Journal of Artificial Intelligence and Digital Economy
Homepage : https://journal.antispublisher.id/index.php/JAIDE
Email : admin@antispublisher.com

e-ISSN : 3032-1077
JAIDE, Vol. 1, No. 7, July 2024
Page 105-125

Trust-Calibrated Explainable Multi-Agent AI for Safe Predictive Cyber Defence in Critical Infrastructure and Cloud-Native Systems

Sajidul Haque Chowdhruy[1], Shakila Akter[2], Md. Golam Mostafa[3]
[1]Washington University of Science and Technology, United States
[2]Lewis University, United States
[3]National University, Bangladesh

DOI : https://doi.org/10.61796/jaide.v1i7.1897

Sections Info

Article history:
Submitted: April 23, 2024
Final Revised: May 11, 2024
Accepted: June 04, 2024
Published: July 20, 2024

Keywords:
Autonomous Cyber Defence
Multi-Agent Systems
Explainable Artificial Intelligence
Trust Calibration
Critical Infrastructure
Cloud-Native Security
Response Risk

ABSTRACT

Objective: *This article proposes TCX-MAD, a trust-calibrated explainable multi-agent defence framework for critical infrastructure and cloud-native systems that places a safety governor between collaborative detection and response execution.* **Method:** *A design-science methodology specifies the architecture, formal decision policy, threat model, public-dataset evaluation plan, and safety-centred metrics.* **Results:** *The framework reports analytical safety properties and an illustrative decision trace rather than fabricated benchmark results because no experimental observations were supplied. Its primary evaluation target is unsafe automated actions prevented without materially increasing valid response latency, supported by detection, disruption, rollback, explanation, and human-approval metrics.* **Novelty:** *TCX-MAD integrates separate threat and response-risk estimation, tiered autonomy, dynamic manipulation-aware agent trust, an explanation-sufficiency gate, and rollback-bounded execution. It reframes autonomous cyber defence as constrained and accountable action selection under dual uncertainty.*

INTRODUCTION

Critical infrastructure operators and cloud-native enterprises face an asymmetry that complicates automation. Attackers can move at machine speed, but defenders must preserve safety, availability, regulatory obligations, and mission continuity while responding. The NIST Cybersecurity Framework (CSF) 2.0 places cybersecurity response within an organization-wide risk-management and governance context [1]. A technically correct detection can still trigger an operationally harmful response: isolating a safety controller may remove visibility, rotating a production identity may interrupt emergency workloads, or scaling down a compromised service may cascade across dependent applications. Consequently, autonomous cyber defence cannot be assessed solely by detection accuracy or mean response time. It must also be assessed by whether the system declines, defers, or safely constrains actions whose expected operational harm exceeds their security benefit.

Multi-agent architectures are attractive because heterogeneous agents can specialize in network flows, identity activity, endpoint behaviour, industrial processes, and cloud workloads. Autonomous cyber-defence reference architectures already distribute sensing, planning, collaboration, execution, and learning across agent functions [2]. Collaboration can improve coverage and distribute reasoning, but it also creates new failure modes. Agents may be miscalibrated, mutually correlated, compromised through poisoned telemetry or malicious inputs, or confidently wrong in the same direction. A

majority vote therefore does not necessarily constitute reliable evidence. In high-consequence environments, the authority to act should depend not only on consensus, but also on the provenance and independence of evidence, the historical reliability of participating agents, the reversibility of the action, and the asset's mission criticality.

The National Institute of Standards and Technology (NIST) AI Risk Management Framework treats validity and reliability, safety, security and resilience, accountability and transparency, and explainability and interpretability as interdependent characteristics of trustworthy AI [3]. In parallel, CSF 2.0 emphasizes cybersecurity governance and the integration of response and recovery into broader risk management [1]. Together, these frameworks strengthen the case for an execution-control layer that translates trustworthiness and cybersecurity-risk principles into measurable pre-action constraints.

This article develops TCX-MAD (Trust-Calibrated Explainable Multi-Agent Defence), a design-science framework for safe predictive cyber defence. It does not claim to be the first multi-agent, explainable, or autonomous cyber-defence architecture. Its narrower contribution is the integration of four controls before response execution: dual risk scoring, tiered autonomy, dynamic agent trust management, and an explanation-sufficiency gate. The intended contribution lies at the intersection of AI-driven cybersecurity, safe and explainable autonomous systems, and critical-infrastructure/cloud-native operations.

The remainder of the article reviews the relevant literature, identifies a bounded research gap, formalizes objectives and research questions, presents the architecture and evaluation protocol, derives analytical results and testable propositions, and discusses implications, limitations, and future empirical work.

The cyber-defence literature increasingly distributes sensing, classification, planning, and response across collaborating agents. The AICA reference architecture formalizes autonomous sensing, planning, collaboration, action execution, and learning functions for cyber-defence agents [2]. CybORG provides simulation and emulation interfaces for training and evaluating autonomous cyber agents [4], while CAGE Challenge 4 extends the experimental setting to a multi-agent reinforcement-learning enterprise scenario [5]. Collectively, these studies demonstrate the feasibility of autonomous and collaborative defensive action, but they primarily emphasize agent capability, policy learning, and environment reward rather than a unified pre-execution safety contract.

Concrete autonomous-cyber-defence platforms illustrate both the progress and the remaining governance gap. CybORG and the CAGE challenges make agent action selection experimentally tractable [4], [5], but their principal abstraction remains the learning or evaluation of defensive policies inside a cyber environment. They do not, by themselves, impose TCX-MAD's combined pre-execution contract of dynamic contextual trust, separately calibrated response risk, machine-verifiable explanation sufficiency, tiered authority, and mandatory rollback controls. This distinction positions TCX-MAD

as a governance layer that can sit above heterogeneous detection and planning agents rather than as a replacement for existing cyber ranges or learning algorithms.

In cloud-native systems, orchestration APIs make defensive actions fast and programmable. Quarantining a pod, revoking a token, applying a network policy, or rolling back an image can be executed within seconds. Yet the same control-plane power increases blast radius. A response agent with excessive privileges or incomplete dependency knowledge can transform a localized compromise into a service outage. NIST incident-response guidance emphasizes that incident handling requires planned capabilities, explicit response procedures, and restoration of services rather than isolated technical containment [6]. TCX-MAD follows this principle by evaluating action risk against mission context before execution.

Explainable AI in cybersecurity has commonly focused on interpreting detection models: identifying influential features, highlighting anomalous events, or explaining why traffic was labelled malicious. Surveys of XAI in cybersecurity report rapid growth in explanation methods for intrusion detection, malware analysis, phishing, and related security tasks, while also identifying persistent challenges in evaluation, usability, and analyst trust [7], [8]. Feature-attribution methods such as LIME [9] and SHAP [10] are widely used to explain individual predictions, but local fidelity or contribution scores do not by themselves justify an operational response action.

For response governance, an explanation must bridge model evidence and operational consequence. NIST's explainable-AI principles emphasize explanation provision, meaningfulness, explanation accuracy, and knowledge limits [11], while broader XAI evaluation research stresses that explanation quality must be assessed against users, tasks, fidelity, and measurable outcomes [12], [13], [14], [15]. Analysts therefore need to know which evidence was accepted, which asset and dependencies are affected, why a specific action dominates alternatives, what uncertainty remains, and how the action can be reversed. TCX-MAD treats explanation quality as an executable policy condition rather than a user-interface accessory: the explanation gate can prevent an otherwise permitted action when required fields are absent, provenance is weak, or predicted operational impact is not stated.

Decades of human-factors research show that trustworthy automation requires appropriate reliance rather than maximal reliance. Misuse occurs when users over-trust automation; disuse occurs when they reject useful assistance; and abuse can arise when systems are deployed without adequate regard for human performance [16]. Trust should therefore be calibrated to the automation system's capabilities and operating context [17]. Levels-of-automation research likewise supports adjustable allocation of functions between humans and machines rather than a single fixed degree of autonomy [18].

These insights are directly relevant to security operations. Permanent human approval for every event is too slow and can produce alert fatigue, but unconditional autonomy is unsafe for consequential actions. Tiered autonomy offers a middle path: low-confidence or low-severity events are observed; defensible cases are recommended;

tightly scoped, reversible actions may be executed under policy; and irreversible or high-impact actions require authorization. Crucially, human approval is not treated as a universal safety guarantee. The system must present an intelligible evidence and impact summary so that the approver can detect uncertainty, disagreement, and possible manipulation [16], [17], [18].

Interacting agents enlarge the attack surface. An adversary may poison telemetry, forge tool output, compromise one agent, induce correlated errors through a shared model or training set, manipulate inter-agent messages, replay stale evidence, or exploit excessive permissions. Autonomous-agent architectures therefore require controlled collaboration, bounded action authority, protected communications, and explicit separation between sensing, reasoning, and execution functions [2]. Formal trust frameworks such as subjective logic also show why uncertainty about evidence sources should be represented explicitly rather than collapsed into an unqualified consensus score [19]. Secure coordination in TCX-MAD consequently requires authenticated messaging, least privilege, provenance, replay protection, policy isolation, and monitoring of behavioural drift.

Reputation or trust scoring is established in distributed systems, but a cyber-defence trust score should not collapse into a simple historical accuracy average. Subjective-logic approaches demonstrate that belief, disbelief, and uncertainty can be represented separately when evidence is incomplete [19]. In cyber defence, accuracy may be high on common events and poor on rare, safety-critical cases; agreement may reflect correlated failure; and a previously reliable agent may become compromised. TCX-MAD consequently combines proper scoring, calibration, consistency with independent evidence, provenance integrity, recency, and adversarial indicators. Trust modifies evidential weight and permissible authority but never overrides hard safety constraints.

Public intrusion datasets support reproducible comparison but are primarily designed to evaluate detection. CICIDS2017 supplies labelled network flows and common attack profiles [20]. TON_IoT integrates network, operating-system, and IoT/IIoT telemetry [21]. Edge-IIoTset provides a realistic IoT/IIoT dataset designed for centralized and federated learning [22]. The HAI datasets use a hardware-in-the-loop industrial-control testbed and support time-series anomaly evaluation [23]. Together, these datasets provide complementary evidence for detection performance across enterprise-network, IoT/IIoT, and industrial-control settings, but they do not directly represent the operational consequences of defensive actions.

These resources do not normally label whether a candidate defensive action would be operationally unsafe. Response-safety evaluation therefore requires a separate, transparent action-impact layer: an executable testbed, digital twin, fault-injection environment, or expert-validated scenario catalogue that maps actions to availability, safety, and recovery consequences. Combining public detection datasets with an openly specified response oracle is essential; otherwise, unsafe-action claims are not reproducible. This detection-response divide is the principal literature gap addressed by TCX-MAD: existing studies provide strong foundations for autonomous agents, XAI,

calibrated human reliance, trust reasoning, and benchmark detection, but these elements are rarely combined as enforceable pre-action conditions.

Existing autonomous cyber-defence frameworks primarily emphasize detection accuracy and response speed, but provide limited mechanisms for jointly evaluating threat severity, response-induced operational risk, agent trustworthiness, and explanation quality before executing defensive actions in critical infrastructure and cloud-native systems.

The gap is not the absence of multi-agent detection, autonomous response, explainability, human oversight, or secure coordination as individual topics. Rather, it is the absence of an integrated and measurable pre-execution contract linking them. Three practical weaknesses follow. First, response selection is often optimized as if containment had no mission cost. Second, agent agreement is often accepted without accounting for calibration, evidence dependence, or compromise indicators. Third, explanations are typically evaluated after a model decision and are rarely enforced as a condition for action authority.

A defensible contribution must therefore be evaluated on a safety–latency frontier: how many unsafe automated actions are prevented, and at what cost to the speed and success of valid incident response? This reframing makes response safety the primary outcome and predictive accuracy a necessary but insufficient condition.

Table 1. Comparison of representative prior approaches and TCX-MAD

Approach	Multi-agent	Dynamic trust	Response risk	Explanation gate	Tiered autonomy	Rollback
XAI for cyber detection [7], [8], [9], [10]	Usually no	No	No	Post-hoc explanation; not an authority gate	No	No
Levels of automation and calibrated reliance [16], [17], [18]	Conceptual	Human trust calibration	Conceptual consequence awareness	No machine-verifiable gate	Yes, conceptual	Not cyber-action specific
TCX-MAD (proposed)	Yes	Yes; contextual, time-varying, manipulation-aware	Yes; calibrated R(u) separate from T	Yes; completeness and fidelity required before A2	Yes; A0–A3	Mandatory TTL, health checks, and tested rollback for A2

The principal objective is to design and evaluate a trust-calibrated, explainable multi-agent framework that prevents harmful autonomous defensive actions while preserving timely containment of genuine incidents.

1. Develop separate, calibrated estimators for threat risk and response-induced operational risk.
2. Define a tiered-autonomy policy that binds action authority to dual risk, asset criticality, reversibility, agent trust, and explanation sufficiency.

3. Design a dynamic agent-trust mechanism resilient to drift, correlated error, unreliable provenance, and manipulation.
4. Specify action-level explanations that are faithful, complete, operationally meaningful, and suitable for human approval.
5. Evaluate the framework using public telemetry datasets plus reproducible cloud/ICS response scenarios, with unsafe-action prevention as the primary outcome.

The study addresses four research questions:

RQ1: Does joint threat–response risk assessment reduce unsafe automated actions relative to threat-only response policies?

RQ2: Can dynamic trust weighting maintain detection utility when one or more agents drift, become unreliable, or are adversarially manipulated?

RQ3: How much additional valid-response latency and human-approval workload is introduced by the safety governor?

RQ4: Do structured action explanations improve analyst decision quality and calibration compared with feature-attribution-only explanations?

RQ4 requires a future controlled human-subject study; the present paper specifies the study design but reports no analyst-participant data.

RESEARCH METHODS

Research design

The study follows design-science research: problem identification; objective definition; artefact design; demonstration in controlled IT, OT, and cloud-native scenarios; evaluation against baselines; and communication of design knowledge. The artefact is not a single detection model. It is a model-agnostic governance layer that accepts calibrated outputs and signed evidence from specialist agents, estimates two kinds of risk, selects an autonomy tier, generates an action explanation, and records a rollback-ready audit trace.

A complete empirical study should preregister data splits, thresholds, scenario definitions, unsafe-action labels, exclusion rules, and statistical tests. All code, policy files, agent versions, random seeds, infrastructure manifests, and response-oracle rules should be archived. Dataset licenses and access restrictions must be verified before redistribution.

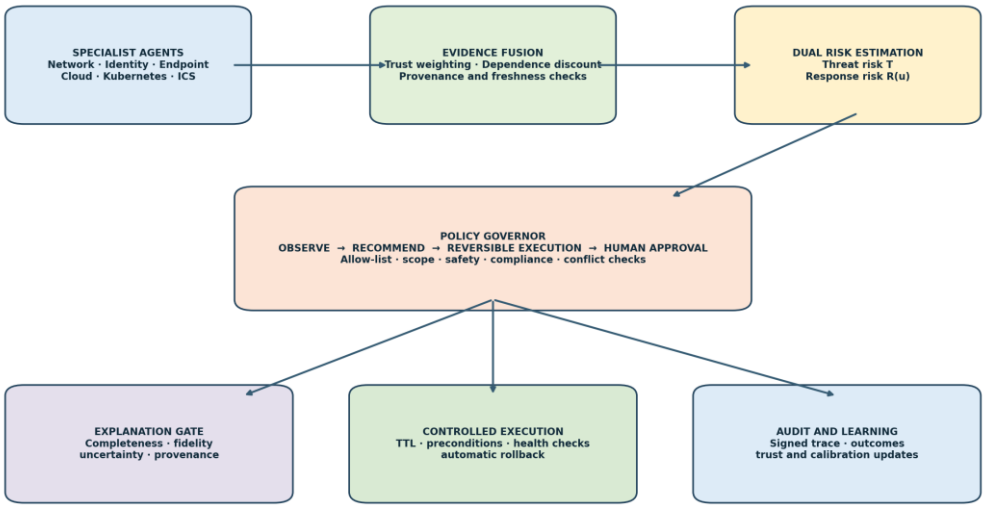

Figure 1. TCX-MAD architecture and pre-execution safety controls
Source: Authors' design. The architecture is a proposed artefact and not an empirical result.

Let an incident candidate at time t be represented by context x_t, asset a, and evidence items $e_1 \ldots e_n$. Specialist agents produce a claim c_i, confidence p_i, evidence provenance v_i, and candidate action set A_i. Typical roles include network, identity, endpoint, cloud-control-plane, Kubernetes-workload, industrial-process, threat-intelligence, and mission-impact agents. A separate policy governor has no detection incentive and cannot be bypassed by planner agents. Agents communicate through authenticated, schema-validated messages; tools are least-privileged; and evidence is bound to time, source, and integrity metadata.

To reduce common-mode failure, the framework records model lineage and data-source dependence. Two agents using the same foundation model or derived telemetry are not treated as independent votes. The evidence-fusion module discounts correlated claims and abstains when provenance or freshness requirements are not satisfied.

Threat risk $T_t \in [0,1]$ estimates the expected security and mission loss if the detected activity is not contained. One implementable form is:

$$T_t = Cal[\sigma(\beta_0 + \beta_1 Pattack + \beta_2 Iasset + \beta_3 Cconfidence + \beta_4 Pprogression + \beta_5 Eexposure)],$$

where Cal denotes post-hoc calibration, Pattack is fused attack probability, Iasset is asset/mission impact, Cconfidence reflects trust-weighted evidence quality, Pprogression estimates likely attack progression, and Eexposure represents vulnerability and exposure. Calibration should be assessed with reliability diagrams, Brier score, and expected calibration error, not accuracy alone.

For a proposed action u, response risk $R(u) \in [0,1]$ estimates expected operational harm:

$$R(u) = Cal[\omega_1 Bblast + \omega_2 Mirreversibility + \omega_3 Ddependency + \omega_4 Ssafety + \omega_5 Uuncertainty + \omega_6 Ccompliance - \omega_7 Rrollback],$$

where blast radius, irreversibility, dependency centrality, physical/process safety, model uncertainty, and compliance impact increase risk, while tested rollback capacity

decreases it. In production, the components must be grounded in asset inventories, service maps, safety constraints, recovery objectives, change windows, and sector-specific rules. A response action is beneficial only when its risk-adjusted avoided loss exceeds its expected action harm and policy constraints are satisfied.

The coefficients β for threat risk and ω for response risk should be estimated on temporally separated, labelled incident and action-outcome data rather than fixed by convenience. A regularized logistic or ordinal regression can estimate β from incident labels and ω from action-harm labels such as service-level-objective violation, unsafe process deviation, failed recovery, or compliance breach. Where harmful actions are rare, class weighting, hierarchical partial pooling across asset classes, and expert elicitation can supply priors, but expert judgments must be documented and tested against held-out scenarios. Post-hoc isotonic or Platt calibration may be fitted on a calibration split; conformal calibration can additionally provide finite-sample risk sets or abstention bounds under stated exchangeability assumptions.

Trust weights w should be constrained to preserve the intended signs and selected by nested cross-validation or Bayesian estimation against agent outcome labels. The decay factor λ controls adaptation versus stability: it should be chosen by replaying abrupt and gradual drift, then minimizing a preregistered loss that penalizes both delayed detection of degradation and false trust alarms. Initial trust priors should be conservative for agents with few validated outcomes. All fitted parameters must be versioned by environment, asset class, and agent role; transport to a new site requires recalibration rather than silent reuse.

The thresholds θT, θR, $\theta \tau$, and θE should be selected jointly on a validation set or digital-twin scenario catalogue. Candidate threshold tuples are evaluated on the unsafe-action/valid-response-latency Pareto frontier subject to hard constraints, including a maximum unsafe-action rate, a maximum approval workload, and the preregistered non-inferiority margin δ for valid-response delay. θE should also enforce mandatory-field completeness before any learned usefulness score is considered. Threshold sensitivity analysis should report outcomes across plausible neighborhoods, and bootstrap or conformal intervals should identify settings whose safety claims are unstable. Independent experts may define unacceptable regions, but final thresholds must be frozen before the test set is opened.

Each agent i receives a time-varying trust score $\tau_i,t \in [0,1]$. A practical exponentially decayed update is:

$$\tau_i,t = clip[(1-\lambda)\tau_i,t-1 + \lambda(w_1 Sproper + w_2 Kcal + w_3 Cind + w_4 Vprov - w_5 Mmanip - w_6 Ddrift)],$$

where Sproper is a proper-scoring-rule performance term, Kcal measures calibration, Cind measures consistency with independently sourced evidence, Vprov scores provenance validity, Mmanip captures compromise or injection indicators, and Ddrift measures distributional or behavioural drift. Trust is contextual: an identity agent may be highly trusted for authentication anomalies but not for physical-process impact. Bayesian credible intervals or conformal bounds should accompany point estimates when sample counts are small.

Table 2. Operational measurement of dynamic-trust components

Term	Operational measure	Window / update	Normalization	Failure signal
S_proper	1 minus normalized Brier or logarithmic loss on adjudicated outcomes	Rolling N labelled predictions; recency weighted	Map loss to [0,1] using preregistered bounds	Worse proper score than role baseline
K_cal	1 minus expected calibration error (ECE), with adaptive bins and reliability slope/intercept	Rolling N predictions; minimum count required	K_cal = 1 - clipped ECE	ECE above limit or slope materially differs from 1
C_ind	Agreement with independently sourced, adjudicated evidence after lineage/dependence discount	Per incident, then exponentially aggregated	Weighted agreement in [0,1]	Agreement exists only among correlated agents
V_prov	Fraction of evidence passing signature, freshness, schema, source, and attestation checks	Per message and per incident	Passed checks / applicable checks	Missing, stale, replayed, or unverifiable evidence
M_manip	Composite rate/severity of injection, spoofing, anomalous tool use, policy-evasion, and attestation alerts	Event driven plus rolling incident window	Calibrated anomaly score in [0,1]	Any high-severity compromise indicator
D_drift	Population stability index, feature-distribution distance, prediction shift, and residual-error change	Scheduled and change-triggered	Ensemble drift score in [0,1]	Sustained threshold breach across dwell period

Trust affects fusion weight and the maximum autonomy an agent can support. Hard controls dominate trust: a high-trust agent cannot authorize an irreversible action, override a safety interlock, or act when explanation and provenance requirements fail. Quarantine, re-attestation, and reduced privilege are triggered when trust drops below policy thresholds or changes abruptly.

Trust accumulation is vulnerable to "reputation farming": a compromised agent may behave correctly on frequent, low-cost cases to build a high score and then exploit that reputation on a rare, high-impact action. TCX-MAD therefore caps the authority contribution of historical trust, weights errors by action consequence, maintains separate trust by task and asset class, requires fresh independent evidence for high-stakes decisions, and applies abrupt penalties or quarantine on manipulation indicators. No accumulated trust can waive the hard A2 gates.

Illustrative trust update (hypothetical). Let $\lambda = 0.20$, *prior* $\tau_{i,t-1} = 0.70$, *positive weighted evidence* $w_1 S_{proper} + w_2 K_{cal} + w_3 C_{ind} + w_4 V_{prov} = 0.78$, *and penalties* $w_5 M_{manip} + w_6 D_{drift} = 0.08$. *Then* $\tau_{i,t} = 0.80(0.70) + 0.20(0.70) = 0.70$. *If a later manipulation alert raises the penalty-adjusted evidence term to 0.10, the next update becomes* $0.80(0.70) + 0.20(0.10) = 0.58$, *before any policy-mandated quarantine override. These values demonstrate the update mechanics only; they are not observed results.*

The governor evaluates T, R(u), asset criticality κ, reversibility q, aggregate trust $\bar{\tau}$, explanation sufficiency E, and policy constraints P. The four autonomy tiers are:

Table 3. TCX-MAD autonomy tiers and authority conditions

Tier	Authority	Typical condition	Example
A0 — Observe	Log, enrich, and continue monitoring	Low threat or insufficient evidence	Increase telemetry and preserve artefacts
A1 — Recommend	Present ranked actions; no machine execution	Moderate threat, uncertain impact, or low trust	Recommend identity review or network segmentation
A2 — Reversible execute	Execute pre-authorized, scoped action with automatic timeout/rollback	High-enough threat; low response risk; adequate trust and explanation	Temporarily quarantine one pod; rate-limit a source; expire a short-lived token
A3 — Human approval	Block execution pending authorized decision	High/critical response risk, irreversible action, safety asset, or failed gate	Isolate a PLC network; revoke a production root; shut down a cluster

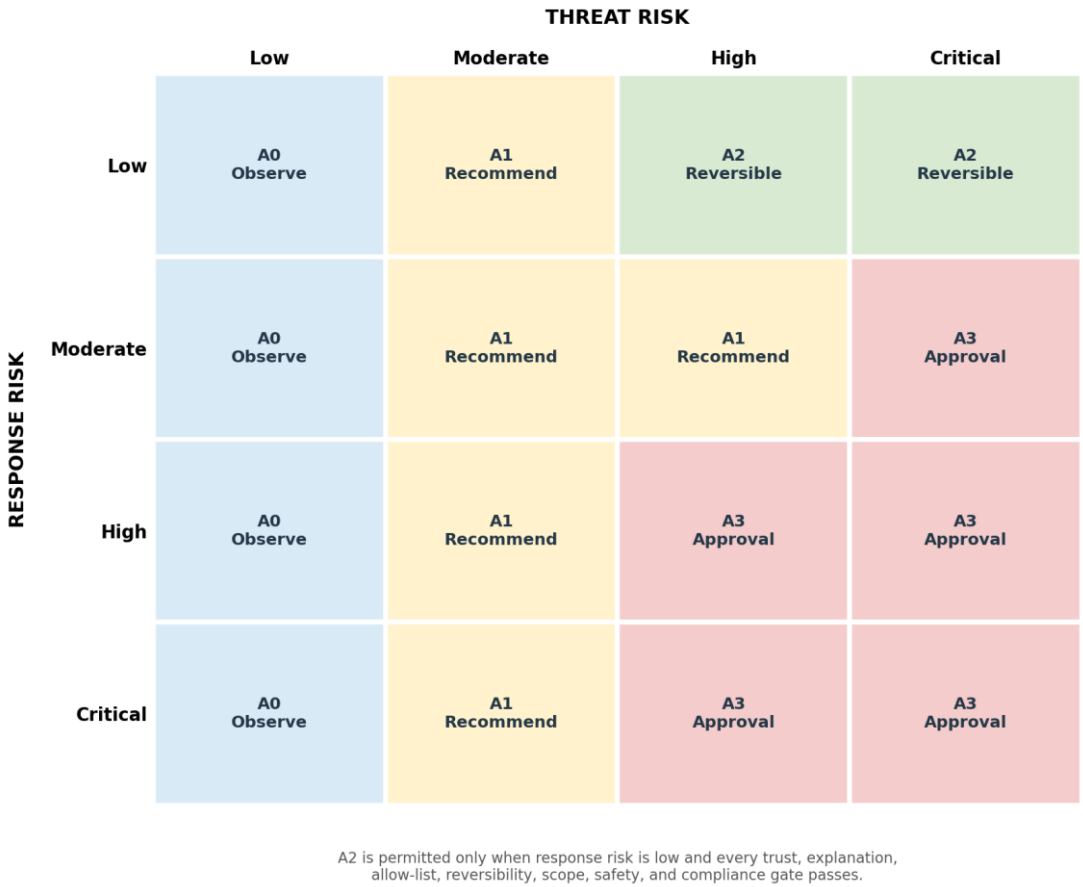

Figure 2. Illustrative policy matrix (actual thresholds must be sector- and asset-specific) Source: Authors' design. A high threat score does not automatically justify a high-risk response.

A conservative reference rule is: A2 is allowed only if $T \geq \theta T$, $R(u) \leq \theta R$, $\bar{\tau} \geq \theta \tau$, $E \geq \theta E$, the action is on an allow-list, reversibility has been tested, affected scope is bounded, and no safety or compliance constraint is violated. Otherwise the governor selects A0, A1, or A3. A2 actions receive a time-to-live, precondition checks, post-action health checks, and automatic rollback. Concurrent actions are serialized or conflict-checked to prevent interacting agents from creating cumulative harm.

To prevent tier flapping near θT, θR, $\theta \tau$, or θE, production policy should use asymmetric enter/exit thresholds, a minimum dwell time, and debouncing over consecutive evaluations. A tier may be raised only after the higher-authority conditions persist for k evaluations, whereas any hard-safety failure downgrades authority immediately.

Algorithm 1. TCX-MAD policy-governor decision logic

Input: context x_t, evidence e, candidate actions U, policies P, agent history H
1 Validate message schema, provenance, freshness, and model/data lineage.
2 Compute contextual agent trust $\tau_{i,t}$ and dependence-discounted aggregate trust $\bar{\tau}$.
3 Fuse evidence and estimate calibrated threat risk T.
4 For each u in U, estimate calibrated response risk $R(u)$, reversibility q, and scope.
5 Generate signed decision card; compute explanation sufficiency E and fidelity checks.
6 If threat/evidence is insufficient: return A0 (observe).
7 If u is irreversible, safety/compliance constrained, or $R(u)$ is high/critical: return A3 (approval).
8 If $T \geq \theta T$ and $R(u) \leq \theta R$ and $\bar{\tau} \geq \theta \tau$ and $E \geq \theta E$ and allow-list/scope/rollback gates pass:
9 return A2 (execute reversibly with TTL, health checks, audit, and rollback).
10 If action is defensible but any soft gate fails: return A1 (recommend).
11 Apply hysteresis to tier increases; apply immediate downgrade on any hard-gate failure.
Output: tier, selected action or abstention, signed explanation, rollback plan, audit record

An action explanation is represented as a signed decision card. Completeness is machine-checkable and fidelity is evaluated against the actual features, rules, trust weights, and counterfactual alternatives used by the governor. The minimum fields are shown in Table 4.

Table 4. Required fields and verification criteria for signed action explanations

Required field	Question answered	Verification criterion
Triggering evidence	What observations caused escalation?	Source, timestamp, provenance, uncertainty, and supporting/contradicting evidence are present
Affected asset	What will be acted on?	Unique asset identity, owner, criticality, dependencies, and current health are resolved
Action rationale	Why this action over alternatives?	Expected avoided loss, alternatives rejected, and policy rule are stated
Operational risk	What could the defence disrupt?	Blast radius, availability/safety impact, uncertainty, and compliance implications are stated
Trust disclosure	Whose judgment was relied upon?	Agent trust values, confidence bounds, and correlation discounts are shown
Rollback and verification	How is harm bounded and success checked?	Timeout, rollback command, health indicators, stop conditions, and responsible human are identified

Explanation sufficiency E is a weighted combination of completeness, fidelity, uncertainty disclosure, actionability, and analyst-rated usefulness. Completeness alone is inadequate: a fluent explanation can be fully populated yet unfaithful. Evaluation

therefore includes perturbation or counterfactual tests to confirm that stated reasons track the governor's actual decision logic.

The evaluation assumes adversaries may manipulate telemetry, exploit shared-model vulnerabilities, inject malicious instructions into logs or tickets, spoof inter-agent messages, compromise one specialist agent, induce distribution shift, replay stale evidence, or exploit a response tool's privileges. It also considers non-adversarial failure: sensor faults, asset-inventory errors, dependency-map staleness, model overconfidence, human delay, and rollback failure. The trusted computing base comprises the policy governor, credential boundary, audit log, policy store, and attestation mechanism; compromise of this entire base is outside the initial experimental scope and is listed as a limitation. The model specifically includes trust-mechanism gaming in which an agent builds reputation on cheap, frequent decisions before attempting a rare high-consequence action.

A reproducible evaluation should use complementary datasets rather than a single benchmark. Network detection can be evaluated with CICIDS2017 [20]; heterogeneous IoT/IIoT telemetry with TON_IoT [21] and Edge-IIoTset [22]; and industrial-process anomalies with HAI [23]. Cloud-native evaluation should add an instrumented Kubernetes testbed that generates versioned audit logs, Falco-compatible runtime events, identity traces, service-level indicators, and ground-truth attacks. Candidate scenarios include credential misuse, malicious deployment, container escape attempt, service-account token abuse, lateral movement, data exfiltration, and supply-chain image replacement.

Table 5. Evaluation evidence sources and response-safety augmentation

Evidence source	Primary role	Key caution	Response-safety augmentation
CICIDS2017 [20]	Network intrusion benchmark	Age, synthetic profiles, and leakage risks	Replay into a segmented testbed; label containment scope and service impact
TON_IoT [21]	Network, OS, and IoT/IIoT telemetry	Cross-source alignment and class imbalance	Map affected services to reversible isolation and rate-limiting actions
Edge-IIoTset [22]	IoT/IIoT and federated settings	Testbed-to-production generalization	Inject device dependency and safety-criticality metadata
HAI [23]	ICS time-series anomalies	Point-adjustment and testbed specificity	Use a digital twin/HIL safety oracle; prohibit unsafe actuator or network actions

Evidence source	Primary role	Key caution	Response-safety augmentation
Kubernetes testbed	Cloud audit, identity, workload, and SLO data	Must be created and versioned reproducibly	Measure outage minutes, error budget, rollback success, and dependency effects

No empirical dataset was supplied with the present manuscript. Accordingly, no performance value in Section 6 is represented as observed. Before journal submission as an empirical paper, the authors must execute the protocol and replace the reporting template with measured results, confidence intervals, and artefact links.

The principal comparison uses identical detector outputs wherever possible so that safety effects are not confounded with predictive model changes. The following baselines are required:

B1—Threat-only automation: action tier depends on threat score, with no response-risk estimator.

B2—Static-trust multi-agent system: fixed or equal agent weights.

B3—Dual-risk without explanation gate: tests whether explanations operate as a real safety control.

B4—Human approval for all actions: safety-oriented latency and workload reference.

B5—TCX-MAD full framework.

Ablations remove response risk, trust updating, evidence-correlation discounting, explanation sufficiency, rollback verification, and manipulation detection one at a time. Stress conditions include label noise, 5–40% poisoned telemetry, one compromised agent, two colluding agents, stale asset dependencies, abrupt workload drift, and delayed human approval. Experiments should be repeated across seeds and incident families, with temporal or scenario-level splits to reduce leakage.

The primary metric is unsafe automated actions prevented without significant valid-response delay. Let Ubase be unsafe actions executed by a baseline and Utcx the number under TCX-MAD. Unsafe-action prevention rate is (Ubase − Utcx)/Ubase. This must be reported beside ΔMRT, the difference in mean or median response time for valid containment. A Pareto analysis identifies settings that reduce unsafe actions while keeping delay below a preregistered operational margin δ.

Table 6. Evaluation metrics and reporting requirements

Metric	Definition / reporting	Why it matters
Precision, recall, F1	Incident- and class-level; macro and weighted variants	Detection utility under imbalance
False-positive rate	Benign events incorrectly escalated	Automation burden and disruption exposure

Metric	Definition / reporting	Why it matters
Mean/median detection time	First malicious event to qualified alert	Predictive timeliness
Mean/median response time	Qualified alert to successful containment	Cost of governance controls
Unsafe-action rate	Unsafe automated actions / automated actions proposed	Primary response-safety failure rate
Unsafe actions prevented	Baseline unsafe actions avoided, with Δresponse time	Distinctive safety–latency outcome
Service-disruption rate	Actions causing SLO/safety threshold violation	Operational consequence
Rollback success/time	Successful restorations and time to recovery	Reversibility validity
Explanation quality	Completeness, fidelity, usefulness, calibration effect	Human oversight quality
Human-approval percentage	Incidents/actions routed to A3	Workload and autonomy balance

Binary safety outcomes should be compared using mixed-effects logistic regression or generalized estimating equations with scenario and dataset as clustering factors. Time outcomes should use survival analysis or robust non-parametric comparisons when skewed. Report absolute risk differences, odds ratios, bootstrap 95% confidence intervals, corrected multiple comparisons, calibration plots, and effect sizes. A non-inferiority margin δ for valid-response delay must be operationally justified before analysis. Analyst studies should be counterbalanced and report inter-rater reliability, decision accuracy, decision time, workload, and trust calibration.

RESULTS AND DISCUSSION

Result

Proposition 1—No unbounded autonomous action. If the allow-list, scope bound, reversibility test, and rollback requirement are enforced by a non-bypassable governor, no action outside the A2 reversible-action envelope can be autonomously executed. This is a policy-enforcement property, not a probabilistic model claim.

Justification: The governor is the sole execution path, and its transition relation contains no A2 edge for actions failing any of the four envelope predicates. Therefore, assuming non-bypassability, an out-of-envelope action cannot reach the executor; the claim follows by construction rather than empirical observation.

Proposition 2—High threat does not erase response risk. Because T and R are evaluated separately, a critical attack can still be routed to human approval when the

proposed action threatens safety or availability. This prevents the common but unsafe inference that greater threat severity always warrants more aggressive autonomy.

Justification: R(u) is evaluated as an independent guard after T is computed. Increasing T cannot make the predicate $R(u) \leq \theta R$ true, so a high or critical response-risk action remains excluded from A2 even at maximal threat.

Proposition 3—Low trust cannot increase authority. When the tier function is monotone non-decreasing in aggregate trust for otherwise fixed inputs, decreasing trust can only preserve or reduce the authorized tier. Hard constraints continue to dominate even at maximum trust.

Justification: Aggregate trust appears only as the lower-bound condition $\bar{\tau} \geq \theta\tau$ and as a weight on evidence; hard constraints do not relax as trust rises. Holding all other inputs fixed, lowering $\bar{\tau}$ can remove eligibility for A2 but cannot create it.

Proposition 4—Explanation failure is fail-safe. When E falls below threshold or required fields are missing, A2 execution is denied and the decision is downgraded to recommendation or escalated for approval. Explanation generation therefore influences authority rather than merely documenting it.

Justification: $E \geq \theta E$ and mandatory-field completion are conjunctive A2 conditions. If either fails, the A2 branch is unreachable and the governor returns A1 or A3, so explanation failure is fail-safe by policy construction.

Proposition 5—Reversible execution is temporally bounded. Every A2 action has a time-to-live and post-action health checks. Failure to meet success or safety conditions triggers rollback or escalation, limiting duration even when initial response-risk estimation is wrong.

Justification: Every A2 transition instantiates a TTL, post-action checks, and a rollback or escalation branch. Thus autonomous authority terminates at the TTL or earlier on a failed health condition, assuming the executor and rollback mechanism operate as specified.

Consider a hypothetical Kubernetes incident in which network and identity agents detect unusual east–west connections and a service-account token used from a new workload. After trust and correlation adjustment, the fused threat risk is $T = 0.86$. The planner proposes (u1) temporary isolation of one pod and (u2) revocation of the namespace-wide service account. The response estimator assigns $R(u1) = 0.18$ because scope is one replica, a healthy replacement exists, and rollback is tested; $R(u2) = 0.72$ because multiple production services share the identity and the dependency map is uncertain. Aggregate contextual trust is 0.81 and explanation sufficiency is 0.93.

Under the illustrative policy, u1 qualifies for A2 reversible execution with a five-minute timeout and health checks. The higher-impact u2 is routed to A3 human approval despite the same high threat score. A threat-only policy might execute both actions. The example shows the mechanism by which TCX-MAD can prevent a potentially disruptive action without delaying a bounded containment step; it does not establish a measured prevention rate.

Table 7. Hypothetical Kubernetes decision trace

Input / outcome	Pod isolation u1	Identity revocation u2
Threat risk T	0.86 (hypothetical)	0.86 (hypothetical)
Response risk R	0.18 (hypothetical)	0.72 (hypothetical)
Reversibility	Tested; automatic timeout	Uncertain; broad dependency
Decision	A2—execute reversibly	A3—request approval
Safety rationale	Bounded scope and healthy replacement	Potential multi-service outage

Table 8. Preregistered empirical evaluation and reporting template

System	F1	FPR	Median response time	Unsafe-action rate	Service disruption	Human approval
B1 Threat-only	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
B2 Static trust	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
B3 No explanation gate	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured
B4 Human approval all	To be measured	To be measured	To be measured	To be measured	To be measured	100% by design
B5 TCX-MAD	To be measured	To be measured	To be measured	To be measured	To be measured	To be measured

Primary inferential statement to complete: 'Compared with B1, TCX-MAD prevented [n/N; %; 95% CI] unsafe automated actions, while changing median valid-response time by [value; 95% CI]. The preregistered non-inferiority margin was [δ], and the observed difference [did/did not] meet it.' This sentence must not be completed until experiments are run.

Discussion

TCX-MAD changes the optimization target of autonomous cyber defence. A conventional pipeline asks whether activity is malicious and which action most rapidly contains it. The proposed framework asks an additional question before execution: whether this system, using this evidence, should be trusted to take this action against this asset now. The shift is especially important in operational technology and tightly coupled cloud platforms, where availability and physical safety are first-class security objectives.

Dual risk scoring makes explicit a trade-off that is often hidden in response playbooks. Threat risk estimates the loss of inaction; response risk estimates the loss created by intervention. Keeping the scores separate preserves interpretability and makes policy review possible. A single composite score could conceal a critical threat paired

with an unacceptable action, producing the appearance of moderate overall risk. The two-dimensional policy instead supports alternative selection: a high-risk response can be rejected while a lower-risk containment action proceeds.

Dynamic trust management addresses the fact that multi-agent consensus is not inherently safe. Historical accuracy, current calibration, evidence independence, provenance, and manipulation indicators contribute different information. The framework's use of contextual trust also avoids giving a generally strong agent authority outside its competence. Nevertheless, trust scores can create false precision. They should be accompanied by uncertainty intervals, minimum evidence counts, drift alarms, and clear governance over who can change weights and thresholds.

The explanation gate connects explainability to control. Most XAI evaluation ends with an analyst's perception of clarity. Here, explanation completeness and fidelity determine whether autonomous authority is granted. This creates an incentive to generate decision-relevant explanations, but it also creates a new attack surface: an agent may produce a persuasive yet unfaithful narrative. Machine-verifiable provenance and counterfactual fidelity tests are therefore necessary, and natural-language fluency should never be used as a proxy for correctness.

Tiered autonomy is also preferable to the false dichotomy between full automation and universal human approval. Pre-authorized, reversible actions can preserve speed for bounded cases, while human attention is reserved for high-impact or ambiguous decisions. The empirical question is whether this routing reduces disruption without overloading analysts. The percentage of incidents requiring approval must therefore be reported alongside safety, delay, and analyst workload.

The governor is not cost-free. Provenance verification, risk inference, trust aggregation, explanation generation, and policy checks add computation and coordination delay before action. That overhead is an explicit empirical quantity under RQ3 and should be decomposed by stage at the median, 95th, and 99th percentiles, including degraded-network and cross-boundary IT/OT conditions.

Operational deployment presupposes a mature asset inventory, current service and process dependency maps, tested rollback procedures, and interfaces to SIEM, SOAR, identity, orchestration, and change-management systems. Sites lacking these foundations should begin in A0/A1 shadow mode and promote only well-instrumented, reversible playbooks to A2. Implementation cost is therefore driven as much by telemetry quality, dependency modelling, and recovery testing as by the agent models themselves.

Alignment with NIST guidance is conceptual rather than a claim of certification. The framework operationalizes AI RMF characteristics through governance, measurement, documentation, and management controls [3], and it reflects the CSF 2.0 emphasis on cybersecurity-risk governance [1]. These standards strengthen the policy relevance of secure coordination, lifecycle assurance, and tested safeguards, but sector regulators and operators must still define acceptable thresholds, accountable roles, and prohibited actions.

CONCLUSION

Fundamental Finding: Safe predictive cyber defence requires more than accurate detection and rapid automation. In critical infrastructure and cloud-native systems, defensive actions can disrupt the same missions they are intended to protect. TCX-MAD addresses this problem through an integrated pre-execution safety governor that jointly evaluates threat risk, response-induced operational risk, dynamic agent trust, and explanation sufficiency. It authorizes only bounded, pre-approved, reversible actions and escalates high-impact, irreversible, low-trust, or poorly explained decisions for human approval. **Implication:** The framework's central empirical test is deliberately safety-centred: how many unsafe automated actions are prevented without materially slowing valid incident response? Public detection datasets, instrumented cloud/ICS testbeds, reproducible response oracles, ablation studies, adversarial stress tests, and human-factor evaluation are required to answer that question. The present article contributes the architecture, formal policy, metrics, and reporting protocol; it does not substitute hypothetical values for experiments. This evidence boundary is essential to the trustworthy-AI principles the framework itself advocates. **Limitation:** The article presents a design artefact and evaluation protocol; no executed benchmark or analyst-study data were provided, so effectiveness is not empirically established; Public intrusion datasets do not directly label operationally unsafe responses. Response-safety claims depend on the validity of the testbed, digital twin, expert labels, and dependency model; Risk and trust scores may appear precise while reflecting uncertain, correlated, or incomplete inputs. Calibration can degrade under distribution shift; A non-bypassable governor and trustworthy audit trail are assumed. Compromise of the trusted computing base could invalidate all higher-level safeguards; Sector-specific safety constraints, legal duties, and acceptable downtime vary; a universal threshold policy would be inappropriate; Human approval can be delayed, biased, or affected by automation bias. Escalation is not automatically safe without usable explanations, training, and accountability; The framework may introduce compute, communication, and orchestration latency, particularly when evidence must be verified across IT and OT boundaries. **Future Research:** Implement TCX-MAD in a reproducible Kubernetes cyber range and an ICS hardware-in-the-loop or high-fidelity digital twin, then publish action-impact labels and policy manifests; Develop causal response-risk estimators that predict downstream service and physical-process effects rather than relying only on static dependency graphs; Investigate conformal risk bounds and distribution-free abstention for threat, response, and trust estimates under shift; Model collusion and shared-model common-mode failures; quantify how evidence dependence should discount consensus; Use formal methods to verify non-bypassability, least privilege, action conflict constraints, rollback invariants, and temporal safety properties; Conduct controlled analyst studies comparing structured decision cards, feature attribution, and raw alerts for decision quality, workload, time, and calibrated reliance; Create sector-specific policy profiles for energy, water, health, transport, finance, and telecommunications, aligned with relevant safety and regulatory obligations; Evaluate privacy-preserving cross-

organization trust learning without exposing sensitive incident data or enabling reputation poisoning.

REFERENCES

[1] C. Pascoe, S. Quinn, and K. Scarfone, "The NIST Cybersecurity Framework (CSF) 2.0," NIST Cybersecurity White Paper NIST CSWP 29, National Institute of Standards and Technology, Gaithersburg, MD, USA, Feb. 2024, doi: 10.6028/NIST.CSWP.29.

[2] A. Kott, P. Théron, M. Drašar, E. Dushku, B. LeBlanc, P. Losiewicz, A. Guarino, L. Mancini, A. Panico, M. Pihelgas, K. Rzadca, and F. De Gaspari, "Autonomous Intelligent Cyber-defense Agent (AICA) Reference Architecture, Release 2.0," arXiv:1803.10664, 2018, doi: 10.48550/arXiv.1803.10664.

[3] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[4] M. Standen, M. Lucas, D. Bowman, T. J. Richer, J. Kim, and D. Marriott, "CybORG: A Gym for the Development of Autonomous Cyber Agents," in Proc. 30th Int. Joint Conf. Artificial Intelligence (IJCAI), 2021, pp. 4957–4959, doi: 10.48550/arXiv.2108.09118.

[5] TTCP CAGE Working Group, "CAGE Challenge 4: A Multi-Agent Reinforcement Learning Cyber Defence Challenge," 2024. [Online]. Available: https://github.com/cage-challenge/cage-challenge-4

[6] P. Cichonski, T. Millar, T. Grance, and K. Scarfone, "Computer Security Incident Handling Guide," NIST Special Publication 800-61 Rev. 2, National Institute of Standards and Technology, Gaithersburg, MD, USA, Aug. 2012, doi: 10.6028/NIST.SP.800-61r2.

[7] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," IEEE Access, vol. 10, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.

[8] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," IEEE Access, vol. 10, pp. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[11] P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, and M. A. Przybocki, "Four Principles of Explainable Artificial Intelligence," NISTIR 8312, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2021, doi: 10.6028/NIST.IR.8312.

[12] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Trans. Interactive Intelligent Systems, vol. 11, nos. 3–4, Art. no. 24, 2021, doi: 10.1145/3387166.

[13] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[14] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," arXiv:1812.04608, 2018, doi: 10.48550/arXiv.1812.04608.

[15] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017, doi: 10.48550/arXiv.1702.08608.
[16] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," Human Factors, vol. 39, no. 2, pp. 230–253, 1997, doi: 10.1518/001872097778543886.
[17] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50_30392.
[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," IEEE Trans. Systems, Man, and Cybernetics—Part A: Systems and Humans, vol. 30, no. 3, pp. 286–297, 2000, doi: 10.1109/3468.844354.
[19] A. Jøsang, Subjective Logic: A Formalism for Reasoning Under Uncertainty. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-42337-1.
[20] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116, doi: 10.5220/0006639801080116.
[21] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems," IEEE Access, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.
[22] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning," IEEE Access, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.
[23] H.-K. Shin, W. Lee, J.-H. Yun, and B.-G. Min, "Two ICS Security Datasets and Anomaly Detection Contest on the HIL-Based Augmented ICS Testbed," in Proc. Cyber Security Experimentation and Test Workshop (CSET), 2021, pp. 36–40, doi: 10.1145/3474718.3474719.

***Sajidul Haque Chowdhruy (Corresponding Author)**
Washington University of Science and Technology, United States
Email: Sajidulhc@gmail.com

Shakila Akter
Lewis University, United States
Email: saktershakila23@gmail.com

Md. Golam Mostafa
National University, Bangladesh
Email: g.mostafabd67@gmail.com