Explainable AI for Trustworthy 6G Edge-Cloud Smart Cities: A Systematic Review

Qaisar Abbas*, Riyad Almakki
College of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic University (IMSIU), 11432, Riyadh, Saudi Arabia

***Abstract*—Smart cities are becoming more interconnected, data-driven, and increasingly autonomous, with 6th-generation communication, edge-cloud computing, Internet of Things infrastructures, digital twins, federated learning, and artificial intelligence supporting them. The same technology that enables low-latency mobility, adaptive energy management, public safety analytics, environmental analytics, and responsive digital governance is creating a hard accountability conundrum. The decisions can be generated by black-box models across multiple computing tiers, take advantage of decentralized learning, and be executed before the human operator has time to review the reasoning. This review focuses on the trustworthy and explainable design of AI in these contexts. In this review, the focus is on the trustworthy and explainable design of AI in such contexts. The review is based on a multi-dimensional taxonomy of explanation timing, scope, target, representation, model dependence, computing location, stakeholder, trustworthiness objective, and urban application. It also outlines a tiered reference architecture that links local explanations at devices, operational explanations at mobile-edge nodes, network-level explanations for 6G orchestration, and strategic explanations for cloud-hosted urban digital twins. The synthesis confirms that the use of attribution techniques (SHAP, LIME, gradient-based saliency, feature importance) is prevalent, and that counterfactual explanations, concept-based explanations, causal explanations, inherently interpretable explanations, and interactive explanations are less developed for use in city-level deployments. Lacking are explanation fidelity for distribution shifts and real-time latency and energy cost, privacy leakage, federated explanation consistency, security against explanation manipulation, interoperability across urban digital twins, and weak human-centered evaluation. Finally, the study provides a research agenda and an engineering checklist of reliable, accountable, private, and sustainable urban intelligence.**

Keywords*—*Explainable artificial intelligence; trustworthy artificial intelligence; smart cities; 6G networks; edge intelligence; cloud computing; federated learning; digital twins; cybersecurity; AI governance; human-centered AI; sustainable urban intelligence

I. Introduction

Smart cities are no longer just technology-supported systems, but rather socio-technical systems engaged in constant interaction between sensing, communication, data analysis, artificial intelligence, infrastructure management, and public governance. Previous studies focused on the use of information and communication technologies for individual services. Recent research defines smart sustainable cities as "adaptive systems that should enhance environmental performance, public welfare, city resilience and inclusion, while being constrained by resources, laws and institutions" [1-5]. It's an important change since urban intelligence no longer exists solely in dashboards or past reporting. It increasingly foresees demand, allocates resources, identifies risks, suggests solutions, and even takes actions in transportation, energy, health, public safety, environment, construction, and digital government.

The architecture of the computing services is also evolving. While centralized cloud platforms offer solutions for storing data at scale, training models, and analyzing domains, they do not necessarily work for applications where latency is a key consideration, where privacy protection is paramount, or where connections are intermittent. Fog and edge computing bring data processing closer to vehicles, cameras, meters, gateways, hospitals, and public infrastructure, which decreases round-trip latency and bandwidth consumption [6–10]. The edge-cloud continuum is not a straightforward technical tuning. It transforms the place, the people, and the records of decision-making, the evidence that is kept, and the accountability shared by device vendors, network operators, providers of cloud services, public agencies, and human decision-makers.

Federated learning and urban digital twins further complicate things. For data that is institutionally distributed or legally sensitive, federated learning offers the potential for collaborative model improvement without the need to centralize all raw data, which is useful [11]. Digital twins are virtual replicas of physical assets and urban processes that enable simulation, prediction, monitoring, and testing of urban operations and policymaking [12]. Combined, the technologies enable distributed learning and reasoning at the system level, yet they pose additional challenges for maintaining model provenance, consistency of explanations, and accountability. The prediction can be based on local sensors, a federated global model, a federated decision about the resources of the network, and a prediction from a digital-twin simulation. Any explanation that addresses just one component of the chain may be technically correct but not operationally understandable.

The proliferation of advanced machine learning and deep learning predictive and decision-making models thus raises the center of trust issue. Engineers, citizens, operators, and policymakers have trouble understanding many high-performing models. Explainable AI tackles this problem by revealing influential features, examples, concepts, rules, counterfactual changes, uncertainty, and causal relationships [13-16]. However, "explainability" does not guarantee "trustworthy behavior. An explanation might be inconsistent,

*Corresponding author.

unfaithful, overly technical, misleading, revealing too much information about privacy, or perhaps unsuitable for its target market. Therefore, the concept of a trustworthy AI calls for a wider range of properties, such as validity, reliability, robustness, security, privacy, fairness, transparency, accountability, safety, and meaningful human oversight [17].

This is likely to be exacerbated by 6th-generation wireless communication. The 6G vision features very high data rates, ultra-low latency, massive connectivity, integrated sensing and communication, non-terrestrial coverage, semantic or goal-oriented communication, as well as the ability to control the entire network using AI [18–20]. AI will be applied to radio-resource management, mobility, network slicing, service placement, anomaly detection, and adaptive orchestration on future networks. The same networks will enable smart-city applications to be connected to autonomous vehicles, drones, sensors, robots, buildings, public-safety systems, and digital twins. The lack of transparency in network and network architecture decisions makes network failures hard to diagnose and challenge. A citizen impacted by the automated urban decision might not even be able to tell you what caused the decision, whether it was because of the application model, biased training data, an approximation on the edge that was resource-limited, an old federated model, or an optimization policy for the networks.

There are existing reviews that provide good foundations, but are fragmented. Methods and evaluation metrics are classified in XAI surveys, but generally deployment is considered a single-model problem [13–16, 38–53]. The computation, communication, privacy, and optimization aspects of edge and federated-learning surveys have been proposed, but rarely define the cross-layer movement of explanations [67–75, 80, 81]. In the context of digital-twin reviews, modeling and interoperability problems and governance issues are addressed without a common description of explanation provenance [76–79]. Although 6G enabling technologies and network requirements are explained in the 6G literature, only recent studies directly link 6G explainability to 6G use cases [20, 82–94]. Smart-city reviews focus on sustainability, services, ethics, security, and energy, but fail to integrate XAI design with distributed edge-cloud and 6G architectures to the fullest extent [95–100].

This review tackles that integration issue. It's not just an explanation algorithm list. It explores what is to be explained, from where explanations should be created, to whom, to which trust objective, how they are assessed, and how they are kept reliable on devices, edge nodes, networks, in the cloud, and in digital twins. The primary contributions are five-fold. The study first develops an up-to-date 100-reference corpus, which is a compilation of the 20 references from the seed manuscript and 80 good-quality reference sources. Second, it establishes a common taxonomy connecting the properties of XAI methods to computing layers, stakeholders, dimensions of trustworthiness, and urban domains. Thirdly, it suggests a reference design of 6G edge-cloud systems with explanations. Fourth, it brings together deployment patterns, threats, and evaluation metrics in transportation, energy, healthcare, public safety, environmental monitoring, and governance. Fifth, it offers a research agenda of concrete technical questions and practical design advice.

The rest of this study is organized as follows: The review protocol and its limitations are described in Section II. The conceptual foundations are defined in Section III. The reference architecture is introduced in Section IV. The taxonomy is given in Section V. Section VI combines application domains. Securing trust, security, privacy, and governance will be explored in Section VII. This is followed by a corpus analysis of trends in maturity in Section VIII. The future research agenda is proposed in Section IX. Section X gives engineering and policy guidelines, with conclusions in Section XI.

II. Review of Protocol

The review process was structured according to PRISMA reporting principles, following the identification, screening, eligibility, and inclusion stages summarized in Fig. 1. Candidate studies retrieved from the selected databases and supplementary sources were screened for relevance, duplicate records were removed, and full-text eligibility was assessed using the predefined inclusion and exclusion criteria. This PRISMA-guided selection process resulted in a final corpus of 100 references, which was retained for qualitative synthesis and descriptive corpus analysis. Fig. 1 provides a transparent account of the study-selection pathway used to construct the final analytical corpus.

A targeted search strategy was employed on major publisher platforms and authoritative databases such as IEEE, ACM, Elsevier, Springer Nature, MDPI, Frontiers, and the National Institute of Standards and Technology. Search concepts were amalgamated into multiple families: explainable AI or interpretable machine learning; trustworthy AI, responsible AI, fairness, accountability, privacy, or human-centered AI; edge intelligence, mobile edge computing, cloud, federated learning, or digital twin; 6G, AI-native network, semantic communication, network slicing, or integrated sensing and communication; and smart city, transportation, energy, healthcare, safety, environment, building, or governance. To boost coverage, backward citation chaining from large surveys and forward-oriented searches for 2023-2026 reviews were carried out.

If it met at least one of four relevance conditions, it was included in the studies. The source had to include some form of foundational or commonly used method for XAI, a rigorous XAI taxonomy/evaluation framework, a treatment of edge-cloud, federated, digital-twin, or 6G infrastructure, or a substantive analysis of AI-enabled smart-city applications, ethics, security, privacy, or governance. Only journal articles, major peer-reviewed conference papers, and authoritative technical frameworks were included. Sources were removed if they were redundant, peripheral to the integrated topic, had insufficient methodological content, or contained promotional claims without technical and/or empirical backing. The update emphasized work since 2020 but included previous foundational studies if they were not included, they would not complete the method taxonomy.

Each source was coded by publication year and primary theme. The detailed qualitative extraction used nine dimensions: contribution type; AI model family; explanation family; explanation scope; computing location; urban application; trustworthiness objective; evaluation approach; and deployment maturity. Because many sources span more than one category, the figures in this study use a single primary-theme code to avoid double counting. These figures should be interpreted as descriptive views of the curated corpus, not as a complete bibliometric map of all publications in the field.

The synthesis combined aggregative and configurative reasoning. Aggregative synthesis counted publication years and primary themes. Configurative synthesis connected technical ideas that are usually studied separately, such as explanation fidelity, network-level orchestration, federated consistency, digital-twin provenance, public-sector accountability, and citizen contestability. The research questions are listed in Table I. The protocol and quality checks are summarized in Table II, while Fig. 1 shows corpus construction.

III. Conceptual Foundations

The term explainable AI is applied unevenly. The property of interpretability is often used to describe a property of a model whose operation can be explained directly, such as a sparse rule list or a constrained generalized additive model. Explanability is typically used to describe information that is produced to aid in understanding a model or a specific output, such as post-hoc explanations that provide users a way to understand an opaque model [13-16, 32, 33]. This division is helpful but not definitive. Mathematically interpretable models can be in urban systems; the decision process is opaque if there are data transformations, service composition, network constraints, human procedures, etc.

A second distinction is with respect to explanation scope. Global explanations define model behavior, decision logic, feature effects, or learned concepts in general. Local explanations: One prediction, recommendation, or action. LIME builds a local surrogate near an instance [21], and Shapley gives an explanation of a prediction based on Shapley-value principles [22]. Gradient-based methods like Grad-CAM or integrated gradients provide insights into regions of the input that are important to the model or features [23, 24]. Perturbation methods estimate the changes in outputs that occur when evidence is perturbed [25]. Another group of methods relates a prediction to training examples via the concept of influence functions and example-based methods [26]. Counterfactual explanations involve highlighting changes that would impact an outcome [27, 46]. Concept-based explanations are decisions based on human-meaningful concepts, instead of just on raw features [28]. These are different families and should not be used interchangeably to answer the same question.

A third difference is between explanation and justification. An explanation should characterize the actual mechanism or evidence that influenced a model. A justification can give a convincing reason, making it seem like a decision was made on a good basis, but not accurately explain the model. The risk is particularly high for generative explanations generated by language models. There is a tendency to conceal low fidelity by writing floppy text. Natural-language explanation interfaces for critical urban applications must be based on a model's evidence, provenance, confidence, and explicitly stated uncertainty. The generated text should be used as a communication layer, rather than the source of the explanation.

There are also multiple aspects of how trustworthy someone is. Technical correctness is required but a reliable system should be reliable in the presence of distribution shift, adversarial behavior, missing sensors, device failures, and organizational change. A security concern is whether models and explanations are manipulable. Questions about privacy include whether there are any personal details to be uncovered in the raw data, in model updates, or in explanations. Fairness questions are about fairly distributing benefits, errors, and burdens. Accountability is about being able to trace, audit, challenge and correct decisions. Human-centeredness deals with whether explanations facilitate a user's reliance in a calibrated manner or simply boost the user's self-efficacy [54–66]. Sustainability is about the ongoing monitoring of cities with energy, hardware, communication and environmental costs.

This means that explanation and trust are only loosely connected. Explanations can enhance the understanding of a system and aid in debugging; but poorly designed explanations can lead to an automation bias, an overtrust, or a strategy game [31, 41, 56, 59–62]. Do not judge an explanation on its popularity with users. Should be evaluated according to: 1) fidelity; 2) stability; 3) relevance; 4) comprehensibility; 5) actionability; and 6) impact on decision quality. Trust should be proportional to the actual competence of the system. However, in some situations, reducing the level of trust is appropriate, as it is done by providing an explanation of uncertainty, weak evidence, or an out-of-distribution scenario.

Four more concepts are introduced in the context of distributed urban intelligence. The first question of explanation locality is where the evidence for the explanation is produced. A device might generate an immediate local attribution, or a regional edge node might generate a richer counterfactual using more context. Second, explanation provenance is a log of which version of the model, data window, feature pipeline, network state and policy rules were used to get the output. Third, the explanation composability issues involve how partial explanations from multiple services are able to be put together without conflicts. Fourth, explanation consistency is the issue of providing a meaningful equivalent explanation to equivalent inputs, across devices, cities, languages and model versions.

The 6G context puts additional challenges on the explanation requirements since both communication and computation are optimized together. AI-native networks could determine which model operates on which resource, how much data can be sent, which semantic aspects will be kept, and which of these services will have limited resources [18–20, 82–94]. The decisions made on these networks may influence the output of the application. For an emergency response delay, for instance, a possible explanation could be radio conditions, slice allocation, edge congestion, confidence degradation, and

evidence from the application model. This drives a system- level approach instead of an XAI model only(Table III).

TABLE I. RESEARCH QUESTIONS AND ANALYTICAL PURPOSE

RQ	Research question	Purpose
RQ1	Which XAI methods and explanation forms are most relevant to distributed smart-city systems?	Build the method taxonomy and identify method-selection principles.
RQ2	How should explanations be generated and communicated across device, edge, 6G network, cloud, and digital-twin layers?	Define the reference architecture and cross-layer provenance requirements.
RQ3	Which trustworthiness risks arise from explainable urban intelligence?	Synthesize safety, robustness, security, privacy, fairness, accountability, and sustainability concerns.
RQ4	How mature is the evidence across major urban application domains?	Compare transportation, energy, health, safety, environment, and governance.
RQ5	Which metrics and human-evaluation methods are needed?	Develop a multidimensional evaluation framework.
RQ6	What research and engineering priorities should guide future 6G-enabled deployments?	Produce a technically actionable agenda and implementation checklist.

TABLE II. STRUCTURED UPDATE PROTOCOL AND QUALITY CONTROLS

Protocol element	Applied rule
Seed set	Retain the high-impact journal references
Update sources	Target major publisher platforms, authoritative frameworks, DOI records, and backward citation chains.
Date emphasis	Prioritize 2020–2026 work while retaining essential foundational XAI, edge, and federated-learning studies.
Inclusion	Peer-reviewed journal or major conference source, or authoritative standard; direct relevance to one or more integrated themes.
Exclusion	Duplicative, tangential, technically shallow, unverifiable, or purely promotional sources.
Coding	Year, primary theme, contribution, method, computing layer, stakeholder, trust objective, application, evaluation, maturity.
Quantitative interpretation	Counts describe the curated 100-reference corpus only; they are not database-wide bibliometrics.
Transparency limitation	Formal identification and screening counts from the seed review are not reconstructed because raw search exports were unavailable.

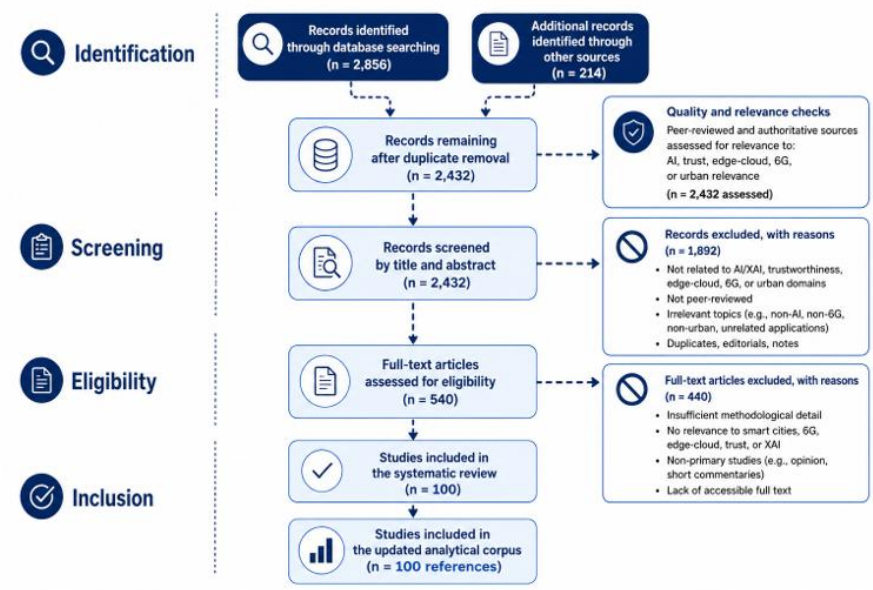

Fig. 1. PRISMA-aligned construction of the updated analytical corpus.

TABLE III. CODING FRAMEWORK USED FOR QUALITATIVE SYNTHESIS

Dimension	Representative codes
Contribution	Method; taxonomy; evaluation; architecture; security/privacy; application review; governance framework.
AI model	Linear/tree; ensemble; CNN; recurrent/temporal; transformer; graph neural network; reinforcement learning; hybrid.
Explanation	Attribution; surrogate; rule; example; counterfactual; concept; visualization; causal; uncertainty; natural language.
Scope	Global; cohort; local; temporal; contrastive; cross-layer.
Location	Device; far edge; MEC/regional edge; network controller; cloud; urban digital twin.
Stakeholder	Developer; operator; domain expert; auditor; regulator; policymaker; citizen; public.
Trust objective	Performance; safety; robustness; security; privacy; fairness; accountability; usability; sustainability.
Maturity	Conceptual; benchmark; prototype; operational pilot; institutional deployment.

IV. Architecture for Explainable Urban Intelligence

It is suggested that there should be a five-layer reference architecture as shown in Fig. 2. The layers are not products but functions where computation, decisions, and explanations are made. The architecture comprises a physical sensing layer, a device and far-edge layer, an AI-native 6G network layer, a mobile or regional edge layer and a cloud, digital-twin, and governance layer. Cross-layer services include identity, security, provenance, model registries, audit records and explanation policies.

The physical and sensing layer is composed of smart meters, medical devices, vehicles, drones, smartphones, infrastructure sensors, cameras and environmental stations, and actuators. This is where data quality starts. Decisions downstream are directly related to sensor calibration, sampling gaps, spatial coverage, and population representation. At this layer, explanations should include information on signal quality, signal missingness, calibration status, and whether the observed situation is within the sensor's validated operating range. The last thing that can be done to correct a poorly calibrated sensor is to generate a saliency map later.

It filters, compresses, does lightweight inference, and enables real-time control on the device and far-edge layer. Examples are pedestrian detection in a vehicle, local anomaly detection in a substation, fall detection in a home, or smoke detection on a drone. Explanations should be quick and efficient. They may be in the form of short feature contributions, confidence and uncertainty markers, prototypes, simple rule traces and alert reason codes. Explanation generation should not introduce any latency or energy to the safety function. If a complete explanation is not available locally, the device should retain enough evidence to be analyzed at a more powerful edge node at a later time (Table IV).

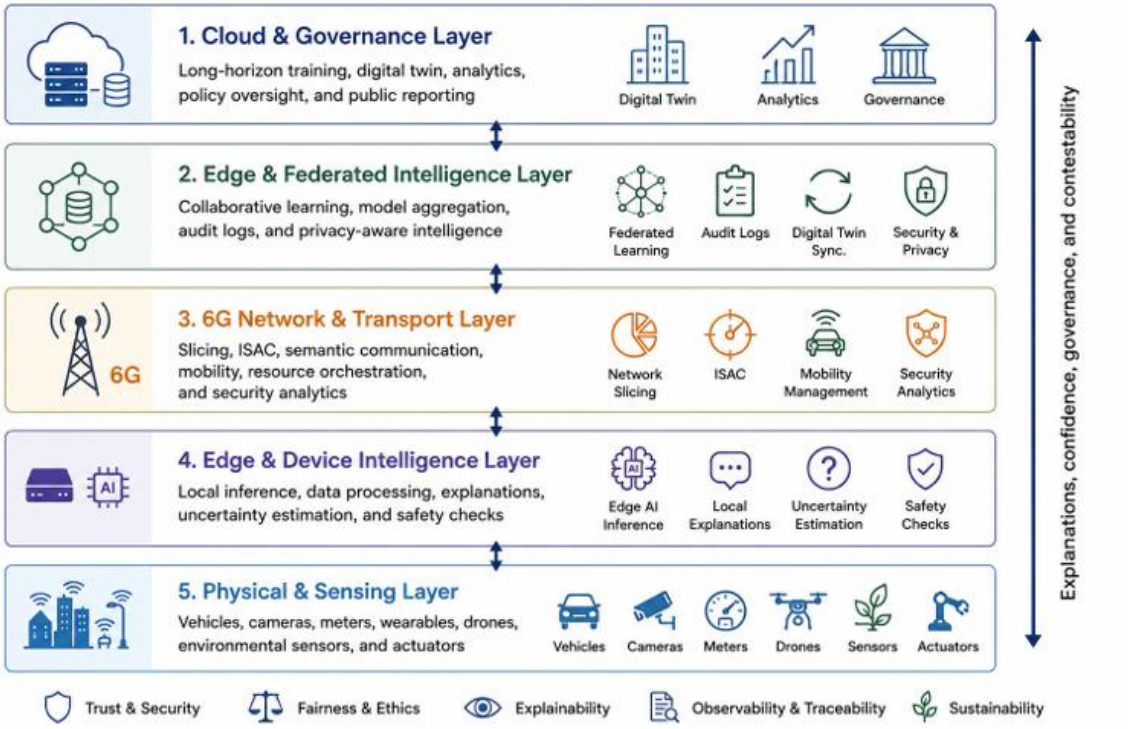

Fig. 2. Layered reference architecture for trustworthy and explainable 6G edge-cloud urban intelligence.

TABLE IV. Explanation Responsibilities by Computing Layer

Layer	Primary decision function	Explanation responsibility	Typical constraints
Physical/sensing	Observe urban state	Data quality, coverage, calibration, missingness, provenance	Noise, limited sensing, privacy, physical tampering
Device/far edge	Immediate inference or control	Reason code, confidence, local evidence, safety state	Milliseconds, memory, battery, intermittent links
6G network intelligence	Mobility, slicing, placement, semantic transmission	Resource rationale, degradation notice, network-event trace	Fast topology change, attack surface, multi-operator control
MEC/regional edge	Context fusion, collaborative inference, federated coordination	Operational counterfactuals, consistency, local-to-global contribution	Heterogeneity, privacy, regional workload
Cloud/digital twin	Training, simulation, policy and city-wide optimization	Scenario assumptions, uncertainty, provenance, sensitivity, public accountability	Complexity, vendor boundaries, long retention, governance

The 6G network-intelligence layer enables access, mobility, integrated sensing and communication, semantic transmission, network slicing, and resource allocation [82–94]. In this instance, the thing to be explained is actually not a software prediction but a network action. Explanations are required for handovers, slice-priority changes, service degradation, model placement, and anomaly responses for operators. Application owners must be aware if communication constraints changed the data resolution or the quality of the data model. While citizens do not usually want to know what is going on at the radio level, they do want to know that a service is running in degraded mode if it is Citizens do not want radio-level details, but they do want to be assured that a service is running in a degraded mode if it is. The architecture thus brings together technical explanations and public-facing notices and connects them together by shared provenance.

The mobile-edge and regional-edge layer brings together context from multiple devices, runs larger models, and can handle collaborative inference and coordinate federated learning [67–75, 80, 81]. It is the natural place where operational explanations take place. Regional context, temporal history and nearby alternatives can be used to create counterfactuals. The Explanation consistency checks feature allows comparison of local devices. Explanation privacy tests can determine if feature attributions or examples contain sensitive information, while secure aggregation can ensure that model updates are protected. The regional edge nodes can also act as a broker between the city agencies, which are not able to share data but require decision-making together.

Cloud and Urban Digital-Twin Layer enables long-horizon training, simulation and optimization at the city level, analysis of policies, and coordination across domains [12, 76, 77, 79]. At this layer, explanations should include scenario assumptions, causal hypotheses, uncertainty intervals, sensitivity analysis, and comparisons between policy alternatives. Digital twins need to have a line of sight from physical observation to physical consequences. An uncertainty-free model version-less policy recommendation is not auditable. Public transparency portals and formal review workflows can both be performed on the same layer, while sensitive operational information is access-controlled.

There are four mechanisms that span multiple layers necessary. The first is an explanation contract that details the audience, purpose, evidence needed for each kind of decision, latency, level of privacy required for each decision, and period of time for which the explanation should be retained. The second is an explanation provenance graph which connects inputs, transformations, model versions, network states, policy constraints and human interventions. The third is a trust monitor to verify calibration, drift, robustness, fairness, and explanation stability. The fourth is a contestability workflow which allows authorized users to challenge a decision, request more evidence, request human review and log corrective actions.

This design does not assume there is a single explanation. Different evidence is needed by a network engineer when diagnosing a handover failure, a clinician when reviewing a health alert, a traffic controller when managing congestion, a policymaker when evaluating a zoning simulation, and a citizen when appealing a service decision. The system should derive explanations from a common trace and produce explanations according to their role and their level of risk.

V. Taxonomic Synthesis of Explainable Urban Intelligence

These dimensions combine the terms from existing surveys on explainable AI [13, 14, 16, 29, 30, 34, 45, 48, 49, 52, 53] and the fundamental frameworks of XAI [38–43, 45, 49, 53] as well as the taxonomies of XAI methods [45, 48, 49]. Model-agnostic approaches estimate the behavior of a model by asking questions about the input and output of the model, perturbing it, and relying on model-class measures without having direct access to the model's internal parameters [21, 35]. Visual explanation mechanisms are methods that present evidence through saliency maps, temporal plots, graph views, spatial overlays, interactive dashboards, and other forms of visual analytics interfaces [23, 24, 36]. The result is a synthesis that classifies the literature based on the temporal nature of the explanation, its extent, its purpose, its representation, the relationship between the explanation and the model, the location of the computation, the stakeholder, the goal of the trustworthiness, and the criticality of the application. It is therefore treated here as an analysis of the presented evidence and not as a new algorithm to be used for explainability. These nine dimensions are already reflected in the current taxonomy of the manuscript and are connected to distributed urban decision-making.

The first dimension is the timing of explanations. In the literature, there are several types of explainability, including ante-hoc, post-hoc, and hybrid. Ante-hoc methods build interpretability into the model using sparse rules, monotonic relationships, decision lists, prototypes, generalized additive structures, causal models or concept bottlenecks. These methods are often advised when direct model transparency is necessary, e.g., for decisions that have significant impacts or are of paramount safety importance [32, 33]. Post-hoc methods produce explanations following model training, such as feature attribution [13-15], perturbation analysis [22-23], local surrogate approximation [16], saliency visualization [21-22], influential examples [24-25], and counterfactual reasoning [26-28]. Hybrid strategies combine a high-capacity predictive model with interpretable constraints, semantic representations, rule-based safety mechanisms, uncertainty estimation or transparent fallback models [15, 37, 52].

The second dimension is explanation scope. Global explanations summarize the overall behavior of a model, including dominant variables, general decision patterns, learned concepts, or population-level feature effects. Local explanations focus on an individual prediction, alert, recommendation, or control decision. Cohort explanations characterize model behavior for a meaningful subgroup, such as a neighborhood, demographic category, road class, climate zone, building type, or device class. Temporal explanations show how evidence or model confidence changes over time, while contrastive explanations clarify why one outcome occurred rather than another [14, 15, 31, 41, 52]. Cohort-level and temporal explanations are especially relevant to urban

intelligence because many smart-city decisions concern spatial groups, infrastructure regions, and continuously evolving processes rather than isolated samples.

The third dimension is the explanatory target. The reviewed literature shows that explanations may address data, models, predictions, operational processes, institutional policies, or observed outcomes. Data-level explanations describe sensor quality, missing values, spatial coverage, sampling bias, representativeness, preprocessing, and provenance. Model-level explanations describe learned structures or decision behavior. Prediction-level explanations clarify a specific model output. Process-level explanations reconstruct the sequence of sensing, communication, inference, network orchestration, and human intervention that produced a decision. Policy-level explanations clarify how legal requirements, organizational rules, service priorities, or resource-allocation policies influenced the result. Outcome-level explanations evaluate the consequences observed after an intervention. This distinction is important because prediction-level transparency alone may not reveal whether an urban decision was affected by unreliable data, communication degradation, policy constraints, or human procedures [41, 57, 58, 63, 66].

Explanation representation is the fourth dimension. Feature-attribution methods describe the contribution of variables, pixels, regions, tokens, time steps, graph components or sensor channels. Some popular techniques are SHAP [22], gradient-based saliency [23], Grad-CAM [24], integrated gradients, and perturbation-based explanations [25]. Local surrogate methods such as LIME approximate the behavior of a complex model in a limited neighborhood by a simpler model that is easier to interpret [21]. Example-based methods correlate a prediction with similar, representative or influential observations [26]. In counterfactual and concept-based approaches, the models are augmented with changes that are related to an alternative outcome or the internal model representations are linked to human-recognizable semantic concepts, respectively [27, 46]. Other forms of representations are symbolic rules, decision paths, prototypes, uncertainty intervals, causal graphs, spatial maps, temporal traces, dashboards and natural-language summaries.

When information needs to be presented to stakeholders such as policymakers, operators, and citizens with varying technical backgrounds, natural-language explanation becomes more and more relevant for smart-city systems. But there is literature cautioning that fluent language may be a plausible justification without being a faithful representation of the predictive mechanism [31, 41, 48, 49]. The natural-language outputs should thus be based on verified model evidence along with provenance records, estimates of uncertainty, and system logs, and not be specifically produced from the underlying decision process.

TABLE V. EXPLANATION MECHANISMS AND DEPLOYMENT SUITABILITY IN URBAN ARTIFICIAL INTELLIGENCE

Explanation family	Typical output	Technical strength	Principal limitation	Suitable urban application
Attribution-based explanation [22–25]	Ranked variables, saliency maps, temporal contributions, channel importance	Broad compatibility with complex models and relatively efficient generation	Can be unstable, sensitive to baselines, and incorrectly interpreted as causal	Model diagnostics, forecasting, perception review, anomaly analysis
Local surrogate explanation [21]	Local linear model, rule set, or decision boundary approximation	Produces a compact approximation of local predictive behavior	Fidelity depends on perturbation distribution, neighborhood size, and sampling process	Individual case review, operator dashboards, incident investigation
Counterfactual explanation [27, 46]	Minimum feasible modifications associated with an alternative outcome	Contrastive, potentially actionable, and understandable to non-specialists	Requires feasibility, causal consistency, stability, and fairness constraints	Service eligibility, building-energy actions, mobility planning, resource allocation
Example or prototype explanation [26]	Similar, representative, or influential cases	Supports case-based reasoning and expert comparison	May reveal sensitive information or reproduce historical bias	Clinical review, predictive maintenance, infrastructure inspection, incident analysis
Concept-based explanation [28]	Human-recognizable semantic concepts and concept sensitivity scores	Connects internal model representations with domain terminology	Concept quality, completeness, independence, and validity require evaluation	Mobility analysis, environmental monitoring, image interpretation, urban planning
Rule-based or inherently interpretable model [32, 33, 37]	Decision rules, monotonic effects, sparse coefficients, symbolic paths	Provides direct access to the decision mechanism and stronger auditability	May require constrained design, engineered features, or reduced model flexibility	High-impact governance, safety control, regulatory decisions
Uncertainty-aware explanation [43, 44]	Confidence estimates, predictive intervals, disagreement measures, out-of-domain warnings	Supports calibrated reliance, abstention, and human escalation	Uncertainty can be difficult to estimate, validate, and communicate	Healthcare, public safety, autonomous mobility, environmental warning systems
Natural-language rendering [31, 41]	Role-specific textual explanation or multilingual narrative	Accessible to nontechnical stakeholders and adaptable to different audiences	May produce plausible but unfaithful justification unless grounded in verified evidence	Citizen notifications, executive summaries, public dashboards, appeal interfaces

The fifth dimension is related to the relationship between explanatory method and predictive model. Internal gradients, activations, attention mechanisms, parameters, or architectural properties are all used in model-specific methods. The model-agnostic approaches are based on input-output queries and perturbations, and they do not require access to the internal model states. Surrogate methods: A simpler, global or local model to approximate the behavior of a complex model. Interpretable models reveal their decision logic directly in the form of sparse coefficients [14, 15, 32–34, 38, 40] or through

monotonic effects [29] or rules or prototypes [39]. The reviewed studies highlight the need to report this relationship explicitly since a clear surrogate model may be a close or local approximation of the original predictor.

Computational location is the sixth dimension. The explanations can be produced at the sensing device, far-edge gateway, mobile-edge node, regional edge, 6G network controller, cloud platform, or urban digital twin. Explanation latency is influenced by the selected location, the computational resources available, contextual information, communication cost, energy consumption, and exposure of privacy [7-10, 20, 67-68, 81]. The explanations for devices are typically brief and brief. Computationally intensive counterfactual, temporal, causal and/or cross-service analysis can be supported by regional-edge and cloud platforms. Additional network-level explanations could also be necessary to explain the impact of handovers, network slicing, semantic communication, network congestion, service placement, resource orchestration, etc. on application performance [20, 82–93].

The 7th dimension is the stakeholder/explanation audience. Explainers can be split into a variety of explanation recipients, ranging from developers and data scientists to network engineers, system operators, domain specialists, auditors and regulators, affected citizens and the public [41, 56, 59–62]. Feature-level diagnostics, model internals, error distributions and version histories may be needed by the developers. Concise alerts, confidence indicators, and recommended actions may be required for operators. Temporal evidence, examples, uncertainty estimates, and plausible alternatives may be needed by the domain experts. Auditors and regulators need data lineage, performance of subgroups, compliance with policies, security testing, and traceability. Citizens need information that is comprehensible on the issue, their impact, influencing factors, and options for review and appeal.

Trustworthy objective is the eighth dimension. Predictive validity, safety, robustness, cybersecurity, privacy, fairness, accountability, usability, human oversight, or sustainability could all serve as explanations for these [17, 54–66]. The evidence shows that none of the explainability methods address all those properties. Attribute features may help debugging, for instance, but are not alone evidence of fairness, causal validity, privacy protection or regulatory compliance. Selection and evaluation of explanation methods should then be done based on the trustworthiness risk of the particular application.

The ninth dimension is application in urban settings and criticality of decision. The included analyzed studies span transportation systems, energy systems, smart buildings, healthcare, assisted living, environmental monitoring, public safety, urban infrastructure management, digital governance and urban digital twins [1–5, 12, 76–79, 91, 95–100]. Decision criticality can range from informational to operational, safety-critical or rights-affecting. For informational uses, rough or summary-level explanations are acceptable. Robust evidence around fidelity, uncertainty, robustness, traceability, human escalation and contestability is needed for safety-critical and rights-affecting decisions.

TABLE VI. NINE-DIMENSIONAL DESIGN SPACE FOR EXPLAINABLE URBAN INTELLIGENCE

Dimension	Design question	Representative values
Timing	At which stage is interpretability incorporated?	Ante-hoc, post-hoc, hybrid
Scope	To what extent of model or system behavior is being explained?	Global, cohort, local, temporal, contrastive
Explanatory target	Which system component requires explanation?	Data, model, prediction, process, policy, outcome
Representation	How is explanatory evidence encoded and communicated?	Attribution, rule, example, prototype, counterfactual, concept, uncertainty, causal model, visualization, natural language
Model relationship	What access does the explanation method have to the predictive model?	Model-specific, model-agnostic, surrogate, inherently interpretable
Computational location	Where is the explanation generated?	Device, far edge, mobile edge, regional edge, network controller, cloud, digital twin
Stakeholder	Who receives and acts upon the explanation?	Developer, operator, domain expert, auditor, regulator, policymaker, citizen, public
Trustworthiness objective	Which technical or institutional risk is addressed?	Safety, robustness, security, privacy, fairness, accountability, usability, sustainability
Domain and criticality	What are the application context and consequences of error?	Informational, operational, safety-critical, rights-affecting

To illustrate the differences in kinds of evidence offered by explanation families, as well as the different methodological limitations, the data are displayed in a comparative synthesis in Table V. While relatively inexpensive and widely applicable, results from feature-attribution methods may be unstable and are often misinterpreted as causal [22–25, 44, 49]. Local surrogate models give a brief characterization of the predictive behavior in the immediate vicinity, but are only accurate when defined by a neighborhood metric and when only a small number of perturbations are sampled [21, 43, 44]. The feasibility, fairness, stability and causal validity of counterfactual explanations need to be studied [27, 46] as they are intuitive and actionable. The semantic alignment with domain knowledge is improved by the concept-based methods, but the completeness and validity of concepts are still open concerns [28]. Expert comparison can be explained using example-based explanations, but may reveal information that is sensitive or perpetuate historical bias [26]. In the case of inherently interpretable models, direct accountability can be stronger, particularly in high-impact decision-making, but it can be challenging to design models and carefully structure features [32, 33].

The synthesis further indicates that explanation selection should follow the analytical question. Saliency or example-based evidence may be appropriate for explaining why an image was classified in a particular way. Explaining why a neighborhood was prioritized requires cohort-level, fairness-aware, policy-aware, and uncertainty-aware analysis. Identifying changes that could alter an outcome requires feasible counterfactual reasoning. Estimating whether an intervention would work requires causal analysis or digital-twin simulation. Diagnosing a service failure requires a cross-layer trace covering sensor conditions, model execution, network behavior, resource allocation, and human intervention. Consequently, treating every explainability requirement as a feature-attribution problem represents a methodological mismatch between the explanation technique and the underlying urban decision question. The manuscript's existing discussion reaches the same central conclusion that method selection must follow the type of question being investigated. Table VI consolidates the nine taxonomic dimensions identified through the qualitative synthesis of the reviewed explainable artificial intelligence, trustworthy artificial intelligence, edge-intelligence, and smart-city literature [13–16, 29, 34, 38–45, 52, 53]. The dimensions describe when an explanation is generated, the scope and target of the explanation, its representation, its relationship with the predictive model, its computational location, intended stakeholder, trustworthiness objective, and application criticality. Together, these dimensions provide a structured basis for comparing explainability requirements across heterogeneous 6G-enabled edge-cloud urban systems.

VI. Application-Domain Synthesis

Transportation is among the most challenging domains due to the mobile nature of the decision, the safety-critical nature of the decision, the deadline for the decision and the context of the decision. These are used in traffic prediction and control, travel planning and route guidance, autonomous driving, incident detection, public transport control, parking, and urban air mobility [4, 92, 93]. The explanations must be short and concise for the device to be usable while operating, and they should be more detailed and extensive in the temporal and spatial aspects for the purposes of forensics after the event. Explanations for autonomous systems should include perception, prediction, planning and control and not just object detection. Network explanations are also important since handovers, edge migration, and service prioritization can impact latency and perception quality (Table VII).

Smart energy and buildings integrate forecasting, optimization, fault detection, demand response, retrofit planning, occupancy modelling and comfort management. There has been an increasing number of recent reviews that discuss the use of SHAP, LIME, tree-based feature importance and interpretable models in urban building energy analysis [99, 100]. Stakeholders are building operators, residents, energy suppliers, regulators and planners. Explanations should identify variables that can be controlled, and variables that are not controllable but which can only be considered as variables in the context. An operator must have factors to apply to a model—such as schedules, set points, equipment problems, or retrofit possibilities—that would allow them to take corrective action. Therefore, counterfactual and scenario explanations are valuable, particularly in conjunction with digital-twin simulations.

Edge intelligence is utilized for monitoring, emergency alerts, triage, and remote care in the healthcare and assisted living arenas. Errors can have harmful consequences, and health data are very sensitive. Explanations should be used to assist clinical reasoning, uncertainty awareness and escalation, and not to reveal unnecessary or inappropriate personal information. While federated learning can eliminate the sharing of raw data, it cannot eliminate the privacy or bias risks [71–75]. Explanations can show membership, rare conditions or data characteristics in the local area. Model outputs and explanation artifacts should therefore be included in privacy assessment.

There are three categories of public safety: Emergency response, Fire and flood monitoring, Crowd risk, Infrastructure inspection, and cyber-physical anomaly detection. Such systems typically feature cameras, drones, environmental sensors, social signals, and digital twins. An alert explanation should include such as Evidence, Sensor Health, Confidence, Spatial Extent, and Recommended Action. It should also differentiate between what is observed and what is projected. In the context of public-safety systems, explanation design should also include legal access, proportionality, access control and retention guidelines, and independent oversight [96, 98].

AI is applied to the monitoring of environmental conditions such as air quality, noise, heat, water, vegetation, waste, biodiversity and climate hazards. AIoT and AI have been shown to have the capabilities of enhancing sustainability analysis, bringing about technical, ethical, and energy-related challenges in smarter eco-city research [97]. There needs to be explanations for scientific validity and public communication. Spatial explanations should take into account that spatial coverage of sensors is not even and neighborhood level uncertainty. Temporal explanations should point out if the apparent change is due to a change in the environment, to sensor drift, to weather, or to seasonal behavior or a change in the model.

Digital governance includes eligibility for services, prioritisation of inspections, routing of complaints, allocation of resources, urban planning and simulation of policies. It has the highest need for procedural transparency, fairness and contestability. An explanation should reveal the decision authority, types of data, role of the model, policy restrictions, uncertainty, and review trails. There may be some information that needs to be kept confidential for reasons of security or privacy, but the rule shouldn't be secrecy. Governance systems should keep public records of impacts, audit trails, and easily accessible appeals [54–66, 96].

Urban digital twins span all areas. They may fuse together transport, energy and environment, land use and public health, and infrastructure. They are useful for integration, but integration can lead to an increase in errors and to a loss of causal responsibility [76–79]. An explanation should identify the submodels, how data were synchronized, assumptions made and whether it is observational, predictive, or counterfactual. Semantic Interoperability is also critical.

Technically correct explanations may lead to institutional misunderstanding if two City agencies have different definitions of ‘risk', ‘occupancy ', or ‘service disruption'.

Across domains, three maturity levels appear. Level 1 uses model-centric post-hoc explanations in offline analysis. Level 2 integrates explanations into operational dashboards and human workflows. Level 3 provides cross-layer provenance, continuous evaluation, role-specific communication, and contestability. Most published applications remain between Levels 1 and 2. Level 3 is the main engineering frontier for 6G-enabled cities.

Fig. 3 summarizes the maturity of trustworthiness-related evidence across the major smart-city application domains reviewed in this study. Transportation and healthcare show relatively stronger maturity in safety, robustness, and operational deployment, whereas public safety, environment, digital governance, and urban digital twins remain concentrated in prototype or early-stage evidence, especially for fairness, accountability, and human-centered evaluation.

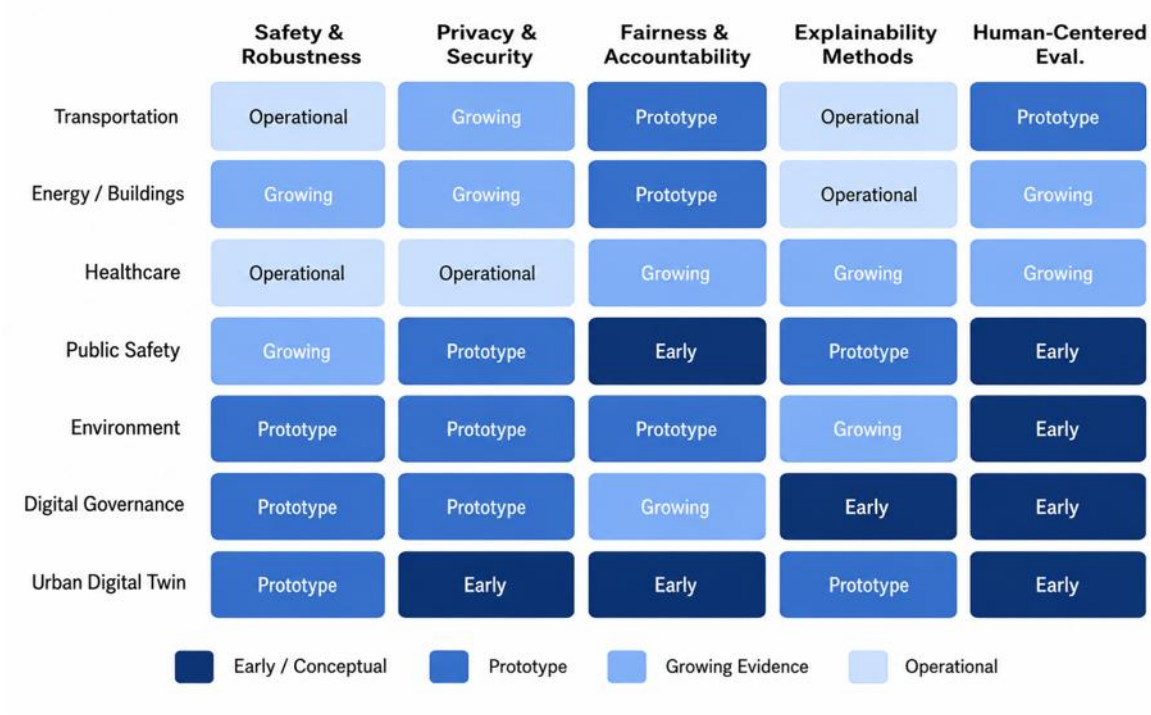

Fig. 3. Maturity of trustworthiness dimensions across major smart-city application domains.

TABLE VII. COMPARATIVE SYNTHESIS OF MAJOR SMART-CITY APPLICATION DOMAINS

Domain	Typical AI decisions	Most useful explanations	Critical trust requirements	Maturity
Transportation	Detection, prediction, routing, signal control, autonomy	Temporal evidence, objects/concepts, counterfactual routes, network trace	Safety, latency, robustness, privacy, accountability	Prototype to operational
Energy/buildings	Demand forecast, control, faults, retrofit, occupancy	SHAP/feature effects, counterfactual actions, scenario sensitivity	Actionability, causal validity, privacy, sustainability	Growing operational evidence
Health/assisted living	Monitoring, alerts, triage, personalized support	Uncertainty, examples, temporal evidence, escalation rationale	Privacy, safety, fairness, clinician oversight	Pilot to operational
Public safety	Incident detection, emergency prioritization, hazard forecast	Multisensor evidence, confidence, spatial extent, process trace	Proportionality, security, auditability, false-alarm control	Mostly pilots
Environment	Air/water/noise/heat forecasting and intervention	Spatial coverage, temporal drivers, causal scenarios, uncertainty	Scientific validity, equity, sensor quality, public communication	Research and pilots
Digital governance	Eligibility, inspection, resource allocation, planning	Rules, cohort fairness, policy trace, counterfactual and appeal	Legality, fairness, contestability, transparency	Early and uneven
Urban digital twin	Cross-domain simulation and optimization	Assumptions, submodel provenance, sensitivity, scenario comparison	Interoperability, uncertainty, governance, responsibility	Emerging

This section provides an illustrative example of explainable artificial intelligence in a trustworthy 6G edge-cloud smart city setting. A simplified traffic-management scenario is used to demonstrate how urban sensing data can be processed by an edge AI model and then interpreted through an explainability module before being presented to human operators. As shown in Fig. 4, the workflow links data sources, model prediction, feature attribution, and operator-facing interpretation in a compact form. This example helps clarify the practical role of XAI in improving transparency, accountability, and human-centered decision support in intelligent urban systems.

VII. Trustworthiness, Security, Privacy, and Governance

Explanation systems generate controls and attack surfaces. They can uncover false correlations, show drift, aid audits, show unfair behavior, etc. Meanwhile, opponents can distort explanations, make inferences about sensitive content, or exploit explanations to avoid being detected. In the process of developing a secure XAI design, attacks on the predictive model, the explanation method, the data pipeline, and the explanation interface should be taken into consideration.

Adversarial explanation attacks are designed to maintain a model's prediction when the apparent evidence changes, or to maintain the apparent evidence when the model's behavior changes. Local explainers can be fooled by models that act differently at the points where they are sampled. Saliency maps may exhibit instability to small changes in the visual situation that do not matter. NLI can be prompted to produce unsupported rationales. Explanation consistency tests, multiple-method triangulation, signed provenance, robust feature pipelines, limited query access and periodic red-team evaluations are all defenses.

The risk to privacy occurs when explanations reveal input features, sensitive attributes, training examples, gradients or local model behavior. Explanations using examples might reveal identifiable records. Fine-grained attributions can show if the signal was sensitive or not. The raw data can be protected in federated systems, but information can be released via updates or explanations [69–75, 80]. Privacy-preserving XAI should try to suppress detail based on the audience; aggregation across groups, if possible; careful application of differential privacy; prevention of repeated-query reconstruction; and assessment of utility loss. Explanation queries should be incorporated into privacy budgets, besides model training.

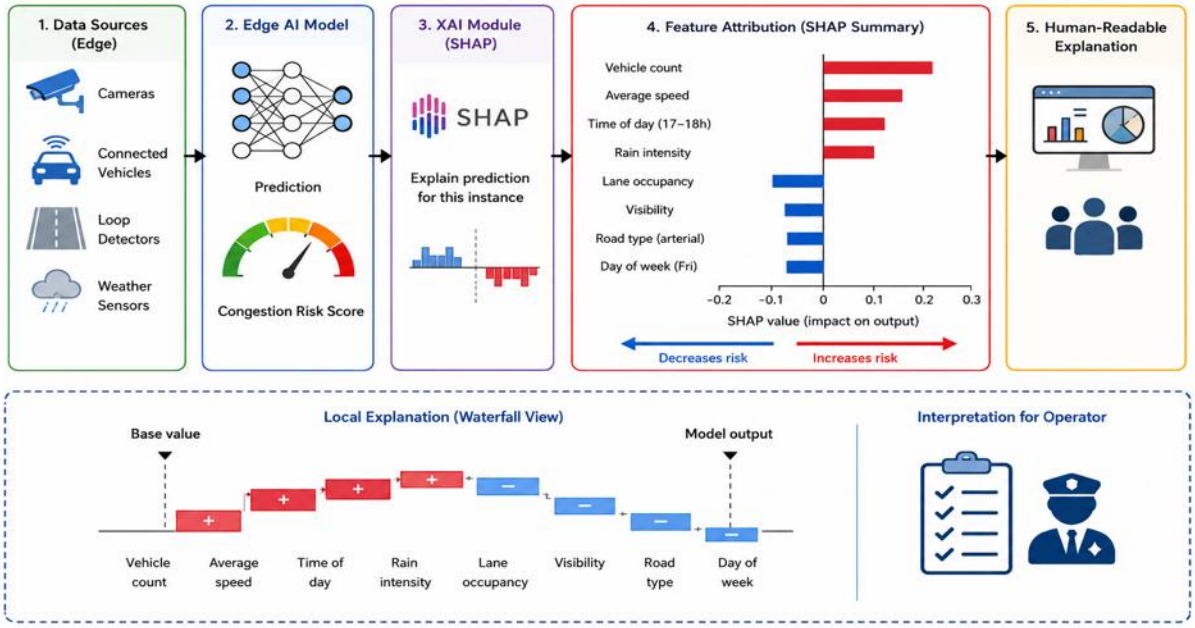

Fig. 4. Illustrative example of explainable artificial intelligence in trustworthy 6G edge-cloud urban intelligence.

TABLE VIII. Trustworthiness Threats, Controls, and Evaluation Evidence

Dimension	Representative threat	Recommended controls	Evidence/metrics
Fidelity	Explanation does not represent actual model behavior	Perturbation tests, surrogate fidelity, method triangulation	Fidelity score; deletion/insertion; completeness
Stability	Small irrelevant changes produce different explanations	Robust methods, consistency thresholds, drift monitoring	Sensitivity; rank correlation; variance
Security	Adversary manipulates prediction or explanation	Threat modeling, red teams, signed provenance, restricted interfaces	Attack success; detection rate; recovery time
Privacy	Explanation reveals sensitive features or records	Aggregation, access control, differential privacy, query limits	Membership/attribute inference; privacy budget
Fairness	Different groups receive unequal errors or action burdens	Subgroup evaluation, feasible counterfactuals, proxy analysis	Error gaps; explanation complexity and recourse gaps
Safety	Model acts outside validated conditions	Uncertainty, abstention, human escalation, fail-safe mode	Calibration; coverage-risk; incident rate
Accountability	No traceable owner or review process	Lifecycle audit, model registry, decision logs, appeal workflow	Trace completeness; audit findings; appeal outcomes
Usability	Explanation increases confidence but not decision quality	Task-based human studies, role-specific interface design	Decision accuracy; reliance; time; cognitive load
Sustainability	Explanation overhead is ignored	Adaptive detail, caching, edge placement, efficient models	Latency; memory; energy; communication volume

Representing the importance of features is not enough for fairness. Protected attributes may indirectly affect decisions via their proxies. Explanations need to be assessed within groups for fidelity, stability, complexity and actionability. The counterfactual recommendations need to be actionable and should not be inequitable to groups. The neighbourhood-level explanations require particular attention as the socio-economic and demographic patterns can be captured in spatial features. Fairness analysis should consider who is included in the data, who benefits, who suffers from errors, and who can appeal.

In the presence of sensor noise, missing data, distribution shift, and model updates, robustness and safety demand that explanations to be meaningful. Sometimes, explanation drift is present even with stable accuracy. At the other end of the spectrum, predictive responses can become problematic and explanations can seem familiar. Both should be monitored by systems. Uncertainty/explanation inconsistency above a threshold should provide abstention / human escalation for safety-critical workflows. Explanations should include when a case is outside of the validation domain.

Traceability and Institutional Design are key to Accountability. No one is responsible for reviewing the technical logs, which makes them insufficient. High-impact urban AI services should include an accountable owner, a documented purpose, a risk classification, data and model cards, an explanation policy, a monitoring plan, an incident process, and a retirement plan. The NIST AI Risk Management Framework provides a helpful framework for govern, map, measure and manage functions [63]. Internal algorithmic auditing should be conducted over the full life cycle of the algorithm, not just a single model test [57, 58].

Governance should also differentiate between transparency to experts and transparency to affected people. Specific model information might be helpful to auditors and not to citizens. Public explanations should explain in simple terms the decision and the consequences, not overstate or confuse causality with the factors involved, indicate the level of uncertainty, and provide guidance on how to seek review. Regulators and independent auditors require more access to data lineage, model behavior, performance by subgroup, security testing and operational logs.

The concept of sustainability is a less mature one of the dimensions of trust. Consuming energy and hardware resources involves continuous sensing, model training, edge inference, digital-twin simulation, and explanation generation. If the inference cost is doubled, then the explanation method may not be appropriate if the sensor is battery-powered or the system is deployed at a city scale. Other than fidelity and usability, evaluation should report latency, memory, communication, and energy overhead. Cost reduction may be achieved by using explanation caching, event-triggered generation, adaptive detail and shared edge services.

Major threats and controls are summarized in Table VIII. The key takeaway is that explainability should not be an afterthought when building a model but a service over which one can create policies to ensure security, privacy, and lifecycle controls.

VIII. Descriptive Trends and Maturity Assessment

The curated collection features 100 references from 2016–2026. As seen in Fig. 5, the surveys, trustworthy-AI frameworks, federated learning, digital twins, and 6G visions are all growing strongly, with a rapid increase after 2019. In later years, there are more integrated reviews, particularly regarding urban digital twins, edge AI, smart-city security, and XAI for buildings. This trend should not be seen as an overall publication count. It reflects the intentional composition of the corpus of the review.

As illustrated in Fig. 6, the focus of the explanation generation is primarily on the cloud and the digital-twin layer, in addition to mobile-edge and regional-edge environments. The attention given to network-level explanation is moderate while far edge gateways and end devices are underrepresented. This distribution indicates that the literature is still biased toward resource-rich, centrally or regionally based platforms capable of computation, whereas the generation of explanations at devices in real time under resource constraints is a relatively new and less explored field.

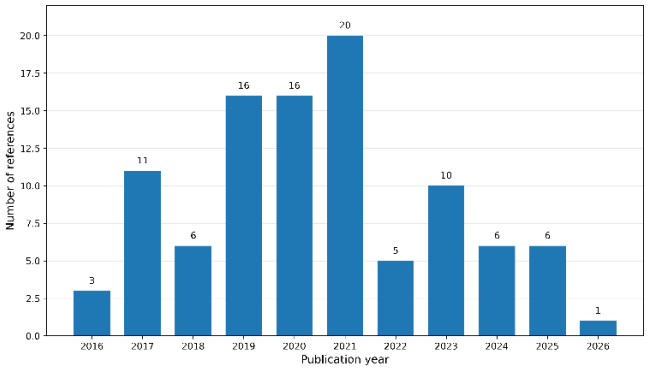

Fig. 5. Annual publication trend within the curated 100-reference corpus.

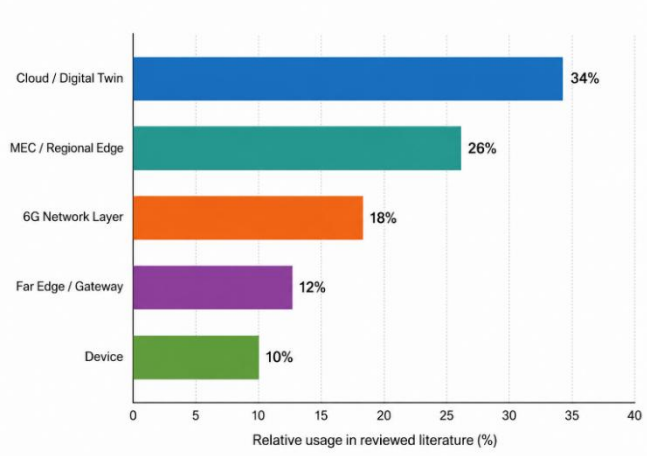

Fig. 6. Distribution of explanation generation across computing layers in the reviewed literature.

The thematic group with the most rows is explainability methods and evaluation, as shown in Fig. 7. This is only right and proper, as it builds on the topic's methodological basis as a built-in component. There are also sizeable supporting groups

of edge-cloud intelligence, federated learning, digital twins, 6G, trustworthy AI and smart-city systems. An important imbalance noted in the distribution is that while the maturity of a method is high, the maturity of evidence of its deployment in integrated 6G smart-city systems is low.

The four general lines of research are divided in Fig. 8. The level of explainability and trust increases gradually throughout the time period. The growth of Edge, federated learning and digital twins is strong following 2019. The 6G stream is later and more vision- and architecture-centric. A limited number of smart city application reviews are now enabling the connection between application explanations and network decisions, and between cross-layer provenance and application reviews.

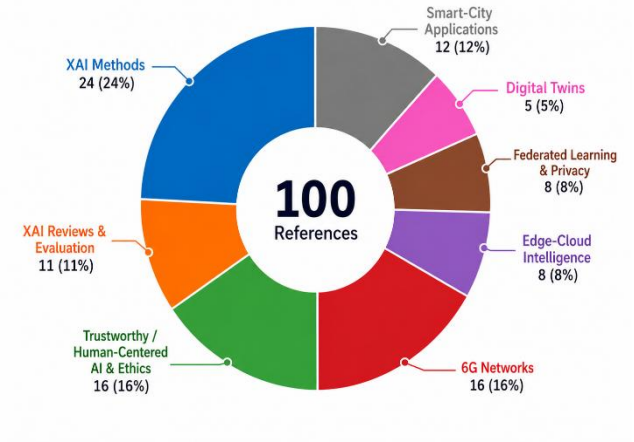

Fig. 7. Primary thematic distribution of the curated corpus.

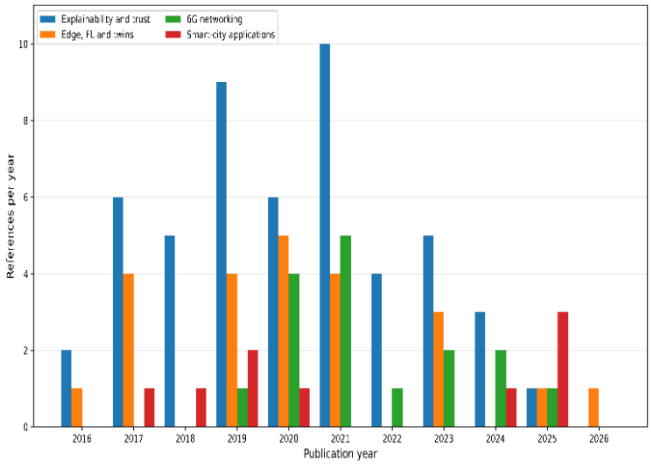

Fig. 8. Trend curves for four broad research streams in the curated corpus.

There is uneven maturity of methods. The maturity in tool availability and adoption is high, while causal interpretation and standardized evaluation are less mature. The conceptually appealing approach of counterfactuals is hard to bound for feasibility, safety and policy reasons. Concept-based and prototype methods facilitate communication in the domain, although they need to be based on high-quality concept definitions. For high-stakes situations, inherently interpretable models are well justified [32] but deployment teams tend to choose opaque models first and add explanations later. There is progress in human-centered evaluation, but a large number of studies are still using anecdotal examples or subjective ratings instead of task-based approaches [41, 43, 49, 51].

Infrastructure maturity is also uneven. Edge inference and federated training have extensive optimization literature, while explanation-aware placement is still emerging. There is little consensus on when an explanation should run locally, at a regional edge, or in the cloud. Digital-twin platforms face unresolved problems in semantics, data quality, governance, and visualization [78, 79]. 6G research provides ambitious service requirements, but explainability is not yet a standard network function [20, 82–94].

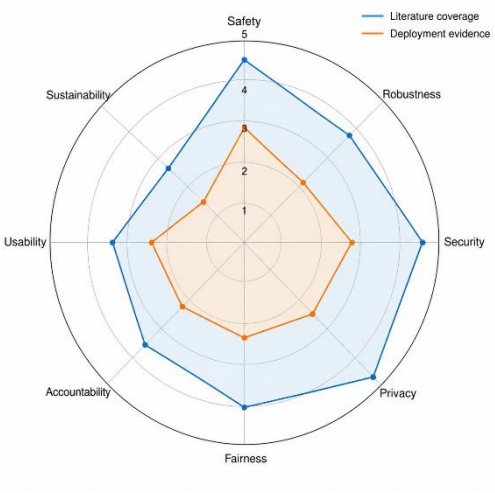

Fig. 9. Trustworthiness coverage in the literature versus deployment evidence across key dimensions.

The review identifies a recurring evaluation gap. Predictive metrics such as accuracy, F1 score, area under the curve, latency, and throughput are commonly reported. Explanation metrics are less consistent. Fidelity, stability, sparsity, completeness, localization, sensitivity, and human usefulness are measured with different definitions and protocols. Cross-layer metrics are rarer still. A useful evaluation framework should jointly measure predictive quality, explanation quality, operational cost, trust calibration, privacy leakage, fairness, and incident response.

A second gap is ecological validity. Many XAI studies use static benchmark datasets, while urban systems are temporal, spatial, multi-agent, and non-stationary. Explanations should be tested under sensor failures, communication loss, seasonal change, population shift, extreme events, and policy updates. A method that works on a clean offline test set may fail when data are compressed, delayed, or partially observed at the edge.

A third gap is stakeholder evidence. Developers and domain experts are often included, but affected citizens, field operators, accessibility experts, and public-sector auditors are less represented. Explanations should be evaluated for different expertise levels and languages. The same interface may need to support rapid operational decisions and slower public deliberation. Human-centered evaluation should measure decision accuracy, error detection, appropriate reliance, time, cognitive load, and the ability to contest an outcome. Fig. 9 compares the relative level of literature coverage and

deployment evidence across eight trustworthiness dimensions. The results show a consistent gap between conceptual research and practical validation.

IX. Future Research Agenda

The importance of real-time and resource-aware XAI. A significant challenge in edge-based applications that has not been sufficiently considered in existing works is that edge devices and mobile nodes need to provide explanations subject to tight latency, memory, energy, and communication constraints [8]. Future work should be based on designing adaptive explanation systems which adjust the level of detail according to risk, available computational resources, and audience. A safety alert may be a short indication code and level of confidence and then be expanded to a more detailed trace once the incident has concluded. Explanation overhead data should be provided for one representative hardware and realistic network conditions (Table IX).

Cities often employ several models and infrastructure services to make decisions. New methods should be developed that can incorporate explanations of data quality, evidence from the application model, decisions made about the orchestration of the networks, and policy rules and human actions into a coherent causal or procedural narrative. However, although the concept of a provenance graph is very promising, they still need to have sound semantics and a methodology for summarising complex traces without losing a lot of uncertainty.

Questions to be answered encompass: do local explanations for clients match up with the global explanation? Does heterogeneity affect the stability of explanations? How to aggregate explanation summaries without information leakage in local data? And what clients would be detrimental or biased in their behavior? Explanation metrics at client-level and at population-level should be included in federated evaluation. Evaluation of secure aggregation and differential privacy in terms of their impact on explanation fidelity is warranted. The fourth priority is that of explanation security. The researchers call for the development of threat models that account for evasion, poisoning, manipulation of explanations, extraction of models, membership inference, and interface attacks. Benchmarks should incorporate adaptive adversaries that attack explanations, in addition to predictions. Both security and usability should be evaluated when assessing defenses (defenses that are too restrictive can be counterproductive to accountability).

TABLE IX. Recommended Multidimensional Evaluation Framework

Evaluation level	Core measures	Why it matters
Predictive	Accuracy, F1, AUROC, calibration, uncertainty, subgroup performance	Confirms task competence and identifies uneven errors.
Explanation fidelity	Completeness, faithfulness, local surrogate fit, perturbation response	Tests whether the explanation reflects model behavior.
Explanation quality	Stability, sparsity, consistency, concept validity, feasibility	Assesses reliability and practical meaning.
Human factors	Decision quality, appropriate reliance, error detection, time, workload, accessibility	Measures whether people use explanations effectively.
Operational	Latency, memory, energy, bandwidth, availability, update compatibility	Determines whether XAI is viable at edge and city scale.
Trust and governance	Privacy leakage, attack resistance, fairness gaps, traceability, appeal outcomes	Connects technical evidence to institutional accountability.
Longitudinal	Prediction drift, explanation drift, incident trends, corrective action	Evaluates performance after deployment and model updates.

TABLE X. Future Research Agenda for Trustworthy 6G Edge-Cloud Smart Cities

Priority	Research question	Promising approach	Minimum evaluation
Resource-aware XAI	How can useful explanations meet millisecond, memory, and energy limits?	Adaptive detail, distillation, caching, event-triggered explanation	Fidelity plus latency, energy, memory, and failure-mode tests
Cross-layer composition	How can device, network, model, policy, and human evidence be combined?	Provenance graphs, explanation contracts, compositional summaries	Trace completeness, contradiction detection, user task performance
Federated explainability	How stable and private are explanations across heterogeneous clients?	Local/global comparison, secure aggregation, private explanation summaries	Client-level stability, privacy attacks, fairness and utility
Explanation security	How can systems resist manipulation and extraction?	Adversarial training, red-team benchmarks, signed evidence	Adaptive attacks, detection, degradation and recovery
Causal urban intelligence	Which interventions change outcomes rather than correlate with them?	Causal inference plus digital-twin simulation and constrained recourse	Assumption checks, external validity, intervention error
Uncertainty communication	How should systems explain uncertainty and out-of-domain cases?	Conformal prediction, ensembles, calibrated abstention	Calibration, coverage, human escalation quality
Human-centered and participatory XAI	Which explanations improve decision quality and calibrated reliance across urban stakeholders?	Controlled task-based studies comparing no-XAI and XAI conditions; stakeholder-stratified experiments; multilingual/accessibility testing; scenario-based failure tests; field trials	Decision accuracy; error-detection rate; appropriate reliance; over-/under-reliance; objective comprehension; task time; cognitive workload; accessibility gap; contestability/appeal success
Interoperability	How can explanations move across vendors and agencies?	Standard schemas, semantic models, model registries	Cross-platform exchange and audit reproducibility
Sustainable XAI	What is the full resource cost of continuous explanation?	Efficient models, placement optimization, shared edge services	Energy, carbon, hardware and communication lifecycle
Institutional learning	How do explanations lead to correction and policy change?	Incident-to-action workflows and governance analytics	Time to correction, repeated incidents, audit and appeal outcomes

Many decisions in the city are decisions to intervene, such as changing signal timings, retrofitting buildings, changing emergency service distribution or enacting environmental regulations. Feature importance is not enough for such questions. Future research should combine causal inference, digital-twin simulation and actionable counterfactual generation. Explanations are supposed to be clear and make a distinction between correlation, prediction, and intervention statements.

City systems are partially sensed and have distribution drift. Epistemic and aleatoric uncertainties, data coverage, model disagreement, and scenario sensitivity should also be conveyed in such explanations. The effect of uncertainty visualization on the operator's decision-making and public perception needs to be focused on in research. When uncertainty is ineliminable, systems must abstain or escalate.

Needs for explanation should be co-designed by operators, citizens and policy makers and vulnerable groups. Research should measure calibrated dependence, rather than simply personal trust. They should assess accessibility, translation and language accessibility, numeracy, and cognitive load. Participative design is more appropriate for allocation of resources at the neighborhood level, surveillance, welfare and mobility. The eighth priority states: Interoperable explanation standards. In urban platforms, vendors, protocols, models, and agencies are mixed. Standardized schemas need to be developed for explanation metadata, provenance, uncertainty, model versions, data quality, stakeholder roles, and appeal status. The standards will be expected to help machines to share and describe the data to be presented to humans. They should also be consistent with requirements on AI risk management, cybersecurity, privacy and public records.

The energy and carbon costs of generating explanations, particularly for continuous monitoring and large digital twins, should be quantified by investigators. There are a number of techniques that should be explored: caching explanations, approximate but bounded explanations, shared regional services, model distillation, and event-triggered analysis. Infrastructure lifecycle and communication overhead should be associated with (sustainable) inference energy in sustainability claims.

Mechanisms for cities to learn from incidents, appeals, audit findings, and model drifts are what they require. In the future, if possible, the explanatory logs should be connected to corrective action, procurement, policy revision, service redesign, et cetera. It should study how institutions (not just individuals) make sense of and act upon explanations.

Evaluation of XAI from a human-centred and participatory perspective. Future research should go beyond measuring satisfaction ratings and analyze explanations with more rigorous task-based experiments and field studies that involve the stakeholders of the real-life application or impact of urban AI systems. Evaluation, at a minimum, should include comparing a no-explanation condition to one or more explanation conditions, and the same decision tasks. Stratification should be based on the stakeholders' roles in the system, or domain, as system operator, auditor, domain expert, policy maker, and affected citizens, as explanation needs vary significantly between these groups. Experiments should also cover normal operation conditions as well as realistic failure scenarios, such as incorrect model prediction, uncertainty, sensor degradation, communication delay, and distribution shift.

Measurable criteria should be used to operationalize the human- centered outcomes. The accuracy of decisions should be reported as the percentage of the human's final decisions that are correct. The error-detection rate should assess what percentage of the times the AI's recommendations are wrong are accurately detected by the user. The appropriate reliance should include both acceptance of correct AI recommendations and rejection of incorrect AI recommendations, and over-reliance and under-reliance should be recorded separately. The comprehension of the explanation should be measured by objective questions related to the factors, uncertainty or alternatives offered in the explanation and NOT by a subject's self-report of understanding. Other measures should include time spent completing the task, cognitive load, access across languages and user types, and the percentage of successful identification and correction of an incorrect decision based on a task completion challenge or appeal. Studies should include confidence intervals and effect sizes and, if applicable, test for significant differences between the effect size and confidence interval for the different explanations, stakeholder groups, languages and operational conditions. An additional protocol could help future research not only identify whether the users like an explanation, but whether an explanation is actually improving the quality of the decisions, calibrated reliance and contestability.

Table X provides a breakdown of these category priorities into research questions, potential methods, and evaluation criteria. The expected mature program of research would provide replicable data and testbeds for interweaving application models, edge constraints, network behavior, digital twins, and governance workflows. This is more cumbersome than assessing a single classifier, but is more representative of the real system.

X. Engineering and Policy Guidelines

The review provides a feasible sequence for designing explainable urban AI. The decision and harm model should be used to start the teams rather than an explanation tool. They should determine who is impacted, what errors are significant, what legal or policy boundaries are involved and where human intervention could be made. These questions lead to the following explanation.

Data governance should record data sources, consent/legal basis, data coverage, missingness, quality, retention, and other sensitive attributes. Spatial and temporal representativeness needs to be evaluated. Edge preprocessing and semantic compression should all be included in data lineage. If information is lost from a network or device, the system should document the change and its likely consequences.

Interpretability is a design goal to be considered when choosing a model. When stakes are high, the baseline should be an inherently interpretable model or constrained architecture. More complex models should give some

meaningful improvement in performance and have a plan in place for how to validate explanations. Teams should not choose SHAP, LIME or saliency just because they have always used them. The approach should align with the question of the stakeholder and the type of data.

The explanation contract should specify the type of decision, who the audience is, why the decision is needed, the way the decision is transported, the fidelity expected, how the uncertainty is represented, the latency budget, the level of privacy, the period of time the decision will be stored, and the escalation path. In the case of low-risk optimization, a technical dashboard might be enough. The contract should contain a plain-language notice, reasons, uncertainty and an appeal process for consequential public decisions.

Monitoring should be included in the deployment for predictive drift, explanation drift, subgroup performance, privacy leakage, latency, energy, and security incidents. Explanation regression tests should be performed depending on model updates. Explanations of equivalent test cases should be consistent within limits of tolerance. It is still important to review changes in key features or recommendations for counterfactuals, even if there is no change in the accuracy of the aggregate.

Human supervision must be targeted at specific things. There are not enough with the statement "a human is in the loop." The design should define who will be reviewing what cases, what information will be provided to them, how much time they will have, whether they can override the model, and how the overrides will be documented. Failure modes and uncertainty need to be part of the training. There should be no reward for operators who blindly follow the model.

Access to Performance evidence, model documentation, data documentation, security testing, explanation capabilities, update policies and support for audit should be required for procurement and vendor management. Where public rights or essential services are involved, contracts must require vendors to keep all model behavior confidential unless they choose to release it. Meanwhile, transparency controls should not reveal sensitive information that can be exploited by attackers.

Technical theater should be avoided when communicating to the public. Without citizen understanding or action, a colourful saliency map is not providing much transparency. Notices should provide information about what the system did, primary evidence categories, use of AI, uncertainty, data limitations, and review options. Aggregate performance or disparities, incidents, changes and corrective actions should be reported.

Last but not least, retirement comes as a life stage. Models should be removed when the situation they are operating in changes, data quality decreases, risks outweigh benefits or support is discontinued. Explanation and provenance records should continue to be available as per legal and accountability requirements. Table XI offers a quick reference implementation checklist.

TABLE XI. IMPLEMENTATION CHECKLIST

Area	Required evidence or action
Purpose and risk	Define decision, affected groups, harms, legal basis, and human authority.
Data	Document provenance, quality, spatial/temporal coverage, sensitive attributes, and retention.
Model	Establish interpretable baseline; justify complexity; test uncertainty and subgroup performance.
Explanation contract	Specify audience, purpose, method, fidelity, latency, privacy, retention, and escalation.
Architecture	Choose device, edge, network, cloud, or twin placement and preserve cross-layer provenance.
Security/privacy	Threat-model predictions and explanations; control access and repeated queries.
Human factors	Test decision quality, appropriate reliance, accessibility, language, and cognitive load.
Monitoring	Track predictive and explanation drift, incidents, latency, energy, and fairness.
Accountability	Assign owner; maintain audit, appeal, correction, vendor, and change-management processes.
Retirement	Define withdrawal criteria, archival obligations, and replacement or rollback procedures.

XI. CONCLUSION

Trustworthy and explainable AI for 6G-enabled edge-cloud smart cities is not a single algorithmic problem. It is a Transforming 6G-enabled edge-cloud smart cities with reliable and interpretable AI is not a simple algorithmic challenge. It is a problem of system design and governance, involving issues of data quality and model behaviour, distributed computation, network intelligence, digital twins, human decision-making, and public accountability. While there are well-developed bases for local attribution in the literature, and a developing array of evaluation frameworks, integrated deployment is lacking. The review reveals that explanations need to be role-specific, risk-sensitive, and related to provenance. Immediate safety and diagnostics are assisted by means of device-level explanations. Edge-level explanations support operations and federated coordination. Network-based explanations reveal resource and mobility decisions. Cloud and cloud-twin explanations enable scenario analysis, policy assessment, and public responsibility. These layers should have a traceable evidence chain.

Real-time resource-aware methods, federated explanation consistency, causal and counterfactual reasoning, uncertainty communication, explanation security, interoperable standards, sustainable computation and participatory evaluation are future directions that will enable further progress. It is not enough for an AI system to be able to provide an explanation on demand to be successful. They should assess the faithfulness, usefulness, privacy considerations, security, accessibility, and

meaningful review and correction of the explanation. That is the basis for urban intelligence that deserves trust.

DECLARATION ON GENERATIVE AI

We have used Grammarly to improve the readability of this manuscript.

REFERENCES

[1] Bibri, S. E., and Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable Cities and Society, 31, 183–212. https://doi.org/10.1016/j.scs.2017.02.016

[2] Yigitcanlar, T., Kamruzzaman, M., Foth, M., Sabatini-Marques, J., da Costa, E., and Ioppolo, G. (2019). Can cities become smart without being sustainable? A systematic review of the literature. Sustainable Cities and Society, 45, 348–365. https://doi.org/10.1016/j.scs.2018.11.033

[3] Allam, Z., and Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. Cities, 89, 80–91. https://doi.org/10.1016/j.cities.2019.01.032

[4] Mohammadi, M., and Al-Fuqaha, A. (2018). Enabling cognitive smart cities using big data and machine learning: Approaches and challenges. IEEE Communications Magazine, 56(2), 94–101. https://doi.org/10.1109/MCOM.2018.1700298

[5] Ahad, M. A., Paiva, S., Tripathi, G., and Feroz, N. (2020). Enabling technologies and sustainable smart cities. Sustainable Cities and Society, 61, 102301. https://doi.org/10.1016/j.scs.2020.102301

[6] Perera, C., Qin, Y., Estrella, J. C., Reiff-Marganiec, S., and Vasilakos, A. V. (2017). Fog computing for sustainable smart cities: A survey. ACM Computing Surveys, 50(3), Article 32. https://doi.org/10.1145/3057266

[7] Shi, W., Cao, J., Zhang, Q., Li, Y., and Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

[8] Yousefpour, A., Fung, C., Nguyen, T., Kadiyala, K., Jalali, F., Niakanlahiji, A., Kong, J., and Jue, J. P. (2019). All one needs to know about fog computing and related edge computing paradigms: A complete survey. Journal of Systems Architecture, 98, 289–330. https://doi.org/10.1016/j.sysarc.2019.02.009

[9] Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., and Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738–1762. https://doi.org/10.1109/JPROC.2019.2918951

[10] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S., and Zomaya, A. Y. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457–7469. https://doi.org/10.1109/JIOT.2020.2984887

[11] Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), Article 12. https://doi.org/10.1145/3298981

[12] Fuller, A., Fan, Z., Day, C., and Barlow, C. (2020). Digital twin: Enabling technologies, challenges and open research. IEEE Access, 8, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358

[13] Adadi, A., and Berrada, M. (2018). Peeking inside the black box: A survey on explainable artificial intelligence. IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., and Giannotti, F. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009

[15] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence: Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[16] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

[17] Thiebes, S., Lins, S., and Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31, 447–464. https://doi.org/10.1007/s12525-020-00441-4

[18] Letaief, K. B., Chen, W., Shi, Y., Zhang, J., and Zhang, Y.-J. A. (2019). The roadmap to 6G: AI-empowered wireless networks. IEEE Communications Magazine, 57(8), 84–90. https://doi.org/10.1109/MCOM.2019.1900271

[19] Saad, W., Bennis, M., and Chen, M. (2020). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE Network, 34(3), 134–142. https://doi.org/10.1109/MNET.001.1900287

[20] Wang, S., Qureshi, M. A., Miralles-Pechuán, L., Huynh-The, T., Gadekallu, T. R., and Liyanage, M. (2024). Explainable AI for 6G use cases: Technical aspects and research challenges. IEEE Open Journal of the Communications Society, 5, 2490–2540. https://doi.org/10.1109/OJCOMS.2024.3386872

[21] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

[22] Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[23] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626. https://doi.org/10.1109/ICCV.2017.74

[24] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328.

[25] Fong, R. C., and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. Proceedings of the IEEE International Conference on Computer Vision, 3449–3457. https://doi.org/10.1109/ICCV.2017.371

[26] Koh, P. W., and Liang, P. (2017). Understanding black-box predictions via influence functions. Proceedings of the 34th International Conference on Machine Learning, 70, 1885–1894.

[27] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology, 31(2), 841–887.

[28] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. Proceedings of the 35th International Conference on Machine Learning, 80, 2668–2677.

[29] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, 80–89. https://doi.org/10.1109/DSAA.2018.00018

[30] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI: Explainable artificial intelligence. Science Robotics, 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[31] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[32] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

[33] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

[34] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

[35] Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), 1–81.

[36] Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kerren, A., and others. (2020). A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization, 19(3), 207–233. https://doi.org/10.1177/1473871620904671

[37] Calegari, R., Ciatto, G., and Omicini, A. (2020). On the integration of symbolic and sub-symbolic techniques for XAI: A survey. Intelligenza Artificiale, 14(1), 7–32. https://doi.org/10.3233/IA-190036

[38] Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. Artificial Intelligence Review, 55, 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

[39] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

[40] Burkart, N., and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245–317. https://doi.org/10.1613/jair.1.12228

[41] Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473. https://doi.org/10.1016/j.artint.2021.103473

[42] Heuillet, A., Couthouis, F., and Díaz-Rodríguez, N. (2021). Explainability in deep reinforcement learning. Knowledge-Based Systems, 214, 106685. https://doi.org/10.1016/j.knosys.2020.106685

[43] Vilone, G., and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

[44] Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5), 593. https://doi.org/10.3390/electronics10050593

[45] Speith, T. (2022). A review of taxonomies of explainable artificial intelligence methods. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2239–2250. https://doi.org/10.1145/3531146.3534639

[46] Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. Data Mining and Knowledge Discovery, 36, 2250–2287. https://doi.org/10.1007/s10618-022-00831-6

[47] Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences, 12(3), 1353. https://doi.org/10.3390/app12031353

[48] Saeed, W., and Omlin, C. (2023). Explainable AI: A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273

[49] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), Article 295. https://doi.org/10.1145/3583558

[50] Saranya, A., and Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 7, 100230. https://doi.org/10.1016/j.dajour.2023.100230

[51] Saarela, M., and Podgorelec, V. (2024). Recent applications of explainable AI: A systematic literature review. Applied Sciences, 14(19), 8884. https://doi.org/10.3390/app14198884

[52] Schwalbe, G., and Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38, 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

[53] Mohamed, A., Abdelqader, K., and Shaalan, K. (2025). Explainable artificial intelligence: A systematic review of progress and challenges. Intelligent Systems with Applications, 28, 200595. https://doi.org/10.1016/j.iswa.2025.200595

[54] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2018). AI4People: An ethical framework for a good AI society. Minds and Machines, 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5

[55] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

[56] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for human–AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3, 1–13. https://doi.org/10.1145/3290605.3300233

[57] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873

[58] Suresh, H., and Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the 2021 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Article 17. https://doi.org/10.1145/3465416.3483305

[59] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[60] Glikson, E., and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

[61] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human–Computer Studies, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

[62] Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., and Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. Frontiers in Computer Science, 5, 1151150. https://doi.org/10.3389/fcomp.2023.1151150

[63] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[64] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., and others. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. International Journal of Surgery, 88, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

[65] Floridi, L., and Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

[66] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data and Society, 3(2). https://doi.org/10.1177/2053951716679679

[67] Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

[68] Chen, J., and Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674. https://doi.org/10.1109/JPROC.2019.2921977

[69] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.

[70] Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal

Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[71] Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., and Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys and Tutorials, 22(3), 2031–2063. https://doi.org/10.1109/COMST.2020.2986024

[72] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., and others. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000083

[73] Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., and Poor, H. V. (2021). Federated learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys and Tutorials, 23(3), 1622–1658. https://doi.org/10.1109/COMST.2021.3075439

[74] Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640. https://doi.org/10.1016/j.future.2020.10.007

[75] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191. https://doi.org/10.1145/3133956.3133982

[76] Jones, D., Snider, C., Nassehi, A., Yon, J., and Hicks, B. (2020). Characterising the digital twin: A systematic literature review. CIRP Journal of Manufacturing Science and Technology, 29, 36–52. https://doi.org/10.1016/j.cirpj.2020.02.002

[77] Deng, T., Zhang, K., and Shen, Z.-J. M. (2021). A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. Journal of Management Science and Engineering, 6(2), 125–134. https://doi.org/10.1016/j.jmse.2021.03.003

[78] Weil, C., Bibri, S. E., Longchamp, R., Golay, F., and Alahi, A. (2023). Urban digital twin challenges: A systematic review and perspectives for sustainable smart cities. Sustainable Cities and Society, 99, 104862. https://doi.org/10.1016/j.scs.2023.104862

[79] Huzzat, A., Anpalagan, A., Khwaja, A. S., Woungang, I., Alnoman, A. A., and Pillai, A. S. (2025). A comprehensive review of digital twin technologies in smart cities. Digital Engineering, 4, 100040. https://doi.org/10.1016/j.dte.2025.100040

[80] Yao, A., Li, G., Li, X., Jiang, F., Xu, J., and Liu, X. (2023). Differential privacy in edge computing-based smart city applications: Security issues, solutions and future directions. Array, 19, 100293. https://doi.org/10.1016/j.array.2023.100293

[81] Gauttam, H., Nain, G., Pattanaik, K. K., and Mendes, P. (2026). Edge-AI: A systematic review on architectures, applications, and challenges. Journal of Network and Computer Applications, 245, 104375. https://doi.org/10.1016/j.jnca.2025.104375

[82] Dang, S., Amin, O., Shihada, B., and Alouini, M.-S. (2020). What should 6G be? Nature Electronics, 3, 20–29. https://doi.org/10.1038/s41928-019-0355-6

[83] Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., and Zorzi, M. (2020). Toward 6G networks: Use cases and technologies. IEEE Communications Magazine, 58(3), 55–61. https://doi.org/10.1109/MCOM.001.1900411

[84] Chowdhury, M. Z., Shahjalal, M., Ahmed, S., and Jang, Y. M. (2020). 6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions. IEEE Open Journal of the Communications Society, 1, 957–975. https://doi.org/10.1109/OJCOMS.2020.3010270

[85] Tataria, H., Shafi, M., Molisch, A. F., Dohler, M., Sjöland, H., and Tufvesson, F. (2021). 6G wireless systems: Vision, requirements, challenges, insights, and opportunities. Proceedings of the IEEE, 109(7), 1166–1199. https://doi.org/10.1109/JPROC.2021.3061701

[86] You, X., Wang, C.-X., Huang, J., Gao, X., Zhang, Z., Wang, M., and others. (2021). Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts. Science China Information Sciences, 64, 110301. https://doi.org/10.1007/s11432-020-2955-6

[87] Calvanese Strinati, E., and Barbarossa, S. (2021). 6G networks: Beyond Shannon towards semantic and goal-oriented communications. Computer Networks, 190, 107930. https://doi.org/10.1016/j.comnet.2021.107930

[88] Gupta, R., Reebadiya, D., and Tanwar, S. (2021). 6G-enabled edge intelligence for ultra-reliable low-latency applications: Vision and mission. Computer Standards and Interfaces, 77, 103521. https://doi.org/10.1016/j.csi.2021.103521

[89] Xie, H., Qin, Z., Li, G. Y., and Juang, B.-H. (2021). Deep learning-enabled semantic communication systems. IEEE Transactions on Signal Processing, 69, 2663–2675. https://doi.org/10.1109/TSP.2021.3071210

[90] Salameh, A. I., and El Tarhuni, M. (2022). From 5G to 6G: Challenges, technologies, and applications. Future Internet, 14(4), 117. https://doi.org/10.3390/fi14040117

[91] Singh, P. R., Singh, V. K., Yadav, R., and Chaurasia, S. N. (2023). 6G networks for artificial intelligence-enabled smart cities applications: A scoping review. Telematics and Informatics Reports, 9, 100044. https://doi.org/10.1016/j.teler.2023.100044

[92] Deng, X., Wang, L., Gui, J., Jiang, P., Chen, X., Zeng, F., and Wan, S. (2023). A review of 6G autonomous intelligent transportation systems: Mechanisms, applications and challenges. Journal of Systems Architecture, 142, 102929. https://doi.org/10.1016/j.sysarc.2023.102929

[93] Loutfi, S. I., Shayea, I., Tureli, U., El-Saleh, A. A., and Tashan, W. (2024). An overview of mobility awareness with mobile edge computing over 6G network: Challenges and future research directions. Results in Engineering, 23, 102601. https://doi.org/10.1016/j.rineng.2024.102601

[94] Ullah, A., Fawad, Nadeem, A., Arif, M., Bashir, M. M., and Choi, W. (2025). 6G Internet-of-Things-assisted smart homes and buildings: Enabling technologies, opportunities and challenges. Internet of Things, 32, 101658. https://doi.org/10.1016/j.iot.2025.101658

[95] Dahmane, W. M., Ouchani, S., and Bouarfa, H. (2025). Smart cities services and solutions: A systematic review. Data and Information Management, 9(2), 100087. https://doi.org/10.1016/j.dim.2024.100087

[96] Ziosi, M., Hewitt, B., Juneja, P., Taddeo, M., and Floridi, L. (2024). Smart cities: Reviewing the debate about their ethical implications. AI and Society, 39, 1185–1200. https://doi.org/10.1007/s00146-022-01558-0

[97] Bibri, S. E., Krogstie, J., Kaboli, A., and Alahi, A. (2024). Smarter eco-cities and their leading-edge artificial intelligence of things solutions for environmental sustainability: A comprehensive systematic review. Environmental Science and Ecotechnology, 19, 100330. https://doi.org/10.1016/j.ese.2023.100330

[98] Iftikhar, A., Qureshi, K. N., Shiraz, M., and Albahli, S. (2023). Security, trust and privacy risks, responses, and solutions for high-speed smart cities networks: A systematic literature review. Journal of King Saud University - Computer and Information Sciences, 35(9), 101788. https://doi.org/10.1016/j.jksuci.2023.101788

[99] Darvishvand, L., Kamkari, B., Huang, M. J., and Hewitt, N. J. (2025). A systematic review of explainable artificial intelligence in urban building energy modeling: Methods, applications, and future directions. Sustainable Cities and Society, 128, 106492. https://doi.org/10.1016/j.scs.2025.106492

[100] Haghighat, M., MohammadiSavadkoohi, E., and Shafiabady, N. (2025). Applications of explainable artificial intelligence and interpretable artificial intelligence in smart buildings and energy savings in buildings: A systematic review. Journal of Building Engineering, 107, 112542. https://doi.org/10.1016/j.jobe.2025.112542.