Review



# Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience

Qiuyue Liao , Yue Chen , Shuangjiang He , Ruiqi Wang , Wei Xu , Weishen Chu *

Posted Date: 8 December 2025

doi: 10.20944/preprints202512.0667.v1

Keywords: explainable AI; trustworthy AI; trustworthy machine learning; explainability; wireless network privacy; 5G security; zero trust architecture

 doi:10.20944/preprints202512.0667.v1



*Review*

# Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience

**Qiuyue Liao [1], Yue Chen [2], Shuangjiang He [3], Ruiqi Wang [4], Wei Xu [5] and Weishen Chu [6,*]**

[1] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA
[3] Information Technology Program, University of the Cumberlands, Williamsburg, KY 40769, USA
[4] College of Graduate and Professional Studies, Trine University, Angola, IN 46703, USA
[5] Independent Researcher, Los Altos, CA 94024, USA
[6] Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA
* Correspondence: weishenchu2014@u.northwestern.edu

**Abstract**

Explainable artificial intelligence (XAI) plays a central role in strengthening security, privacy, and trust in AI-driven 5G and future 6G networks. In this review, we first refine the concepts of transparency and interpretability, and introduce the notions of marginal transparency and marginal interpretability to describe the diminishing returns that arise from progressively deeper disclosure of model internals. We then survey key XAI methods, including LIME, SHAP, interpretable neural networks, and federated, privacy-preserving techniques, and assess their suitability for wireless resource management, intrusion detection, and regulatory auditing in next-generation networks. Building on these foundations, we outline a 2025–2030 research roadmap that integrates XAI into Zero Trust architectures, edge intelligence, and self-explaining 6G systems. Across these layers, we argue that explainability should be built in as a design-time requirement, enabling wireless infrastructures that are not only high performance but also auditable, accountable, and resilient.

**Keywords:** explainable AI; trustworthy AI; trustworthy machine learning; explainability; wireless network privacy; 5G security; zero trust architecture

## 1. Introduction

Artificial intelligence is now deeply embedded in wireless networks [1]. In 5G systems, and in the designs that are shaping 6G, machine learning is applied almost everywhere: resource allocation, channel prediction, beamforming, intrusion detection, and even service orchestration [2–5]. Engineers rely on these models because they process massive and noisy data far more quickly than traditional methods [6]. Without them, it would be hard to meet the demands of modern applications such as smart factories, connected vehicles, or real-time immersive media.

But the strength of these systems is also their weakness. A scheduling algorithm driven by deep reinforcement learning may boost throughput, yet no operator can clearly explain why certain users were favored [7]. An anomaly detector can flag malicious packets, but analysts may struggle to see what features the model relied on. In practice, this opacity is risky [8]. When wireless networks are used for surgery, traffic safety, or industrial automation, a lack of clarity about system behavior is not a minor issue — it can undermine trust and even safety [9].

This is where explainable AI, or XAI, enters the discussion. XAI does not replace advanced models; it adds a layer of transparency. It shows which traffic attributes triggered an alarm, or why a slice of spectrum was allocated to one device instead of another. In this way, it helps engineers verify performance, regulators ensure compliance, and users gain confidence in the system. The legal side also matters. The European GDPR stresses a "right to explanation," and telecom regulators have

 doi:10.20944/preprints202512.0667.v1



begun asking how AI decisions in networks can be audited [10–15]. Recent advances in explainable representation learning and privacy-aware model understanding — such as camera-aware graph consistency, self-feedback feature enhancement, multi-hop RAG for financial compliance, and structured privacy-policy understanding — further highlight the growing need for transparency in modern AI systems [16–19].

Concrete research examples confirm the need. Basaran and Dressler proposed XAInomaly, which explains anomaly detection inside O-RAN while still meeting low-latency targets [20]. Uccello and Nadjm-Tehrani tested SHAP and other attribution methods in 5G intrusion detection, showing that some explanations are sparse and stable while others are not [21]. Work by Kaur and Gupta on IoT security in 6G points in a similar direction: explanations help reveal which device activities may indicate attacks [22]. These studies do not solve every problem, but they illustrate a path forward.

The present review takes these developments as a starting point. We revisit the core notions of transparency and interpretability, and then adapt them to wireless systems. Two new ideas — marginal transparency and marginal interpretability — are introduced to capture how additional layers of explanation add less and less value in practice. We also examine common methods such as LIME, SHAP, decision trees, and interpretable neural networks, but with an emphasis on how they apply to wireless resource management and network security. Finally, we suggest a roadmap for 2025–2030 that looks toward federated edge intelligence, Zero Trust networking, and eventually self-explaining 6G systems. Figure 1 summarizes this challenge–solution landscape by illustrating how black-box AI models interact with the 5G/6G ecosystem, where opacity can lead to risky decisions, and how an XAI layer restores transparency for engineers, regulators, and end users.

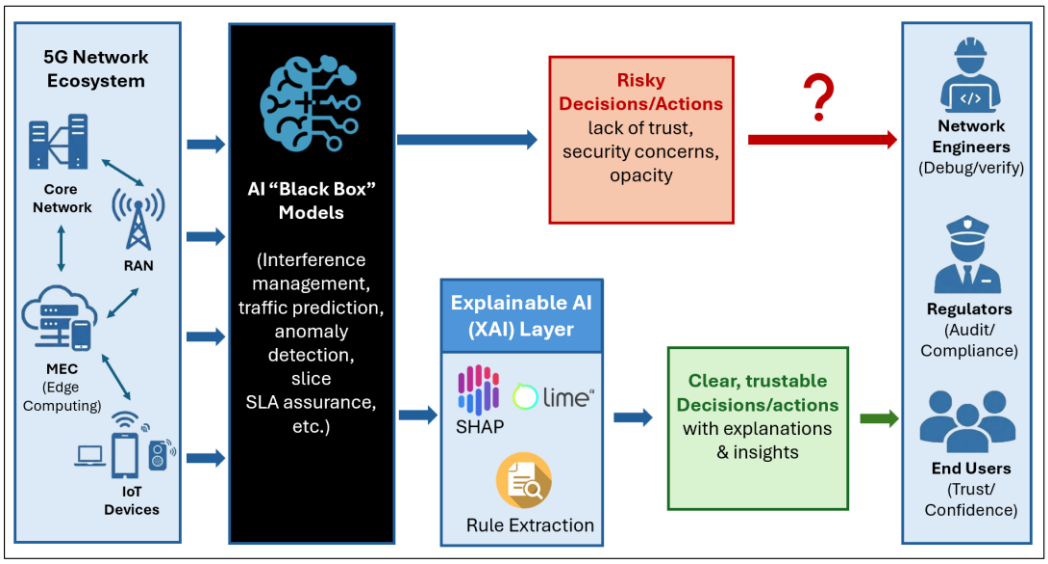


**Figure 1.** Conceptual framework of AI-driven 5G and 6G network operations and the role of explainable artificial intelligence (XAI). AI "black box'' models support tasks such as interference management, traffic prediction, and anomaly detection. However, the opacity of these models may result in risky or unverifiable decisions. The XAI layer uses methods including SHAP, LIME, and rule extraction to provide clear and interpretable insights that help engineers, regulators, and end users achieve transparent and trustworthy network operations.

## 2. Key Concepts

The discussion of explainable AI in wireless networks often begins with two related ideas: **transparency** and **interpretability**. These terms are sometimes used interchangeably, but in practice they capture different aspects of how a system communicates its reasoning [23]. In 5G/6G contexts — where spectrum allocation, handover management, and security decisions may occur in milliseconds — the distinction is more than academic. It defines how engineers, regulators, and even end users perceive the trustworthiness of the system [11,24].

 doi:10.20944/preprints202512.0667.v1



### 2.1. Transparency

Transparency refers to how clearly the internal operation of an AI model can be exposed. In communication systems, this might mean revealing how input features such as channel quality indicators, user mobility patterns, or interference levels are combined to produce a scheduling decision. A transparent model does not merely announce the outcome; it also shows what data were considered and how intermediate logic shaped the final result.

For example, an AI-based beamforming controller that highlights which antennas and channel states influenced its decision provides far greater transparency than one that outputs a precoding vector with no explanation [4]. As Van der Waa and colleagues have argued in other domains, visibility into the process is often as important as the outcome itself [25]. The same principle applies here: without transparency, operators may hesitate to trust algorithms even when performance looks strong. In the fast-changing 5G/6G environment, that hesitation can slow adoption or force costly manual overrides.

### 2.2. Interpretability

Interpretability, on the other hand, is less about raw access to model internals and more about whether humans can make sense of them. Wireless systems generate high-dimensional data: hundreds of radio features, traffic metrics, and temporal patterns. Even if a model exposes every weight and activation, the result may remain unintelligible. Interpretability requires explanations that match human cognitive limits.

One common approach is to simplify outputs into feature attributions. For instance, in intrusion detection for 5G cores, SHAP values can highlight that abnormal port usage and sudden packet bursts jointly drove the alarm [21]. In scheduling, LIME may show that poor channel conditions were the decisive factor for assigning extra resources [26]. Such explanations help network engineers and analysts see the reasoning without digging into raw matrices or gradient maps. In practice, interpretability makes the difference between a model that is transparent but unreadable and one that is genuinely useful.

### 2.3. Marginal Transparency

The notion of *marginal transparency* borrows from economics. It asks: how much extra clarity do we gain by adding one more layer of disclosure? In wireless AI, the first few steps — such as showing feature importance or providing simplified flow diagrams — often yield major benefits. Operators quickly see whether spectrum allocation aligns with policy or whether a handover was triggered by signal strength.

But as disclosure continues, the returns diminish. Publishing every hyperparameter of a neural scheduler, or every gradient in an O-RAN anomaly detector, may add little to practical understanding. Engineers may already have enough visibility to trust the system, and further detail simply overwhelms. The idea of marginal transparency reminds us that explainability is not binary. Each extra disclosure matters, but not all matter equally, and designers must decide how far is enough.

### 2.4. Marginal Interpretability

*Marginal interpretability* extends the same logic to explanations. Early explanations — such as feature attributions from SHAP or counterfactual examples showing why a different resource allocation was denied — provide large gains in human understanding. Analysts can validate model behavior and check for fairness or compliance.

Yet as explanations become more technical or complex, the usefulness drops. Showing intricate interactions between dozens of radio features, or long mathematical proofs of why a neural precoder behaves a certain way, may confuse rather than clarify. In 6G, where network slicing and edge learning will only add complexity, this problem is likely to intensify. The principle of diminishing

 doi:10.20944/preprints202512.0667.v1



returns is clear: more explanation is not always better. Effective XAI for wireless must balance depth of detail with the limits of human comprehension. The diminishing returns of explainability in wireless systems is illustrated in Figure 2 .

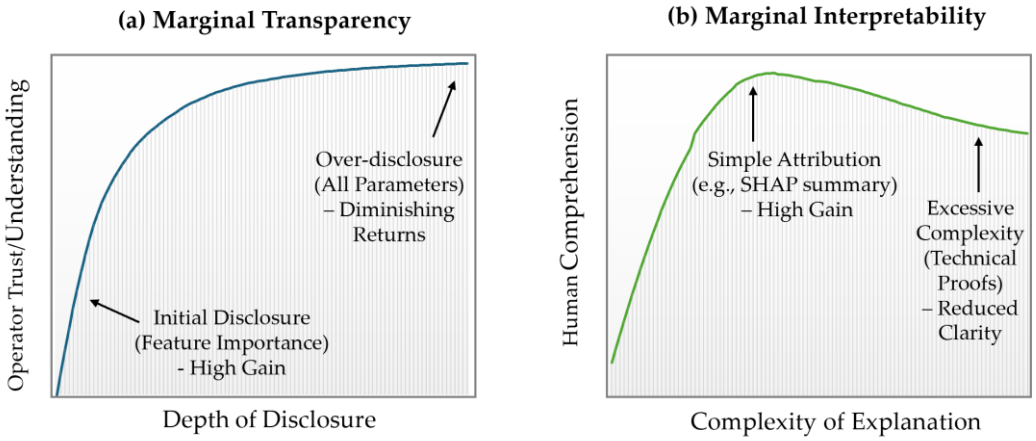


**Figure 2.** The diminishing returns of explainability in wireless systems. (a) Marginal Transparency: Illustrates that initial disclosures of model details yield significant trust gains, but publishing virtually every parameter offers diminishing practical value to operators (b) Marginal Interpretability: Shows that while simplified explanations enhance human understanding, overly complex or technical explanations can exceed cognitive limits and reduce clarity. Effective XAI must balance detail with comprehension limits.

## 3. Key Methods

Explainable AI in wireless networks is not an abstract concept; it is tied to specific tools and techniques. While the broader machine learning community often divides methods into model-agnostic and model-specific, the wireless domain gives these categories a particular flavor. When the goal is to make sense of spectrum allocation, interference control, or anomaly detection in 5G/6G, the choice of method affects not only interpretability but also latency, scalability, and operational trust.

### *3.1. Open-Source Toolkits*

Several open-source toolkits as listed in Table 1 have made XAI more practical for communication systems. IBM's AIX360, for example, provides a range of interpretability methods that can be adapted to wireless scheduling or QoS prediction. Alibi from SeldonIO has been used to test counterfactual explanations in network intrusion tasks. OmniXAI, with its graphical interface, allows engineers to quickly explore feature contributions without coding every detail. Others, such as Explabox or Xplique, focus on robustness checks, which can be crucial when adversaries try to trick intrusion detection systems.

The point here is not that one library solves everything. Rather, these toolkits show that explainability has matured to the point where wireless researchers do not have to build methods from scratch. With small adjustments — for instance, mapping SHAP outputs to radio metrics instead of tabular finance data — existing tools can already be useful.

**Table 1.** Representative open-source explainable AI (XAI) toolkits and their main functionalities.

| Toolkit | Developer / Maintainer | Main Functions | Reference |
|---|---|---|---|
| AIX360 (AI Explainability 360) | IBM Research | Comprehensive suite of local and global interpretability algorithms (e.g., LIME, SHAP, rule- | **[27]** |

 doi:10.20944/preprints202512.0667.v1



| | | | |
|---|---|---|---|
| | | based, contrastive explanations) | |
| Alibi | SeldonIO | Implements model-agnostic methods such as counterfactuals, anchors, and adversarial detectors | **[28]** |
| OmniXAI | Salesforce Research | Unified interface for post-hoc explanation (SHAP, LIME, Grad-CAM, Integrated Gradients) | **[29]** |
| Explabox | University of Glasgow / OpenXAI Community | Framework for testing robustness and reproducibility of explanations | **[30]** |
| Xplique | Inria / Sorbonne University | Lightweight library for explanation visualization and robustness evaluation | **[31]** |

### *3.2. Model-Agnostic Methods*

#### 3.2.1. Local Interpretable Model-Agnostic Explanations (LIME)

LIME works by building simple surrogate models around specific predictions [26]. In wireless networks, this might mean explaining why a base station decided to hand off a user at a given moment. The method perturbs the input — such as varying the reported signal-to-noise ratio or mobility pattern — and checks how the model's output changes. From there, a small linear model approximates the local decision boundary.

To be specific, the model explains predictions by approximating a complex model $f$ with an interpretable model $g$ in the neighborhood of the instance being explained. Let $x \in \mathbb{R}^d$ be the input, and $x' \in \{0,1\}^{d'}$its interpretable representation (e.g., words in text or super-pixels in images). The explanation is obtained by solving:

$$\xi(x) = \underset{g \epsilon G}{\operatorname{argmin}}\ L(f, g, \pi_x) + \Omega(g)$$

Where $L(f, g, \pi_x)$ measures how unfaithful $g$ is in approximating $f$ near $x$, $\pi_x(g)$ defines locality, and $\Omega(g)$ penalizes complexity to keep $g$ interpretable.

This technique has clear value. A network operator can see, for example, that a handover was triggered mainly by declining SINR rather than sudden cell congestion [12]. However, as several studies note, the stability of LIME explanations depends on how the perturbations are designed. In practice, too much randomization may yield different stories from one run to another [32,33].

#### 3.2.2. Shapley Additive Explanations (SHAP)

SHAP approaches the problem from game theory [34]. Each feature — channel state, interference level, traffic load — is treated as a "player" contributing to the model's decision. By distributing credit fairly across all combinations, SHAP provides both local and global views.

To be specific, The method unifies additive feature attribution approaches under a single framework. Let $f(x)$ be the original model and $x \in \mathbb{R}^M$ the input with $M$ features. SHAP explains $f(x)$ through an additive surrogate model:

$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z'_i$$

Where $z' \in \{0,1\}^M$ represents the presence or absence of features, and $\emptyset_i$ is the contribution of feature $i$.

In intrusion detection for 5G cores, SHAP can reveal that abnormal port activity consistently carries high importance, while mobility features matter less [21]. For spectrum allocation, global

 doi:10.20944/preprints202512.0667.v1



SHAP values may show that long-term traffic patterns dominate decisions more than instantaneous measurements [12]. The strength of SHAP is its consistency; its weakness is computational cost. In large-scale wireless systems with thousands of users and features, calculating exact Shapley values may be impractical. Engineers often need approximations or specialized variants to keep explanations within real-time limits.

### *3.3. Model-Specific Methods*

Not all models are black boxes. Some are inherently interpretable, and in wireless communications, these so-called "glass-box" models remain relevant.

#### 3.3.1. Decision Trees

Decision trees remain popular in scenarios where clarity is critical [35]. In a spectrum-sharing system, a tree might branch first on signal strength, then on interference, and finally on mobility. Each step is visible, and operators can follow the logic from root to leaf. The trade-off is well known: shallow trees are easy to interpret but may be inaccurate, while deep trees capture complex patterns but become harder to read. Random forests improve accuracy but at the expense of transparency.

#### 3.3.2. Interpretable Neural Networks

Researchers have tried to make neural networks less of a mystery. One approach is to embed attention mechanisms that highlight which features drive predictions [36]. In a beamforming model, for instance, attention weights can indicate which antennas were decisive for selecting a transmission pattern [37]. Another is to design architectures that enforce simple, interpretable structures, even if they remain deep [38]. These efforts matter because standard deep learning, though powerful, often leaves operators with results they cannot easily audit.

#### 3.3.3. Large Language Models in Networking

More recently, large language models (LLMs) have begun appearing in wireless research [39]. They are used for tasks like log analysis, configuration recommendation, or even natural-language interaction with network operators. LLMs can generate their own textual explanations, which at first glance looks like perfect interpretability. The reality is more complex. Such explanations may be fluent but not faithful — the model may "sound right" while missing the actual reasoning. Work on mechanistic interpretability, which tries to uncover internal circuits and representations in transformers, is an attempt to bridge that gap. For wireless applications, the challenge is to combine the accessibility of natural language with the reliability needed in critical infrastructure.

## 4. Taxonomy of Explainable AI

When researchers talk about the "taxonomy" of XAI, they usually mean a way of grouping methods by how and when explanations are generated [11,40]. In wireless communication systems, this classification becomes especially relevant. A 5G base station running intrusion detection has different needs from a 6G edge node balancing latency and privacy. The taxonomy presented here does not change the fundamentals, but it highlights how these categories matter in practice for next-generation networks.

### *4.1. Ante-Hoc vs. Post-Hoc Approaches*

Ante-hoc methods are interpretable by design. Models such as decision trees or rule-based systems fall into this category [41,42]. In wireless networks, they are often used where regulatory compliance is strict, such as access control policies or spectrum-sharing agreements [43]. A rule-based anomaly detector that blocks traffic when packet size exceeds a threshold is not glamorous, but it is transparent and auditable.

 doi:10.20944/preprints202512.0667.v1



Post-hoc methods, by contrast, take black-box models as given and then generate explanations. In a 5G intrusion detection system built on deep learning, saliency maps or counterfactuals may explain why certain traffic was flagged [44,45]. These methods are widely used but raise questions of faithfulness: does the explanation truly reflect the model, or is it only an approximation [46]? In adversarial environments, such as networks exposed to spoofing or jamming, this gap can be dangerous. Explanations that "look right" may still mislead analysts.

### *4.2. Local vs. Global Explanations*

The second axis distinguishes local and global explanations. Local explanations focus on individual cases. A practical example is fraud detection in mobile payments: an alert might be traced to a sudden location change combined with unusual transaction size [26]. Engineers can see why that one case was flagged.

Global explanations, on the other hand, describe the overall model behavior. For instance, in traffic classification across a 6G core, global SHAP values might show that long-term flow duration and packet entropy are the dominant factors. Such insights help regulators or operators understand whether the system systematically favors or penalizes certain types of traffic [43]. In compliance audits, global transparency is essential; without it, even a perfectly accurate system may not pass review.

### *4.3. Privacy-Preserving Explainability*

Wireless systems carry sensitive data. Location traces, mobility patterns, and device identifiers can all reveal private information. Explanations themselves risk leaking this data. A model might, for example, reveal that a specific user's movement pattern triggered a handover, unintentionally exposing personal details [47,48].

To address this, researchers explore privacy-preserving approaches. Differentially private explanations add controlled noise to protect identities [49,50]. Federated explainability generates explanations locally at edge devices, without sharing raw data [51]. These methods are still developing, but in multi-cloud or edge scenarios — common in 6G — they may prove indispensable [52]. The challenge is clear: explanations should enhance trust, not create new privacy risks.

### *4.4. Visual Taxonomy in Wireless Applications*

One way to make this taxonomy concrete is to map methods against their practical roles in wireless communication. Figure 3 illustrates how ante-hoc, post-hoc, local, and global explanations align with different operational and regulatory needs in 5G and 6G systems, while also indicating where privacy-preserving techniques become necessary. Ante-hoc models like decision trees are applied in spectrum policy enforcement [53]. Post-hoc tools such as LIME and SHAP are widely used in anomaly detection and fraud prevention. Local explanations support operational tasks, for example helping an operator decide whether to block a suspicious session [54]. Global explanations serve audits and long-term policy evaluations. Privacy-preserving methods are critical in IoT and edge computing, where data cannot easily leave the device [55].

Such a taxonomy is not rigid. In practice, methods overlap. A federated system may combine local and global views, or a hybrid approach may use both interpretable models and post-hoc visualizations [53]. What matters is that designers choose the right balance for the problem at hand. In 5G/6G, where latency, scale, and security interact in complex ways, no single axis of classification is enough.

 doi:10.20944/preprints202512.0667.v1



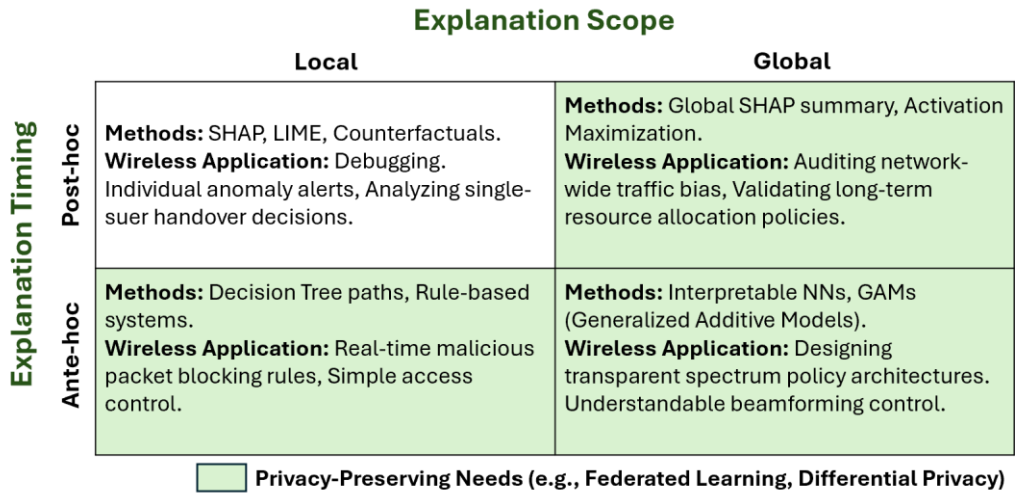


**Figure 3.** Visual taxonomy of explainable artificial intelligence (XAI) methods in wireless communication systems. The taxonomy organizes techniques along two axes, explanation timing and explanation scope, and links them with representative wireless applications. Ante-hoc and post-hoc approaches are shown in relation to local and global explanations, such as debugging individual anomaly alerts or auditing network-wide resource allocation policies. The taxonomy also identifies privacy-preserving needs, including federated learning and differential privacy, which are increasingly important in edge and multi-cloud 5G and 6G deployments.

## 5. Research Roadmap for 2025–2030+

Explainability in wireless networks is still a moving target. While existing methods provide useful insights, the requirements of 5G and future 6G systems create a much tougher environment. Networks are expected to manage spectrum dynamically, serve billions of IoT devices, and meet sub-millisecond latency demands. In such a setting, explanations cannot be an afterthought. They must be fast, scalable, and meaningful to engineers and regulators alike. Looking ahead, the research trajectory for 2025–2030 can be described in three overlapping stages: foundations, integration, and global standardization, as shown in Figure 4.

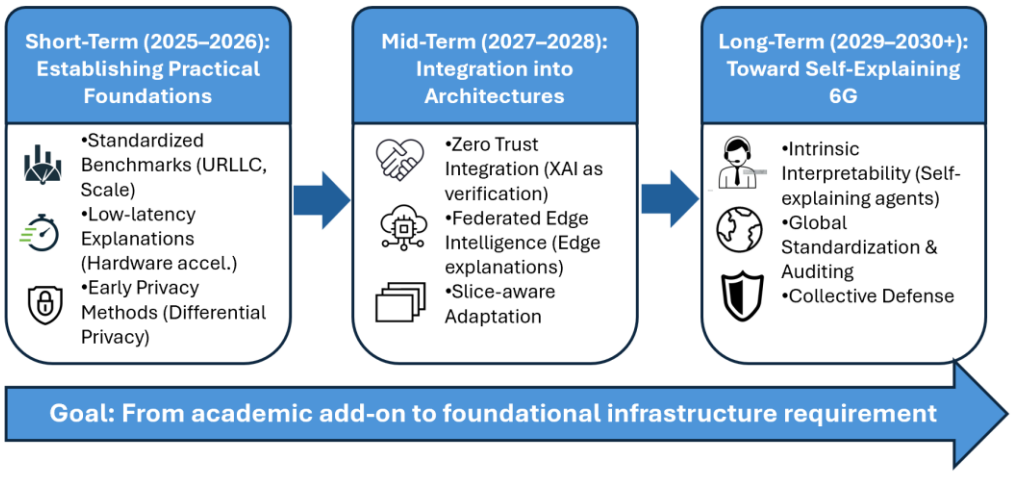


**Figure 4.** Roadmap for integrating explainable artificial intelligence (XAI) into next-generation wireless networks. The short-term phase focuses on establishing practical foundations such as standardized benchmarks, low-latency explanations, and early privacy methods. The mid-term phase highlights architectural integration including Zero Trust verification, federated edge intelligence, and slice-aware adaptation. The long-term phase

 doi:10.20944/preprints202512.0667.v1



moves toward intrinsic interpretability, global standardization, and collective defense, ultimately aiming for self-explaining 6G systems.

### *5.1. Short-Term (2025–2026): Establishing Practical Foundations*

In the immediate future, two priorities stand out. First, the community needs standardized evaluation benchmarks. Current XAI metrics — fidelity, comprehensibility, or stability — are often defined in abstract terms. For wireless systems, benchmarks must reflect real-world constraints: Does the explanation arrive quickly enough for URLLC? Does it scale to thousands of users in a cell? Can it be audited under telecom regulation? Without such criteria, different methods cannot be fairly compared.

Second, explanations must become low-latency. A spectrum allocation decision explained a second later is of little use if packets have already been dropped. Work such as XAInomaly [1] shows that lightweight variants of SHAP can be tuned for near real-time operation in O-RAN. This line of research will likely expand, focusing on approximations and hardware acceleration to bring explanation delays down to the millisecond range.

Privacy is also a short-term issue. Since user mobility and traffic traces are sensitive, early efforts in differentially private and federated explanations will be critical. In healthcare IoT or connected vehicles, an explanation that leaks individual identity could be more harmful than helpful.

### *5.2. Mid-Term (2027–2028): Integration into Wireless Architectures*

By the later 2020s, explainability will need to embed itself into the structure of 5G/6G systems rather than remain an external add-on. The most obvious step is integration with Zero Trust Architectures (ZTA). Wireless operators are moving toward ZTA to enforce continuous verification of devices and users. Here, XAI provides the justification layer: why a device was denied access, or why a trust score was downgraded. Without such explanations, Zero Trust policies may be viewed as arbitrary.

Another area is federated edge intelligence. With millions of IoT devices producing data, centralized training is impractical. Federated learning allows models to be trained locally and aggregated globally, but explanations must follow the same path. That means interpretable outputs generated at the edge, consumable both by local operators and by central regulators. Research into adaptive explanations — where detail levels shift depending on the recipient, from technicians to policymakers — will also gain traction.

Finally, network slicing presents a new frontier. Different slices serve very different needs (emergency services vs. consumer video vs. industrial automation). Explanations must adapt accordingly, showing not only why resources were allocated but also how slice-level priorities shaped the decision.

### *5.3. Long-Term (2029–2030+): Toward Self-Explaining 6G Systems*

Looking further ahead, the vision is for networks that explain themselves. Instead of relying solely on post-hoc tools, the architectures of 6G may embed interpretability into their very design. Self-explaining agents could justify routing, beamforming, or security decisions in natural language while still being auditable against standards.

Global standardization will also come to the forefront. Just as ISO/IEC 27001 unified cybersecurity practices, a comparable framework for explainability in wireless AI is likely to emerge. Such standards would define not only performance requirements but also how explanations must be logged, audited, and shared across borders. This becomes especially important in roaming scenarios or multinational service providers, where consistency of trust is essential.

Another long-term theme is collective defense through explainability. Imagine 6G security systems in different countries exchanging not raw traffic but interpretable alerts, allowing

 doi:10.20944/preprints202512.0667.v1



collaborative detection of global attacks without breaching data sovereignty. This vision is ambitious, but the groundwork is already visible in early federated and privacy-preserving research.

### *5.4. Synthesis*

The path ahead is not linear but layered. In the short term, benchmarks and latency improvements will make XAI usable. In the mid-term, integration with Zero Trust, federated edge intelligence, and slicing will make it indispensable. In the long term, intrinsic interpretability and global standards may turn XAI into a foundation of wireless trust itself. By 2030, explainability should not be viewed as a luxury for academics but as a baseline requirement for operating the world's most critical communication infrastructure.

## 6. Applications of Explainable AI in 5G/6G

The move toward 5G and 6G has transformed wireless networks into highly intelligent and data-driven systems. With this transformation, questions of trust, accountability, and security are no longer optional. Explainable AI (XAI) provides a set of tools to address these concerns. In this section, we highlight several domains where explainability can make a tangible difference in next-generation wireless communication.

### *6.1. Transparent Resource Allocation*

AI is increasingly used for dynamic spectrum management and beamforming control. These tasks require fast decisions, but operators still want to know why one user or slice was prioritized over another. XAI methods can provide clarity [54]. For example, a SHAP-based explanation might show that a particular user received additional bandwidth because of persistent low signal-to-noise ratio combined with urgent latency requirements. Such transparency helps operators verify that allocation policies align with service-level agreements and fairness expectations. Without these insights, automatic decisions may be viewed as arbitrary, which could limit trust and adoption.

### *6.2. Privacy-Preserving IoT and Edge Learning*

In 6G environments, billions of IoT devices—from wearables to home sensors—connect through mobile edge computing (MEC) infrastructures. These devices continuously produce sensitive data such as location traces, activity patterns, and biometric signals. When AI models analyze such data at the edge, the explanations they generate may inadvertently reveal private information. For example, an explanation stating "abnormal motion detected on Device X" or highlighting a distinctive biometric pattern could expose a user's identity or behavior.

To mitigate this risk, emerging approaches seek to design privacy-preserving explainability mechanisms that operate directly on edge devices. As shown in Figure 5, attention-driven federated learning provides a practical pathway: each device performs local training and generates attention-based explanations internally, ensuring the raw data and sensitive attribution signals never leave the device.

This approach, demonstrated by Pande et al. [55], allows local attention weights to identify which segments of the input—whether ECG windows, motion patterns, or sensor readings—contributed most to a prediction, offering interpretable feedback while maintaining strict data locality.

After local training, only attention-weighted model updates are transmitted to fog nodes and the global server for aggregation. This hierarchical process avoids the transfer of raw sensor data and reduces the risk of explanation leakage, while still enabling the global model to improve using contributions from heterogeneous IoT clients. In healthcare IoT or smart-home environments, this combination of federated learning and on-device explainability provides meaningful insights without compromising confidentiality.

 doi:10.20944/preprints202512.0667.v1



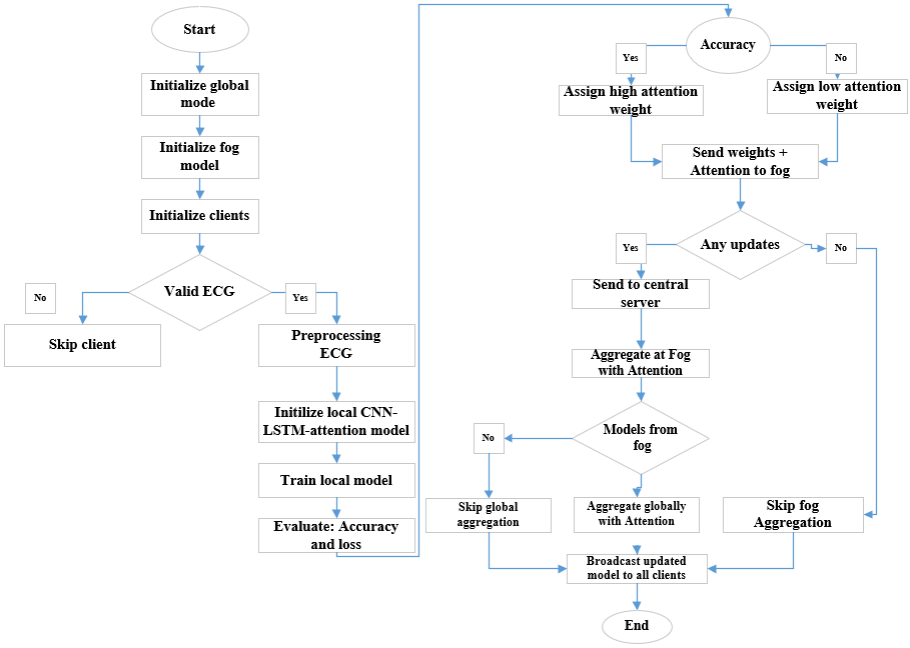


**Figure 5.** Flowchart representation of the attention-driven federated learning framework. The process includes client-side ECG validation, local CNN–LSTM attention model training, attention-based weighting of updates, fog-level aggregation, and final global aggregation before broadcasting the updated model to all clients. [55].

The key challenge is to maintain this balance: explanations must be informative enough for debugging, monitoring, and regulatory auditing, yet private enough to prevent reconstruction of individual behaviors. Attention-based federated architectures offer a promising foundation for achieving trustworthy and privacy-preserving edge intelligence in future 6G networks.

### *6.3. Regulatory Compliance and Trustworthy AI*

Telecom operators face increasing regulatory pressure as AI-driven automation becomes embedded in network management. Europe's General Data Protection Regulation (GDPR) introduces safeguards for individuals affected by automated decision-making, including provisions commonly interpreted as a form of a "right to explanation" for algorithmic outputs. Similar expectations are emerging in other jurisdictions—ranging from consumer-protection requirements in the United States to transparency obligations in Asia-Pacific telecom regulations. For mobile networks, this means that automated actions such as credit-based access assignment, prioritization of emergency traffic, fraud detection, or QoS throttling may require justification in human-interpretable terms.

Explainable AI (XAI) provides a concrete mechanism to meet these obligations. Through model-agnostic attribution methods or rule-based summaries, operators can show how a trust score was computed, why a specific request was flagged as suspicious, or which features influenced a throttling decision. In regulatory investigations, explanations can demonstrate proportionality, fairness, and non-discrimination—criteria emphasized in emerging AI governance frameworks such as the EU Artificial Intelligence Act [56].

Beyond compliance, explanation mechanisms support internal governance. Telecom operators increasingly integrate explanation logs into their operational pipelines, maintaining a parallel record of model behavior alongside performance metrics and network events. Such logs can help validate policy alignment, identify unintended bias, and offer defensible evidence in audits or legal disputes. As 5G and 6G systems rely more heavily on automated decision engines for spectrum allocation, mobility prediction, and anomaly detection, trustworthy AI practices will become as essential as traditional security and reliability requirements [57].

 doi:10.20944/preprints202512.0667.v1



Ultimately, explainability is not only a regulatory checkbox but a foundational capability for building public trust. Operators who can clearly articulate why an automated decision occurred will be better positioned to meet both legal obligations and user expectations in an increasingly algorithmic telecommunications ecosystem.

### *6.4. Adversarial Robustness in Wireless Systems*

Wireless AI models are increasingly exposed to adversarial manipulation. Attackers can inject poisoned samples, distort I/Q sequences, or craft modulation-shaped perturbations that fool deep learning–based classifiers. These attacks often remain visually imperceptible but can drastically shift a model's decision boundary. Explainable AI offers a way to expose such hidden vulnerabilities by revealing how a classifier internally responds to perturbed wireless signals.

As illustrated in Figure 6 of Dong et al. [58], the SHAP-AFT framework computes Shapley values over received wireless sequences and identifies destructive feature points—time–frequency positions whose contributions become negative under adversarial perturbation. These negative Shapley regions indicate that the model is being pushed toward incorrect decisions. By visualizing how these contribution patterns shift, operators can detect when a classifier begins reacting abnormally to adversarial noise rather than to meaningful modulation structures.

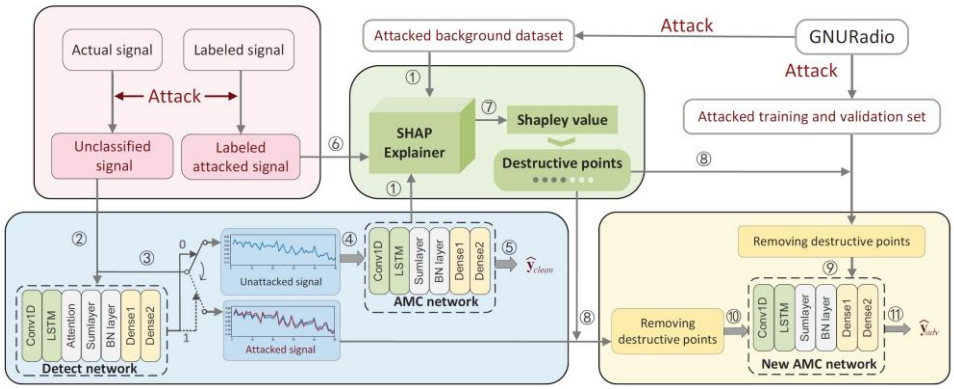


**Figure 6.** Workflow of SHAP-assisted adversarial detection and model refinement for automatic modulation classification. The attacked signals are processed by a detection network, and SHAP is used to identify destructive points in the AMC model. These points are removed to retrain an improved AMC network with enhanced robustness. [58].

In practice, such explainability cues function as early warning indicators. For example, a saliency map that should highlight the characteristic temporal transitions of a modulation format may instead become dominated by attacker-induced high-frequency jitter. When negative Shapley contributions expand or relocate, it signals internal model degradation even before a notable increase in misclassification rates.

Yet explanations themselves may also be manipulated by sophisticated attackers. This makes robust explainability essential. The defense strategy illustrated in Figure 6 integrates explanation analysis into the model-adaptation loop: destructive feature points are identified and removed, and the AMC classifier is fine-tuned on refined adversarial samples [58]. This creates redundancy—if multiple explanation signals (e.g., gradient-based saliency, perturbation sensitivity, Shapley contribution maps) consistently flag anomalies, it becomes significantly more difficult for adversarial traffic to forge plausible yet deceptive explanations.

For wireless security teams, such explainability-enhanced defense mechanisms build trust by revealing why the model fails and *how* adversarial perturbations propagate. In dense 5G/6G settings—where jamming, spoofing, and crafted waveforms may co-occur—this layered interpretability becomes a critical component of robust wireless AI.

 doi:10.20944/preprints202512.0667.v1



### *6.5. Human-in-the-Loop Network Management*

Despite advances in automation, human expertise remains central to running mobile networks. Operators often prefer to cross-check AI-driven recommendations with their own judgment. XAI supports this by making AI outputs interpretable. In O-RAN, for example, an anomaly detection xApp might flag abnormal traffic, while explanations show which flow features caused the alert. Analysts can then validate or override the system's decision.

This interaction creates a feedback loop: humans correct mistakes, and models improve with retraining. Over time, such collaboration increases both system accuracy and operator trust. In critical services — emergency communications, industrial control, or connected vehicles — this synergy between human insight and machine speed may prove indispensable.

### *6.6. Synthesis*

Applications of XAI in 5G/6G demonstrate that explainability is not a theoretical luxury but a practical necessity. From spectrum allocation and IoT privacy to regulatory compliance and adversarial defense, interpretability shapes how next-generation networks will be deployed and trusted. The lesson is straightforward: high performance is essential, but without explanations, even the most accurate models may be rejected in practice.

## 7. Conclusion

This review has argued that in 5G and emerging 6G networks, explainability is not an optional add-on to AI-driven functions but a core requirement for trust, safety, and compliance. We began by clarifying transparency and interpretability, then introduced the notions of marginal transparency and marginal interpretability to capture the diminishing returns of ever-deeper disclosure. These concepts highlight that "more explanation" is not always better: wireless operators, regulators, and users need the *right* level of insight at the *right* time, not an unfiltered view of every weight, gradient, or hyperparameter. Building on this foundation, we surveyed model-agnostic tools such as LIME and SHAP, inherently interpretable models like decision trees and structured neural networks, and emerging uses of large language models in network operations. We further organized the space through a taxonomy that distinguishes ante-hoc vs. post-hoc and local vs. global explanations, while emphasizing that privacy-preserving mechanisms—federated and differentially private explainability—will be essential as sensitive mobility and IoT data move closer to the edge.

The research roadmap from 2025 to 2030+ suggests a phased but overlapping progression: first, establishing realistic benchmarks and low-latency explanation pipelines; next, embedding explainability into Zero Trust architectures, federated edge intelligence, and slice-aware control planes; and finally, moving toward self-explaining 6G systems governed by global standards for logging, auditing, and sharing explanations. Concrete applications already show how XAI can make a difference today: transparent resource allocation, privacy-preserving IoT and MEC, regulatory reporting under GDPR-style regimes, adversarially robust modulation classification, and human-in-the-loop O-RAN analytics. Across these domains, the pattern is consistent: high predictive performance is necessary, but without explanations that are timely, faithful, and privacy-aware, AI-driven wireless functions will face resistance from operators, regulators, and end users.

Taken together, these insights point to a simple but demanding conclusion: future wireless networks must be designed as *explainable-by-default* infrastructures. XAI should not merely justify individual decisions after the fact; it should shape how algorithms, protocols, and architectures are conceived in the first place. If the community can align methods, metrics, and standards around this goal, explainability will evolve from a niche research topic into a foundational pillar of trustworthy 5G/6G systems—enabling networks that are not only faster and more intelligent, but also auditable, accountable, and worthy of the critical roles they will play in digital society.

 doi:10.20944/preprints202512.0667.v1



## Reference


1. Wang, C.-X.; Di Renzo, M.; Stanczak, S.; Wang, S.; Larsson, E.G. Artificial intelligence enabled wireless networking for 5G and beyond: Recent advances and future challenges. *IEEE Wireless Communications* **2020**, *27*, 16-23.
2. Shaer, I. Network Resource and Performance Optimization in Autonomous Systems: A Connected Vehicles and Autonomous Networks Perspective. The University of Western Ontario (Canada), 2020.
3. Huang, H.; Yang, J.; Huang, H.; Song, Y.; Gui, G. Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system. *IEEE Transactions on Vehicular Technology* **2018**, *67*, 8549-8560.
4. Brilhante, D.d.S.; Manjarres, J.C.; Moreira, R.; de Oliveira Veiga, L.; de Rezende, J.F.; Müller, F.; Klautau, A.; Leonel Mendes, L.; P. de Figueiredo, F.A. A literature survey on AI-aided beamforming and beam management for 5G and 6G systems. *Sensors* **2023**, *23*, 4359.
5. Hussain, F.; Hussain, R.; Hassan, S.A.; Hossain, E. Machine learning in IoT security: Current solutions and future challenges. *IEEE Communications Surveys & Tutorials* **2020**, *22*, 1686-1721.
6. Jiang, W.; Han, B.; Habibi, M.A.; Schotten, H.D. The road towards 6G: A comprehensive survey. *IEEE Open Journal of the Communications Society* **2021**, *2*, 334-366.
7. Zhang, C.; Patras, P.; Haddadi, H. Deep learning in mobile and wireless networking: A survey. *IEEE Communications surveys & tutorials* **2019**, *21*, 2224-2287.
8. Samek, W.; Wiegand, T.; Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* **2017**.
9. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* **2020**, *32*, 4793-4813.
10. Gunning, D.; Aha, D. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* **2019**, *40*, 44-58.
11. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82-115.
12. Zuo, Y.; Guo, J.; Gao, N.; Zhu, Y.; Jin, S.; Li, X. A survey of blockchain and artificial intelligence for 6G wireless communications. *IEEE Communications Surveys & Tutorials* **2023**, *25*, 2494-2528.
13. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), 2018; pp. 80-89.
14. Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* **2017**, *38*, 50-57.
15. Liu, X.; Huang, D.; Yao, J.; Dong, J.; Song, L.; Wang, H.; Yao, C.; Chu, W. From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI). *AI* **2025**, *6*, 285.
16. Liu, Z.; Pang, B.; Sun, F.; Li, Q.; Zhang, Y. Camera-Aware Graph Consistency based Unsupervised Domain Adaptive Person Re-identification. In Proceedings of the Journal of Physics: Conference Series, 2025; p. 012027.
17. Liu, Z.; Feng, H.; Sun, F.; Wang, Q.; Tian, F. Research on Pedestrian Re-identification Methods Based on Local Feature Enhancement Using Self-Feedback Human Analysis. In Proceedings of the Journal of Physics: Conference Series, 2025; p. 012028.
18. Huang, X.; Lin, Z.; Sun, F.; Zhang, W.; Tong, K.; Liu, Y. A Multi-Hop Retrieval-Augmented Generation Framework for Intelligent Document Question Answering in Financial and Compliance Contexts. **2025**.
19. Yu, Y.; Sun, F.; Sun, A. Multi-Granularity Adapter Fusion with Dynamic Low-Rank Adaptation for Structured Privacy Policy Understanding. In Proceedings of the 2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), 2025; pp. 481-484.
20. Basaran, O.T.; Dressler, F. XAInomaly: Explainable, interpretable and trustworthy AI for xURLLC in 6G open-RAN. In Proceedings of the 2024 3rd International Conference on 6G Networking (6GNet), 2024; pp. 93-101.

 doi:10.20944/preprints202512.0667.v1




21. Uccello, F.; Nadjm-Tehrani, S. Investigating Feature Attribution for 5G Network Intrusion Detection. *arXiv preprint arXiv:2509.10206* **2025**.
22. Kaur, N.; Gupta, L. Securing the 6G–IoT Environment: A Framework for Enhancing Transparency in Artificial Intelligence Decision-Making Through Explainable Artificial Intelligence. *Sensors* **2025**, *25*, 854.
23. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31-57.
24. Burkart, N.; Huber, M.F. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* **2021**, *70*, 245-317.
25. Van Der Waa, J.; Verdult, S.; Van Den Bosch, K.; Van Diggelen, J.; Haije, T.; Van Der Stigchel, B.; Cocu, I. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI* **2021**, *8*, 640647.
26. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016; pp. 1135-1144.
27. Arya, V.; Bellamy, R.K.; Chen, P.-Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* **2019**.
28. Klaise, J.; Van Looveren, A.; Vacanti, G.; Coca, A. Alibi explain: Algorithms for explaining machine learning models. *Journal of Machine Learning Research* **2021**, *22*, 1-7.
29. Yang, W.; Le, H.; Laud, T.; Savarese, S.; Hoi, S.C. Omnixai: A library for explainable ai. *arXiv preprint arXiv:2206.01612* **2022**.
30. Robeer, M.; Bron, M.; Herrewijnen, E.; Hoeseni, R.; Bex, F. The Explabox: Model-Agnostic Machine Learning Transparency & Analysis. *arXiv preprint arXiv:2411.15257* **2024**.
31. Fel, T.; Hervier, L.; Vigouroux, D.; Poche, A.; Plakoo, J.; Cadene, R.; Chalvidal, M.; Colin, J.; Boissin, T.; Bethune, L. Xplique: A deep learning explainability toolbox. *arXiv preprint arXiv:2206.04394* **2022**.
32. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020; pp. 180-186.
33. Kumar, I.E.; Venkatasubramanian, S.; Scheidegger, C.; Friedler, S. Problems with Shapley-value-based explanations as feature importance measures. In Proceedings of the International conference on machine learning, 2020; pp. 5491-5500.
34. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
35. Quinlan, J.R. C4. 5: programs for machine learning; Elsevier: 2014.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
37. Jabbarvaziri, F.; Lampe, L. Attention-Based Deep Learning for Hybrid Beamforming in OFDM Systems with Phase Noise. *IEEE Transactions on Wireless Communications* **2025**.
38. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **2019**, *1*, 206-215.
39. Zhou, H.; Hu, C.; Yuan, D.; Yuan, Y.; Wu, D.; Chen, X.; Tabassum, H.; Liu, X. Large language models for wireless networks: an overview from the prompt engineering perspective. *IEEE Wireless Communications* **2025**.
40. Adadi, A.; Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* **2018**, *6*, 52138-52160.
41. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **2019**, *116*, 22071-22080.
42. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016; pp. 2921-2929.

 doi:10.20944/preprints202512.0667.v1




43. Guo, W. Explainable artificial intelligence for 6G: Improving trust between human and machine. *IEEE Communications Magazine* **2020**, *58*, 39-45.
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017; pp. 618-626.
45. Mothilal, R.K.; Sharma, A.; Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020; pp. 607-617.
46. Lakkaraju, H.; Bastani, O. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020; pp. 79-85.
47. Chakkappan, G.; Morshed, A.; Rashid, M.M. Explainable AI and Big Data Analytics for Data Security Risk and Privacy Issues in the Financial Industry. In Proceedings of the 2024 IEEE Conference on Engineering Informatics (ICEI), 2024; pp. 1-9.
48. Shokri, R.; Strobel, M.; Zick, Y. On the privacy risks of model explanations. In Proceedings of the Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021; pp. 231-241.
49. Harder, F.; Bauer, M.; Park, M. Interpretable and differentially private predictions. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020; pp. 4083-4090.
50. Tao, Y.; Gilad, A.; Machanavajjhala, A.; Roy, S. Differentially private explanations for aggregate query answers. *The VLDB Journal* **2025**, *34*, 20.
51. Renda, A.; Ducange, P.; Marcelloni, F.; Sabella, D.; Filippou, M.C.; Nardini, G.; Stea, G.; Virdis, A.; Micheli, D.; Rapone, D. Federated learning of explainable AI models in 6G systems: Towards secure and automated vehicle networking. *Information* **2022**, *13*, 395.
52. Nguyen, T.T.; Huynh, T.T.; Ren, Z.; Nguyen, T.T.; Nguyen, P.L.; Yin, H.; Nguyen, Q.V.H. Privacy-preserving explainable AI: a survey. *Science China Information Sciences* **2025**, *68*, 111101.
53. Senevirathna, T.; La, V.H.; Marcha, S.; Siniarski, B.; Liyanage, M.; Wang, S. A survey on XAI for 5G and beyond security: Technical aspects, challenges and research directions. *IEEE Communications Surveys & Tutorials* **2024**, *27*, 941-973.
54. Khan, N.; Abdallah, A.; Celik, A.; Eltawil, A.M.; Coleri, S. Explainable AI-aided Feature Selection and Model Reduction for DRL-based V2X Resource Allocation. *IEEE Transactions on Communications* **2025**.
55. Pande, P.; Babu, B.M.; Bhargav, P.; Roy, T.D.; Muniyandy, E.; El, T.D.Y.A.B.; Ebiary, D.V. Attention-Driven Hierarchical Federated Learning for Privacy-Preserving Edge AI in Heterogeneous IoT Networks.
56. Veale, M.; Borgesius, F.Z. Demystifying the draft EU artificial intelligence act. *arXiv preprint arXiv:2107.03721* **2021**.
57. Sun, Y.; Peng, M.; Zhou, Y.; Huang, Y.; Mao, S. Application of machine learning in wireless networks: Key techniques and open issues. *IEEE Communications Surveys & Tutorials* **2019**, *21*, 3072-3108.
58. Dong, P.; Wang, J.; Gao, S.; Zhou, F.; Wu, Q. Explainable Deep Learning Based Adversarial Defense for Automatic Modulation Classification. *arXiv preprint arXiv:2509.15766* **2025**.