Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Explainable and Risk-Aware AI for Autonomous Cyber Defense and Reliable Incident Response Across U.S. Public and Enterprise Digital Services: The XR-ACD Framework

Muhammad Saad khawar[1], Fawad Mir[2] and Jawad Yaqoob Mir[3] ✉
[123]*Master of Science in Information Technology, Washington University of Science and Technology (WUST), Alexandria, Virginia, US*
Corresponding Author: Jawad Yaqoob Mir, **E-mail**: jawadmir488@gmail.com

| ABSTRACT

The scale, velocity and complexity of cyber threats are growing, and with them there are new challenges for security teams tasked with defending public and enterprise digital services. Artificial intelligence (AI) can speed up the discovery, investigation and response process to threats, but a second risk is often overlooked: autonomous response may disrupt service, impact critical dependencies, or have irreversible effects if the assessment is wrong. The paper presents a conceptual framework for an autonomous cyber defense process, Explainable and Risk-Aware Autonomous Cyber Defense (XR-ACD), which combines explainable threat assessment, explicit action-risk evaluation, risk-proportional autonomy, and continuous governance. While some methods rely mainly on the size of the threat or on the confidence of the model to trigger automatic responses, XR-ACD differentiates between the risk of the threat and the risk of responding to the threat and considers both risks before allowing autonomous action. Action risk is represented with asset criticality, blast radius, dependency impact, operational impact, uncertainty and reversibility. These factors are linked to four response modalities: autonomous execution, bounded or supervised autonomy, human-in-the-loop decision making and human authorization. The framework also includes user-friendly explanations, policy limitations, audit reports, outcome tracking, and feedback-driven risk reassessment. In a municipal water-treatment IT/OT scenario, the same detected threat can trigger various response authorities, depending on the operational consequences of candidate actions. The paper proposes a comparative evaluation methodology that includes threat-severity-only automation, human-led response, XR-ACD without explainability and the complete XR-ACD configuration, for future empirical validation. The criteria for evaluation are response latency, containment effectiveness, inappropriate high-impact action rate, disruption to operations, human intervention, quality of explanation, and trust calibration. The paper provides a structured framework for safer autonomous cyber defense and lays the groundwork for empirical analysis of risk-proportional autonomy in public and enterprise digital services.

| KEYWORDS

Explainable AI; Autonomous Cyber Defense; Incident Response; Risk-Aware AI; Human-AI Collaboration.

| ARTICLE INFORMATION

ACCEPTED: 02 June 2026 **PUBLISHED:** 19 August 2026 **DOI:** 10.32996/fcsai.2026.5.9.19

1. Introduction

Digital services are essential for the functioning of today's government, business, utilities, financial services, healthcare, and other critical service providers. As technology becomes ever more interconnected, including through cloud platforms, operational technology (OT), application programming interfaces (APIs), endpoints and distributed network infrastructure, opportunities for digital transformation have grown along with the potential risk of cyber incidents. Meanwhile, adversaries are increasingly operating at machine speed, with massive amounts of malicious activity that can stress traditional, analyst-centric security operations.

AI and ML have thus emerged as vital parts of cybersecurity over the years. AI-based systems can sift through huge amounts of telemetry, detect abnormal behavior, categorize threats, integrate and correlate diverse evidence, and aid in incident investigation. But in the context of cybersecurity, AI brings a key trade-off between decision support and autonomous actions.

However, an accurate model is not necessarily sufficient to determine whether a specific defensive action is safe to take automatically.

The distinction matters most in high-impact environments. While isolating a compromised workstation might be the right measure in a ransomware attack, disconnecting an entire operational network could be detrimental to a critical service. Likewise, blocking a suspicious external IP address could be fairly reversible, while blocking a shared authentication service may impact hundreds of legitimate users and dependent applications. As a result, the risk of the incident itself and the risk of the proposed response should not be treated as the same.

The latest cybersecurity guidance has shifted towards a focus on risk management, governance, and ongoing response. The NIST Cybersecurity Framework (CSF) 2.0 organizes cybersecurity activities around six functions: Govern, Identify, Protect, Detect, Respond, and Recover, with a clear emphasis on governance as a key cybersecurity function [2]. NIST SP 800-61 Revision 3 also stresses embedding incident response in the enterprise's broader cybersecurity risk management program, instead of focusing on incident response as a standalone technical process [3]. At the same time, the NIST AI Risk Management Framework (NIST AI RMF) lists validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy, and fairness as some of the key features of trustworthy AI [1].

These developments suggest that autonomous cyber defense should not be focused only on detection accuracy or response speed. Autonomous defense must also take into account the suitability of the proposed action for the protected asset, the reversibility of the effects it can generate, how uncertainty should impact the decision, and when human authorization is needed. In this paper, we present a framework called **Explainable and Risk-Aware Autonomous Cyber Defense (XR-ACD)** to address this problem. XR-ACD is based on four principles. First, the system must be able to continuously incorporate heterogeneous cybersecurity evidence. Second, explanations of AI-generated threat assessments should be presented in a manner that is easily understood by humans. Third, all candidate responses need to be subjected to a clear action-risk analysis and not simply based on the threat level. Fourth, the degree of autonomy should be proportional to both threat risk and action risk.
The resulting principle can be summarized as:

$$\text{Safe Autonomy} = f(\text{Threat Risk}, \text{Action Risk}, \text{Uncertainty}, \text{Policy Context})$$

rather than:

$$\text{Autonomy} = f(\text{Threat Severity})$$

This is the key difference in the proposed framework.

1.1 Problem Statement

Many of the current AI-based cybersecurity solutions center on enhancing detection, classification, prediction, or response automation. Explainable AI (XAI) research has been directed towards the "black box" nature of cybersecurity models, and autonomous cyber-defense research has focused on autonomous decision-making and response. But these are typically the subject of separate studies.

XAI can provide explanations for **why the system believes an event is malicious**, but this is not necessarily the reason for a specific response. On the other hand, an autonomous defense agent can choose a response quickly without thoroughly considering the operational impact of the response.

Thus, the resulting research problem is:

How can AI-based cyber defense determine an appropriate level of response autonomy by jointly considering threat risk, action risk, uncertainty, operational context, explainability, and governance constraints?

1.2 Research Questions

The study addresses four research questions:

RQ1: How can explainable AI be integrated into autonomous cyber-defense decision-making so that security operators can understand the evidence underlying threat and response decisions?

RQ2: How can the operational risk of a proposed defensive action be explicitly modeled independently of the severity of the detected threat?

RQ3: What is the relationship between threat risk and action risk, and how can this relationship be used to determine an appropriate level of autonomy while maintaining human oversight for high-impact or uncertain decisions?

RQ4: How can the resulting framework be empirically evaluated with respect to response speed, containment effectiveness, operational disruption, inappropriate autonomous actions, explainability, and human intervention?

1.3 Contributions

The main contributions of this paper are:

1. **An integrated XR-ACD framework** that links threat detection, explainability, action-risk assessment, autonomy selection, execution, monitoring, and governance.
2. **An explicit action-risk model** that takes into account asset criticality, blast radius, dependency impact, operational impact, uncertainty, and reversibility.
3. **A risk-proportional autonomy mechanism** that differentiates autonomous, bounded/supervised, human-in-the-loop, and human-authorized response mechanisms.
4. **A threat-risk/action-risk decision matrix** formalizing the rationale for considering both the threat and the consequences of the proposed action when determining response authority.
5. **A scenario-based demonstration** of the operational implications of risk-aware response in an IT/OT water-treatment environment.
6. **A comparative empirical evaluation methodology** for future testing of the framework against threat-severity-only automation, human-led response, and non-explainable variants.

The paper intentionally presents a framework and validation methodology rather than claiming empirical performance that has not yet been experimentally measured.

2. Background and Related Work

2.1 AI for Cybersecurity and Autonomous Defense

AI has been applied to a wide range of cybersecurity tasks, including intrusion detection, malware identification, anomaly detection, vulnerability assessment, threat intelligence, and security operations. ML and deep learning can process large amounts of high-dimensional and heterogeneous telemetry and identify patterns that may not be obvious through static rules.

Autonomous cyber defense, however, is more complex than threat classification. A defense agent must operate in an environment that is constantly changing, work with incomplete information, choose appropriate actions, and account for the consequences of those actions. Previous work on automated cyber defense distinguishes fixed, hard-coded automation from more autonomous learning-based approaches and highlights the need for realistic training environments to move beyond static playbooks [14]. A more recent systematic review also identifies autonomous response as an emerging research area, with challenges including explainability, continuous adaptation, realistic environments, and real-world deployment [4].

This distinction is important for XR-ACD because the framework is concerned not only with whether a threat can be identified, but also with whether the system should be allowed to act on that assessment and at what level of authority.

2.2 Explainable AI in Cybersecurity

As ML models used in cybersecurity become more complex, the need to understand their outputs also increases. Security operators often need to know why an event has been classified as malicious before deciding whether to investigate, escalate, contain, or remediate it.

Previous research has examined XAI applications in intrusion detection, malware detection, anomaly detection, and other cybersecurity tasks. Charmet et al. (2022) highlight the importance of explainability in helping operators interpret large numbers of security alerts and assess model outputs within operational constraints [5]. Capuano et al. (2022) similarly identify opacity as an important issue in AI-based cybersecurity decisions [6], while Zhang et al. (2022) review the use of XAI methods across different cybersecurity tasks [12].

Several approaches can be used to generate explanations, including local surrogate methods, feature-attribution methods such as SHAP, rule-based explanations, and counterfactual explanations [5]. A broader taxonomy of XAI concepts and methods also covers transparent models, post-hoc explanations, and responsible-AI principles that provide a wider foundation for these

approaches [16]. Explainability and interpretability are also identified by NIST as characteristics of trustworthy AI, including principles for explainable AI [7].

However, explaining why an event is considered a threat is not necessarily the same as explaining why a particular response should be taken.

For autonomous cyber defense, an explanation should therefore help answer at least two questions:

1. Why is this event considered a threat?
2. Why is the proposed response considered appropriate and safe?

XR-ACD treats explanation as part of the decision process rather than simply as a visualization added after the decision has already been made.

2.3 Human-AI Collaboration and Cyber Defense

Human expertise remains important in security operations because cybersecurity decisions often depend on organizational context, mission requirements, business processes, and potential consequences that may not be directly represented in telemetry.

Research on human-AI teaming in cybersecurity highlights the value of combining AI's computational capabilities with human contextual reasoning, particularly in situations where the consequences of a decision may be significant [8].

XR-ACD is therefore not intended to remove humans from incident response. Instead, it proposes **risk-proportional human involvement**. Actions with lower operational risk and greater reversibility may be suitable for autonomous execution, while actions with higher operational consequences or greater uncertainty should require stronger human oversight.

2.4 Cybersecurity Governance and Risk Management

NIST CSF 2.0 provides a common cybersecurity risk-management structure that can be applied across different organizations and technology environments, including IT, IoT, OT, cloud, and AI systems. Its six functions - Govern, Identify, Protect, Detect, Respond, and Recover - emphasize that cybersecurity is both a technical and organizational risk-management activity [2].

Similarly, the NIST AI RMF is organized around the functions Govern, Map, Measure, and Manage and emphasizes continuous risk management throughout the AI lifecycle [1]. CISA's AI roadmap also highlights the need for incident-response capabilities, continuous evaluation, and responsible deployment of AI for cyber defense [9].

At the cyber-defense decision level, XR-ACD applies these governance principles by treating policies, risk tolerances, auditability, human oversight, and outcome monitoring as integral parts of autonomous response.

2.5 Research Gap

The literature reviewed above covers several closely related areas, but they are not fully integrated:

Table 1: Research gap

Research area	Primary strength	Remaining limitation
AI-based threat detection	Rapid identification and classification	Does not determine whether intervention is safe
Explainable AI for cybersecurity	Makes model outputs more understandable	Usually focuses on explaining predictions rather than response authority
Autonomous cyber defense	Enables machine-speed response	Requires stronger controls for uncertainty and operational consequences
Cybersecurity governance	Defines risk and organizational controls	Does not directly specify how AI should select response autonomy
XR-ACD (proposed)	Integrates detection, explanation, action risk, autonomy, and governance	Requires empirical validation

The gap addressed by this paper is therefore not the absence of AI-based detection or autonomous response. Rather, it is the lack of an integrated decision process that explicitly considers the risk of the defensive action itself before autonomous authority is granted.

2.6 IT/OT and ICS Security Context

Because the scenario in Section 9 involves an operational technology (OT) environment, it is important to ground the discussion in existing OT and ICS security guidance rather than assume that IT and OT risks are interchangeable. NIST SP 800-82 Revision 3 provides guidance for securing operational technology and industrial control systems and notes that OT environments have reliability, safety, and availability requirements that can differ from those of traditional IT systems [13]. MITRE also provides a dedicated ATT&CK matrix for ICS that documents adversarial tactics and techniques relevant to industrial control environments [19].

XR-ACD does not attempt to redefine OT risk criteria from first principles. Instead, these existing OT/ICS requirements and security guidance are treated as important inputs to the Mission and Risk Context layer described in Section 4.2.

3. Problem Formulation and Design Objectives

3.1 System Model

Let a cybersecurity environment at time (t) be represented by:

$$E_t = \{T_t, A_t, C_t, P_t, U_t\}$$

where:

- T_t represents observed threat evidence;
- A_t represents affected assets;
- C_t represents operational and dependency context;
- P_t represents organizational policies and risk tolerances;
- U_t represents uncertainty associated with the available evidence and model outputs.

The AI system receives heterogeneous telemetry:

$$X_t = \{L_t, N_t, E_t, I_t, K_t\}$$

where L_t denotes system logs, N_t network telemetry, E_t endpoint telemetry, I_t threat intelligence, and K_t contextual security knowledge.

The system produces a threat assessment:

$$TR_t = f_T(X_t, C_t)$$

where TR_t is threat risk.

Candidate response actions are represented as:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$$

Examples include blocking an external IP address, terminating a malicious process, isolating an endpoint, restricting an account, segmenting a network, or disconnecting an IT/OT connection.

The key design constraint is that the system should not immediately execute the action associated with the highest threat score. Instead, each candidate action a_i is evaluated using an action-risk function:

$$\mathrm{AR}(a_i) = f_R(a_i, C_t, U_t, P_t)$$

The final autonomy decision is:

$$\mathrm{AL}(a_i) = f_A(TR_t, \mathrm{AR}(a_i), U_t, P_t)$$

where AL denotes the permitted autonomy level.

3.2 Design Objectives

XR-ACD is designed around six objectives:

O1 - Explainability. The system should provide human-readable evidence explaining both the threat assessment and the selection of a response.

O2 - Risk Awareness. The system should distinguish the severity of the cyber threat from the potential consequences of the defensive action.

O3 - Proportional Autonomy. The level of autonomy should increase only when the risk associated with the response is acceptable within organizational policy.

O4 - Operational Safety. The framework should explicitly consider asset criticality, dependencies, blast radius, reversibility, uncertainty, and operational impact.

O5 - Human Oversight. High-impact or highly uncertain decisions should remain subject to human review or authorization.

O6 - Continuous Governance. Response outcomes should be monitored and incorporated into subsequent risk assessment and policy refinement.

4. The Proposed XR-ACD Framework

4.1 Framework Overview

XR-ACD consists of five functional layers:

1. Mission and Risk Context
2. Unified Telemetry and Threat Intelligence
3. Explainable AI Engine
4. Risk-Aware Action Planner
5. Autonomous Execution and Governance

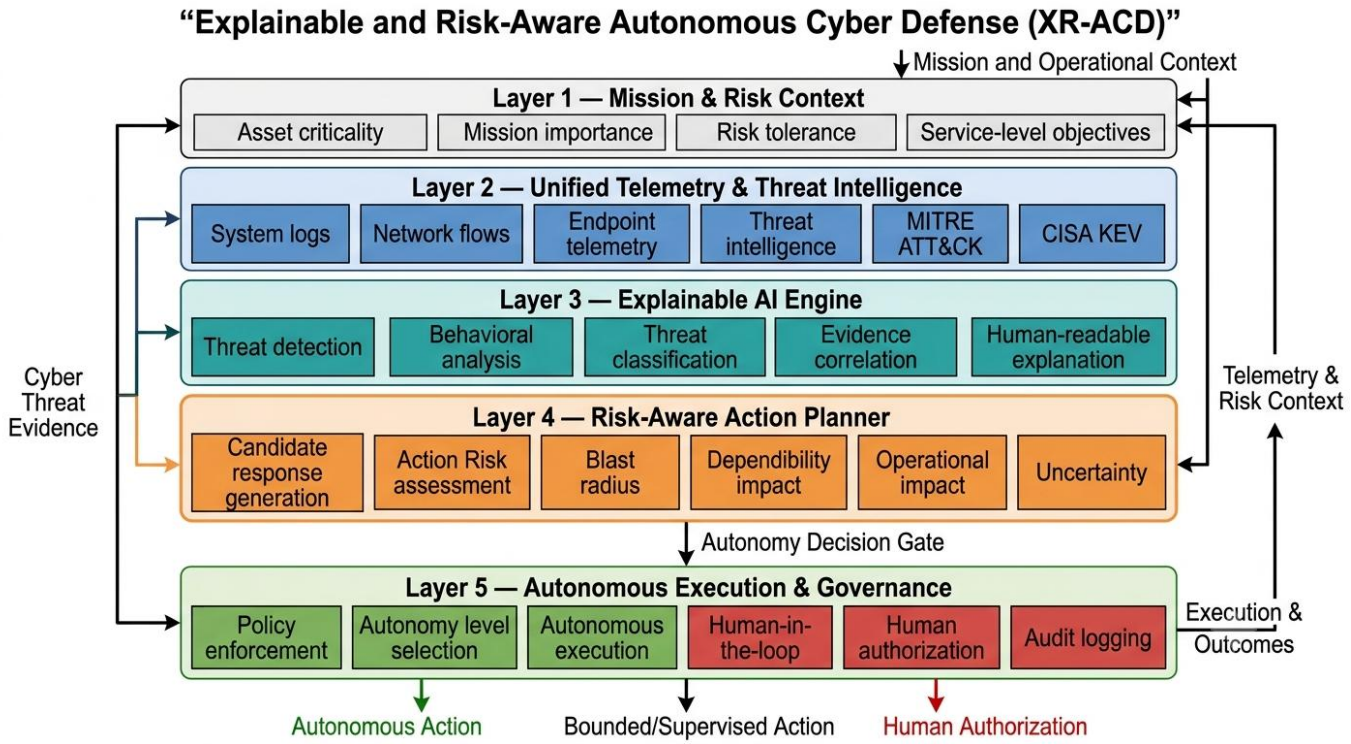

Figure 1. XR-ACD architecture integrating mission context, heterogeneous security telemetry, explainable AI, risk-aware action planning, and governed autonomous execution.

The framework is designed as a closed-loop system rather than a one-way detection pipeline.

4.2 Layer 1 - Mission and Risk Context

The first layer provides the context needed to understand the operational consequences of cyber actions. It includes:

- Asset criticality
- Mission importance
- Organizational risk tolerance
- Service-level objectives
- Regulatory and policy constraints
- Operational dependencies
- Acceptable downtime
- Predefined response authorities

This layer is important because the same technical action can have very different consequences in different environments. For example, isolating an employee workstation may have limited organizational impact, whereas isolating a control-system gateway could interrupt a physical process.

4.3 Layer 2 - Unified Telemetry and Threat Intelligence

XR-ACD aggregates multiple sources of evidence rather than relying on a single detector. Potential inputs include:

- System logs
- Network flows
- Endpoint telemetry
- Authentication events
- Vulnerability information
- Threat intelligence
- MITRE ATT&CK mappings
- CISA Known Exploited Vulnerabilities (KEV)
- Application and service context

MITRE ATT&CK provides a common representation of adversarial tactics, techniques, and procedures based on observed real-world activity [10]. CISA's KEV catalog provides an authoritative source of vulnerabilities known to have been exploited in the wild and can be incorporated into vulnerability prioritization [11].

The purpose of this layer is not to make the final response decision. Its role is to provide sufficiently rich evidence for the subsequent analysis.

4.4 Layer 3 - Explainable AI Engine

The Explainable AI Engine performs four primary functions: (1) threat detection, (2) behavioral analysis, (3) evidence correlation, and (4) human-readable explanation.

An explanation may include triggering events, the most influential indicators, relevant behavioral patterns, associated ATT&CK techniques, confidence or uncertainty, supporting threat-intelligence evidence, and contradictory evidence. For example:

> **Threat assessment**: High-confidence ransomware behavior. **Supporting evidence** includes abnormal encryption activity, unusual outbound traffic, and lateral-movement indicators. **Associated behaviors** include file-encryption and lateral-movement techniques. **Uncertainty**: Moderate because encryption telemetry is incomplete.

This type of explanation gives an analyst a concise evidence trail rather than simply presenting a numerical probability.

The framework does not require a specific XAI algorithm. Depending on the underlying AI model and operational requirements, SHAP, LIME, inherently interpretable models, rule-based explanations, counterfactual reasoning, or combinations of these approaches may be used. Prior cybersecurity XAI research indicates that multiple explanation approaches can be relevant to different security tasks [5].

5. Risk-Aware Action Planning and Autonomy

5.1 Threat Risk

Threat risk estimates the potential severity of the detected cyber incident. A conceptual threat-risk function can be represented as:

$$TR = w_T S + w_L L + w_I I + w_E E$$

where S = threat severity; L = likelihood or confidence; I = potential impact; E = evidence strength; and w_s, w_l, w_i, w_e are configurable weights.

The equation is intentionally conceptual. The weights require empirical calibration and should vary according to organizational context.

5.2 Action Risk

Threat risk alone is insufficient for deciding whether an action should be autonomous. XR-ACD therefore introduces:

$$AR(a) = w_C C + w_B B + w_D D + w_O O + w_U U + w_I I_r$$

where C = asset criticality; B = blast radius; D = dependency impact; O = operational impact; U = uncertainty; I_r = irreversibility; and wc, wB, wD, w_o, w_u, w_i are context-dependent weights.

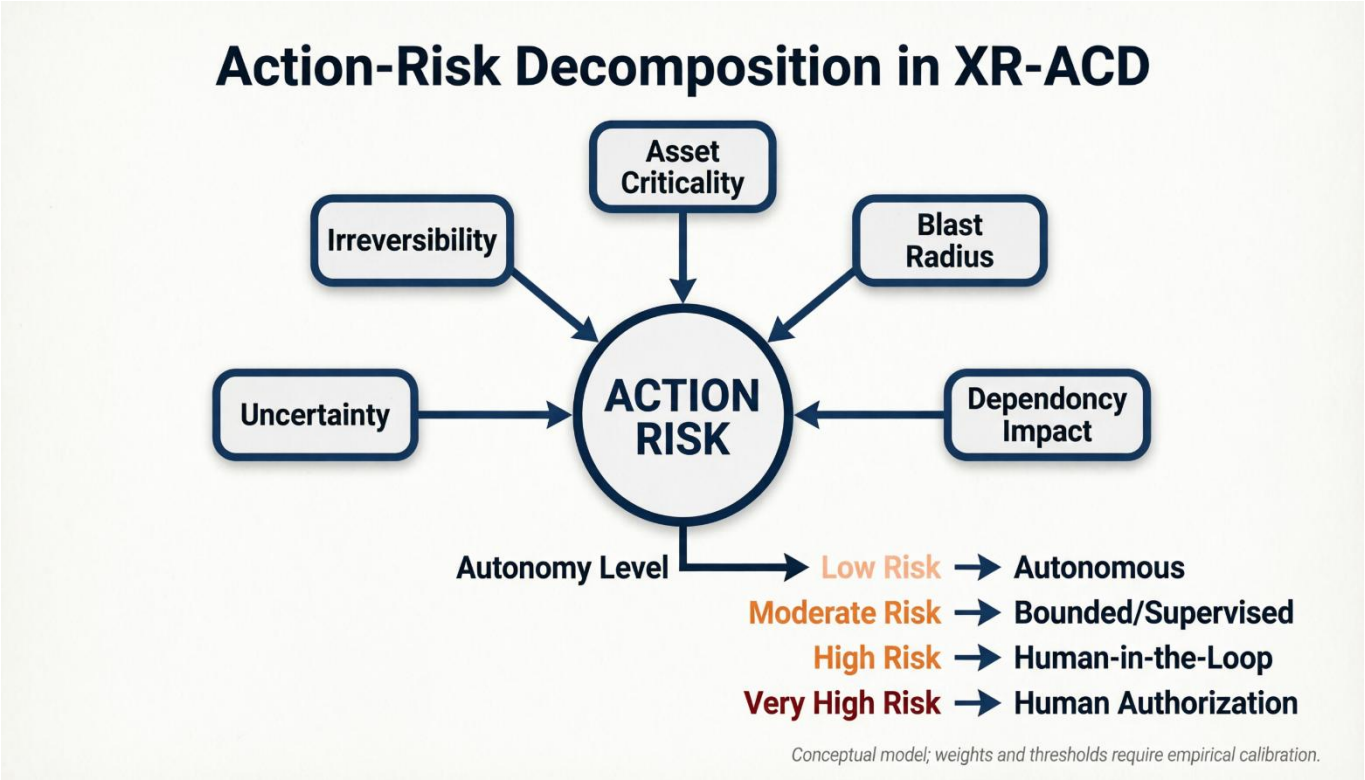

Figure 2. *Action-risk decomposition in XR-ACD. Action risk is determined by the operational consequences and uncertainty associated with a candidate defensive action.*

The key principle is that action risk is associated with the response, not simply with the incident.

5.3 Asset Criticality

Asset criticality represents the importance of the affected asset to organizational or mission objectives. Possible categories include a low-criticality user endpoint, a business application, shared infrastructure, an identity service, a public-facing service, or a critical operational system. Higher asset criticality increases the potential consequence of an incorrect response.

5.4 Blast Radius

Blast radius estimates how many systems, users, services, or operational processes could be affected by an action. For example:

$$B(a) = (\text{potentially affected entities}) / (\text{relevant entities})$$

can provide a normalized conceptual representation. A response affecting one compromised endpoint has a substantially smaller blast radius than a response affecting an entire network segment.

5.5 Dependency Impact

Modern digital services depend on shared authentication systems, databases, APIs, network gateways, DNS, identity providers, cloud services, and other infrastructure. An action can therefore cause indirect failures even when the directly targeted asset is malicious. Dependency impact represents the estimated consequence of disrupting those relationships.

5.6 Operational Impact

Operational impact represents the effect of the response on organizational services and mission activities. This factor becomes particularly important in public services and IT/OT environments, where cybersecurity decisions can affect physical processes or essential services.

5.7 Uncertainty

Uncertainty represents incomplete telemetry, conflicting evidence, model uncertainty, or ambiguity regarding the consequences of an action. The framework treats uncertainty as a risk amplifier. If two actions have otherwise identical consequences, the action supported by stronger evidence should generally receive greater autonomous authority.

5.8 Irreversibility

A reversible action can usually be undone with limited consequence, such as temporary IP blocking, temporary process termination, or short-lived account restrictions. More difficult actions include shutting down critical services, disconnecting operational networks, modifying production configurations, deleting data, or disabling shared infrastructure. Higher irreversibility should therefore increase action risk.

6. Autonomy Decision Model

6.1 Threat Risk × Action Risk

XR-ACD maps threat risk and action risk to a decision matrix.

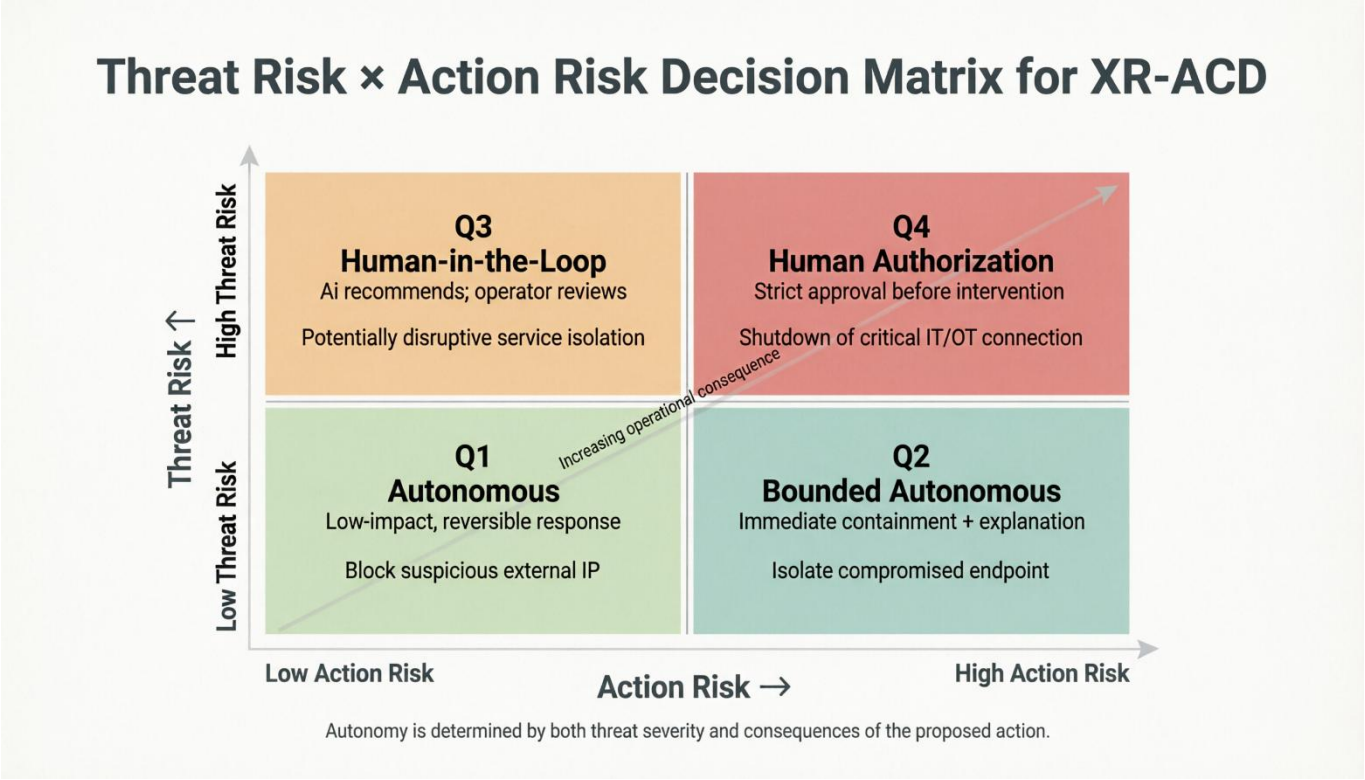

Figure 3. Threat-risk and action-risk decision matrix for selecting risk-proportional autonomy.

The matrix produces four broad response modes.

Q1 - Low Threat Risk / Low Action Risk: Autonomous execution
Suitable for low-impact, reversible actions such as blocking a clearly malicious external address.

Q2 - Low Threat Risk / High Action Risk: Bounded or supervised execution
The system may prepare the action and perform limited containment, but stronger safeguards are required because the action itself has significant consequences.

Q3 - High Threat Risk / Low Action Risk: Human-in-the-loop
The incident is serious, but the response may be relatively safe. AI can rapidly recommend and potentially prepare the response while maintaining human review.

Q4 - High Threat Risk / High Action Risk: Human authorization
The system must not independently execute the action. Human authorization is required before high-consequence intervention.

This produces the fundamental principle:

$$\text{Autonomy Level} = f(\textit{Threat Risk}, \textit{Action Risk})$$

rather than:

$$\text{Autonomy Level} = f(\textit{Threat Risk})$$

6.2 Autonomy Levels

The AL-0 to AL-3 scale below adapts a longer tradition of levels-of-automation models from human factors research, which classify automated systems by how much of the sense-decide-act loop is delegated to the machine versus retained by the human operator [25]. XR-ACD narrows that general scale to the specific decisions relevant to cyber response.

Table 2: Levels of autonomy with example

Level	Mode	Typical characteristics	Example
AL-0	Advisory	AI recommends; human decides	Recommend endpoint isolation
AL-1	Human-Approved	AI prepares action; human authorizes	Disable privileged account
AL-2	Bounded/Supervised	AI executes within predefined limits	Isolate endpoint with rollback
AL-3	Autonomous	Low-risk and reversible action	Block confirmed malicious IP

The exact thresholds are organizational policy parameters rather than universal constants.

7. Risk-Aware Decision Workflow

The complete operational process is shown in Figure 4.

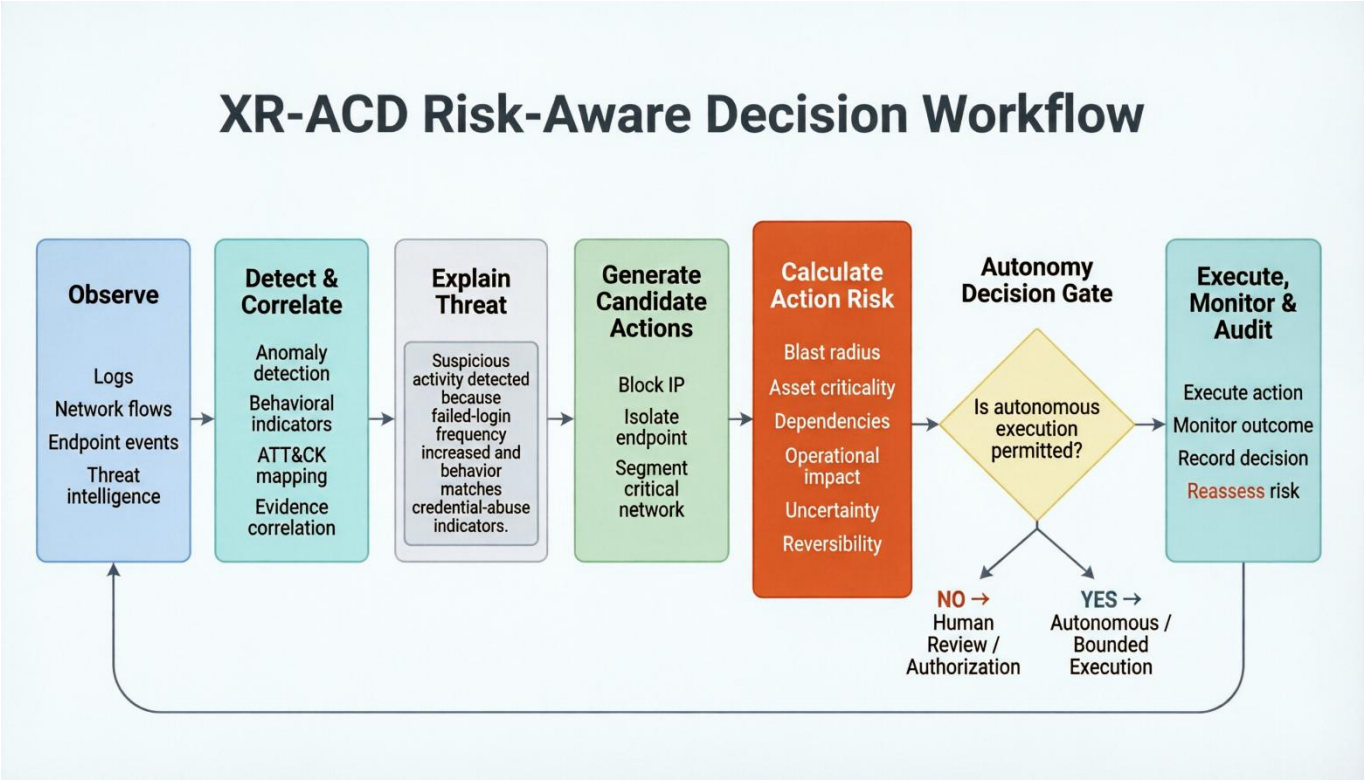

Figure 4. XR-ACD risk-aware decision workflow from observation and detection through explanation, action-risk assessment, autonomy selection, execution, monitoring, and reassessment.

The workflow consists of seven stages.

Stage 1 - Observe. Collect logs, network flows, endpoint events, threat intelligence, and contextual information.

Stage 2 - Detect and Correlate. Identify anomalous behavior and correlate multiple evidence sources.

Stage 3 - Explain Threat. Generate an explanation of why the activity is considered suspicious or malicious.

Stage 4 - Generate Candidate Actions. Construct one or more potential responses.

Stage 5 - Calculate Action Risk. Each candidate action receives an action-risk assessment.

Stage 6 - Autonomy Decision Gate. The system compares threat risk, action risk, uncertainty, and policy constraints.

Stage 7 - Execute, Monitor, and Audit. The selected response is executed according to its permitted autonomy level. The outcome is then monitored.

For example, at Stage 4:

$$\mathcal{A} = \{ a_1 = \text{Block IP},\ a_2 = \text{Isolate endpoint},\ a_3 = \text{Segment network},\ a_4 = \text{Disconnect IT/OT link} \}$$

If the environment changes, the system reassesses the risk. This feedback mechanism is essential because cyber incidents are dynamic. A response that was safe at time t may become unsafe at time t + 1.

8. Governance and Continuous Feedback

Autonomous cyber defense cannot be treated as a static automation pipeline.

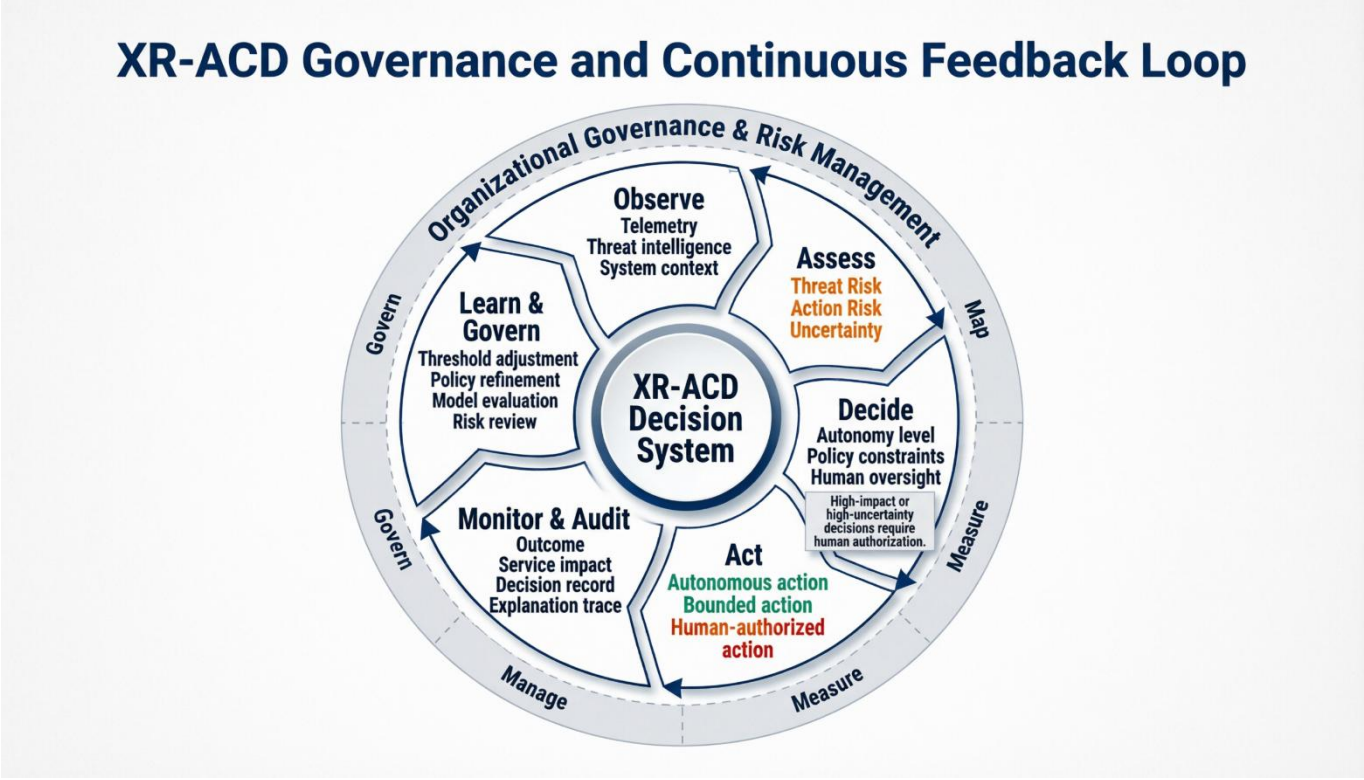

Figure 5. XR-ACD governance and continuous feedback loop.

The governance loop follows:

$$\text{Observe} \rightarrow \text{Assess} \rightarrow \text{Decide} \rightarrow \text{Act} \rightarrow \text{Monitor} \rightarrow \text{Learn} \rightarrow \text{Govern}$$

The governance layer maintains policy constraints, autonomy thresholds, decision records, explanation traces, outcome measurements, human approvals, and service-impact observations. This design aligns with the broader risk-management principle that AI risk management should be continuous throughout the system lifecycle [1].

8.1 Policy Refinement

Repeated incidents can reveal that certain actions are consistently safe or unsafe. For example, an organization may determine that blocking externally originated connections associated with high-confidence malicious infrastructure is sufficiently low-risk for autonomous execution. Conversely, modifying routing rules on a critical operational network may always require human authorization. Policies can therefore evolve while remaining subject to governance.

8.2 Auditability

Every significant response should generate an auditable decision record containing:

1. Incident identifier
2. Evidence used
3. Threat assessment
4. Explanation
5. Candidate actions
6. Action-risk assessment
7. Selected autonomy level
8. Policy constraint
9. Execution result
10. Post-action outcome

This provides a basis for incident review, accountability, model evaluation, and future calibration.

9. Scenario Demonstration – Municipal Water-Treatment IT/OT Environment

For illustration, consider a municipal water-treatment environment with an enterprise IT network and an operational technology (OT) network.

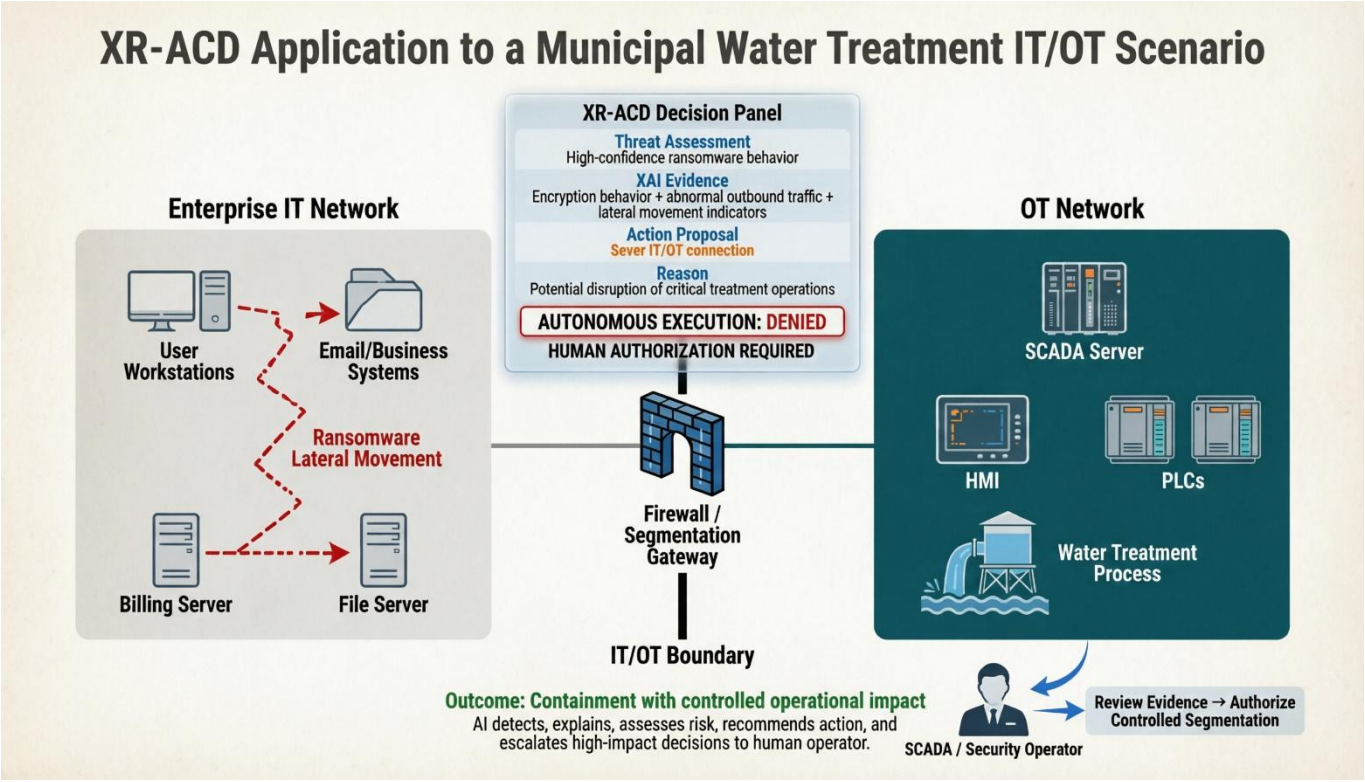

Figure 6. *Application of XR-ACD to a municipal water-treatment IT/OT environment.*

The enterprise environment contains user workstations, business systems, billing services, and file servers. The OT environment includes SCADA servers, human-machine interfaces (HMIs), programmable logic controllers (PLCs), and the physical water-treatment process.

Suppose that XR-ACD detects activity consistent with ransomware and lateral movement within the enterprise network. The evidence includes abnormal file-encryption activity, unusual outbound traffic, lateral-movement indicators, and suspicious authentication activity. The Explainable AI Engine produces a high-confidence threat assessment. However, XR-ACD does not immediately execute the strongest available response. Instead, it generates multiple candidate actions.

9.1 Candidate Action 1: Block Suspicious External IP

The first response is to block an external IP address associated with the suspicious activity. The action has a limited blast radius, low dependency impact, high reversibility, and relatively low operational impact. Therefore, $AR(a_1) \rightarrow$ Low. If the threat evidence is sufficiently strong, autonomous execution may be permitted.

9.2 Candidate Action 2: Isolate Compromised Endpoint

The second response isolates a workstation exhibiting malicious behavior. The action has greater operational impact because the workstation becomes unavailable to its user, but the blast radius remains relatively small. Therefore, $AR(a_2) \rightarrow$ Moderate. The framework may assign bounded or supervised autonomy.

9.3 Candidate Action 3: Sever the IT/OT Connection

The third response is to sever the connection between the enterprise IT network and the OT network. This may reduce the possibility of lateral propagation, but it may also affect critical treatment operations. The action has high asset criticality, potentially large operational impact, high dependency impact, substantial irreversibility during an active process, and significant consequences if incorrectly executed. Therefore, $AR(a_3) \rightarrow$ High.

Even though the threat is severe, XR-ACD does not automatically execute this action. Instead, $AL(a_3)$ = Human Authorization. The system provides the operator with the threat evidence, explanation, action-risk assessment, expected operational consequences, recommended response, and the reason autonomous execution is denied. The operator can then authorize a controlled segmentation action if appropriate. This example illustrates the central contribution of XR-ACD:

> ***High threat risk does not automatically justify high autonomous authority.***

10. Proposed Empirical Validation Methodology

The current paper does not claim experimental performance results. Instead, it defines an empirical methodology through which XR-ACD can be evaluated.

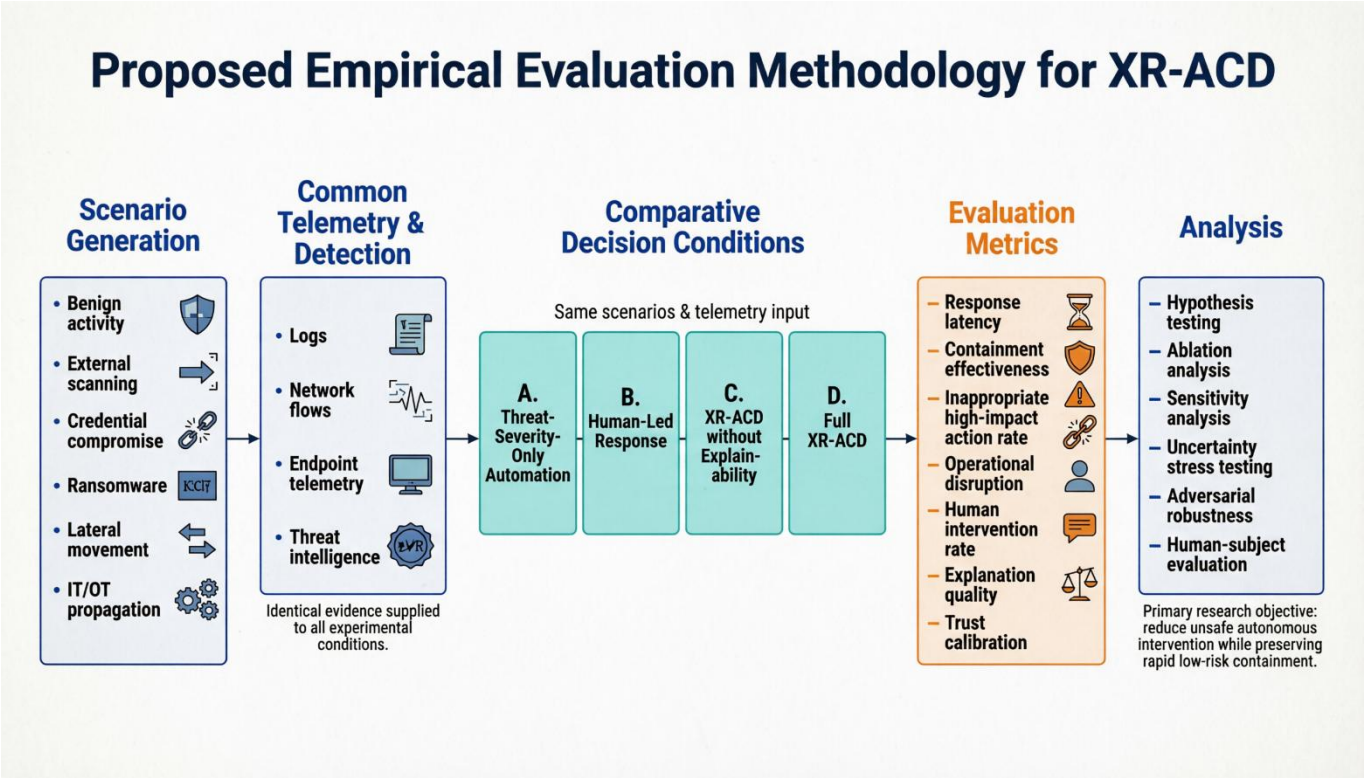

Figure 7. *Proposed empirical evaluation methodology for XR-ACD.*

The methodology uses identical scenarios and telemetry inputs across four experimental conditions.

10.1 Experimental Conditions

Condition A – Threat-Severity-Only Automation. The system selects responses primarily from threat severity or confidence without explicit action-risk assessment.

Condition B – Human-Led Response. AI assists with detection, but response decisions remain primarily manual.

Condition C – XR-ACD Without Explainability. The risk-aware autonomy mechanism is enabled, but human-readable explanations are removed.

Condition D – Full XR-ACD. The complete framework is enabled: threat assessment, explainability, action-risk assessment, autonomy selection, policy constraints, human oversight, and monitoring.

This design allows the individual contributions of risk awareness and explainability to be investigated.

10.2 Scenario Classes

Future experiments should include representative scenarios such as benign activity, external scanning, credential compromise, ransomware, lateral movement, privilege escalation, and IT/OT propagation. Each scenario should be represented using identical telemetry across experimental conditions. Existing autonomous cyber-defense research environments, such as the CybORG simulation and emulation gym for training and evaluating autonomous blue-team and red-team agents, offer a practical starting point for implementing these scenario classes without building a bespoke range from scratch [15].

10.3 Evaluation Metrics

The most important metric for XR-ACD is not simply detection accuracy.

Table 3: Evaluation metrics

Metric	Purpose
Response latency	Measures speed of intervention
Containment effectiveness	Measures reduction in attack propagation
Inappropriate high-impact action rate	Measures unsafe autonomous behavior
Operational disruption	Measures collateral service impact
Human intervention rate	Measures dependence on human oversight
Explanation quality	Measures usefulness of AI reasoning
Trust calibration	Measures whether operator trust corresponds to system reliability
Risk calibration	Measures whether predicted risk corresponds to observed consequences

A system could achieve excellent detection accuracy while still being unsafe if it frequently executes inappropriate high-impact actions. Therefore, the central safety metric should be:

$$IHR = (\text{inappropriate high-impact autonomous actions}) / (\text{total autonomous actions})$$

where IHR denotes the inappropriate high-impact response rate. A successful risk-aware system should aim to reduce IHR without producing unacceptable degradation in containment speed.

10.4 Hypotheses

The proposed methodology enables the following hypotheses:

- **H1 – Response Safety.** XR-ACD will produce a lower inappropriate high-impact action rate than threat-severity-only automation.
- **H2 – Operational Impact.** XR-ACD will reduce unnecessary operational disruption compared with unrestricted autonomous response.
- **H3 – Response Efficiency.** XR-ACD will maintain lower response latency than fully human-led response for low-risk incidents.
- **H4 – Explainability.** The full XR-ACD configuration will produce better explanation quality and operator understanding than the non-explainable configuration.

- **H5 – Risk Calibration.** Action-risk scores will correlate with observed operational consequences more strongly than threat severity alone.

These hypotheses should be tested experimentally rather than treated as established findings in the current paper.

11. Discussion

One of the main arguments of XR-ACD is that cyber-defense decisions involve two separate questions: **How dangerous is the incident?** and **How dangerous is the proposed response?** These two questions do not always lead to the same decision. A severe ransomware incident may justify immediate containment, but that does not mean every available containment action should receive the same level of autonomy. The distinction matters most where IT systems interact with physical processes or essential public services, since an incorrect response can cause significant damage. XR-ACD treats explainability as more than a usability feature. An explanation can serve as an intermediate control between AI inference and autonomous intervention:

$$\text{Evidence} \rightarrow \text{Threat Assessment} \rightarrow \text{Explanation} \rightarrow \text{Candidate Action} \rightarrow \text{Action Risk} \rightarrow \text{Autonomy}$$

This provides multiple opportunities for human verification. A security operator can challenge the evidence, the threat interpretation, the candidate response, the action-risk estimate, or the permitted autonomy level. Such a design can reduce the likelihood that an opaque model directly controls high-impact infrastructure. XR-ACD deliberately avoids both extremes. Fully manual response may be too slow to keep up with machine-speed threats, while fully autonomous response can become unsafe when actions carry large or irreversible consequences. The framework instead proposes a sliding scale:

$$\text{More Consequence} \Rightarrow \text{More Oversight}$$
$$\text{More Uncertainty} \Rightarrow \text{Less Autonomy}$$

This creates a continuum between automation and human control. The same framework can be applied across different organizational environments because the risk model is context-dependent. For enterprise services, action-risk factors may include business-service availability, user productivity, customer-facing applications, identity infrastructure, and financial systems. For public or critical services, additional factors may include public safety, service continuity, physical processes, emergency operations, and public infrastructure dependencies.

The framework does not require identical thresholds across these environments. Instead,

$$\text{Policy}(\text{public}) \neq \text{Policy}(\text{enterprise})$$

may be appropriate even when
$\text{Framework}(\text{public}) = \text{Framework}(\text{enterprise})$.

This allows the architecture to remain consistent while organizational risk tolerances differ.

12. Limitations and Future Research

The present work has several limitations. First, XR-ACD is currently a conceptual framework rather than a fully validated operational system. The mathematical formulations define the structure of the risk model but do not establish universally optimal weights or thresholds. Second, action risk is context-dependent. Asset criticality, service dependencies, and acceptable operational disruption vary between organizations. Therefore, a single global action-risk threshold would be inappropriate. Third, explainability does not automatically guarantee correctness. An explanation may be incomplete, misleading, or vulnerable to manipulation. The security of the explanation mechanism itself must therefore be considered. Fourth, cybersecurity environments are adversarial. Attackers may deliberately manipulate telemetry, exploit model weaknesses, or attempt to induce unsafe defensive responses. Such vulnerabilities have been demonstrated for deep learning models, including in the foundational work of Papernot et al. [18], and have since been addressed through detailed attack-and-mitigation taxonomies such as those developed by NIST [17]. Robustness against adversarial inputs and explanation manipulation should therefore form part of future evaluation. Fifth, autonomous cyber defense requires realistic experimental environments. Laboratory datasets may not adequately represent operational dependencies, incomplete telemetry, changing adversary behavior, or real service consequences.

Future work should therefore focus on five areas: (1) empirical calibration of threat-risk and action-risk weights; (2) development of reproducible cyber-range experiments; (3) comparison of XAI methods for response justification; (4) human-subject evaluation of operator trust and decision quality; and (5) adversarial testing of the complete decision pipeline. Recent autonomous cyber-defense research similarly identifies realistic environments, continuous adaptation, and explainability as important challenges for practical deployment [4].

13. Conclusion

This paper presented XR-ACD, an explainable and risk-aware framework for autonomous cyber defense and incident response in both public and enterprise digital-service environments. The core problem it addresses is straightforward but easy to overlook in practice: the severity of a detected threat does not, by itself, determine whether a particular defensive action is safe to execute autonomously. XR-ACD therefore treats threat risk and action risk as separate quantities, with the appropriate level of autonomy determined by both rather than by either one alone. The framework brings together heterogeneous telemetry, threat intelligence, explainable AI, candidate response generation, action-risk assessment, autonomy selection, policy enforcement, human oversight, audit logging, and continuous outcome monitoring. The municipal water-treatment scenario used throughout the paper illustrates how this principle can work in practice: low-impact and reversible actions may be contained autonomously and quickly, while actions that could disrupt critical IT/OT operations remain subject to human authorization. The paper does not claim that XR-ACD has already outperformed existing approaches in an experiment. It has not yet been experimentally validated. Instead, the paper provides a structured methodology for future testing, covering response latency, containment effectiveness, inappropriate autonomous actions, operational disruption, human intervention, explanation quality, and risk calibration. More broadly, reliable autonomous cyber defense should not simply mean an AI system that can act without a human in the loop. A more useful definition of reliable autonomy is the ability to act independently when the likely consequences are acceptable, reduce autonomy as risk increases, and transfer the decision to a human once uncertainty or operational impact crosses a predefined threshold.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] 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., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[2] Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. IEEE Access, 10, 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

[3] Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial intelligence for cybersecurity: A literature survey. Annals of Telecommunications, 77, 789–812. https://doi.org/10.1007/s12243-022-00926-7

[4] Cybersecurity and Infrastructure Security Agency. (2023). 2023–2024 CISA roadmap for artificial intelligence. U.S. Department of Homeland Security.

[5] Cybersecurity and Infrastructure Security Agency. (n.d.). Known exploited vulnerabilities catalog. U.S. Department of Homeland Security. Retrieved August 9, 2026, from https://www.cisa.gov/known-exploited-vulnerabilities-catalog

[6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

[7] Executive Order 14028. (2021). Improving the Nation's Cybersecurity. Federal Register, 86(93), 26633–26647.

[8] Foley, M., Highnam, K., Hicks, C., & Mavroudis, V. (2022). Autonomous network defence using reinforcement learning. In Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22) (pp. 1252–1254). https://doi.org/10.1145/3488932.3527286

[9] Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) program. AI Magazine, 40(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[10] International Electrotechnical Commission. (2013). Industrial communication networks - Network and system security - Part 3-3: System security requirements and security levels (IEC 62443-3-3:2013).

[11] Joint Task Force. (2020). Security and privacy controls for information systems and organizations (NIST Special Publication 800-53, Rev. 5). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-53r5

[12] Kaloroumakis, P. E., & Smith, M. J. (2021). Toward a knowledge graph of cybersecurity countermeasures (Technical Report). The MITRE Corporation. https://d3fend.mitre.org/resources/D3FEND.pdf

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

[14] 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

[15] MITRE. (n.d.). ATT&CK for industrial control systems. Retrieved August 9, 2026, from https://attack.mitre.org/matrices/ics/

[16] MITRE. (n.d.). MITRE ATT&CK. Retrieved August 9, 2026, from https://attack.mitre.org/
[17] National Institute of Standards and Technology. (2012). Guide for conducting risk assessments (NIST Special Publication 800-30, Rev. 1). https://doi.org/10.6028/NIST.SP.800-30r1
[18] Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2025). Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile (NIST Special Publication 800-61 Rev. 3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r3
[19] Nguyen, T. T., & Reddi, V. J. (2023). Deep reinforcement learning for cyber security. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3779–3795. https://doi.org/10.1109/TNNLS.2021.3121870
[20] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In 2016 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 372–387). IEEE. https://doi.org/10.1109/EuroSP.2016.36
[21] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354
[22] Pascoe, C., Quinn, S., & Scarfone, K. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST Cybersecurity White Paper 29). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29
[23] Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K. K., Broniatowski, D. A., & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence (NISTIR 8312). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312
[24] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
[25] Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207
[26] Sarker, I. H., Janicke, H., Mohammad, N., Watters, P., & Nepal, S. (2023). AI potentiality and awareness: A position paper from the perspective of human-AI teaming in cybersecurity. arXiv. https://arxiv.org/abs/2310.12162
[27] Standen, M., Lucas, M., Bowman, D., Richer, T. J., Kim, J., & Marriott, D. (2021). CybORG: A gym for the development of autonomous cyber agents. arXiv. https://doi.org/10.48550/arXiv.2108.09118
[28] Stouffer, K., Zimmerman, T., Tang, C., Lubell, J., Cichonski, J., & McCarthy, J. (2023). Guide to operational technology (OT) security (NIST Special Publication 800-82, Rev. 3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r3
[29] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
[30] Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2025). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025
[31] Vyas, S., Hannay, J., Bolton, A., & Burnap, P. (2023). Automated cyber defence: A review. arXiv. https://doi.org/10.48550/arXiv.2303.04926
[32] Vyas, S., Mavroudis, V., & Burnap, P. (2025). Towards the deployment of realistic autonomous cyber network defence: A systematic review. ACM Computing Surveys, 58(1), Article 5, 1–36. https://doi.org/10.1145/3729213
[33] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
[34] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10, 93104–93139.