Intelligent Systems and Robotics

E-ISSN: 3123-9803

Research Article

Explainable Artificial Intelligence Methods for Autonomous Robot Decision Making: A Multi Agent Framework with Safety Assurance and Ethical Constraint Optimization

Harry Setya Hadi [1]*, Nicodemus Rahanra [2]

1 Universitas Ekasakti, Indonesia xmoensen@gmail.com
2 Universitas Satya Wiyata Mandala, Indonesia nicorh73@gmail.com
* Corresponding Author : Harry Setya Hadi

Abstract: Autonomous decision-making systems increasingly rely on complex artificial intelligence models to operate in dynamic and safety-critical environments. While these models provide strong predictive capabilities, their black-box nature limits transparency, trust, and accountability. This study proposes a structured research methodology for integrating Explainable Artificial Intelligence (XAI) into autonomous decision making systems. The research adopts a conceptual–analytical approach to develop an explainability oriented framework that embeds transparency across perception, decision-making, and action execution stages. The methodology includes literature-driven problem identification, conceptual framework construction, classification and mapping of XAI methods, and formulation of explainability evaluation criteria. The results demonstrate that effective explainability in autonomous systems requires a hybrid integration strategy, combining in model transparency with post-hoc explanation mechanisms. A structured mapping of XAI techniques to autonomous system components and a conceptual decision flow diagram are presented to illustrate explainability integration. The findings highlight that layered and context-aware explainability enhances system interpretability, supports human oversight, and improves safety relevance without compromising autonomous operation. This study contributes a reusable methodological foundation for the design and evaluation of explainable autonomous systems, offering practical guidance for future empirical validation and real-world deployment in safety-critical applications.

Keywords: Autonomous decision making; Explainable artificial intelligence; Safety-critical systems; Transparent AI; Trustworthy AI.

Received: November 20, 2025
Revised: Desember 30, 2025
Accepted: January 14, 2026
Published: January 18, 2026
Curr. Ver.: January 20, 2026

1. Introduction

Autonomous robotic systems have experienced rapid and substantial development over the past decade, particularly within safety critical domains such as transportation, industrial manufacturing, public services, and healthcare. These systems integrate artificial intelligence (AI), machine learning, sensing, and control technologies to enable decision-making and action execution with minimal or no human intervention. Their adoption promises significant improvements in operational efficiency, productivity, and safety, while simultaneously introducing complex challenges related to reliability, security, ethics, and regulatory compliance [1], [2]. In the transportation sector, autonomous technologies have emerged as a transformative force. Autonomous vehicles and robot-assisted mobility systems are designed to optimize traffic flow, reduce human error, and enhance mobility services across urban and industrial environments [3], [4]. Beyond passenger transportation, autonomous robots are increasingly deployed in last-mile delivery and logistics operations, where they contribute to improved efficiency, reduced operational costs, and minimized human-related errors. Despite these advantages, transportation-focused autonomous systems must operate

under highly dynamic and unpredictable conditions, making safety assurance and robust decision-making critical concerns [2].

In industrial contexts, autonomous robots play a vital role in manufacturing, material handling, quality inspection, and hazardous operations. Advances in AI and learning-based approaches allow robots to adapt to complex environments, learn from operational data, and collaborate with human workers [1]. By assuming dangerous and repetitive tasks, autonomous robots contribute significantly to workplace safety and risk reduction. However, their growing autonomy also raises concerns regarding system dependability, fault tolerance, and secure interaction within interconnected industrial ecosystems [5]. Public service applications further demonstrate the societal impact of autonomous robotic systems. In search-and-rescue missions, autonomous robots operate in disaster-stricken or hazardous environments where human access is limited or unsafe. Similarly, in healthcare settings, autonomous mobile service robots are increasingly employed for tasks such as medical equipment transportation, logistics support, and surgical assistance. Compliance with safety standards, including those governing human–robot interaction, is essential to ensure safe and trustworthy deployment in these sensitive environments [6].

Despite their expanding adoption, autonomous robotic systems face persistent challenges related to safety, reliability, cybersecurity, and ethical governance. Autonomous decision-making systems must function correctly in unstructured environments while remaining resilient to cyber threats and system failures [7], [8]. Furthermore, ethical considerations and regulatory frameworks are increasingly important to ensure transparency, accountability, and social acceptance of autonomous technologies across safety-critical domains [1], [5]. Addressing these challenges is essential for the sustainable and responsible integration of autonomous robots into mission-critical applications. Deep learning has become a dominant paradigm for extracting patterns and making predictions from complex, high-dimensional data, enabling breakthroughs across healthcare, finance, and geospatial analytics. Yet, many high-performing deep neural networks remain difficult to understand, often being characterized as black-box models because their internal reasoning is opaque to users and stakeholders [9], [10]. This opacity is not merely a theoretical limitation: in high-stakes settings, stakeholders must be able to justify and audit automated decisions, especially when model outputs influence clinical actions, financial risk, or public safety [11], [12].

A primary challenge of black-box deep learning lies in its intrinsic complexity. Multilayer nonlinear transformations, distributed representations, and feature hierarchies make it difficult to trace how specific inputs lead to particular outputs, even when prediction accuracy is high [9]. In computer vision, for example, models may attend to spurious correlations or background artifacts while still achieving strong benchmark performance, motivating extensive research into explainers and visualization-based methods [10]. Similar challenges arise in geospatial AI, where models used for satellite imagery and GeoAI tasks often require interpretability to ensure reliability under dataset shift, uncertain sensing conditions, and diverse spatial contexts [13], [14]. The consequences of opacity become especially pronounced in safety- and mission-critical domains. In medical diagnosis, deep learning systems can support detection and classification tasks, but clinicians frequently require interpretable evidence to validate whether model behavior aligns with biomedical reasoning and patient-specific factors [12]. Recent applied studies illustrate this demand: layer-attribution approaches have been used to interpret diabetic foot ulcer image classification, providing insight into which regions contribute most to decisions [15], while post-hoc local interpretability techniques have been explored to explain autism diagnosis models [16]. Beyond healthcare, interpretability also matters in critical infrastructure, such as power system transient stability assessment, where interpretable deep learning with tree regularization can support both strong performance and clearer decision rationale for operators [17]. In transportation analytics, explainability methods applied to graph neural networks for road traffic forecasting highlight the importance of understanding feature and structural contributions in complex spatiotemporal predictions [18].

Black-box behavior also raises ethical and reliability concerns. When models are trained on biased or unrepresentative datasets, their opaque decision rules may propagate unfairness and discrimination without clear pathways for detection, contestation, or accountability [9],

[11]. In domains such as finance where AI systems support credit scoring, fraud detection, or trading decisions lack of transparency can undermine trust, complicate governance, and amplify risks associated with model error or misuse [19]. These concerns strengthen the case for explainable AI (XAI) not only as a usability enhancement, but as a core requirement for trustworthy deployment. Regulatory developments further intensify this need. Emerging governance frameworks increasingly demand transparency, accountability, and risk controls for high-impact AI. As a result, explainability is becoming a practical necessity for compliance and auditability rather than an optional add-on [9]. In response, research has expanded across post-hoc and design-time strategies. Post-hoc explainability aims to interpret trained models by attributing importance to inputs, features, or internal activations particularly common in computer vision and clinical imaging using visualization and relevance propagation approaches [10], [20]. For instance, combining Grad-CAM with Layer-Wise Relevance Propagation (LRP) can provide complementary perspectives on convolutional neural network (CNN) decisions, supporting both localization and attribution-based reasoning [20]. Meanwhile, design-time (ante-hoc) transparency embeds interpretability into model architecture or learning constraints from the outset, which can reduce dependence on potentially fragile post-hoc explanations and improve reliability [9], [14] .

A complementary direction is hybrid modeling, which seeks to preserve deep learning's predictive strength while improving transparency by coupling neural networks with more interpretable structures, such as tree-based regularization or surrogate models [9], [17]. This line of work reflects a growing consensus: for deep learning to be responsibly adopted in high-stakes domains, performance must be accompanied by interpretable and reliable justification that can be examined by experts and aligned with domain constraints [11], [12]. Accordingly, the development and evaluation of XAI methods across post-hoc explainers, ante-hoc transparency, and hybrid interpretability remain central to advancing trustworthy deep learning in real-world decision-making systems [9], [10].

2. Literature Review

Autonomous Robot Decision Making Architectures

Autonomous robots operating in dynamic and uncertain environments require advanced decision-making architectures capable of perceiving environmental states, reasoning under uncertainty, and executing appropriate actions in real time. Traditional robotic systems relied on rule-based or modular pipelines that separated perception, planning, and control. However, increasing task complexity and environmental variability have motivated the development of more integrated and adaptive decision-making frameworks [21].

One prominent direction involves the synthesis of cognitive architectures and multi-agent systems. In such approaches, autonomous robot intelligence is modeled as a collection of interacting agents, often inspired by neurocognitive principles. Individual neurons or functional units are represented as rational software agents that cooperate to optimize local objective functions, enabling emergent goal-directed behavior under partial observability and uncertainty [21]. These multi-agent neurocognitive architectures aim to improve adaptability and robustness by distributing reasoning processes rather than relying on centralized decision modules.

Despite these advances, many autonomous decision-making systems still rely heavily on black-box learning models, particularly deep neural networks. While these models offer strong representational power, their opaque internal reasoning poses challenges for reliability, safety assurance, and human oversight especially in safety-critical robotic applications.

Limitations of Black-Box Approaches in Autonomous Robotics

Black-box approaches in autonomous robot decision making exhibit several well-documented limitations. One major issue concerns generalization. Deep learning models

trained on limited or task-specific datasets may fail to perform reliably when confronted with unseen environmental conditions, dynamic obstacles, or rare events [22]. In real-world robotic deployment, such failures may necessitate human intervention, retraining, or transfer learning strategies, undermining full autonomy.

Another limitation arises from fragmented feature transmission in conventional modular robotic architectures. Systems that strictly separate perception, mapping, and decision-making stages often suffer from information loss and insufficient environmental modeling. As a result, downstream decision modules may operate on incomplete or distorted representations of the environment, leading to suboptimal or unsafe actions [23]. These limitations highlight the need for architectures that more tightly integrate perception and reasoning.

Furthermore, the opacity of black-box decision models complicates system verification and validation. In autonomous robots, understanding why a particular action was selected is essential for debugging, safety certification, and human–robot interaction. Without interpretable reasoning processes, trust in autonomous decision-making remains limited, particularly in applications involving close human collaboration or mission-critical tasks.

Emerging Solutions for Explainable and Robust Decision Making

To address the shortcomings of black-box approaches, several emerging solutions have been proposed in the literature. One influential trend is the adoption of end-to-end learning architectures that jointly optimize perception and decision making. Deep reinforcement learning (DRL) frameworks, combined with spatiotemporal attention mechanisms and Transformer-based architectures, have demonstrated improved obstacle avoidance, temporal reasoning, and policy generalization in autonomous navigation tasks [22], [24]. By learning continuous mappings from sensory inputs to actions, these architectures reduce fragmentation and enhance responsiveness in dynamic environments.

Another complementary approach emphasizes explainable artificial intelligence (XAI). Integrating XAI techniques into robotic decision-making systems enables the generation of human-understandable explanations for autonomous actions. Such explanations can take textual, symbolic, or visual forms, facilitating transparency and increasing operator trust [25]. XAI-based situation recognition frameworks have shown particular promise in environments with partially unlabeled data, where explainability supports both learning efficiency and situational understanding.

Situational awareness (SA) has also emerged as a critical component of autonomous robot decision making. SA frameworks extend beyond traditional simultaneous localization and mapping (SLAM) by integrating sensing, spatial perception, semantic understanding, and state estimation into a unified representation of the environment [23]. Enhanced situational awareness enables robots to anticipate environmental changes, assess risks, and select context-appropriate actions more effectively.

Finally, recent research has explored hybrid cognitive and learning-based models that combine the adaptability of deep learning with structured reasoning mechanisms. By embedding reasoning models, attention structures, or explainability constraints into learning architectures, these approaches aim to balance performance, interpretability, and robustness [21]. Such hybrid frameworks are increasingly viewed as a promising direction for achieving trustworthy and scalable autonomous robot decision making.

Concept and Rationale of XAI

Explainable Artificial Intelligence (XAI) refers to methods and system designs that make AI decision-making processes understandable to human users. The motivation for XAI is largely driven by the widespread use of complex models especially deep learning whose internal reasoning is often opaque, limiting transparency, trust, and accountability [26], [27]. XAI aims to clarify how and why a model produces a specific output, including decision criteria, influential features, and potential failure modes, thereby reducing the "black-box"

effect and supporting responsible adoption in real-world settings [28], [29]. Beyond technical transparency, XAI is also shaped by human factors such as perceived usefulness, explanation satisfaction, and user expertise, which jointly influence trust and acceptance [30].

Taxonomies of XAI Methods

The literature commonly classifies XAI methods along multiple dimensions, helping researchers select techniques aligned with goals and constraints. A frequent categorization distinguishes explanations by purpose (pre-model, in-model, post-model), scope (local vs. global), and usability (model-agnostic vs. model-specific) [26], [29]. These categories are practically important: local explanations support case-by-case justification, while global explanations help audit general behavior and identify systemic bias [26]. Model-agnostic techniques can be applied broadly across algorithms, whereas model-specific methods can exploit internal structures (e.g., gradients, activations, attention) to offer richer insights at the cost of generality [13].

Widely Used Techniques: LIME and SHAP

Among model-agnostic, post-hoc approaches, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are widely adopted because they provide feature-level contributions that are relatively easy to communicate to end users. Comparative reviews emphasize their practical utility for explaining individual predictions and supporting debugging, validation, and stakeholder communication [26], [31]. LIME constructs a local surrogate model around a specific prediction to approximate complex behavior in an interpretable form, while SHAP uses Shapley-value principles from cooperative game theory to estimate fair feature attributions [26], [28]. However, surveys also highlight limitations such as instability under perturbations, sensitivity to feature correlations, and computational overhead in large-scale or real-time settings [26], [32].

XAI for Vision Models: Grad-CAM and Attribution Families

In computer vision, explainability frequently relies on attribution-based methods that highlight regions of an image contributing to a decision. Grad-CAM remains a canonical technique for convolutional neural networks, generating saliency-like visualizations by leveraging gradient information to locate influential image regions [13]. Reviews of XAI in computer vision further distinguish explainers into attribution-based, activation-based, perturbation-based, and Transformer-based families, reflecting whether explanations rely on input attributions, internal neuron responses, input modifications, or attention patterns [33] . This taxonomy is increasingly relevant because modern vision models are shifting toward Transformer architectures, where self-attention can support more global interpretability narratives compared with purely local saliency maps [33].

Another commonly cited attribution approach is Integrated Gradients, which assigns feature importance by integrating gradients along a path from a baseline input to the actual input. It is often discussed as a principled gradient-based method that can complement or validate other attribution techniques [26]. In addition, research has explored generating textual explanations for predicted images, moving beyond heatmaps to more human-readable rationales useful for non-technical stakeholders [34].

XAI in Safety-Critical and Autonomous Systems

XAI is particularly critical in safety- and mission-critical domains where decisions must be traceable for accountability and risk management. In healthcare, XAI supports clinical interpretability, model auditing, and safer adoption of AI-assisted decisions; surveys emphasize interpretability as essential for trust, ethics, and error analysis [26], [35]. In autonomous driving, explainability contributes to safer deployment by clarifying perception and planning decisions, improving human trust, and enabling post-incident analysis. Systematic reviews in intelligent transportation emphasize that explainable autonomous driving must address both technical transparency and human-centered explanation quality [36].

Within autonomous vehicle research, studies propose hybrid XAI frameworks that combine multiple explanation types (e.g., feature attribution plus rule-like summaries) to enhance transparency and trust under real driving constraints [37]. Broader treatments of XAI for autonomous vehicles also document challenges such as real-time explainability requirements, multi-modal sensor fusion, and the need to explain sequential decisions rather than single predictions [38]. Related work in unmanned aerial vehicles (UAVs) highlights XAI's role in navigation and obstacle detection in complex urban environments, where interpretability can improve reliability and operator confidence [39].

Evaluation Issues, Challenges, and Future Directions

Despite significant progress, XAI research faces persistent challenges. A frequently cited issue is the trade-off between accuracy and interpretability: highly expressive models may yield strong performance but are harder to explain meaningfully, while simpler models can be interpretable yet less accurate [26], [32]. Another concern is scalability, since many explanation methods impose computational costs that limit applicability in real-time or large-scale deployments, especially in autonomous systems where decisions occur continuously [36], [40]. A third challenge is standardization the field lacks universally accepted evaluation metrics and benchmarking frameworks for explanation quality, fidelity, and usefulness across diverse user groups and domains [26], [28].

Consequently, current research trends emphasize developing more efficient explainers, aligning explanations with human cognitive needs, and establishing standardized evaluation protocols. Reviews also point toward integrating XAI directly into system design (rather than relying only on post-hoc explainers), improving explanation robustness, and ensuring that XAI supports ethical and trustworthy AI deployment [41], [42]. The growing body of domain-focused work including education, healthcare, computer vision, and transportation suggests that future XAI methods will increasingly be tailored to context, user roles, and operational constraints while maintaining fidelity to underlying model behavior [33], [43]

Artificial Intelligence and Autonomous Decision Making Systems

Artificial Intelligence (AI) has significantly transformed the development of autonomous systems, particularly in robotics and intelligent decision-making environments. Autonomous robots rely on AI algorithms to interpret sensory information, analyze environmental conditions, and perform actions without continuous human intervention. These systems combine perception, reasoning, and learning mechanisms to enable adaptive responses in dynamic environments. The integration of machine learning, distributed computing, and intelligent frameworks has enhanced the ability of autonomous systems to process large amounts of data and generate real-time decisions. Research on hybrid AI models and distributed learning environments demonstrates how intelligent systems can detect patterns, adapt to changing conditions, and optimize decision processes across complex computational infrastructures [44], [45].

Recent studies also emphasize the importance of integrating AI with modern technological infrastructures such as cloud computing, blockchain, and trusted execution environments. These technologies support secure data exchange, scalable processing, and reliable decision support systems. Frameworks combining machine learning with blockchain-based architectures have been proposed to enhance system integrity, transparency, and resilience in digital environments [46]. Such developments provide an important theoretical foundation for autonomous robotic systems that require both computational intelligence and reliable infrastructure to support safe and trustworthy decision making.

Explainable Artificial Intelligence in Intelligent Systems

Explainable Artificial Intelligence (XAI) refers to methods and techniques that enable AI systems to provide understandable explanations for their decisions or predictions. As AI systems become increasingly complex, especially those using deep learning architectures, transparency and interpretability become essential requirements. In autonomous robotic environments, explainability helps developers and users understand the reasoning behind

system actions, ensuring that decisions are not only accurate but also interpretable and accountable.

Explainability is particularly important in safety-critical systems where autonomous decisions may directly affect human safety or operational stability. Transparent decision processes enable auditing, debugging, and validation of AI models. Research addressing digital governance, secure technology adoption, and responsible AI implementation highlights the importance of transparency and accountability in modern intelligent systems [47] . In this context, XAI serves as a bridge between complex AI models and human understanding, enabling trust and acceptance of autonomous robotic technologies in real-world applications.

Furthermore, integrating explainability with secure and adaptive AI frameworks enhances the reliability of intelligent systems. Studies focusing on AI-driven security frameworks and hybrid machine learning models demonstrate how complex systems can maintain both high performance and transparency in decision processes [45]. These insights support the theoretical premise that explainable models are essential for building trustworthy autonomous systems.

Multi Agent Systems for Distributed Autonomous Decision Making

Multi-agent systems (MAS) represent a computational paradigm in which multiple intelligent agents interact and collaborate to achieve system objectives. Each agent operates independently while sharing information and coordinating actions with other agents within a distributed environment. This architecture is particularly relevant for autonomous robots operating in complex and dynamic scenarios where centralized decision making may be inefficient or impractical.

Distributed AI frameworks enable agents to perform local reasoning while contributing to global system goals. In robotic systems, this approach allows agents to distribute tasks, coordinate navigation strategies, and share environmental information. Studies on federated learning and distributed detection systems illustrate how decentralized intelligence can improve scalability, adaptability, and resilience in large-scale computing environments [44]. These principles can be extended to robotic environments where multiple agents collaborate to interpret environmental data and optimize collective decisions.

Hybrid distributed architectures also improve system robustness by allowing agents to operate even when certain components fail or communication delays occur. This distributed resilience is essential for autonomous robotic applications such as swarm robotics, industrial automation, and smart infrastructure systems. By combining multi-agent coordination with intelligent learning models, robotic systems can achieve higher levels of autonomy and adaptability.

Safety Assurance in Autonomous Robotic Systems

Safety assurance is a critical requirement for autonomous systems, particularly when robots operate in environments involving human interaction or high-risk industrial processes. Safety assurance refers to the mechanisms that ensure a system behaves predictably and avoids actions that may cause harm or operational failures. These mechanisms include system monitoring, anomaly detection, redundancy, and secure architecture design.

Recent research highlights the importance of integrating security and reliability mechanisms within intelligent systems to ensure continuous and safe operations. Hybrid security models, such as zero-trust architectures and AI-based threat detection systems, have been proposed to maintain system stability in complex computing environments [44] . Although originally designed for cloud infrastructures, these principles can be applied to autonomous robotic systems to protect decision processes from data manipulation, cyber threats, or unexpected environmental disturbances.

In addition, secure infrastructure frameworks combining machine learning, blockchain technology, and trusted execution environments provide additional layers of protection for intelligent systems [46]. By integrating such mechanisms into robotic control architectures, autonomous systems can ensure that their decision-making processes remain safe, reliable, and resistant to external disruptions.

Ethical Constraints and Responsible AI in Autonomous Systems

The rapid advancement of autonomous technologies has raised important ethical considerations regarding how intelligent systems should make decisions. Ethical constraint optimization refers to the incorporation of ethical principles, social norms, and human safety priorities into the decision-making algorithms of AI systems. Instead of optimizing purely for efficiency or performance, autonomous systems must also consider fairness, accountability, and societal impact.

Research on sustainable digital culture and responsible technology development emphasizes the importance of aligning technological innovation with social values and governance frameworks [45]. Ethical AI frameworks encourage the development of systems that prioritize human safety, transparency, and fairness in automated decision processes. In autonomous robotics, ethical constraints may guide robots to prioritize human well-being, avoid harmful actions, and comply with regulatory or social expectations.

Integrating ethical constraints into optimization models ensures that autonomous decisions remain aligned with human values while maintaining operational efficiency. This approach also strengthens public trust in intelligent systems by demonstrating that AI technologies operate within well-defined ethical boundaries.

IoT and Sensor Driven Perception in Autonomous Robots

Autonomous robots rely heavily on sensor data to perceive and understand their surrounding environments. Technologies such as Internet of Things (IoT), embedded systems, and sensor networks enable robots to collect real-time environmental information that supports intelligent decision making. IoT-based monitoring systems demonstrate how distributed sensors can gather environmental data, transmit information to intelligent systems, and support automated responses [48].

Research on IoT-based security systems and sensor-driven automation also illustrates the role of integrated sensors in enabling real-time monitoring and control [49], [50]. In autonomous robotics, such sensor networks serve as the foundation for perception modules that allow robots to detect obstacles, recognize objects, and adapt to environmental changes.

When combined with explainable AI techniques, sensor-based perception can provide transparent explanations for robotic actions. For example, a robot may explain that a specific maneuver was performed due to sensor detection of an obstacle or environmental hazard. This integration between perception data and explainable decision models enhances system transparency and reliability.

3. Research Method

Research Design

This study employs a conceptual analytical research design aimed at developing an explainable artificial intelligence (XAI) oriented framework for autonomous decision-making in safety-critical systems. The research focuses on structuring methodological principles rather than implementing or benchmarking a specific algorithm. The primary objective is to formulate a coherent methodological approach that integrates explainability as an inherent component of autonomous decision-making processes.

The methodology is organized into four main stages: (1) literature-driven problem identification, (2) conceptual framework construction, (3) explainability method classification

and integration, and (4) formulation of evaluation criteria for explainable autonomous systems.

Literature-Driven Problem Identification

The first stage involves identifying key challenges associated with black-box decision-making models in autonomous systems. These challenges include limited transparency, difficulties in generalization, reduced human trust, and constraints on safety assurance. The analysis focuses on how these issues emerge across perception, reasoning, and action execution stages in autonomous decision pipelines. The findings from this stage are used to define methodological requirements for explainability integration.

Conceptual Framework Construction

Based on the identified challenges, a conceptual framework for XAI-enabled autonomous decision making is developed. The framework models the autonomous system as a layered structure consisting of perception, representation, decision-making, action execution, and explanation components. Explainability is treated as a core functional element rather than an auxiliary feature, allowing explanations to be generated alongside or directly from the decision-making process.

The framework emphasizes human-centered design, ensuring that explanations are aligned with the cognitive needs of system operators, developers, and other stakeholders. This design supports transparency, accountability, and informed human oversight in safety-critical environments.

Explainability Method Classification and Integration

In the third stage, explainability methods are classified according to their role and applicability within the autonomous decision-making framework. The classification follows three dimensions: purpose, scope, and usability. Methods are categorized as pre-model, in-model, or post-model explanations; as local or global explanations; and as model-agnostic or model-specific techniques.

These categories are then mapped to different stages of the autonomous decision pipeline. Post-model explainability is used to analyze and audit system behavior, while in-model transparency mechanisms are integrated into the decision-making layer for critical actions requiring immediate interpretability. This hybrid integration strategy is designed to balance decision accuracy with interpretability and operational feasibility.

Explainability Integration Strategy

The proposed methodology adopts a hybrid explainability integration strategy that combines internal transparency mechanisms with external explanation modules. Internal mechanisms provide insight into model reasoning during decision generation, while external modules translate these insights into human-understandable representations. Explanations are designed to be delivered in multiple formats, such as visual, numerical, or textual, depending on system context and user requirements.

This strategy enables continuous monitoring of autonomous decisions and supports post-event analysis, system debugging, and safety validation without disrupting real-time system performance.

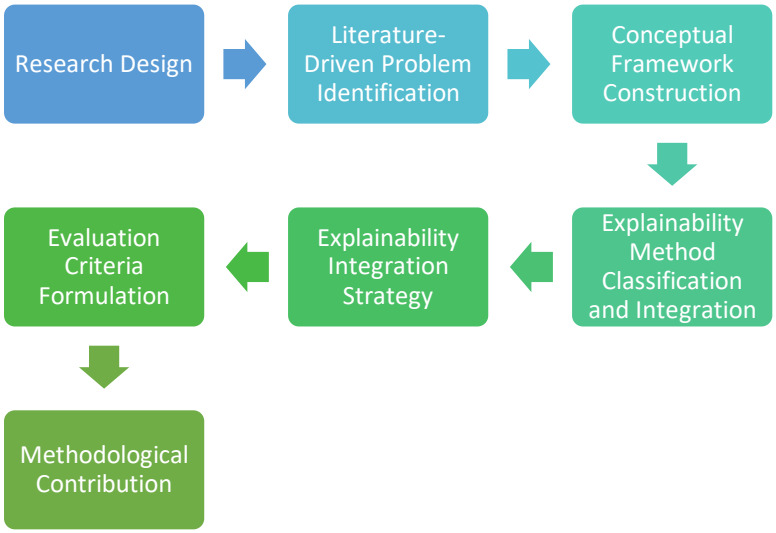
Evaluation Criteria Formulation

Instead of empirical performance testing, this study defines conceptual evaluation criteria for assessing the effectiveness of explainable autonomous decision-making systems. The criteria include explanation fidelity, transparency, interpretability, operational suitability, and safety relevance. These criteria provide a structured basis for future empirical validation and comparative studies.

Methodological Contribution

The proposed methodology contributes a structured and explainability-centered approach to autonomous decision-making research. By embedding explainability into the methodological design, the approach supports the development of autonomous systems that are not only effective but also transparent, trustworthy, and aligned with safety and ethical requirements.

Table 1. Research Framework and Methodological Flow



4. Results and Discussion

Overview of the Research Outcomes

This section presents the results derived from the proposed research methodology, which focuses on the conceptual integration of Explainable Artificial Intelligence (XAI) into autonomous decision-making systems. The results are presented in three forms: (1) structured methodological outcomes, (2) a tabular mapping of XAI methods to decision-making stages, and (3) a conceptual diagram illustrating the explainability-oriented decision flow. These results collectively demonstrate how explainability can be systematically embedded into autonomous systems to enhance transparency, trust, and safety.

Structured Outcomes of the Proposed Methodology

The primary outcome of this research is a methodological framework that integrates XAI as a core component of autonomous decision making. The framework emphasizes layered decision processing, hybrid explainability integration, and human-centered explanation delivery. From the methodological analysis, three key outcomes are identified:

1. Autonomous decision-making processes can be decomposed into explainability-aware stages.

2. Different XAI techniques serve distinct functional roles depending on decision criticality and system constraints.

3. Explainability evaluation must consider safety relevance and operational feasibility, not only interpretability.

Mapping XAI Methods to Autonomous Decision-Making Stages

Table of XAI Integration Results

Table 1 summarizes the mapping between autonomous system components and appropriate XAI techniques based on purpose, scope, and usability.

Table 2. Mapping of XAI Methods to Autonomous Decision-Making Components

Autonomous System Stage	Decision Function	Type of Explainability	Scope	Expected Outcome
Perception	Sensor data interpretation	Attribution-based (e.g., visual saliency)	Local	Transparency of environmental perception
Representation	Feature abstraction	Activation-based	Global	Understanding internal feature representations
Decision-Making	Action selection	In-model / Hybrid	Local & Global	Explainable policy reasoning
Action Execution	Control and navigation	Post-hoc explanation	Local	Accountability of executed actions
System Monitoring	Audit and validation	Model-agnostic	Global	Safety verification and trust

Explanation of Table 1

Table 1 demonstrates that explainability requirements vary across system stages. Early stages such as perception benefit from local, attribution-based explanations that clarify sensory interpretation, while higher-level decision-making stages require hybrid approaches capable of explaining both individual actions and overall policy behavior. This mapping confirms that a single XAI technique is insufficient for autonomous systems, reinforcing the need for a structured, multi-level integration strategy.

Conceptual Diagram of the XAI-Based Decision Flow

Introductory Description of the Diagram

To further illustrate the research results, a conceptual diagram is introduced to visualize the explainability-enabled autonomous decision-making flow. The diagram highlights how explainability mechanisms interact with core system components and support continuous human understanding.

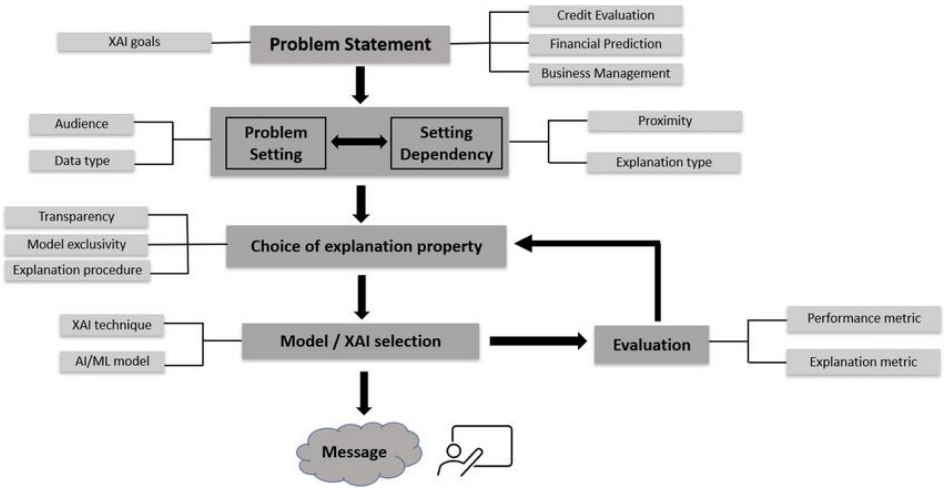


Figure 1. Conceptual Diagram of the XAI-Based Autonomous Decision-Making Framework

Explanation of the Diagram

The diagram illustrates a layered autonomous system architecture consisting of perception, representation, decision-making, and action execution modules. An explainability layer operates in parallel with the decision-making process, capturing internal reasoning signals and transforming them into interpretable outputs. This structure enables real-time transparency during operation and post-event analysis for safety auditing. The diagram also shows feedback loops that allow explanations to inform system refinement and human oversight.

Discussion

Interpretation of the Methodological Results

The results indicate that integrating XAI into autonomous decision-making systems requires a system-level perspective rather than isolated explanation tools. The tabular mapping and conceptual diagram together reveal that explainability must be aligned with functional roles, decision risk levels, and operational constraints. This finding supports the view that explainability is not a single feature but a distributed capability across the autonomous system lifecycle.

Implications of Table-Based Findings

The mapping presented in Table 1 demonstrates that explainability effectiveness depends on selecting techniques appropriate to each system stage. For example, attribution-based explanations are well-suited for perception tasks but are insufficient for explaining long-term decision policies. Conversely, global explanations are critical for system validation and regulatory compliance but may lack actionable detail for real-time decisions. This reinforces the importance of hybrid explainability strategies that combine multiple techniques to address diverse stakeholder needs.

Implications of the Conceptual Diagram

The conceptual diagram highlights how explainability can be embedded without disrupting autonomous operation. By positioning explainability as a parallel layer rather than a post-processing add-on, the framework supports continuous transparency while preserving system performance. This design is particularly relevant for safety-critical applications, where operators must understand system behavior both during operation and after incidents.

Contribution to Autonomous and XAI Research

The presented results contribute a structured methodological perspective to XAI research in autonomous systems. Unlike application-specific studies, this work provides a reusable framework that can guide future system design, empirical validation, and regulatory alignment. The results also demonstrate how explainability can support trust, accountability, and ethical deployment without compromising autonomy.

Limitations and Future Research Directions

While the results offer a comprehensive methodological foundation, they are conceptual in nature and require empirical validation in real-world or simulated environments. Future research should implement the proposed framework in specific autonomous domains and evaluate explainability effectiveness using human-centered and safety-oriented metrics.

5. Comparison

Compared to conventional autonomous decision-making approaches that predominantly rely on black-box learning models, the proposed XAI-oriented framework offers a more structured and transparent integration of explainability across the entire decision pipeline. Traditional approaches typically treat explainability as an optional, post-hoc component, applied only after decisions are made, which limits its usefulness for real-time understanding and safety assurance. In contrast, the proposed framework embeds explainability as a core methodological element, aligning specific XAI techniques with distinct stages of perception, representation, decision making, and action execution. This layered integration enables both local and global explanations, supporting immediate operational insight as well as system-level accountability. Moreover, while many existing methods focus on optimizing predictive performance in isolation, the proposed approach explicitly balances interpretability, operational feasibility, and safety relevance, making it more suitable for safety-critical autonomous systems that require human oversight and regulatory compliance.

6. Conclusions

This study presents a structured methodological framework for integrating Explainable Artificial Intelligence (XAI) into autonomous decision-making systems, addressing the limitations of black-box models in safety-critical environments. By positioning explainability as a core component rather than a post-hoc addition, the proposed approach enhances transparency, trust, and accountability across the autonomous decision pipeline.

The results demonstrate that effective explainability in autonomous systems requires a layered integration strategy, where different XAI techniques are aligned with specific system stages and decision risks. The mapping of explainability methods and the conceptual decision-flow diagram highlight that no single explanation technique is sufficient; instead, hybrid and multi-level explainability is necessary to support both real-time understanding and system-level validation.

From a methodological perspective, the proposed framework contributes a reusable foundation for future research and system development in explainable autonomous intelligence. It supports human-centered oversight, facilitates safety assurance, and aligns with emerging ethical and regulatory expectations without compromising autonomous operation.

Although the findings are conceptual in nature, they provide clear directions for empirical validation and practical implementation. Future work should focus on deploying the framework in simulated or real-world autonomous systems, evaluating explanation quality using human-centered metrics, and refining scalability for real-time applications. Overall, this study underscores the critical role of XAI in advancing trustworthy and responsible autonomous decision-making systems.

References

[1] S. Hemalatha, K. V. S. V. T. Reddy, T. V Rao, T. Ramaswamy, N. M. Pillai, and G. K. Mohan, "Advancing autonomous systems: A review of emerging trends in robotics," *J. Eur. des Systèmes Autom.*, vol. 58, no. 5, pp. 913–921, 2025, doi: 10.18280/jesa.580505.

[2] C. Thames and Y. Sun, "A survey of artificial intelligence approaches to safety and mission-critical systems," in *Integrated Communications, Navigation and Surveillance Conference*, 2024. doi: 10.1109/ICNS60906.2024.10550712.

[3] A. Sevgi Ostim, A. M. Kadio\uglu Ostim, and A. Durmu\cs Ostim, "Using of robotic systems in transportation," in *Lecture Notes in Networks and Systems*, Springer, 2025, pp. 458–468. doi: 10.1007/978-3-031-81799-1_42.

[4] U. A. Usmani, A. Happonen, and J. Watada, "Revolutionizing transportation: Advancements in robot-assisted mobility systems," in *Lecture Notes in Networks and Systems*, Springer, 2023, pp. 603–619. doi: 10.1007/978-981-99-4932-8_55.

[5] S. Akhai, M. Abbass, P. Kaur, and T. Kaur, "Digital transformation across generations: Robotics and AI in action," in *Impacts of Digital Technologies Across Generations*, IGI Global, 2025, pp. 23–39. doi: 10.4018/979-8-3693-6366-9.ch002.

[6] K. Thamrongaphichartkul, N. Worrasittichai, T. Prayongrak, and S. Vongbunyong, "Development of autonomous mobile service robot with safety standard for cart moving applications in hospitals," in *AIP Conference Proceedings*, 2024, p. 70006. doi: 10.1063/5.0205057.

[7] S. Katzenbeisser, I. Polian, F. Regazzoni, and M. Stottinger, "Security in autonomous systems," in *Proceedings of the European Test Workshop*, 2019. doi: 10.1109/ETS.2019.8791552.

[8] M. Hamad and S. Steinhorst, "Security challenges in autonomous systems design," in *Communications in Computer and Information Science*, Springer, 2025, pp. 142–154. doi: 10.1007/978-3-031-81981-0_13.

[9] E. Şahin, N. N. Arslan, and D. Özdemir, "Unlocking the black box: An in-depth review on interpretability, explainability, and reliability in deep learning," *Neural Comput. Appl.*, vol. 37, no. 2, pp. 859–965, 2025, doi: 10.1007/s00521-024-10437-2.

[10] V. Buhrmester, D. Münch, and M. Arens, "Analysis of explainers of black box deep neural networks for computer vision: A survey," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 4, pp. 966–989, 2021, doi: 10.3390/make3040048.

[11] F. Emmert-Streib, O. Yli-Harja, and M. Dehmer, "Explainable artificial intelligence and machine learning: A reality rooted perspective," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 6, p. e1368, 2020, doi: 10.1002/widm.1368.

[12] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimed. Syst.*, vol. 28, no. 6, pp. 2335–2355, 2022, doi: 10.1007/s00530-022-00960-4.

[13] X. Cheng *et al.*, "Explainability in GeoAI," in *Handbook of Geospatial Artificial Intelligence*, CRC Press, 2023, pp. 177–200. doi: 10.1201/9781003308423-9.

[14] S. B. Kasetty and K. Rajakumar, "Understanding satellite image processing and black box models with ante-hoc and post-hoc explanations in deep learning," in *Proceedings of the 2024 10th International Conference on Communication and Signal Processing (ICCSP)*, 2024, pp. 848–853. doi: 10.1109/ICCSP60870.2024.10544196.

[15] Z. M. Arkah, B. Pontes, and C. Rubio, "Interpretation of Diabetic Foot Ulcer Image Classification Using Layer Attribution Algorithms," in *Lecture Notes in Networks and Systems*, 2024, pp. 13–22. doi: 10.1007/978-3-031-75013-7_2.

[16] D. S. Sujana and D. P. Augustine, "Explaining autism diagnosis model through local interpretability techniques—A post-hoc approach," in *2023 International Conference on Data Science, Agents and Artificial Intelligence (ICDSAAI)*, 2023. doi: 10.1109/ICDSAAI59313.2023.10452575.

[17] C. Ren, Y. Xu, and R. Zhang, "An interpretable deep learning method for power system transient stability assessment via tree regularization," *IEEE Trans. Power Syst.*, vol. 37, no. 5, pp. 3359–3369, 2022, doi: 10.1109/TPWRS.2021.3133611.

[18] J. García-Sigüenza, F. Llorens-Largo, L. Tortosa, and J. F. Vicent, "Explainability techniques applied to road traffic forecasting using graph neural network models," *Inf. Sci. (Ny).*, vol. 645, p. 119320, 2023, doi: 10.1016/j.ins.2023.119320.

[19] B. Oliveira and C. C. Leal, "AI in finance: Applications and challenges," in *Challenges and Opportunities in the Artificial Intelligence Era*, Springer, 2025, pp. 79–107. doi: 10.1007/978-3-031-85272-5_6.

[20] A. Mishra and M. Malhotra, "A dual approach with Grad-CAM and layer-wise relevance propagation for CNN models explainability," in *Communications in Computer and Information Science*, Springer, 2025, pp. 116–129. doi: 10.1007/978-3-031-80842-5_10.

[21] K. Bzhikhatlov, O. Nagoeva, M. Anchokov, and D. Makoeva, "Methods and algorithms (modeling of reasoning) to synthesize intellectual behavior of autonomous mobile robots and program complexes based on received reasoning models," in *Studies in Computational Intelligence*, vol. 477, Springer, 2024, pp. 87–98. doi: 10.1007/978-3-031-76516-2_7.

[22] Y. Zhou and W. Zhang, "End-to-end robot intelligent obstacle avoidance method based on deep reinforcement learning with spatiotemporal transformer architecture," *Front. Neurorobot.*, vol. 19, p. 1646336, 2025, doi: 10.3389/fnbot.2025.1646336.

[23] H. Bavle, J. L. Sanchez-Lopez, C. Cimarelli, A. Tourani, and H. Voos, "From SLAM to situational awareness: Challenges and survey," *Sensors*, vol. 23, no. 10, p. 4849, 2023, doi: 10.3390/s23104849.

[24] B. Abdelkader, N. Emira, and E. Nadjib, "From perception to action: Transformer-enhanced deep reinforcement learning for autonomous robot navigation," in *Proceedings of the 7th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2025. doi: 10.1109/PAIS66004.2025.11126486.

[25] S. Lokhande, J. Dailey, Y. Liu, S. Connolly, and H. Xu, "A novel explainable AI based situation recognition for autonomous robots with partial unlabeled data," in *Proceedings of SPIE*, 2023, p. 1254606. doi: 10.1117/12.2664016.

[26] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, p. 128111, 2024, doi: 10.1016/j.neucom.2024.128111.

[27] T. H. Sardar, S. Das, and B. K. Pandey, "Explainable AI (XAI): Concepts and theory," in *Medical Data Analysis and Processing using Explainable Artificial Intelligence*, CRC Press, 2023, pp. 1–18. doi: 10.1201/9781003257721-1.

[28] M. H. Azam, M. H. Hasan, N. Y. Murad, and E. A. B. Patah, "Transparency in AI: A review of explainable artificial intelligence techniques," in *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2024. doi: 10.1109/ICCUBEA61740.2024.10774981.

[29] B. Mohammed, "A review on explainable artificial intelligence methods, applications, and challenges," *Indones. J. Electr. Eng. Informatics*, vol. 11, no. 4, pp. 1007–1024, 2023, doi: 10.52549/ijeei.v11i4.5151.

[30] G. M. Alarcon and S. M. Willis, "Explaining explainable artificial intelligence: An integrative model of objective and subjective influences on XAI," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2023, pp. 1095–1104.

[31] S. Bhatnagar and R. Agrawal, "Understanding explainable artificial intelligence techniques: A comparative analysis for practical application," *Bull. Electr. Eng. Informatics*, vol. 13, no. 6, pp. 4451–4455, 2024, doi: 10.11591/eei.v13i6.8378.

[32] A. Singhal, P. Pratap, K. K. Dixit, and K. Kathuria, "Advancements in explainable AI: Bridging the gap between model complexity and interpretability," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 2024, pp. 675–680. doi: 10.1109/ICDT61202.2024.10489277.

[33] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A comprehensive review of explainable artificial intelligence (XAI) in computer vision," *Sensors*, vol. 25, no. 13, p. 4166, 2025, doi: 10.3390/s25134166.

[34] B. P. Sheela and H. Girisha, "An explainable artificial intelligence (XAI) framework for deep learning based classification to generate textual explanations on predicted images," *Int. J. Intell. Eng. Syst.*, vol. 17, no. 6, pp. 651–662, 2024, doi: 10.22266/ijies2024.1231.50.

[35] M. Thamer and Z. N. Sultani, "Explainable AI in the medical field: A survey on machine learning interpretability and use cases," *Al-Nahrain J. Sci.*, vol. 28, no. 4, pp. 188–206, 2025, doi: 10.22401/ANJS.28.4.15.

[36] A. Kuznietsov, B. Gyevnar, C. Wang, S. Peters, and S. V Albrecht, "Explainable AI for safe and trustworthy autonomous driving: A systematic review," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 12, pp. 19342–19364, 2024, doi: 10.1109/TITS.2024.3474469.

[37] R. K. Shinde, K. D. Shinde, and H. Mehta, "A hybrid explainable AI framework for enhancing trust and transparency in autonomous vehicles," in *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2025. doi: 10.1109/ESCI63694.2025.10987990.

[38] K. Malik, M. Sharma, S. Deswal, U. Gupta, D. Agarwal, and Y. O. B. Al Shamsi, *Explainable artificial intelligence for autonomous vehicles: Concepts, challenges, and applications*. CRC Press, 2024. doi: 10.1201/9781003502432.

[39] S. Javaid, M. A. Khan, H. Fahim, B. He, and N. Saeed, "Explainable AI and monocular vision for enhanced UAV navigation in smart cities: Prospects and challenges," *Front. Sustain. Cities*, vol. 7, p. 1561404, 2025, doi: 10.3389/frsc.2025.1561404.

[40] S. S. Malwade and S. J. Budhavale, "Exploring explainable AI: Current trends, challenges, techniques and its applications," in *ACM International Conference Proceeding Series*, 2023, p. 85. doi: 10.1145/3647444.3647912.

[41] M. Kopzhasarova and D. Kozhamzharova, "Explainable AI (XAI): Techniques, applications, and challenges," in *CEUR Workshop Proceedings*, 2025.

[42] A. Apicella, L. Di Lorenzo, F. Isgrò, A. Pollastro, and R. Prevete, "Strategies to exploit XAI to improve classification systems," in *Communications in Computer and Information Science*, Springer, 2023, pp. 147–159. doi: 10.1007/978-3-031-44064-9_9.

[43] G. Türkmen, "The Review of Studies on Explainable Artificial Intelligence in Educational Research," *J. Educ. Comput. Res.*, vol. 63, no. 2, pp. 277–310, 2025, doi: 10.1177/07356331241310915.

[44] D. Danang, S. Siswanto, W. Aryani, and P. Wibowo, "Hybrid Federated Ensemble Learning Approach for Real-Time Distributed DDoS Detection in IIoT Edge Computing Environment," *J. Eng. Electr. Informatics*, vol. 5, no. 1, pp. 9–17, 2025, doi:

10.55606/jeei.v5i1.5099.

[45] D. Danang, A. B. Santoso, and M. U. Dewi, "CICA Framework: Harnessing CSR, AI, and Blockchain for Sustainable Digital Culture," *Int. J. Adv. Comput. Sci. \& Appl.*, vol. 16, no. 11, 2025.

[46] D. Danang, T. Wahyono, I. Sembiring, T. Wellem, and N. H. Dzulkefly, "An Adaptive Framework Integrating ML Blockchain and TEE for Cloud Security," in *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, 2025, pp. 1–7.

[47] D. Danang, H. Haryani, Q. Aini, F. A. Ramahdan, and J. Edwards, "Empowering digital literacy through blockchain based alphasign for secure and sustainable e-governance," 2025.

[48] D. Danang, N. D. Setiawan, and E. Siswanto, "Pemanfaatan Teknologi Internet of Things untuk Monitoring Kualitas Air Sungai di Wilayah Perkotaan," *J. New Trends Sci.*, vol. 2, no. 1, pp. 23–34, 2024.

[49] E. Muhadi, S. Sulartopo, D. Danang, D. Sasmoko, and N. D. Setiawan, "Rancang bangun sistem keamanan ruang persandian menggunakan RFID dan sensor PIR berbasis IOT," *Router J. Tek. Inform. dan Terap.*, vol. 2, no. 1, pp. 8–20, 2024.

[50] M. K. Umam, D. Danang, E. Siswanto, and N. D. Setiawan, "Rancangan Bangun Otomasi Air Suling Daun Cengkeh Berbasis Arduino," *Repeater Publ. Tek. Inform. dan Jar.*, vol. 2, no. 2, pp. 1–10, 2024.