to engaged users while providing concise summaries for others [53][54].

Beyond cognitive traits, demographic factors such as age, education and cultural background influence how explanations are perceived. Younger participants, who have grown up with digital technologies, may be more comfortable interacting with AI and may demand less justification, whereas older participants may require more extensive explanations to build trust. Domfeh et. al. [5] have shown that collectivist cultures value group consensus and may prefer explanations that emphasise fairness and social implications, whereas individualist cultures may focus on personal benefit and effectiveness. Domain expertise also plays a role: novices might require high-level analogies, whereas domain experts might expect explanations that use technical terminology and align with professional reasoning. Incorporating user modelling and personalisation into XAI systems can improve relevance and reduce cognitive overload, but it raises privacy and ethical questions about profiling users and adapting persuasive strategies [55][56].

3.4 Trust, algorithmic aversion and overreliance

While explanations can increase trust, they may also produce undesirable effects. *Algorithmic aversion* refers to the tendency of users to discount algorithmic advice after observing errors, even when it outperforms human judgement. Explanations may exacerbate aversion by making errors more salient and by revealing complexities that undermine confidence [57][58]. For example, if a feature attribution explanation highlights a spurious correlation (e.g., zip code influences medical diagnosis), users may distrust the model altogether rather than focusing on the overall pattern. Conversely, users may develop *overreliance* when explanations are too persuasive or when they erroneously believe that the model's reasoning is infallible. Overreliance is particularly dangerous in safety-critical contexts, where blind acceptance can lead to fatal errors.

The *white-box paradox* posits that exposing the internal workings of a model can create a halo effect: users might trust the model more simply because they see an explanation, even if the explanation is misleading or incorrect. Misleading explanations can arise because the explanation module and the prediction module are separate

components—one may be correct while the other is wrong. For instance, an AI system may produce an accurate prediction for a patient's risk of disease but accompany it with a low-quality explanation due to a bug or approximate method. Users might accept the prediction uncritically because of the apparent transparency, leading to *explainability washing*. Designers must therefore calibrate the persuasiveness of explanations to avoid both overreliance and unwarranted scepticism. Strategies include highlighting model uncertainty, providing counterexamples that illustrate failure cases, and educating users on the limitations of AI [59][60].

4 Empirical Evidence on XAI and Human Decision Making

4.1 Improving task performance and decision quality

Suffian et. al [6] hypothesised that providing explanations would automatically enhance decision quality. A randomised experiment involving more than 2 000 participants evaluated whether heatmap-based XAI improves human–AI collaboration in manufacturing and medical tasks. The results showed that explanations increased task performance by 7.7% points in the manufacturing task and by 4.7 % points in the medical task [61]. Participants also reported higher trust and found the AI easier to work with. Other experiments in image classification have shown that visual saliency maps help participants identify errors in AI predictions and improve accuracy in adversarial settings, though the effect size decreases as tasks become more complex [62][63].

However, a subsequent meta-analysis found that explanation type contributed little to performance differences, and that studies with lower risk of bias reported smaller effect sizes [64][65]. This meta-analysis aggregated results across tasks, including credit scoring, medical imaging and natural language processing, and noted that heterogeneity among experimental designs made it difficult to draw universal conclusions. Gambetti et. al. [7] found that an XAI-based clinical decision support system improved diagnostic accuracy compared with existing scores but that clinicians still trusted the traditional Centor score more and demanded additional testing. This underscores that trust and reliance may not align: participants may benefit from AI advice but still distrust the system due to lack of familiarity or concerns about liability [66][67][68].

Moreover, performance gains from explanations may be domain-dependent. In tasks with well-understood causal structures, such as diagnosing a simple disease, explanations may reinforce existing knowledge and improve accuracy. In complex socio-technical systems, such as predicting recidivism, explanations may not significantly improve performance and could even mislead if they oversimplify the model's reasoning. The interplay between explanation quality, domain complexity and user expertise therefore warrants careful empirical investigation [69][70].

4.2 Cognitive load and explanation complexity

The relationship between explanation complexity and cognitive load has been explored in several recent studies. An empirical study with prospective physicians compared local explanation types and found that complex explanations increased cognitive load and sometimes decreased performance [71]. In this study, participants assessed patient risk using explanations based on SHAP values, decision rules and narrative text. Although narrative explanations were easier to read, they sometimes lacked specificity, while technical explanations increased mental load. A separate study measured cognitive load and task time across explanation formats; rule-based and counterfactual explanations improved performance but were harder to understand. The authors suggested that counterfactuals require mental simulation of alternative states, which is cognitively demanding, but provide actionable insights that can improve decisions [72][73][74].

Another line of work investigates the role of explanation length, format and modality [75]. Long textual explanations may overwhelm users, whereas concise visual explanations may omit important nuances. Interactive explanations that allow users to expand or collapse details could strike a balance. Multi-modal explanations—combining text, charts and examples—may cater to different learning styles. When cognitive resources are exhausted, users may ignore explanations or misinterpret them, leading to poor decisions. Therefore, explanation designers should consider cognitive load when choosing the modality, depth and timing of explanations, possibly by monitoring user attention and adapting accordingly [76][77].

4.3 Trust calibration and imperfect explanations

Recent research has begun to investigate scenarios where explanations themselves may be incorrect or uninformative. In a mixed-methods study, participants were exposed to AI advice accompanied by correct or incorrect explanations [78]. The study showed that participants were easily misled by flawed explanations and often failed to detect the inaccuracies, leading to poor decisions and misplaced trust. This highlights the *fragility* of human judgement when the explanation is accepted at face value. Similarly, a user study observed that misleading explanations paired with accurate advice produced a *halo effect*, causing users to trust the AI more than warranted. The effect persisted even when participants were trained to question the advice, suggesting that the presence of an explanation can disarm skepticism [79][80].

These findings underscore the need for mechanisms to detect and signal the quality of explanations. A recent proposal introduced *User-centric Low-quality Explanation Rejector (ULER)*, which learns to abstain from making predictions when it cannot provide a satisfactory explanation. ULER relies on human ratings of explanations and follows guidelines from government agencies suggesting that AI systems should recognise their limitations [81][82]. Other approaches include estimating explanation uncertainty, evaluating the stability of feature attribution under perturbations, and cross-validating explanations using multiple methods. For instance, comparing SHAP values with counterfactual explanations can reveal inconsistencies; if they point in opposite

directions, the explanation may be unreliable. Ensemble explanation frameworks may provide more robust insights by aggregating different explanation outputs [83].

4.4 Evaluation of human-centred XAI

A systematic review of human-centred XAI applications identified 73 studies across domains such as healthcare, finance, education and criminal justice. The review highlighted that evaluation measures vary widely, including subjective metrics (trust, satisfaction), objective metrics (accuracy, time) and behavioural metrics (choices, reliance). Beyond these, some studies consider *calibrated trust*, defined as the degree to which reliance on AI matches its actual performance, and *fairness perceptions*, reflecting whether users believe the AI treats different groups equitably [84][85]. Another survey focusing on clinical decision support systems emphasised that adoption is hindered by cognitive load and misalignment with clinical reasoning. The review called for deeper stakeholder engagement and user studies to evaluate explanations in real-world contexts. For example, in clinical settings, explanations might need to reference evidence hierarchies or connect to established guidelines to be accepted by clinicians [86].

Research on smart home environments also emphasised that existing XAI methods are tailored for developers rather than general users, urging human-centred approaches that consider user needs, cognitive abilities and accessibility [87]. Evaluations should not only measure whether users understand the explanation but also whether it leads to better decisions and supports fairness and accountability [88]. For instance, an explanation that is understandable but encourages discriminatory decisions is ethically unacceptable. Inclusion of diverse participants in user studies is therefore critical for ensuring that explanations do not disproportionately benefit or harm certain groups. To this end, frameworks for *fairness-aware XAI evaluation* are being developed, which assess whether explanations systematically differ across demographic groups and whether they mitigate or exacerbate biases [89][90].

4.5 The effect of human–AI collaboration patterns

Human–AI interaction patterns influence how explanations are used. A systematic review of AI use in disaster management categorised interactions into decision support systems, task and resource coordination, trust and transparency, and simulation and training. Decision support systems often present AI recommendations with or without explanations; task and resource coordination systems require dynamic communication among multiple stakeholders; trust and transparency interventions use XAI to build situational awareness and justify resource allocation; simulation and training platforms employ XAI to teach users how AI behaves under various conditions [91][92]. The review noted that while AI can improve situational awareness and decision making, limitations in scalability, interpretability and interoperability hinder adoption. For example, a disaster response system may rely on multiple heterogeneous AI models (for weather prediction, traffic routing and communication), each requiring different explanations. Aligning explanations across models and presenting them in a unified interface remains challenging [93].

Moreover, research on explainable decision systems for smart homes argued that current XAI methods seldom provide actionable explanations for lay users and emphasised the need for co-designed systems that integrate human–computer interaction (HCI) principles [94]. The emerging theme is that explanations should be integrated into collaborative workflows and support shared decision making rather than being treated as post-hoc add-ons. For instance, in team settings, explanations might need to support group decision making by highlighting how AI suggestions align with team goals and by providing rationales accessible to all members. Additionally, transitions of control between humans and AI should be explicit; when the AI takes over, users should understand why, and when control returns to humans, they should know which factors led to the AI's recommendation. Designing for collaboration requires considering not only individual cognitive processes but also group dynamics, communication protocols and organisational culture [95].

5 Designing and Evaluating Human-Centred XAI Systems

5.1 Evaluation metrics for explanations

Evaluating XAI methods remains an open challenge. Traditional metrics focus on *fidelity* (the extent to which explanations accurately reflect model behaviour) and *interpretability* (how easily humans can understand an explanation) [96]. Fidelity can be quantified by measuring the correlation between explanation scores and true feature contributions or by assessing how well a surrogate model approximates the original model. Interpretability is more subjective; it is often assessed via user studies or cognitive metrics [97]. Additional metrics include *completeness* (whether the explanation accounts for all important factors), *consistency* (whether similar instances yield similar explanations), *stability* (whether small perturbations in input produce similar explanations) and *compositionality* (whether explanations can be composed to explain complex decisions) [98].

User-centric metrics include trust, satisfaction, perceived fairness, reliance, mental workload and *calibration* (how accurately users assess model reliability). However, a recent preprint argues that current evaluations are fragmented and fail to account for the multidimensional nature of explanations [99]. It proposes a normalised evaluation framework that considers the data, model, prediction and user dimensions. Under this framework, an explanation is evaluated not only on its faithfulness to the model but also on its ability to answer stakeholder questions and support decision outcomes. Combining feature attribution with counterfactual explanations may yield holistic insights, but comparative evaluation across methods remains difficult [100]. These various metrics are summarised in Table 1, which outlines their definitions and typical measurement strategies.

Table 1: Evaluation metrics for XAI and their typical measurements.

Metric	Description	Typical measurement
Fidelity	Degree to which the explanation accurately reflects the behaviour of the underlying model.	Correlation between explanation scores and true feature contributions; surrogate model accuracy.
Interpretability	Ease with which humans can understand the explanation.	User studies assessing comprehension, time taken to interpret, or subjective ratings of clarity.
Completeness	Extent to which the explanation accounts for all factors that influence the prediction.	Evaluated by adding features suggested by the explanation and measuring change in prediction or by testing coverage of known causal factors.
Consistency	Whether similar instances yield similar explanations.	Statistical measures of variation across explanations for perturbed inputs; ranking similarity metrics.
Stability	Robustness of explanations to small perturbations in input or model parameters.	Distance metrics between explanations for neighbouring points; sensitivity analyses.
Compositionality	Ability of explanations to be combined to explain complex decisions involving multiple components.	Assessment of whether explanations for subcomponents can be composed to explain the overall decision; hierarchical evaluation.
Fairness	Degree to which explanations do not reinforce or introduce biases across demographic groups.	Analysis of explanation content across subgroups; fairness metrics such as demographic parity or equalised odds applied to explanations.
Actionability	Ability of explanations to provide suggestions that users can act upon to achieve desired outcomes.	User surveys on perceived helpfulness; measurement of changes in user behaviour or outcomes after receiving the explanation.
Efficiency	Computational cost and scalability of generating explanations.	Time complexity, memory usage and throughput; feasibility of real-time operation.

Evaluation frameworks should also consider fairness and ethical dimensions. For example, an explanation that attributes high importance to sensitive attributes (e.g., race) may raise concerns about discrimination, even if the model itself does not directly use these attributes. Fairness-aware evaluation measures whether explanations inadvertently reveal biased correlations or reinforce stereotypes. Another emerging metric is *actionability*, which assesses whether the explanation provides information that users can act upon. For instance, a counterfactual explanation might suggest that a loan applicant could improve their credit score by reducing debt [101]. Actionability is crucial for empowering users to seek recourse and for complying with regulations such as the GDPR. Finally, computational efficiency must be considered, as some explanation techniques (e.g., SHAP) are computationally expensive, limiting their usability in real-time systems.

Table 2: Categories of XAI methods, examples, advantages and limitations.

Category	Examples	Advantages	Limitations
Intrinsic local	Decision trees, rule-based models	Transparent and directly interpretable; no post-hoc approximation	Often less accurate than complex models; may not capture non-linear patterns
Intrinsic global	Generalised additive models (GAMs)	Capture global trends while remaining interpretable; can model smooth functions	Limited flexibility; require careful tuning
Post-hoc feature attribution	LIME, SHAP, integrated gradients	Model-agnostic; highlight important features; easy to visualise via heatmaps	May be unstable and lack faithfulness; ignore feature interactions
Post-hoc example-based	Prototypes, exemplar selection	Provide tangible examples; align with human case-based reasoning	Risk of cherry-picking examples; may not capture general patterns
Post-hoc rule-based	Anchors, decision rules	Provide succinct logical rules; easy to understand	Rules may oversimplify and be brittle; generation can be computationally expensive

Category	Examples	Advantages	Limitations
Counterfactual	Minimal changes to flip prediction	Answer "what-if" questions; actionable for users seeking to change outcomes	May propose unrealistic changes; solving optimisation can be complex
Surrogate models	Decision-tree surrogate, linear regression	Global interpretability; summarise black-box behaviour	Approximate only; may misrepresent complex decision boundaries

Table 2 emphasises that each method involves trade-offs among accuracy, interpretability and faithfulness. Evaluators must consider the target audience and context when selecting appropriate methods.

5.2 Guidelines for human-centred XAI design

Based on the literature review, we propose the following guidelines for designing human-centred XAI systems:

a) **Stakeholder analysis**: Identify the different stakeholders (decision subjects, domain experts, regulators, developers) and their information needs. Tailor explanation content accordingly.
b) **Alignment with cognitive models**: Map explanation types to human question categories (why, why not, how, what if) and to cognitive processes. Provide explanations that support causal and counterfactual reasoning.
c) **Adaptive explanations**: Allow users to request more detailed or alternative explanations. Adaptive systems can modulate explanation complexity based on user expertise, mental workload or decision style (buckpassing vs vigilance).
d) **Trust calibration**: Balance persuasiveness and scepticism to avoid overreliance and algorithmic aversion. Provide confidence information and highlight uncertainties.
e) **Explanation quality control**: Incorporate mechanisms like ULER to abstain when explanations are likely to be misleading or low quality. Use user feedback to improve explanation modules.
f) **Integrate into workflows**: Embed explanations within the decision process rather than as a post-hoc add-on. Support interactive exploration where users can ask questions and test hypotheses.
g) **Evaluate holistically**: Use multidimensional evaluation metrics covering fidelity, interpretability, usefulness, fairness and cognitive load. Conduct user studies in realistic environments to assess decision quality, not just perceived understanding.

5.3 Graphical representation of evaluation results

Figures 4 and 5 already illustrated aggregated patterns across studies. Such visual summaries can help researchers and practitioners compare results across explanation types and conditions. However, caution is warranted: aggregated data may mask important nuances, such as the influence of individual differences or domain-specific constraints. Nonetheless, graphical representation of evaluation results is a valuable tool for communicating findings and facilitating cross-study comparisons [102][103].

6 Challenges and Limitations

6.1 Misleading and imperfect explanations

XAI methods are not infallible. Feature attribution methods such as LIME and SHAP approximate black-box models and may produce explanations that are unstable or unfaithful [104]. When explanations are incorrect or misleading, they can cause users to make erroneous decisions. The *XAI halo effect* describes a phenomenon where users attribute greater credibility to AI advice simply because it is accompanied by an explanation. Addressing these issues requires rigorous evaluation of explanation quality and mechanisms to signal when the explanation should not be trusted [105].

6.2 White-box paradox and overreliance

The white-box paradox arises when revealing too much internal information leads to overreliance. Even accurate explanations may create a false sense of understanding, leading users to accept AI advice without critical scrutiny. Overreliance can have severe consequences in high-stakes settings, e.g., automated hiring or medical diagnosis [106]. Designers should calibrate explanation detail, provide counterexamples and encourage users to interrogate AI outputs. They should also train users to recognise when the AI may be uncertain or wrong.

6.3 Regulatory and ethical considerations

Regulations such as the GDPR emphasise the right to explanations for algorithmic decisions. New AI Acts in Europe, the US and Asia propose risk-based frameworks that require providers to demonstrate the transparency, robustness and governance of AI systems [107]. For instance, the European AI Act classifies AI systems into prohibited, high-risk and low-risk categories and mandates human oversight and explainability for high-risk applications. The US National Institute of Standards and Technology (NIST) released the AI Risk Management Framework, which includes "explainability and interpretability" as a key characteristic of trustworthy AI. Complying with these frameworks requires developers to document the data provenance, model assumptions and limitations and to provide explanations that are both technically accurate and legally meaningful [108].

From an ethical perspective, explanations may influence users' decisions and behaviours. Misleading or biased explanations can be weaponised for persuasion or manipulation, particularly in domains like advertising or political campaigning. Privacy concerns arise when explanations reveal sensitive attributes or training examples that may lead to re-identification of individuals. Ensuring that explanations do not inadvertently leak sensitive

information requires techniques such as differential privacy and data masking. Fairness considerations extend to whether explanations themselves are fair: a system might generate different explanations for users in different demographic groups, possibly reinforcing stereotypes or discrimination. Ethical XAI design therefore must incorporate principles of transparency, accountability, privacy and non-discrimination and should be informed by interdisciplinary collaboration among legal scholars, ethicists, social scientists and technologists [109].

Another regulatory challenge involves intellectual property and model secrecy. Companies may be reluctant to disclose the inner workings of proprietary models due to competitive concerns. Regulatory frameworks need to strike a balance between promoting transparency and protecting trade secrets. Recent proposals suggest using third-party auditors or "algorithmic inspectors" who can access model internals under confidentiality agreements and verify compliance without making proprietary details public. Ultimately, effective governance of XAI requires clear standards, enforcement mechanisms and avenues for redress when AI decisions harm individuals [110].

7 Future Directions

7.1 Interactive and conversational XAI

Future research should explore interactive systems that allow users to ask follow-up questions, request counterfactuals or drill down into specific features. Such systems could employ natural language interfaces and dialogue management to facilitate conversational explanations, providing a more engaging user experience. For example, an applicant seeking a loan could ask: "What aspects of my credit report contributed most to this decision?" The system might reply with a feature-importance explanation, after which the applicant could follow up with "What changes would improve my likelihood of approval?" to trigger a counterfactual explanation. By supporting this kind of dialogue, the system adapts to the user's evolving information needs and clarifies ambiguous points [111][112].

Conversational explanations also require multimodal interfaces. Visualisations (such as heatmaps or timelines), interactive sliders and narrative text can complement verbal explanations. The integration of large language models (LLMs) with traditional XAI techniques offers exciting possibilities: LLMs can translate technical explanation content into fluent language, summarise complex reasoning and generate analogies or metaphors tailored to user contexts. However, ensuring that LLM-generated explanations remain faithful to the underlying model's logic is critical to avoid hallucinations or oversimplification. Maintaining the chain of reasoning may help align the narrative with the model's operations and allow users to trace how specific features influenced the prediction [113].

Interactive XAI should not be limited to user queries; it should also proactively detect confusion or misunderstanding. Physiological signals like eye tracking, pupil dilation or voice hesitations can indicate when a user is puzzled. The system could then offer clarification or ask whether additional detail is needed. User feedback loops can help explanation modules learn which types of explanations are most helpful, which are confusing or overwhelming and how to adapt to different decision styles (e.g., analytical vs intuitive). Incorporating reinforcement learning techniques, the system could optimise its explanation policy based on feedback signals such as user satisfaction, task performance and trust calibration.

7.2 Integration with cognitive models

Advances in cognitive science and neuroscience offer opportunities to design explanations that align more closely with how humans reason. Mapping explanation types onto cognitive processes can guide the development of dynamic explanations tailored to users' mental models. For example, the dual-process theory posits that humans engage in both fast, intuitive processing and slow, analytical reasoning. Explanations may need to appeal to both systems: simple heuristics for quick understanding and detailed causal models for deeper analysis. Adaptive interfaces could present a succinct explanation first and allow the user to expand for more detail as needed [114].

Integration with cognitive models also extends to computational frameworks such as ACT-R and SOAR, which simulate human cognitive architectures. Embedding explanation modules into these architectures enables researchers to test how different explanation styles influence cognitive load, memory and problem solving. Neuroscientific techniques like EEG and fMRI can measure neural correlates of explanation comprehension, offering objective insights into cognitive processes. For instance, a mental-workload study using EEG found that high decision difficulty leads to overreliance on AI and that explanations may not mitigate this effect. Understanding such relationships could inform the design of explanations that adapt to mental workload and cognitive state.

Cognitive models can also inspire new types of explanations. Mental models research suggests that people often construct causal chains and storylines to understand events. Narrative explanations that weave features into a coherent story may be more effective than lists of feature weights. Analogical reasoning theory highlights the power of drawing parallels between unfamiliar and familiar situations; explanations could leverage analogies to link AI decisions to users' everyday experiences. As cognitive science advances, XAI should continue to incorporate insights into memory, attention, bias and reasoning to design explanations that not only inform but also resonate with human cognition.

7.3 Cross-domain evaluation frameworks

There is a pressing need for standardised benchmarks and evaluation frameworks applicable across domains. Current evaluation efforts are often fragmented: computer vision studies use ImageNet-based tasks and heatmaps,

whereas natural-language processing studies use textual explanations, making direct comparisons challenging. A cross-domain framework should include diverse tasks, data types (tabular, text, images, time series) and user populations. For each task, ground-truth explanations derived from domain experts or synthetic data can serve as references to assess fidelity. Benchmark suites like the Explanation Bank have been proposed for specific domains; expanding them to a wider set of tasks would facilitate comparative evaluation.

Crowd-sourcing platforms, citizen science initiatives and collaborative networks (e.g., open science communities) can enable large-scale user studies. Participants from different cultural backgrounds, educational levels and professions can provide feedback on explanation understandability, usefulness and fairness. However, generalising results across populations requires careful experimental design to control for confounding factors and ensure representation. Meta-analysis techniques can aggregate results across studies, identify moderator variables (e.g., explanation type, domain, user expertise) and generate generalisable insights [115].

Standardisation also involves defining protocols for reporting user studies. Researchers should clearly specify the AI model, explanation method, task description, participant demographics, evaluation metrics and statistical analyses. Repositories of code, data and instructions for replicating experiments should be made publicly available to promote transparency and reproducibility. Journals and conferences can encourage or mandate such reporting standards. Over time, cross-domain evaluation frameworks will enable researchers to compare methods, identify best practices and accelerate progress toward effective XAI.

7.4 Ethics, bias mitigation and fairness in explanations

Future work must address how explanations themselves may encode biases or cause harm. Explanations may highlight features that are correlated with protected characteristics (e.g., race or gender) even if the model does not use these attributes directly, thereby enabling inference of sensitive information. Developing fair XAI requires measuring and mitigating biases not only in model predictions but also in explanation content. Techniques such as counterfactual fairness and SHAP-value adjustments can help remove spurious associations from explanations. Researchers must also consider how explanations interact with existing cognitive biases: confirmation bias may lead users to accept explanations that align with prior beliefs and dismiss those that contradict them.

Fairness research emphasises intersectionality—the idea that individuals may belong to multiple protected categories and that biases can compound. XAI evaluation should therefore assess fairness across intersectional groups and consider normative questions: Should explanations mention sensitive attributes at all? Should they emphasise systemic factors beyond individual features? Participatory design with affected communities can help answer these questions and ensure that explanations empower rather than disadvantage minoritised populations [103][104].

Ethical considerations extend to transparency about the limitations of explanations. Systems should disclose when the explanation is an approximation, when it may be unstable and what assumptions underpin it. In some cases, withholding an explanation may be more ethical than providing a misleading one. Mechanisms for redress and contestability should be integrated: users should have avenues to challenge decisions and explanations and to provide feedback that leads to model improvement. Finally, global governance frameworks should ensure that companies and governments are accountable for the fairness and transparency of their AI systems [105].

7.5 Explainability beyond AI predictions

Beyond explaining predictions, XAI could help people understand the limitations of AI systems. Mechanisms like ULER can abstain from providing predictions when explanations are likely to be unreliable. Future systems may present meta-information about data quality, model training conditions and uncertainties, enabling users to decide when to rely on AI. Explainability could extend beyond single predictions to encompass the entire lifecycle of AI systems. For instance, dataset documentation (data cards or model cards) can summarise the sources, sampling procedures, biases and limitations of data used to train the model. Model cards can describe architectural choices, hyper-parameters, performance across different subgroups and known failure modes. Exposing such meta-information fosters transparency and allows users to assess whether the model is appropriate for their context [78].

Moreover, explanations could help users understand when the AI is likely to fail. Uncertainty estimation techniques (e.g., Bayesian neural networks, conformal prediction) can quantify confidence in predictions and highlight regions of the input space where the model is undertrained. Explanations can then be conditioned on uncertainty: they could state, “The model is only 60 % confident in this prediction because the data are outside the training distribution,” prompting the user to exercise caution. Integrating explanation and uncertainty information can support robust decision making, particularly in safety-critical contexts [82].

Finally, the notion of *meta-explanation*—explaining the explanation—has gained attention. Users may want to understand how the explanation itself is generated, what algorithm or heuristic is used, and why it should be trusted. Providing a transparent account of the explanation algorithm can build meta-trust and encourage sceptical evaluation. Future XAI systems could include modules that answer meta-questions such as “Why did the explanation focus on these features?” or “How stable is this explanation if we perturb the input slightly?” This meta-layer fosters deeper engagement and empowers users to critically assess the explanatory process itself.

8 CONCLUSION

Explainable AI represents a pivotal step toward integrating AI systems into human decision making. Recent research emphasises that the value of explanations depends on alignment with human cognitive processes,

stakeholder needs and task contexts. While explanations can enhance performance and trust, they also carry risks of overreliance, algorithmic aversion and ethical pitfalls. Human-centred design, adaptive explanations and multidimensional evaluation frameworks are necessary to harness the benefits of XAI. Future developments should explore interactive explanations, cognitive alignment, cross-domain evaluation and fairness considerations. By addressing these challenges, XAI can support informed, accountable and equitable human–AI collaboration.

REFERENCES

[1] Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. *Scientific reports*, *14*(1), 31150.
[2] Cau, F. M., & Spano, L. D. (2025). Exploring the Impact of Explainable AI and Cognitive Capabilities on Users' Decisions. *arXiv preprint arXiv:2505.01192*.
[3] Shajalal, M., Boden, A., Stevens, G., Du, D., & Kern, D. R. (2024, July). Explaining AI Decisions: Towards Achieving Human-Centered Explainability in Smart Home Environments. In *World Conference on Explainable Artificial Intelligence* (pp. 418-440). Cham: Springer Nature Switzerland.
[4] Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence*, *7*, 1456486.
[5] Adjei Domfeh, E., & Dancy, C. L. (2025). Human-AI Use Patterns for Decision-Making in Disaster Scenarios: A Systematic Review. *arXiv e-prints*, arXiv-2509.
[6] Suffian, M. (2023, July). Explainable AI assisted decision-making and human behaviour. In *International Conference on Computing, Intelligence and Data Analytics* (pp. 376-385). Cham: Springer Nature Switzerland.
[7] Garg, P., Dixit, A., & Sethi, P. (2022). Ml-fresh: novel routing protocol in opportunistic networks using machine learning. *Computer Systems Science & Engineering, Forthcoming*. Tech Science Press.
[8] Yadav, P. S., Khan, S., Singh, Y. V., Garg, P., & Singh, R. S. (2022). A Lightweight Deep Learning-Based Approach for Jazz Music Generation in MIDI Format. *Computational Intelligence and Neuroscience*, *2022*.
[9] Soni, E., Nagpal, A., Garg, P., & Pinheiro, P. R. (2022). Assessment of Compressed and Decompressed ECG Databases for Telecardiology Applying a Convolution Neural Network. *Electronics*, *11*(17), 2708.
[10] Pustokhina, I. V., Pustokhin, D. A., Lydia, E. L., Garg, P., Kadian, A., & Shankar, K. (2021). Hyperparameter search based convolution neural network with Bi-LSTM model for intrusion detection system in multimedia big data environment. *Multimedia Tools and Applications*, 1-18.
[11] Khanna, A., Rani, P., Garg, P., Singh, P. K., & Khamparia, A. (2021). An Enhanced Crow Search Inspired Feature Selection Technique for Intrusion Detection Based Wireless Network System. *Wireless Personal Communications*, 1-18.
[12] Garg, P., Dixit, A., Sethi, P., & Pinheiro, P. R. (2020). Impact of node density on the qos parameters of routing protocols in opportunistic networks for smart spaces. *Mobile Information Systems*, *2020*.
[13] Upadhyay, D., Garg, P., Aldossary, S. M., Shafi, J., & Kumar, S. (2023). A Linear Quadratic Regression-Based Synchronised Health Monitoring System (SHMS) for IoT Applications. Electronics, 12(2), 309.
[14] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). CNN-BI-LSTM-CYP: A deep learning approach for sugarcane yield prediction. *Sustainable Energy Technologies and Assessments*, *57*, 103263.
[15] Saini, P., Nagpal, B., Garg, P., & Kumar, S. (2023). Evaluation of Remote Sensing and Meteorological parameters for Yield Prediction of Sugarcane (Saccharumofficinarum L.) Crop. Brazilian Archives of Biology and Technology, 66, e23220781.
[16] Beniwal, S., Saini, U., Garg, P., & Joon, R. K. (2021). Improving performance during camera surveillance by integration of edge detection in IoT system. *International Journal of E-Health and Medical Communications (IJEHMC)*, *12*(5), 84-96.
[17] Garg, P., Dixit, A., & Sethi, P. (2019). Wireless sensor networks: an insight review. *International Journal of Advanced Science and Technology*, *28*(15), 612-627.
[18] Sharma, N., & Garg, P. (2022). Ant colony based optimization model for QoS-Based task scheduling in cloud computing environment. *Measurement: Sensors*, 100531.
[19] Kumar, P., Kumar, R., & Garg, P. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks.
[20] Raj, G., Verma, A., Dalal, P., Shukla, A. K., & Garg, P. (2023). Performance Comparison of Several LPWAN Technologies for Energy Constrained IOT Network. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(1s), 150-158.
[21] Garg, P., Sharma, N., & Shukla, B. (2023). Predicting the Risk of Cardiovascular Diseases using Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(2s), 165-173.
[22] Patil, S. C., Mane, D. A., Singh, M., Garg, P., Desai, A. B., & Rawat, D. (2024). Parkinson's Disease Progression Prediction Using Longitudinal Imaging Data and Grey Wolf Optimizer-Based Feature Selection. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 441-451.

[23] Gudur, A., Pati, P., Garg, P., & Sharma, N. (2024). Radiomics Feature Selection for Lung Cancer Subtyping and Prognosis Prediction: A Comparative Study of Ant Colony Optimization and Simulated Annealing. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 553-565.
[24] Khan, A. (2024). Optimisation Methods Based on Soft Computing for Improving Power System Stability. *J. Electrical Systems*, *20*(6s), 1051-1058.
[25] Sharma, K. K., Verma, P. K., & Garg, P. (2024). IoT-Enabled Energy Management Systems For Sustainable Energy Storage: Design, Optimization, And Future Directions. *Frontiers in Health Informatics*, *13*(8).
[26] Gupta, S., Yadav, N., Singh, K., & Garg, P. (2025). APPLICATIONS OF SIMULATIONS AND QUEUING THEORY IN SUPERMARKET. *Reliability: Theory & Applications*, *20*(1 (82)), 135-140.
[27] Beniwal, S., Garg, P., Rajpal, R., Sharma, N., & Mittal, H. K. (2025). Fusion of Opportunistic Networks with Machine Learning: Present and Future. *Metallurgical and Materials Engineering*, *31*(1), 311-324.
[28] Garg, P. (2025). Explainable AI & Model Interpretability in Healthcare: Challenges & Future Directions. *EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR*, *46*(1), 104-133.
[29] Rani, P. (2025). From Data to Diagnosis: Unleashing AI and 6G in Modern Medicine. *EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR*, *46*(1), 69-103.
[30] Dixit, A., Garg, P., Sethi, P., & Singh, Y. (2020, April). TVCCCS: Television Viewer's Channel Cost Calculation System On Per Second Usage. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012046). IOP Publishing.
[31] Sethi, P., Garg, P., Dixit, A., & Singh, Y. (2020, April). Smart number cruncher–a voice based calculator. In *IOP Conference Series: Materials Science and Engineering* (Vol. 804, No. 1, p. 012041). IOP Publishing.
[32] S. Rai, V. Choubey, Suryansh and P. Garg, "A Systematic Review of Encryption and Keylogging for Computer System Security," *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 2022, pp. 157-163, doi: 10.1109/CCiCT56684.2022.00039.
[33] L. Saraswat, L. Mohanty, P. Garg and S. Lamba, "Plant Disease Identification Using Plant Images," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 79-82, doi: 10.1109/CCiCT56684.2022.00026.
[34] L. Mohanty, L. Saraswat, P. Garg and S. Lamba, "Recommender Systems in E-Commerce," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 114-119, doi: 10.1109/CCiCT56684.2022.00032.
[35] C. Maggo and P. Garg, "From linguistic features to their extractions: Understanding the semantics of a concept," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 427-431, doi: 10.1109/CCiCT56684.2022.00082.
[36] N. Puri, P. Saggar, A. Kaur and P. Garg, "Application of ensemble Machine Learning models for phishing detection on web networks," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 296-303, doi: 10.1109/CCiCT56684.2022.00062.
[37] R. Sharma, S. Gupta and P. Garg, "Model for Predicting Cardiac Health using Deep Learning Classifier," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 25-30, doi: 10.1109/CCiCT56684.2022.00017.
[38] Varshney, S. Lamba and P. Garg, "A Comprehensive Survey on Event Analysis Using Deep Learning," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 146-150, doi: 10.1109/CCiCT56684.2022.00037.
[39] Dixit, A., Sethi, P., Garg, P., & Pruthi, J. (2022, December). Speech Difficulties and Clarification: A Systematic Review. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 52-56). IEEE.
[40] Garg, P., Dixit, A., Sethi, P., & Pruthi, J. (2023, December). Strengthening Smart City with Opportunistic Networks: An Insight. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 700-707). IEEE.
[41] Rana, S., Chaudhary, R., Gupta, M., & Garg, P. (2023, December). Exploring Different Techniques for Emotion Detection Through Face Recognition. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 779-786). IEEE.
[42] Mittal, K., Srivastava, K., Gupta, M., & Garg, P. (2023, December). Exploration of Different Techniques on Heart Disease Prediction. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)* (pp. 758-764). IEEE.
[43] Gautam, V. K., Gupta, S., & Garg, P. (2024, March). Automatic Irrigation System using IoT. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 100-103). IEEE.
[44] Ramasamy, L. K., Khan, F., Joghee, S., Dempere, J., & Garg, P. (2024, March). Forecast of Students' Mental Health Combining an Artificial Intelligence Technique and Fuzzy Inference System. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 85-90). IEEE.
[45] Rajput, R., Sukumar, V., Patnaik, P., Garg, P., & Ranjan, M. (2024, March). The Cognitive Analysis for an Approach to Neuroscience. In *2024 International Conference on Automation and Computation (AUTOCOM)* (pp. 524-528). IEEE.
[46] Dixit, A., Sethi, P., Garg, P., Pruthi, J., & Chauhan, R. (2024, July). CNN based lip-reading system for visual input: A review. In *AIP Conference Proceedings* (Vol. 3121, No. 1). AIP Publishing.

[47] Bose, D., Arora, B., Srivastava, A. K., & Garg, P. (2024, May). A Computer Vision Based Framework for Posture Analysis and Performance Prediction in Athletes. In *2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE)* (pp. 942-947). IEEE.
[48] Singh, M., Garg, P., Srivastava, S., & Saggu, A. K. (2024, April). Revolutionizing Arrhythmia Classification: Unleashing the Power of Machine Learning and Data Amplification for Precision Healthcare. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 516-522). IEEE.
[49] Kumar, R., Das, R., Garg, P., & Pandita, N. (2024, April). Duplicate Node Detection Method for Wireless Sensors. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 512-515). IEEE.
[50] Bhardwaj, H., Das, R., Garg, P., & Kumar, R. (2024, April). Handwritten Text Recognition Using Deep Learning. In *2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT)* (pp. 506-511). IEEE.
[51] Gill, A., Jain, D., Sharma, J., Kumar, A., & Garg, P. (2024, May). Deep learning approach for facial identification for online transactions. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 715-722). IEEE.
[52] Mittal, H. K., Dalal, P., Garg, P., & Joon, R. (2024, May). Forecasting Pollution Trends: Comparing Linear, Logistic Regression, and Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 411-419). IEEE.
[53] Malik, T., Nandal, V., & Garg, P. (2024, May). Deep Learning-Based Classification of Diabetic Retinopathy: Leveraging the Power of VGG-19. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 645-651). IEEE.
[54] Srivastava, A. K., Verma, I., & Garg, P. (2024, May). Improvements in Recommendation Systems Using Graph Neural Networks. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 668-672). IEEE.
[55] Aggarwal, A., Jain, D., Gupta, A., & Garg, P. (2024, May). Analysis and Prediction of Churn and Retention Rate of Customers in Telecom Industry Using Logistic Regression. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 723-727). IEEE.
[56] Mittal, H. K., Arsalan, M., & Garg, P. (2024, May). A Novel Deep Learning Model for Effective Story Point Estimation in Agile Software Development. In *2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)* (pp. 404-410). IEEE.
[57] Shukla, S. M., Magoo, C., & Garg, P. (2024, November). Comparing Fine Tuned-LMs for Detecting LLM-Generated Text. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (pp. 1-8). IEEE.
[58] Kumar, B., IQBAL, M., Parmer, R., Garg, P., Rani, S., & Agrawal, A. (2025, March). The Role of AI in Optimizing Healthcare Appointment Scheduling. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 881-887). IEEE.
[59] Kumar, B., Garg, V., Ahmed, K., Garg, P., Choudhary, S., & Baniya, P. (2025, March). Enhancing Healthcare with Blockchain: Innovations in Data Privacy, Security, and Interoperability. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 932-938). IEEE.
[60] Raj, V., Prakash, B. K., Kumar, A., & Garg, P. (2024, December). Optimize the Time a Mercedes-Benz Spends on the Test Bench Using Stacking Ensemble Learning. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 445-450). IEEE.
[61] Kaushik, N., Kumar, H., Raj, V., & Garg, P. (2024, December). Proactive Fault Prediction in Microservices Applications Using Trace Logs and Monitoring Metrics. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 410-415). IEEE.
[62] Kumar, A. A., Sri, C. V., Bohara, K. S. K., Setia, S., & Garg, P. (2024, December). Capnivesh: Financing Platform for Startups. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 261-265). IEEE.
[63] Bhandari, P., Setia, S., Kumar, K., & Garg, P. (2024, December). Optimizing Cross-Platform Development with CI/CD and Containerization: A Review. In *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)* (pp. 175-180). IEEE.

[64] Chaudhary, A., & Garg, P. (2014). Detecting and diagnosing a disease by patient monitoring system. *International Journal of Mechanical Engineering And Information Technology*, *2*(6), 493-499.
[65] Malik, K., Raheja, N., & Garg, P. (2011). Enhanced FP-growth algorithm. *International Journal of Computational Engineering and Management*, *12*, 54-56.
[66] Garg, P., Dixit, A., & Sethi, P. (2021, May). Link Prediction Techniques for Opportunistic Networks using Machine Learning. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.
[67] Garg, P., Dixit, A., & Sethi, P. (2021, April). Opportunistic networks: Protocols, applications & simulation trends. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*.
[68] Garg, P., Dixit, A., & Sethi, P. (2021). Performance comparison of fresh and spray & wait protocol through one simulator. *IT in Industry*, *9*(2).

[69] Malik, M., Singh, Y., Garg, P., & Gupta, S. (2020). Deep Learning in Healthcare system. *International Journal of Grid and Distributed Computing*, *13*(2), 469-468.
[70] Gupta, M., Garg, P., Gupta, S., & Joon, R. (2020). A Novel Approach for Malicious Node Detection in Cluster-Head Gateway Switching Routing in Mobile Ad Hoc Networks. *International Journal of Future Generation Communication and Networking*, *13*(4), 99-111.
[71] Gupta, A., Garg, P., & Sonal, Y. S. (2020). Edge Detection Based 3D Biometric System for Security of Web-Based Payment and Task Management Application. *International Journal of Grid and Distributed Computing*, *13*(1), 2064-2076.
[72] Kumar, P., Kumar, R., & Garg, P. (2020). Hybrid Crowd Cloud Routing Protocol For Wireless Sensor Networks.
[73] Garg, P., & Raman, P. K. Broadcasting Protocol & Routing Characteristics With Wireless ad-hoc networks.
[74] Garg, P., Arora, N., & Malik, T. Capacity Improvement of WI-MAX In presence of Different Codes WI-MAX: Speed & Scope of future.
[75] Garg, P., Saroha, K., & Lochab, R. (2011). Review of wireless sensor networks-architecture and applications. IJCSMS International Journal of Computer Science & Management Studies, 11(01), 2231-5268.
[76] Yadav, S., &Garg, P. Development of a New Secure Algorithm for Encryption and Decryption of Images.
[77] Dixit, A., Sethi, P., & Garg, P. (2022). Rakshak: A Child Identification Software for Recognizing Missing Children Using Machine Learning-Based Speech Clarification. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-15.
[78] Shukla, N., Garg, P., & Singh, M. (2022). MANET Proactive and Reactive Routing Protocols: A Comparison Study. International Journal of Knowledge-Based Organizations (IJKBO), 12(3), 1-14.
[79] Arya, A., Garg, P., Vellanki, S., Latha, M., Khan, M. A., & Chhbra, G. (2024). Optimisation Methods Based on Soft Computing for Improving Power System Stability. *Journal of Electrical Systems*, *20*(6s), 1051-1058.
[80] Garg, P. (2025). Cloud security posture management: Tools and techniques. Technix International Journal for Engineering Research, 12(3).
[81] Tyagi, P., Sharma, S., Srivastava, A., Rajput, N. K., Garg, P., & Kumari, M. (2025). AI in Healthcare:Transforming Medicine with Intelligence. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p4
[82] Bhatt, M., Parmar, R., Arsalan, M., & Garg, P. (2025). Generative AI: Evolution, Applications, Challenges And Future Prospects. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p6
[83] Saraswat, P., Garg, P., & Siddiqui, Z. (2025). AI & the Indian Stock Market: A Review of Applications in Investment Decision. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p10
[84] Sharma, S., Mittal, S., Tevatia, R., Tyagi, V. K., Garg, P., & Kapoor, S. (2025). Unlocking Workforce Potential: AI-Powered Predictive Models for Employee Performance Evaluation. Ind Emerging Developments (G-CARED 2025), New Delhi, India. https://doi.org/10.63169/GCARED2025.p21
[85] Shrivas, N., Kalia, A., Roy, R., Sharma, S., Garg, P., & Agarwal, G. (2025). OSINT: A Double-edged Sword. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p22
[86] Aditi, Garg, P., & Roy, B. (2025). A System of Computer Network: Based On Artificial Intelligence. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p24
[87] Parmar, R., Kapoor, S., Saifi, S., & Garg, P. (2025). Case Study on Intelligent Factory Systems for Improving Productivity and Capability in Industry 4.0 with Generative AI. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p28
[88] Singh, R., Sharma, R., Kumar, R., Nafis, A., Siddiqui, M. A. M., & Garg, P. (2025). Detection of Unauthorize Construction using Machine Learning: A Review. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p30
[89] Kapoor, S., Singh, V., Sharma, S., Garg, P., & Ankita (2025). A Bridge between Blockchain and Decentralized Applications Web3 and Non-Web3 Crypto Wallets. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p35
[90] Verma, M., Sharma, S., Garg, P., & Singh, A. (2025). The Hidden Dangers of Prototype Pollution: A Comprehensive Detection Framework. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p36
[91] Sharma, A., Sharma, S., Garg, P., & Bhardwaj, P. (2025). LockTalk: A Basic Secure Chat Application. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India.
[92] Arora, K., Bawane, R., Gupta, C., Ahmed, K., & Garg, P. (2025). Detection and Prevention of Cyber Attack and Threat using AI. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p38

[93] Dhruv, Rahman, A. A., Rai, A., Siddiqui, M. A. M., Garg, P., & Yadav, D. (2025). Easeviewer: An Esports Production Tool. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p46

[94] Lakshita, Mehwish, Nazia, Ahmed, K., & Garg, P. (2025). Emerging Trend in Computational Technology: Innovations, Applications, and Challenges. In *First Global Conference on AI Research and Emerging Developments (G-CARED 2025)*, New Delhi, India. https://doi.org/10.63169/GCARED2025.p51

[95] Chauhan, S., Singh, M., & Garg, P. (2021). Rapid Forecasting of Pandemic Outbreak Using Machine Learning. *Enabling Healthcare 4.0 for Pandemics: A Roadmap Using AI, Machine Learning, IoT and Cognitive Technologies*, 59-73.

[96] Gupta, S., & Garg, P. (2021). An insight review on multimedia forensics technology. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 27.

[97] Shrivastava, P., Agarwal, P., Sharma, K., & Garg, P. (2021). Data leakage detection in Wi-Fi networks. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 215.

[98] Meenakshi, P. G., & Shrivastava, P. (2021). Machine learning for mobile malware analysis. *Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms*, *11*, 151.

[99] Garg, P., Pranav, S., & Prerna, A. (2021). Green Internet of Things (G-IoT): A Solution for Sustainable Technological Development. In *Green Internet of Things for Smart Cities* (pp. 23-46). CRC Press.

[100] Nanwal, J., Garg, P., Sethi, P., & Dixit, A. (2021). Green IoT and Big Data: Succeeding towards Building Smart Cities. In *Green Internet of Things for Smart Cities* (pp. 83-98). CRC Press.

[101] Gupta, M., Garg, P., & Agarwal, P. (2021). Ant Colony Optimization Technique in Soft Computational Data Research for NP-Hard Problems. In *Artificial Intelligence for a Sustainable Industry 4.0* (pp. 197-211). Springer, Cham.

[102] Magoo, C., & Garg, P. (2021). Machine Learning Adversarial Attacks: A Survey Beyond. *Machine Learning Techniques and Analytics for Cloud Security*, 271-291.

[103] Garg, P., Srivastava, A. K., Anas, A., Gupta, B., & Mishra, C. (2023). Pneumonia Detection Through X-Ray Images Using Convolution Neural Network. In *Advancements in Bio-Medical Image Processing and Authentication in Telemedicine* (pp. 201-218). IGI Global.

[104] Gupta, S., & Garg, P. (2023). 14 Code-based post-quantum cryptographic technique: digital signature. *Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity*, 193.

[105] Prakash, A., Avasthi, S., Kumari, P., & Rawat, M. (2023). PuneetGarg 18 Modern healthcare system: unveiling the possibility of quantum computing in medical and biomedical zones. Quantum-Safe Cryptography Algorithms and Approaches: Impacts of Quantum Computing on Cybersecurity, 249.

[106] Gupta, S., & Garg, P. (2024). Mobile Edge Computing for Decentralized Systems. *Decentralized Systems and Distributed Computing*, 75-88.

[107] Gupta, M., Garg, P., & Malik, C. (2024). Ensemble learning-based analysis of perinatal disorders in women. In Artificial Intelligence and Machine Learning for Women's Health Issues (pp. 91-105). Academic Press.

[108] Malik, M., Garg, P., & Malik, C. (2024). Artificial intelligence-based prediction of health risks among women during menopause. *Artificial Intelligence and Machine Learning for Women's Health Issues*, 137-150.

[109] Garg, P. (2024). Prediction of female pregnancy complication using artificial intelligence. In *Artificial Intelligence and Machine Learning for Women's Health Issues* (pp. 17-35). Academic Press.

[110] Pokhrel, L., Arsalan, M., Rani, P., Garg, P., & Pinheiro, P. R. (2026). AI-Powered Healthcare Solutions: Bridging the Medical Gap in Underserved Communities Worldwide. In *Applied AI and Computational Intelligence in Diagnostics and Decision-Making* (pp. 57-86). IGI Global Scientific Publishing.

[111] Kapoor, S., Parmar, R., Sharma, N., Garg, P., & Singh, N. J. (2026). AI and Computational Intelligence in Healthcare: An Introductory Guide. In *Applied AI and Computational Intelligence in Diagnostics and Decision-Making* (pp. 1-26). IGI Global Scientific Publishing.

[112] Pokhrel, L., Kumar, A., Garg, P., Anand, N., & Singh, N. (2026). AI and IoT in Global Health: Ethical Lessons From Pandemic Response. In *Development and Management of Eco-Conscious IoT Medical Devices* (pp. 367-394). IGI Global Scientific Publishing.

[113] Parmar, R., Singh, A., Garg, P., Sharma, T., & Pinheiro, P. R. (2026). Blockchain for Ethical Supply Chains: Transparency in Medical IoT Manufacturing. In *Development and Management of Eco-Conscious IoT Medical Devices* (pp. 337-366). IGI Global Scientific Publishing.

[114] Gupta, S., Garg, P., Agarwal, J., Thakur, H. K., & Yadav, S. P. (2024). Federated learning based intelligent systems to handle issues and challenges in IoVs (Part 1). Bentham Science Publishers. https://doi.org/10.2174/9789815313031124030l

[115] Gupta, S., Garg, P., Agarwal, J., Thakur, H. K., & Yadav, S. P. (2025). Federated learning based intelligent systems to handle issues and challenges in IoVs (Part 2). Bentham Science Publishers. https://doi.org/10.2174/9789815322224125030l